LREC 2014
Reykjavik

NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Held under the Patronage of UNESCO, the United Nations Educational, Scientific and Cultural Organization

MAY 26 – 31, 2014

HARPA CONFERENCE CENTER
REYKJAVIK, ICELAND

CONFERENCE ABSTRACTS

Editors: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis.

Assistant Editors: Sara Goggi, Jérémy Leixa, Hélène Mazo

LREC 2014, NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Title: LREC 2014 Conference Abstracts

Distributed by:
ELRA – European Language Resources Association
9, rue des Cordelières
75013 Paris
France

Tel.: +33 1 43 13 33 33
Fax: +33 1 43 13 33 30

www.elra.info and www.elda.org
Email: info@elda.org and lrec@elda.org

ISBN 978-2-9517408-8-4
EAN 9782951740884

Introduction of the Conference Chair and ELRA President
Nicoletta Calzolari

I wish to express to Mrs. Irina Bokova, Director-General of UNESCO, the gratitude of the Program Committee, of all LREC participants and my personal for her Distinguished Patronage of LREC 2014. Languages – mentioned in the first article of UNESCO Constitution – have been at the heart of UNESCO mission and programmes throughout its history.

I am also especially grateful to Madame Vigdís Finnbogadóttir, UNESCO's Goodwill Ambassador for languages and former President of Iceland (1980-1996), first woman in the world elected as head of state in a democratic election, for the continuous personal support she has granted to LREC since our first visit in Reykjavík in 2012. In her name the Vigdís International Centre for Multilingualism and Intercultural Understanding has been established under the auspices of UNESCO to promote multilingualism and raise awareness of the importance of language as a core element of the cultural heritage of humanity. I quote a sentence from a recent interview where she says: "The land—our nature—and language, those are our national treasures": this tells a lot of why this LREC is in Iceland!

Some figures: all records broken!

LREC 2014, the 9th LREC, with its 1227 submissions, has set a new record! We received 21% more submissions than in 2012. We continue the tradition of breaking our own previous records: out of the 1227 submissions, after the reviewing process by well 970 colleagues, we accepted 745 papers. We also accepted 22 workshops and 9 tutorials. More than 1100 participants have already registered at the beginning of May.

These figures have a meaning. The field of Language Resources and Evaluation is continuously growing. And LREC continues to be – as many say – "the conference where you have to be and where you meet everyone".

Every time I underline the fact that a relatively high acceptance rate (60.7% this time) is for us a reasoned choice. It is important to get a pulse on the situation, to monitor the evolution of the field in the many varieties of approaches and methodologies, and in particular for many different languages. For us, a lexicon in any language is as important as a lexicon in American English. Multilingualism – and equal treatment of all languages – is a feature at the heart of LREC. Other venues promote a sense of exclusivity (also through the equation low acceptance rate and great merit); we always encourage a sense of inclusiveness. This is a typical feature of LREC that makes it a special conference. Quality is not necessarily undermined by a high acceptance rate, but also by the influence of the papers on the community: the ranking of LREC among other conferences in the same area proves this. According to Google Scholar h-index, LREC ranks 4th in Computational Linguistics at a similar level of conferences using much lower acceptance rates, just like the LRE Journal also ranks 4th in the general field of Humanities, Literature and Arts.

LREC 2014 Trends

Language Resources (LRs) being everywhere in Language Technology (LT), LREC is a perfect observation point of the evolution of the field. Looking at all the topics, while building the program and putting all the pieces together, the most striking (even if not surprising) new trend was for me the application of sentiment/opinion discovery/analysis to social media shown by so many papers.

A very rough sketch of LREC 2014 major topics and trends, from my viewpoint, is the following:

- There is a completely new topic:
- Linked Data, also the hot topic of this edition
- Topics that were quite new in 2012 and are now consolidated:
- Social Media, in particular combined with subjectivity, as said above
- Crowdsourcing and Collaborative Construction of LRs
- Other increasing (not the biggest in absolute terms) topics with respect to last LREC are:
- Subjectivity: Sentiments, Emotions, Opinions
- Less-resourced languages, in line with the value we give to safeguarding world's linguistic diversity
- Extraction of Information, Knowledge discovery, Text mining: always a very hot topic
- Computer Aided Language Learning
- Stable Big topics:
- Infrastructural issues and Large projects, and also Standards and Metadata, receive the usual attention by the LREC authors
- Lexicons and Corpora (i.e. the most typical "data"), of many types, modalities and for many purposes and applications: they are the prominent and most crowded topic
- Semantics and Knowledge, in all their variations: from annotation of anaphoric information, to ontologies and WordNets, sense disambiguation, named entities recognition, information extraction, to mention just a few
- Syntax, Grammar and Parsing continues to be a largely represented topic: not solved
- Machine Translation and Multilingualism are areas on which a lot of work is carried out
- Speech and Multimodality keep the same level: good but not enough
- Dialogue and discourse, with contributions from both the Speech and Text communities
- Evaluation is pervasive/everywhere: we are proud to give evidence to its being an essential feature in the LT landscape
- Tools, systems for text analysis and applications are presented in many papers

A usual observation is the relevance of *infrastructural issues* and the attention that LREC – and ELRA – pay to them. They are mostly neglected in other conferences. Infrastructural issues play an important role for the field of LRs and for the LT field at large. But it is a fact that the first to recognise their importance have been people of the LR area. LRs are themselves of infrastructural nature and quite naturally call for attention to these issues. The infrastructural nature of LRs, captured by the term "Resources", was highlighted in the Introduction of Antonio Zampolli to the 1st LREC in Granada in 1998.

The fact that so many topics are represented at LREC means also that all the various LR and LT sub-communities are present at LREC: this increases the LREC impact and gives to LREC the characteristic of being a true melting pot of cultures, and an enabler of new cooperation initiatives.

15th LREC Anniversary

LREC was born in 1998 and on the occasion of its 15th Anniversary, Joseph Mariani has prepared an analysis of all the past LREC Proceedings, rediscovering the dynamism of the field while looking at the major contributors, topics, trends, also comparing them with an analogous survey done for the speech community on the Interspeech conference series. There will also be a Quiz for all the LREC participants and a winner! The survey paper is in the Proceedings as a special paper for the 15th Anniversary and will be presented at the Closing Session.

ELRA and LREC: a tradition of innovations at the service of our community

I am proud to announce a number of recent initiatives of ELRA and LREC that touch topics that are at the forefront of a paradigm shift and together help advance our field and increase confidence in scientific results. As an introduction I use some words of Zampolli in 1998: "The need to preserve, actively promote the use of, and effectively distribute LR, has caused the USA and EU authorities to put in place, respectively, LDC (the Linguistic Data Consortium) and ELRA (the European Language Resources Association)", observing also that their activities "demand regular updating to reflect technical and strategical evolution of their environment". We try to keep with this recommendation.

These innovations – introduced by ELRA and /or LREC – must not be seen as unrelated steps, but as part of a coherent vision, promoting a new culture in our community. We want to encourage also in the field of LT and LRs what is in use in more mature sciences and ensure reproducibility as a normal part of scientific practice. We try thus to influence how our science is organised or should be organised in the future.

I give here a quick picture of some innovations that are critical for the research process and constitute a sort of manifesto for a new kind of sustainability plan around LRs.

LRE Map

The LRE Map (http://www.resourcebook.eu/), started in 2010, is now an established tool, consulted every day and used in other major conferences. At this LREC we have collected by the authors descriptions for more than 1000 resources in more than 150 languages!

Spreading the LR documentation effort across many people, instead of leaving it only in the hands of the LR distribution centres, we also encourage awareness of the importance of metadata and proper documentation. Documenting a LR is the first step towards identifiability, which in its turn is the first step towards reproducibility.

Recognising the value of Linked Data, we just published the LRE Map in LOD (Linked Open Data).

Share your Language Resources and Reproducibility of research results: the vision

After encouraging sharing LR metadata, the next step is sharing the actual content. ELRA has embraced in the last years the notion of "open LRs": we show this also with the "Share your Language Resources" initiative started in this LREC. With it we ask all the authors to consider making background data available with their paper. More than 300 LRs have been made available: a big success for the first experiment! Showing the community commitment to sharing.

LRE Map and Share your LRs must be seen not as isolated initiatives, but as complementary steps towards implementing a new vision of the field. On one side we encourage opening data that could be valuable to others, on the other we try to encourage a sort of cultural change in our community.

Here the *vision*: It must become common practice also in our field that in conferences and journals when you submit a paper you are offered the opportunity to upload the LRs related to your paper. We must unlock the material that lies behind the papers: the adoption of such a policy will make the whole picture clearer. We had to fight in the '90s for concepts like "reusability", which finally led to promoting the need of developing standards in our field (this was still a hot topic in 1998 at the time of the 1st LREC). Now the need for standards is consolidated and we consider it normal, but we need to start another campaign for encouraging more resource sharing. Researchers are not yet sharing very well; they tend to hold back knowledge. I hope that this sharing trend will be more easily embraced by younger colleagues who are familiar with everyday use of social media of all sorts and free ideas sharing: we must port the same attitude in the research environment. This will fundamentally change the way of making science, in a sort of light revolution towards openness of science in all its facets. Hopefully it will diminish the unfortunate phenomenon of reinventing the wheel from time to time, instead of building on your colleagues' findings.

This vision has to do with many important aspects: shifting to a culture of sharing, re-use, reproducibility of research results. If we want to become a mature science we should make data sharing become "normal" practice. Even more important in a data-intensive discipline like LT. The small cost that each of us will pay to document, share, etc. should be paid back benefiting of others' efforts and become worthwhile. This will also lead to a greater opportunity of collaboration, encouraging bigger experiments by larger collaborative teams (something else we should learn from more mature sciences). Moreover, reproducibility encourages trust.

ISLRN

A major achievement of ELRA has been the recent establishment of the *International Standard Language Resource Number* (ISLRN) (http://www.elra.info/Establishing-the-ISLRN.html). It is a unique identifier to be assigned to each LR. Organised and sustained by ELRA, LDC and AFNLP/Oriental-COCOSDA, the ISLRN Portal provides unique identifiers to LRs. LRs in the ELRA

and LDC catalogues have been the first to get an ISLRN (just one if a LR is stored in both catalogues!).

When you publish a LR it can get an ISLRN and thus become a citable product of research.
Data/LR citation must become normal scientific practice also in our field, as it is in others. To make a LR citable can then pave the way to the design of a sort of "impact factor" of LRs. This can become an important incentive for the field, so that researchers can get the credit they deserve also for the LRs they developed.

ISLRN is not only linked to the possibility of getting proper "recognition" for LR developers. It would also enhance experiment replicability, an essential feature of scientific work. It may thus become a very important advance in our field.

META-SHARE sustainability by ELRA
Through these initiatives we try to encourage community efforts towards: documentation of LRs, possibility of identification of LRs, LR sharing, making research results reproducible. There is a lot of buzz these days around these types of topics. As I said above, all these initiatives are closely related and must become integrated with each other.

For them to become common research practices these activities must be well organised and require good mechanisms behind to become possibly a set of related services on a common platform. Pooling together data from all the research described in conference and journal papers will obviously need an infrastructure for distributing research results and such a LR platform must be sustained.

The ELRA Board has decided to support the META-SHARE platform, but META-SHARE – as sustained by ELRA – must in turn be adapted to be able to support these types of initiatives and thus become also a platform for sharing reproducible research results. We must find ways to make these practices as easy as possible and rewarding for the researcher. META-SHARE – in ELRA view – should become also the obvious repository (recognised by the community) where all these types of actions are sustained and where all research results become available, discoverable, identifiable, and citable. ELRA is taking these steps to start enabling to keep track of connected research activities like papers and supporting underlying resources, in an all-inclusive way.

LREC Proceedings in Thomson Citation Index
A great recent achievement for ELRA and LREC has been the fact that the LREC 2010 and LREC 2012 Proceedings have been accepted for inclusion in CPCI (Thomson Reuters Conference Proceedings Citation Index). This is a significant achievement for LREC and it will provide all LREC authors with a deserved recognition. It is for us of great satisfaction, in particular for the benefit it can bring to young colleagues.

ELRA 18th anniversary and NLP12
Coordination is an important issue when infrastructural issues are at stake. None of the actions above can or should be conducted and tackled in isolation.

For this reason we – ELRA – organised, on the occasion of ELRA majority as its 18th anniversary, the first meeting of the major associations/organisations in the field of Language Resources and Technologies, Computational Linguistics, Spoken Language Processing, Big Data and Digital Humanities, the so-called NLP12 (http://www.elra.info/NLP12-Paris-Declaration.html). We started to discuss issues of common interest to coordinate some of the activities and we adopted some common resolutions, such as the encouragement of language resources and tools sharing and promotion of best practices for language resource citation in publications.

Together we should be able to take the necessary steps to better serve the field and the respective communities and to strengthen the bridges between various communities (e.g. Language Technology and Humanities).

ELRA for Open science

I am excited and proud that we – as ELRA and LREC – can contribute to such a (quiet) revolution towards shaping a new type of *open scientific information space* for the future of our field, the Language Resources and Technology future. I have always felt it is our duty to use the means that we have in our hands to try to shape the future of the field, and in this case to play a role in how to change scientific practice and have an impact on the overall scientific enterprise!

Trying to be always forward-looking and to act in a proactive way to serve the field, ELRA continues to be a community-aware association. I would like to work for it to become more also a community-driven association. We would like to discuss with all those who are interested about how to tackle the challenge of truly open research (which is more than open access!) so that we can take the necessary further steps to make this process more efficient, faster and more collaborative.

It is clear that in such a campaign for the cause of reproducibility and open science and for a proper system of attribution and citation – two closely related aspects– we must involve also funding agencies that should help in supporting the necessary policy actions. For sure we will involve in this initiative the NLP12 group. But I strongly believe that the most important change must come from the mind-set of researchers. This is where LREC can help, I hope …

The message that ELRA has for its community, the LREC community, is: We are here to help!

Acknowledgments

In this last part I wish to express my deepest gratitude to all those who made this LREC 2014 possible and hopefully successful.

I first thank the Program Committee members, not only for their dedication in the huge task of selecting the papers, but also for the constant involvement in the various aspects around LREC. A particular thanks goes to Jan Odijk, who has been so helpful in the preparation of the program. To Joseph Mariani for his always wise suggestions. And obviously to Khalid Choukri, who is in charge of so many aspects around LREC.

I thank ELRA and the ELRA Board: LREC is a major service from ELRA to all the community!

A very special thanks goes to Sara Goggi and Hélène Mazo, the two Chairs of the Organising Committee, for all the work they do with so much dedication and competence, and also the capacity to tackle the many big and small problems of such a large conference (not an easy task). They are the two pillars of LREC, without whose commitment for many months LREC would not happen. So much of LREC organisation is on their shoulders, and it is visible to all participants.

A particular expression of gratitude goes to the Local Committee, and especially to Eiríkur Rögnvaldsson (its Chair) and Sigrún Helgadóttir: they have worked with great commitment and enthusiasm for many months for the success of LREC always looking at the best solutions to the many local issues.

All my appreciation goes also to the distinguished members of the Local Advisory Board for their constant support.

Among the Icelanders I wish to mention Guðrún Magnúsdóttir, for a very simple reason: the idea of having LREC in Iceland came out during a lunch that the two of us had together in Berlin!

I express my gratitude to the Sponsors that believe in the importance of our conference, and have helped with financial support. I am grateful to the authorities, and all associations, organisations, companies that have supported LREC in various ways, for their important cooperation.

Furthermore, on behalf of the Program Committee, I praise our impressively large Scientific Committee. They did a wonderful job.

I thank the workshop and tutorial organisers, who complement LREC of so many interesting events.

A big thanks goes to all the LREC authors, who provide the “substance” to LREC, and give us such a broad picture of the field.

I finally thank the two institutions that have dedicated such a great effort to this LREC, as to the previous ones, i.e. ELDA in Paris and ILC-CNR in Pisa. Without their commitment LREC would not have been possible. The last, but not least, thanks are thus, in addition to Hélène Mazo and Sara Goggi, to all the others who have helped and will help during the conference: Victoria Arranz, Paola Baroni, Roberto Bartolini, Irene De Felice, Riccardo Del Gratta, Francesca Frontini, Ioanna Giannopoulou, Johann Gorlier, Olivier Hamon, Jérémy Leixa, Valerie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, Priscille Schneller. You will meet most of them during the conference.

I also hope that funding agencies will be impressed by the quality and quantity of initiatives in our sector that LREC displays, and by the fact that the field attracts all the best groups of R&D from all continents. The success of LREC for us actually means the success of the field of Language Resources and Evaluation.

And lastly, my final words of appreciation are for all the LREC 2014 participants. Now LREC is in your hands. You are the true protagonist of LREC; we have worked for you all and you will make this LREC great. I hope that you discover new paths, that you perceive the ferment and liveliness of the field, that you have fruitful conversations (conferences are useful also for this) and most of all that you profit of so many contacts to organise new exciting work and projects in the field of Language Resources and Evaluation … which you will show at the next LREC.

LREC is not exactly in a Mediterranean location this time, even if all the hot water around gives some Mediterranean flavour! But the tradition of holding LREC in wonderful locations continues, and Reykjavík is a perfect LREC location! I am sure you will like Reykjavík and the friendliness of Icelanders. And I hope that Reykjavík will appreciate the invasion of LRECers!
With all the Programme Committee, I welcome you at LREC 2014 in such a wonderful country as Iceland and wish you a fruitful Conference.

Enjoy LREC 2014 in Reykjavík!

Nicoletta Calzolari

Chair of the 9th International Conference on Language Resources and Evaluation and ELRA President

Message from ELRA Secretary General and ELDA Managing Director
Khalid Choukri

Welcome to this LREC 2014, the 9th edition of one of the major events in language sciences and technologies and the most visible service of ELRA to the community.

ELRA, the **European Language Resource Association**, is very proud to organize LREC 2014 under the auspices of **UNESCO** (the United Nations Educational, Scientific and Cultural Organization), through the patronage of Her Excellency Madame Irina Bokova, UNESCO's Director General, and of Madame Vigdís Finnbogadóttir, former President of the Republic of Iceland and UNESCO Goodwill Ambassador for Languages.

I would like to express my heartfelt thanks to Her Excellencies Madame Irina Bokova and Madame Vigdís Finnbogadóttir for their patronage and support, assuring them of the community continuous efforts to address the common concerns and the crucial challenges, we all share.

It is an important symbol and a path for ELRA that strongly advocates for the preservation of languages, all languages, as major components of our cultures and efficient instruments for boosting education, literacy, and reducing the digital divide.

Welcome to Reykjavik, where you will certainly experience a true Mediterranean atmosphere, associated now with LREC, in the very North, standing in the middle between Europe and America. After having organized LREC in areas that identify themselves with largely spoken language families (Roman, Semitic, Turkic languages), we are heading to a country where the language played a special role in particular through the medieval Icelanders' sagas but also "preserving" itself over centuries as well as preserving the Old Norse spoken by the Vikings.

Organizing LREC under the patronage of UNESCO is an important symbol for ELRA that strives to stimulate the emergence of language technologies so they contribute to better education and easy access to our common knowledge, in all languages. Since its foundation, ELRA has been an active contributor, in particular shedding light on under-resourced languages. The first LREC, in 1998, already featured a workshop on "Minority Languages of Europe", a tradition that continues to date, going beyond the initial geographical and geopolitical coverages. Furthermore, several LRECs have seen the organization of specialized workshops and panels dedicated to educational applications. It is our credo to strive and encourage young generations to learn foreign languages, as many as one can handle. Learning foreign languages, including sign languages, is an extraordinary journey in other humans' cultures and traditions.

HLT (Human Language Technologies) should also support such endeavor and help underprivileged communities access the tremendous and wonderful human being world heritage, in particular UNESCO referenced ones. We hope that our community, through the backing of automated translation and other multilingual tools, improves such accessibility. Contributing to the efficiency of our translation and localization experts should support the cross-cultural fertility we all promote.

Dear ELRA Members, Dear LREC participants,

It is a great and renewed pleasure to address the LREC audience for the ninth time and share with you these thoughts and remarks. On the 28th of May 2014, we will also be remembering the first LREC that took place exactly 16 years ago, on May 28th 1998, and the visionaries who felt the need for such forum.

This 9th LREC is a special milestone as it gives the opportunity to celebrate LREC's 15th anniversary and ELRA's majority after 18th year of dedicated activities and services. It is an opportunity to review the activities carried out so far, draw some conclusions, and plans for the years to come. Some of these topics have been discussed at a workshop held on 19-20 November 2013 in Paris, and attended by representatives of the most distinguished organizations active in our field.

Allow me to take you 18 years back and walk together remembering the landscape as it was, at least on the European scene, from the Language Resources and Language Technology perspectives.

Just remember that the web was only in its infancy in 1995, when ELRA was established. The first reviews and surveys of existing resources in Europe were conducted in a set of projects funded by the European Commission. The field was split over three major domains, represented by clearly three different communities, associated to three big Language Resources categories: speech processing (spoken data), written text analysis (textual corpora and general lexica), and terminological resources (specialized dictionaries). The challenge for ELRA was to try and establish bridges between these different communities (hence LREC) but also capitalize on the findings of these projects to consolidate a catalogue of Language Resources (as stated in ELRA's foundation mission).

ELRA came out with its first catalogue of resources in 1996, a simple plain list comprising 30 resources. We were proud to publish such a catalogue (hardcopy) but we realized, with great humility, the huge task in front of us, we immediately understood that listing such resources could not serve the purpose for which ELRA has been set up: to ensure that LRs are used and re-used, possibly repackaged and repurposed. Users still had to negotiate themselves with the right holders, often located in other countries and different legal systems.

Such a mission required understanding the rationales behind data production and inventing new economic models, different from the ones in use including by the other data centers. A major dimension to be understood and managed was the legal issues behind ownerships, copyright and other associated rights. We had to address such issues and clear all legal aspects so that a user could access the LRs through an easy licensing schema. The mission of ELRA shifted from an archiving house of EU-funded project outcomes to a true distribution agency. For the next decade, we consolidated our identification activity and ensured that a large number of resources were catalogued and made available by ELRA to the community at large under fair conditions and easy licensing. ELRA acted as the EU instrument in distributing all LRs that were co-funded by the EC within its R&D frameworks.
We had the feeling that we were moving from scarcity to an organized framework that would help the community access an abundance of LRs.

Acting truly for multilingualism, we had to get accustomed to negotiating and clearing rights in multiple legal systems. The role of ELRA became even more crucial when users realized they could sign a single agreement to license multiple resources provided by a large number of suppliers, from all over the world.

We were (and still are) under no illusion about how good our coverage was. Through our market analysis and surveys, we knew that less than 20% of existing resources were publicly traded, the 80% were not released and not exchanged even when the right holders were public entities funded by tax payers (a few percentages were privately sub-licensed).

On the other hand, the surveys and inquiries received by our helpdesk (that is still in operation) clearly indicated that many needs were not fulfilled at all despite our supply.

To help people disclose what they had in their archives but also get tribute and scientific recognition for the work done to produce LRs and conducted evaluations, ELRA initiated, in1998, this conference:

the Language Resources and Evaluation Conference (LREC), a forum that aimed at bringing together all interested parties. With over 1200 attendees for the last editions (including over 30% of student and young researchers), LREC became one of the major events in our field. LREC focusses on all issues related to LRs and Evaluation of HLTs. It also gives room to specialized events that run as satellite workshops/tutorials to the conference.

ELRA viewed LRECs as important channels to discover existing Language Resources on which the community works but also to help identify gaps and trends. As such, LREC helps consolidate the community while drawing a clear picture of the state of affairs. The paper about "Rediscovering 15 Years of Discoveries in Language Resources and Evaluation" by Joseph Mariani (ELRA former president and current Honorary President) et. al. reviews some of these findings through an analysis of the papers published in the LREC proceedings over the 15 past years.

ELRA also designed LREC to become one of the best places to meet friends and colleagues, to share ideas and visions, and to plan for new collaborations, proposals and projects. As such LREC also contributed to the community building, an essential part of ELRA's mission. As a supplementary contribution, ELRA endorsed the publication of the **Language Resources and Evaluation Journal**[1] by Springer, on the very same topic.

Inspired from the discussions that took place at LREC, ELRA launched its project called "Universal Catalogue" (UC), with the aim to make it an inventory of all existing LRs within our field, either identified by the ELRA team or through input by the community. The UC comprises LR descriptions, independently of whether such resources would be made available or not. The underlying idea was and is to prioritize ELRA's negotiations, taking into account the requests of our members but also help potential users discover existing material before starting heavy production processes and hopefully negotiate directly with the right holders.

While maintaining our efforts devoted to the Universal Catalogue, ELRA took advantage of LREC to establish the LRE Map (Language Resources and Evaluation Map, http://www.resourcebook.eu/): a resource book that associates scientific publications to descriptions of LR and/or tools. LRE Map, an integrated component of the LREC submission system, requires from all LREC contributors to fill in a simple description of the LRs or the Language Tools (LT) mentioned in their submissions. By doing so, ELRA initiated a community-based bottom-up process that helps describe Language Resources (over 4000 unique LRs so far), consolidating the area of language resources. We are very grateful to the other conferences that adopted the LRE Map to collect more data on the existing Language Resources. Over time such "live" inventory of resources and tools, associated with scientific publications, will constitute a very useful knowledge base for the benefit of the community.

A critical issue that we learnt from the cataloguing and distribution activities is the difficulty to associate a unique name with a given LR. We realized that, despite our efforts and those of other data centers, referencing the LR used and/or described in scientific publications is very fuzzy and we see a large variety of names used for the same resources, even by the same author. This inconsistency could not be prevented even by data centers that could and did enforce the use of their identifiers, as part of the licensing agreement (i.e. ELRA)!

It is one of my deepest regrets that the community missed out a great opportunity to set up its own persistent identification system to name the LRs we are handling. The major instrument could have been the DOI system if we did come to a consensus to have one DOI assigner. It is probably too late as many centers and LR owners became DOI assigners and each can assign a different DOI to a LR.

To overcome such issue, the major organizations behind distribution and sharing of Language Resources, decided to introduce an identifier that is independent from Internet (and hence from DOIs), independent from the right-owners as well as from distribution agencies. This was inspired from the publishing community that adopted the ISBN schema, almost half a century ago. Such identifier, referred to as ISLRN, International Standard Language Resource Number (www.islrn.org), will allow

[1] LRE Journal, http://link.springer.com/journal/10579

a unique identification of a resource, independently from where it is stored, whether it is available or not, which licenses it is associated with, etc. ELRA, LDC[2], and AFNLP[3]/O-COCOSDA[4], committed to establish, run and moderate the ISLRN server at no charge for the community. The initiative will be steered by an international committee consisting of representatives of the major players from the NLP12 group. ELRA, LDC, and AFNLP/O-COCOSDA, in partnership with the major organizations within the field, would like to ease the citation of Language Resources and hence better assess the impact factor of each resource (the NLP12 Paris declaration is available at: http://www.elra.info/NLP12-Paris-Declaration.html).

It is clear from the setting up of ISLRN that it does not prevent data centers and resource right holders from using whatever local identifiers including DOI to refer to their resources but it will be more efficient if such identifiers are used in addition to ISLRN. The ELRA Board is discussing how to enforce such an identifier, making it compulsory for all publications at LREC and LRE Journal.

As mentioned above, ELRA celebrated its 18th anniversary on November 18-19-20, 2013, through a workshop and the NLP12 meeting (http://www.elra.info/ELRA-18th-Anniversary.html). The meeting was an excellent opportunity to gather several influential representatives of the community and discuss several pending hot topics that require more coordination and harmonization. In addition to the identification of LRs, including the endorsement of ISLRN proposal, the participants felt a strong need to harmonize the organization of their conferences and later on with those of neighboring domains. We have seen recently many important events running into each other with conflicting plans such as very close deadline dates for Call For Papers, similar dates for submission of abstracts or final manuscripts, similar milestones for the review process, etc. Given that most of the work is freely carried out by the peers (review scheduling and conduct, paper selection, program design, proceeding preparation, etc.), a conflicting planning demanded more efforts to those who had to juggle with more than one event, if they had to submit a final paper, an abstract on some new research, while reviewing other authors' papers, while continuing their usual work!

To avoid this situation, the NLP12 representatives agreed to develop an internal tool that would help the organizers view their plans while visually reviewing other events' planning, and getting some warnings and alarms. It is clear that, given the number of annual events, such conflicts are impossible to resolve, but at least some of the negative effects could be better handled.
Again, this should help better consolidate the activities of the community, improve synergies, and save some efforts.

New initiatives, European Commission debates on Licensing and Copyright
Regarding the licensing activities, ELRA took part to a large stakeholder dialogue in 2013/2014 organized by the EC about "Licences for Europe". ELRA contributed to the activities of a Working Group on "Text and Data Mining". The WG participants represented most of the parties involved in Data/Text mining both from the supply side (providers of data such as publishers, broadcasters, collective management of copyright and related rights organizations, etc.) as well as the demand side (Librarians, archivists, research centers, technology developers, etc.). ELRA, as a representative of the Human Language Technology developers, both from research and industry, brought in its knowledge of the community concerns and expectations. ELRA highlighted the importance of accessing substantial amounts of data to develop and assess performances of new NLP technologies that are the basis of most of today's search and mining applications. More details: http://ec.europa.eu/licences-for-europe-dialogue/en/content/about-site.

In addition to expressing the requirements and expectations of our community, emphasizing the new trends for free and open resources, ELRA advocated for an intermediate solution based on simplifying the access to copyrighted material for research purposes. ELRA argued that the solution for a competitive Europe requires a revision of the copyright regulations, to adopt a clear rule on the fair use

[2] Linguistic Data Consortium (LDC), https://www.ldc.upenn.edu/
[3] AFNLP : Asian Federation of Natural Language Processing, http://www.afnlp.org/
[4] O-Cocosda, see the 2014 meeting announcement at http://saki.siit.tu.ac.th/ococosda2014/

for research purposes of copyrighted language resources. Further to these WG meetings, the EU invited organizations to express their views on the necessary copyright amendments, which ELRA did along these lines. We are looking forward to hearing of the next steps. The contributions are listed at: http://ec.europa.eu/internal_market/consultations/2013/copyright-rules/index_en.htm.

Despite all these consolidation actions, we have also seen a fragmentation of our field. The last few years have seen an extraordinary development of the web (and more globally of the Internet). The culture of open source and free resources shifted from a fashion phenomenon to a strong and a lasting social and economic best practice. Such expansion has encouraged many institutions to establish their own repositories and offer their resources via internal infrastructures.

This trend definitely increases the availability of LRs (particular with the adoption of free/open sources spirit and licenses like Creative Commons) but renders their discoverability more tedious and their identification more complicated. In Europe, it has become affordable, from all points of view, to set up a LR repository (see details at the ELRA helpdesk at this conference) even if many institutions still rely on staff's personal pages to host resources and disseminate the corresponding information. With almost 30 different and independent entry points, META-SHARE is certainly the most sophisticated example of a distributed and networked repository set, with repositories listing as few as 5 resources and others i.e. ELRA with over a thousand. It is still a challenge to bring down the number of different applicable licenses (over 30 now) to the dozen prescribed by META-SHARE and inspired by ELRA and the Creative Commons spirits. Such a network should prevent profusion of unlinked/unrelated repositories.

Such "paradigm" shift boosted the sharing of language resources and tools while impacting the distribution mechanisms. To keep a proactive role with respect to its mission, ELRA has anticipated some of these changes and new tasks (e-commerce meta-share repository, ISLRN assigned to all its resources, e-licensing and e-signature, a LR forum, etc.) are in an advanced stage and announcements of these novelties under preparation.

To support this consolidation requirement and vital need, ELRA is involved in a new EU funded project called MLi (European Multilingual data & services Infrastructure). As a EU support action, MLi is working to deliver the strategic vision and operational specifications needed for building a comprehensive European Multilingual data & services Infrastructure, along with a multiannual plan for its development and deployment, and foster multi-stakeholders alliances ensuring its long term sustainability. We hope to share these visions with the LREC participants on the ELRA and MLi booth at the HLT Project Village that features exhibition booths for many EU projects, at this conference.

Acknowledgments
Finally, I would like to express my deep thanks to our partners and supporters, who throughout the years make LREC so successful.

I would like to thank our Silver Sponsor Holmes Semantic Solutions, and our Bronze sponsors: EML (European Media Laboratory GmbH), IMMI, VoiceBox and K-dictionaries. I also would like to thank the HLT Village participants, we hope that such gathering will offer the projects an opportunity to foster their dissemination and hopefully discuss exploitation plans with the attendees.
I would like to thank the impressive local advisory committee. Its composition of the most distinguished personalities of Iceland denotes the importance of language and language technologies for the country. We do hope that it is a strong sign for the long term commitment of the Icelandic officials.

I would like to thank the LREC Local Committee, chaired by Professor Eiríkur Rögnvaldsson who helped us with all logistic issues, herein in Iceland and Gudrun Magnusdottir, who introduced us to Iceland and for her continuous support.

Finally I would like to warmly thank the joint team of the two institutions that devote so much effort over months and often behind curtains to make this one week memorable: ILC-CNR in Pisa and my own team, ELDA, in Paris. These are the two LREC coordinators and pillars: Sara Gogi and Hélène Mazo and the team: Victoria Arranz, Paola Baroni, Roberto Bartolini, Irene De Felice, Riccardo Del Gratta, Francesca Frontini, Ioanna Giannopoulou, Johann Gorlier, Olivier Hamon, Jérémy Leixa, Valerie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, and Priscille Schneller.

We were very happy, for this LREC, to enjoy the friendly support and efficient help of Sigrún Helgadóttir – Researcher at the Árni Magnússon Institute for Icelandic Studies, to whom I extend my warm thanks.

Now LREC 2014 is yours; we hope that each of you will achieve valuable results and accomplishments. We, ELRA and ILC-CNR staff, are at your disposal to help you get the best out of it.
Once again, welcome to Reykjavik, welcome to LREC 2014

Khalid Choukri
ELRA Secretary General and ELDA Managing Director

Message of the Chair of the Local Organizing Committee
Eiríkur Rögnvaldsson

Dear LREC 2014 Participants,

On behalf of the Local Organising Committee, the Local Advisory Board, and all the participating organisations I would first of all like to express my profound gratitude to UNESCO and to the Director-General Madam Irina Bokova for kindly agreeing to act as patron for the conference. I would also like to thank Madam Vigdís Finnbogadóttir, UNESCO Goodwill Ambassador for languages and former president of Iceland, for her invaluable support.

Iceland is a country with a rich literary heritage and Icelandic is known for having changed less in the course of the last thousand years than most other languages having a documented history. In the conference bag that you received upon arrival at the conference venue you will find a small booklet, a gift from the local organisers. The title of the booklet is *Hávamál – the Sayings of Óðinn*. It contains an English translation of the famous poem Hávamál which is more than one thousand years old. Icelanders today can still read and understand the text of Hávamál as it is preserved in the 13th century manuscript Codex regius – there is no need for translation. We want to keep it that way.

However, we now feel that Icelandic is being threatened by globalisation and modern technology. Our precious language, which the manuscripts have preserved for us for almost thousand years, might be lost in only a few decades, if it does not adapt to current and future information technology. Icelandic language technology is still in its infancy, and the META-NET survey of 30 European languages demonstrated that language technology support for Icelandic is very limited. Therefore, it is both symbolic and extremely important for us to have the LREC conference here in Iceland – the first time the conference is held outside of the Mediterranean area. It will hopefully raise awareness of the importance of language technology, among politicians, policy makers, academics, journalists, and not least the general public.

I would like to thank all who have made it possible to have the conference here in Iceland – the ELRA team in Paris and Pisa, the Programme Committee and all the reviewers, the Local Committees, our conference bureau Congress Reykjavík, the Ministry of Education, Culture and Science, and others who have supported the conference in various ways. I cannot mention all the people involved but I want to single out two names. Guðrún Magnúsdóttir first mentioned to us the idea of having LREC in Iceland, and she convinced the ELRA team that this was possible – and feasible. Sigrún Helgadóttir has been the secretary of the Local Organising Committee and has done a tremendous job in organising and coordinating innumerable things that had to be taken care of.

Let me conclude by welcoming you all to Iceland, to Reykjavík, UNESCO city of literature, and to the Harpa Concert Hall and Conference Centre, winner of the 2013 European Union Prize for Contemporary Architecture – Mies van der Rohe Award. I sincerely hope you will enjoy LREC 2014, and that the conference and your visit as a whole will be an unforgettable experience.

Eiríkur Rögnvaldsson
Chair of the Local Organising Committee

Table of Contents

O1 - NLP Workflows

Wednesday, May 28, 11:35

Chairperson: **Stephanie Strassel** Oral Session

The Alveo Virtual Laboratory: A Web-based Repository API

Steve Cassidy, Dominique Estival, Timothy Jones, Denis Burnham and Jared Burghold

The Human Communication Science Virtual Laboratory (HCS vLab) is an eResearch project funded under the Australian Government NeCTAR program to build a platform for collaborative eResearch around data representing human communication and the tools that researchers use in their analysis. The human communication science field is broadly defined to encompass the study of language from various perspectives but also includes research on music and various other forms of human expression. This paper outlines the core architecture of the HCS vLab and in particular, highlights the web-based API that provides access to data and tools to authenticated users.

A Stream Computing Approach Towards Scalable NLP

Xabier Artola, Zuhaitz Beloki and Aitor Soroa

Computational power needs have grown dramatically in recent years. This is also the case in many language processing tasks, due to overwhelming quantities of textual information that must be processed in a reasonable time frame. This scenario has led to a paradigm shift in the computing architectures and large-scale data processing strategies used in the NLP field. In this paper we describe a series of experiments carried out in the context of the NewsReader project with the goal of analyzing the scaling capabilities of the language processing pipeline used in it. We explore the use of Storm in a new approach for scalable distributed language processing across multiple machines and evaluate its effectiveness and efficiency when processing documents on a medium and large scale. The experiments have shown that there is a big room for improvement regarding language processing performance when adopting parallel architectures, and that we might expect even better results with the use of large clusters with many processing nodes.

ILLINOISCLOUDNLP: Text Analytics Services in the Cloud

Hao Wu, Zhiye Fei, Aaron Dai, Mark Sammons, Dan Roth and Stephen Mayhew

Natural Language Processing (NLP) continues to grow in popularity in a range of research and commercial applications. However, installing, maintaining, and running NLP tools can be time consuming, and many commercial and research end users have only intermittent need for large processing capacity. This paper describes ILLINOISCLOUDNLP, an on-demand framework built around NLPCURATOR and Amazon Web Services' Elastic Compute Cloud (EC2). This framework provides a simple interface to end users via which they can deploy one or more NLPCURATOR instances on EC2, upload plain text documents, specify a set of Text Analytics tools (NLP annotations) to apply, and process and store or download the processed data. It can also allow end users to use a model trained on their own data: ILLINOISCLOUDNLP takes care of training, hosting, and applying it to new data just as it does with existing models within NLPCURATOR. As a representative use case, we describe our use of ILLINOISCLOUDNLP to process 3.05 million documents used in the 2012 and 2013 Text Analysis Conference Knowledge Base Population tasks at a relatively deep level of processing, in approximately 20 hours, at an approximate cost of US$500; this is about 20 times faster than doing so on a single server and requires no human supervision and no NLP or Machine Learning expertise.

The Language Application Grid

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, di Wang, Keith Suderman, Marc Verhagen and Jonathan Wright

The Language Application (LAPPS) Grid project is establishing a framework that enables language service discovery, composition, and reuse and promotes sustainability, manageability, usability, and interoperability of natural language Processing (NLP) components. It is based on the service-oriented architecture (SOA), a more recent, web-oriented version of the "pipeline" architecture that has long been used in NLP for sequencing loosely-coupled linguistic analyses. The LAPPS Grid provides access to basic NLP processing tools and resources and enables pipelining such tools to create custom NLP applications, as well as composite services such as question answering and machine translation together with language resources such as mono- and multi-lingual corpora and lexicons that support NLP. The transformative aspect of the LAPPS Grid is that it orchestrates access to and deployment of language resources and processing functions available from servers around the globe and enables users to add their own language resources, services, and even service grids to satisfy their particular needs.

Praaline: Integrating Tools for Speech Corpus Research

George Christodoulides

This paper presents Praaline, an open-source software system for managing, annotating, analysing and visualising speech corpora. Researchers working with speech corpora are often faced with multiple tools and formats, and they need to work with ever-increasing amounts of data in a collaborative way. Praaline integrates and extends existing time-proven tools for spoken corpora analysis (Praat, Sonic Visualiser and a bridge to the R statistical package) in a modular system, facilitating automation and reuse. Users are exposed to an integrated, user-friendly interface from which to access multiple tools. Corpus metadata and annotations may be stored in a database, locally or remotely, and users can define the metadata and annotation structure. Users may run a customisable cascade of analysis steps, based on plug-ins and scripts, and update the database with the results. The corpus database may be queried, to produce aggregated data-sets. Praaline is extensible using Python or C++ plug-ins, while Praat and R scripts may be executed against the corpus data. A series of visualisations, editors and plug-ins are provided. Praaline is free software, released under the GPL license (www.praaline.org).

O2 - Machine Translation and Evaluation (1)

Wednesday, May 28, 11:35

Chairperson: **Bente Maegaard** Oral Session

Linguistic Evaluation of Support Verb Constructions by OpenLogos and Google Translate

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Susanne Preuß, Kutz Arrieta, Wang Ling, Fernando Batista and Isabel Trancoso

This paper presents a systematic human evaluation of translations of English support verb constructions produced by a rule-based machine translation (RBMT) system (OpenLogos) and a statistical machine translation (SMT) system (Google Translate) for five languages: French, German, Italian, Portuguese and Spanish. We classify support verb constructions by means of their syntactic structure and semantic behavior and present a qualitative analysis of their translation errors. The study aims to verify how machine translation (MT) systems translate fine-grained linguistic phenomena, and how well-equipped they are to produce high-quality translation. Another goal of the linguistically motivated quality analysis of SVC raw output is to reinforce the need for better system hybridization, which leverages the strengths of RBMT to the benefit of SMT, especially in improving the translation of multiword units. Taking multiword units into account, we propose an effective method to achieve MT hybridization based on the integration of semantico-syntactic knowledge into SMT.

MTWatch: A Tool for the Analysis of Noisy Parallel Data

Sandipan Dandapat and Declan Groves

State-of-the-art statistical machine translation (SMT) technique requires a good quality parallel data to build a translation model. The availability of large parallel corpora has rapidly increased over the past decade. However, often these newly developed parallel data contains contain significant noise. In this paper, we describe our approach for classifying good quality parallel sentence pairs from noisy parallel data. We use 10 different features within a Support Vector Machine (SVM)-based model for our classification task. We report a reasonably good classification accuracy and its positive effect on overall MT accuracy.

Machine Translation for Subtitling: A Large-Scale Evaluation

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maucec, Anja Turner and Martin Volk

This article describes a large-scale evaluation of the use of Statistical Machine Translation for professional subtitling. The work was carried out within the FP7 EU-funded project SUMAT and involved two rounds of evaluation: a quality evaluation and a measure of productivity gain/loss. We present the SMT systems built for the project and the corpora they were trained on, which combine professionally created and crowd-sourced data. Evaluation goals, methodology and results are presented for the eleven translation pairs that were evaluated by professional subtitlers. Overall, a majority of the machine translated subtitles received good quality ratings. The results were also positive in terms of productivity, with a global gain approaching 40%. We also evaluated the impact of applying quality estimation and filtering of poor MT output, which resulted in higher productivity gains for filtered files as opposed to fully machine-translated files. Finally, we present and discuss feedback from the subtitlers who participated in the evaluation, a key aspect for any eventual adoption of machine translation technology in professional subtitling.

Machine Translationness: Machine-likeness in Machine Translation Evaluation

Joaquim Moré and Salvador Climent

Machine translationness (MTness) is the linguistic phenomena that make machine translations distinguishable from human translations. This paper intends to present MTness as a research object and suggests an MT evaluation method based on determining whether the translation is machine-like instead of determining its human-likeness as in evaluation current approaches. The method rates the MTness of a translation with a metric, the MTS (Machine Translationness Score). The MTS calculation is in accordance with the results of an experimental study on machine translation perception by common people. MTS proved to correlate well with human ratings on translation quality. Besides, our approach allows the performance of cheap evaluations since expensive resources (e.g. reference translations, training corpora) are not needed. The paper points out the challenge of dealing with MTness as an everyday phenomenon caused by the massive use of MT.

On the Origin of Errors: a Fine-Grained Analysis of MT and PE Errors and their Relationship

Joke Daems, Lieve Macken and Sonia Vandepitte

In order to improve the symbiosis between machine translation (MT) system and post-editor, it is not enough to know that the output of one system is better than the output of another system. A fine-grained error analysis is needed to provide information on the type and location of errors occurring in MT and the corresponding errors occurring after post-editing (PE). This article reports on a fine-grained translation quality assessment approach which was applied to machine translated-texts and the post-edited versions of these texts, made by student post-editors. By linking each error to the corresponding source text-passage, it is possible to identify passages that were problematic in MT, but not after PE, or passages that were problematic even after PE. This method provides rich data on the origin and impact of errors, which can be used to improve post-editor training as well as machine translation systems. We present the results of a pilot experiment on the post-editing of newspaper articles and highlight the advantages of our approach.

O3 - Grammar and Parsing (1)

Wednesday, May 28, 11:35

Chairperson: **Emily M. Bender** Oral Session

Parsing Chinese Synthetic Words with a Character-based Dependency Model

Fei Cheng, Kevin Duh and Yuji Matsumoto

Synthetic word analysis is a potentially important but relatively unexplored problem in Chinese natural language processing. Two issues with the conventional pipeline methods involving word segmentation are (1) the lack of a common segmentation standard and (2) the poor segmentation performance on OOV words. These issues may be circumvented if we adopt the view of character-based parsing, providing both internal structures to synthetic words and global structure to sentences in a seamless fashion. However, the accuracy of synthetic word parsing is not yet satisfactory, due to the lack of research. In view of this, we propose and present experiments on several synthetic word parsers. Additionally, we demonstrate the usefulness of incorporating large unlabelled corpora and a dictionary for this task. Our parsers significantly outperform the baseline (a pipeline method).

Improvements to Dependency Parsing Using Automatic Simplification of Data

Tomáš Jelínek

In dependency parsing, much effort is devoted to the development of new methods of language modeling and better feature settings. Less attention is paid to actual linguistic data and how appropriate they are for automatic parsing: linguistic data can be too complex for a given parser, morphological tags may not reflect well syntactic properties of words, a detailed, complex annotation scheme may be ill suited for automatic parsing. In this paper, I present a study of this problem on the following case: automatic dependency parsing using the data of the Prague Dependency Treebank with two dependency parsers: MSTParser and MaltParser. I show that by means of small, reversible simplifications of the text and of the annotation, a considerable improvement of parsing accuracy can be achieved. In order to facilitate the task of language modeling performed by the parser, I reduce variability of lemmas and forms in the text. I modify the system of morphological annotation to adapt it better for parsing. Finally, the dependency annotation scheme is also partially modified. All such modifications are automatic and fully reversible: after the parsing is done, the original data and

structures are automatically restored. With MaltParser, I achieve an 8.3% error rate reduction.

All Fragments Count in Parser Evaluation

Joost Bastings and Khalil Sima'an

PARSEVAL, the default paradigm for evaluating constituency parsers, calculates parsing success (Precision/Recall) as a function of the number of matching labeled brackets across the test set. Nodes in constituency trees, however, are connected together to reflect important linguistic relations such as predicate-argument and direct-dominance relations between categories. In this paper, we present FREVAL, a generalization of PARSEVAL, where the precision and recall are calculated not only for individual brackets, but also for co-occurring, connected brackets (i.e. fragments). FREVAL fragments precision (FLP) and recall (FLR) interpolate the match across the whole spectrum of fragment sizes ranging from those consisting of individual nodes (labeled brackets) to those consisting of full parse trees. We provide evidence that FREVAL is informative for inspecting relative parser performance by comparing a range of existing parsers.

Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies

Maria Simi, Cristina Bosco and Simonetta Montemagni

Stanford Dependencies (SD) represent nowadays a de facto standard as far as dependency annotation is concerned. The goal of this paper is to explore pros and cons of different strategies for generating SD annotated Italian texts to enrich the existing Italian Stanford Dependency Treebank (ISDT). This is done by comparing the performance of a statistical parser (DeSR) trained on a simpler resource (the augmented version of the Merged Italian Dependency Treebank or MIDT+) and whose output was automatically converted to SD, with the results of the parser directly trained on ISDT. Experiments carried out to test reliability and effectiveness of the two strategies show that the performance of a parser trained on the reduced dependencies repertoire, whose output can be easily converted to SD, is slightly higher than the performance of a parser directly trained on ISDT. A non-negligible advantage of the first strategy for generating SD annotated texts is that semi-automatic extensions of the training resource are more easily and consistently carried out with respect to a reduced dependency tag set. Preliminary experiments carried out for generating the collapsed and propagated SD representation are also reported.

Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian

Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson and Joel C. Wallenberg

This paper presents ongoing work that aims to improve machine parsing of Faroese using a combination of Faroese and Icelandic training data. We show that even if we only have a relatively small parsed corpus of one language, namely 53,000 words of Faroese, we can obtain better results by adding information about phrase structure from a closely related language which has a similar syntax. Our experiment uses the Berkeley parser. We demonstrate that the addition of Icelandic data without any other modification to the experimental setup results in an f-measure improvement from 75.44% to 78.05% in Faroese and an improvement in part-of-speech tagging accuracy from 88.86% to 90.40%.

O4 - Information Extraction and Knowledge Discovery

Wednesday, May 28, 11:35

Chairperson: **Ricardo Baeza-Yates** Oral Session

Boosting Open Information Extraction with Noun-based Relations

Clarissa Xavier and Vera Lima

Open Information Extraction (Open IE) is a strategy for learning relations from texts, regardless the domain and without predefining these relations. Work in this area has focused mainly on verbal relations. In order to extend Open IE to extract relationships that are not expressed by verbs, we present a novel Open IE approach that extracts relations expressed in noun compounds (NCs), such as (oil, extracted from, olive) from "olive oil", or in adjective-noun pairs (ANs), such as (moon, that is, gorgeous) from "gorgeous moon". The approach consists of three steps: detection of NCs and ANs, interpretation of these compounds in view of corpus enrichment and extraction of relations from the enriched corpus. To confirm the feasibility of this method we created a prototype and evaluated the impact of the application of our proposal in two state-of-the-art Open IE extractors. Based on these tests we conclude that the proposed approach is an important step to fulfil the gap concerning the extraction of relations within the noun compounds and adjective-noun pairs in Open IE.

Clinical Data-Driven Probabilistic Graph Processing

Travis Goodwin and Sanda Harabagiu

Electronic Medical Records (EMRs) encode an extraordinary amount of medical knowledge. Collecting and interpreting this knowledge, however, belies a significant level of clinical understanding. Automatically capturing the clinical information is crucial for performing comparative effectiveness research. In this paper, we present a data-driven approach to model semantic dependencies between medical concepts, qualified by the beliefs of physicians. The dependencies, captured in a patient cohort graph of clinical pictures and therapies is further refined into a probabilistic graphical model which enables efficient inference of patient-centered treatment or test recommendations (based on probabilities). To perform inference on the graphical model, we describe a technique of smoothing the conditional likelihood of medical concepts by their semantically-similar belief values. The experimental results, as compared against clinical guidelines are very promising.

A Framework for Compiling High Quality Knowledge Resources From Raw Corpora

Gongye Jin, Daisuke Kawahara and Sadao Kurohashi

The identification of various types of relations is a necessary step to allow computers to understand natural language text. In particular, the clarification of relations between predicates and their arguments is essential because predicate-argument structures convey most of the information in natural languages. To precisely capture these relations, wide-coverage knowledge resources are indispensable. Such knowledge resources can be derived from automatic parses of raw corpora, but unfortunately parsing still has not achieved a high enough performance for precise knowledge acquisition. We present a framework for compiling high quality knowledge resources from raw corpora. Our proposed framework selects high quality dependency relations from automatic parses and makes use of them for not only the calculation of fundamental distributional similarity but also the acquisition of knowledge such as case frames.

Thematic Cohesion: Measuring Terms Discriminatory Power Toward Themes

Clément de Groc, Xavier Tannier and Claude de Loupy

We present a new measure of thematic cohesion. This measure associates each term with a weight representing its discriminatory power toward a theme, this theme being itself expressed by a list of terms (a thematic lexicon). This thematic cohesion criterion can be used in many applications, such as query expansion, computer-assisted translation, or iterative construction of domain-specific lexicons and corpora. The measure is computed in two steps. First, a set of documents related to the terms is gathered from the Web by querying a Web search engine. Then, we produce an oriented co-occurrence graph, where vertices are the terms and edges represent the fact that two terms co-occur in a document. This graph can be interpreted as a recommendation graph, where two terms occurring in a same document means that they recommend each other. This leads to using a random walk algorithm that assigns a global importance value to each vertex of the graph. After observing the impact of various parameters on those importance values, we evaluate their correlation with retrieval effectiveness.

Extracting Information for Context-aware Meeting Preparation

Simon Scerri, Behrang Q. Zadeh, Maciej Dabrowski and Ismael Rivera

People working in an office environment suffer from large volumes of information that they need to manage and access. Frequently, the problem is due to machines not being able to recognise the many implicit relationships between office artefacts, and also due to them not being aware of the context surrounding them. In order to expose these relationships and enrich artefact context, text analytics can be employed over semi-structured and unstructured content, including free text. In this paper, we explain how this strategy is applied and partly evaluated for a specific use-case: supporting the attendees of a calendar event to prepare for the meeting.

P1 - Corpora and Annotation

Wednesday, May 28, 11:35

Chairperson: **Marko Tadić** Poster Session

TaLAPi – A Thai Linguistically Annotated Corpus for Language Processing

AiTi Aw, Sharifah Mahani Aljunied, Nattadaporn Lertcheva and Sasiwimon Kalunsima

This paper discusses a Thai corpus, TaLAPi, fully annotated with word segmentation (WS), part-of-speech (POS) and named entity (NE) information with the aim to provide a high-quality and sufficiently large corpus for real-life implementation of Thai language processing tools. The corpus contains 2,720 articles (1,043,471 words) from the entertainment and lifestyle (NE&L) domain and 5,489 articles (3,181,487 words) in the news (NEWS) domain, with a total of 35 POS tags and 10 named entity categories. In particular, we present an approach to segment and

tag foreign and loan words expressed in transliterated or original form in Thai text corpora. We see this as an area for study as adapted and un-adapted foreign language sequences have not been well addressed in the literature and this poses a challenge to the annotation process due to the increasing use and adoption of foreign words in the Thai language nowadays. To reduce the ambiguities in POS tagging and to provide rich information for facilitating Thai syntactic analysis, we adapted the POS tags used in ORCHID and propose a framework to tag Thai text and also addresses the tagging of loan and foreign words based on the proposed segmentation strategy. TaLAPi also includes a detailed guideline for tagging the 10 named entity categories

Variations on Quantitative Comparability Measures and their Evaluations on Synthetic French-English Comparable Corpora

Guiyao Ke, Pierre-Francois Marteau and Gildas Menier

Following the pioneering work by [**?**], we address in this paper the analysis of a family of quantitative comparability measures dedicated to the construction and evaluation of topical comparable corpora. After recalling the definition of the quantitative comparability measure proposed by [**?**], we develop some variants of this measure based primarily on the consideration that the occurrence frequencies of lexical entries and the number of their translations are important. We compare the respective advantages and disadvantages of these variants in the context of an evaluation framework that is based on the progressive degradation of the Europarl parallel corpus. The degradation is obtained by replacing either deterministically or randomly a varying amount of lines in blocks that compose partitions of the initial Europarl corpus. The impact of the coverage of bilingual dictionaries on these measures is also discussed and perspectives are finally presented.

Using Transfer Learning to Assist Exploratory Corpus Annotation

Paul Felt, Eric Ringger, Kevin Seppi and Kristian Heal

We describe an under-studied problem in language resource management: that of providing automatic assistance to annotators working in exploratory settings. When no satisfactory tagset already exists, such as in under-resourced or undocumented languages, it must be developed iteratively while annotating data. This process naturally gives rise to a sequence of datasets, each annotated differently. We argue that this problem is best regarded as a transfer learning problem with multiple source tasks. Using part-of-speech tagging data with simulated exploratory tagsets, we demonstrate that even simple transfer learning techniques can significantly improve the quality of pre-annotations in an exploratory annotation.

Priberam Compressive Summarization Corpus: A New Multi-Document Summarization Corpus for European Portuguese

Miguel B. Almeida, Mariana S. C. Almeida, André F. T. Martins, Helena Figueira, Pedro Mendes and Cláudia Pinto

In this paper, we introduce the Priberam Compressive Summarization Corpus, a new multi-document summarization corpus for European Portuguese. The corpus follows the format of the summarization corpora for English in recent DUC and TAC conferences. It contains 80 manually chosen topics referring to events occurred between 2010 and 2013. Each topic contains 10 news stories from major Portuguese newspapers, radio and TV stations, along with two human generated summaries up to 100 words. Apart from the language, one important difference from the DUC/TAC setup is that the human summaries in our corpus are *compressive*: the annotators performed only sentence and word deletion operations, as opposed to generating summaries from scratch. We use this corpus to train and evaluate learning-based extractive and compressive summarization systems, providing an empirical comparison between these two approaches. The corpus is made freely available in order to facilitate research on automatic summarization.

Corpus and Evaluation of Handwriting Recognition of Historical Genealogical Records

Patrick Schone, Heath Nielson and Mark Ward

Over the last few decades, significant strides have been made in handwriting recognition (HR), which is the automatic transcription of handwritten documents. HR often focuses on modern handwritten material, but in the electronic age, the volume of handwritten material is rapidly declining. However, we believe HR is on the verge of having major application to historical record collections. In recent years, archives and genealogical organizations have conducted huge campaigns to transcribe valuable historical record content with such transcription being largely done through human-intensive labor. HR has the potential of revolutionizing these transcription endeavors. To test the hypothesis that this technology is close to applicability, and to provide a testbed for reducing any accuracy gaps, we have developed an evaluation paradigm for historical record handwriting recognition. We created a huge test corpus consisting of four historical data collections of four differing genres and three languages. In this paper, we provide the details of these extensive resources which we intend to release to the research

community for further study. Since several research organizations have already participated in this evaluation, we also show initial results and comparisons to human levels of performance.

The SYN-series Corpora of Written Czech

Milena Hnátková, Michal Křen, Pavel Procházka and Hana Skoumalová

The paper overviews the SYN series of synchronic corpora of written Czech compiled within the framework of the Czech National Corpus project. It describes their design and processing with a focus on the annotation, i.e. lemmatization and morphological tagging. The paper also introduces SYN2013PUB, a new 935-million newspaper corpus of Czech published in 2013 as the most recent addition to the SYN series before planned revision of its architecture. SYN2013PUB can be seen as a completion of the series in terms of titles and publication dates of major Czech newspapers that are now covered by complete volumes in comparable proportions. All SYN-series corpora can be characterized as traditional, with emphasis on cleared copyright issues, well-defined composition, reliable metadata and high-quality data processing; their overall size currently exceeds 2.2 billion running words.

Corpus of 19th-century Czech Texts: Problems and Solutions

Karel Kučera and Martin Stluka

Although the Czech language of the 19th century represents the roots of modern Czech and many features of the 20th- and 21st-century language cannot be properly understood without this historical background, the 19th-century Czech has not been thoroughly and consistently researched so far. The long-term project of a corpus of 19th-century Czech printed texts, currently in its third year, is intended to stimulate the research as well as to provide a firm material basis for it. The reason why, in our opinion, the project is worth mentioning is that it is faced with an unusual concentration of problems following mostly from the fact that the 19th century was arguably the most tumultuous period in the history of Czech, as well as from the fact that Czech is a highly inflectional language with a long history of sound changes, orthography reforms and rather discontinuous development of its vocabulary. The paper will briefly characterize the general background of the problems and present the reasoning behind the solutions that have been implemented in the ongoing project.

Extending Standoff Annotation

Maik Stührenberg

Textual information is sometimes accompanied by additional encodings (such as visuals). These multimodal documents may be interesting objects of investigation for linguistics. Another class of complex documents are pre-annotated documents. Classic XML inline annotation often fails for both document classes because of overlapping markup. However, standoff annotation, that is the separation of primary data and markup, is a valuable and common mechanism to annotate multiple hierarchies and/or read-only primary data. We demonstrate an extended version of the XStandoff meta markup language, that allows the definition of segments in spatial and pre-annotated primary data. Together with the ability to import already established (linguistic) serialization formats as annotation levels and layers in an XStandoff instance, we are able to annotate a variety of primary data files, including text, audio, still and moving images. Application scenarios that may benefit from using XStandoff are the analyzation of multimodal documents such as instruction manuals, or sports match analysis, or the less destructive cleaning of web pages.

Constructing and Exploiting an Automatically Annotated Resource of Legislative Texts

Stefan Höfler and Kyoko Sugisaki

In this paper, we report on the construction of a resource of Swiss legislative texts that is automatically annotated with structural, morphosyntactic and content-related information, and we discuss the exploitation of this resource for the purposes of legislative drafting, legal linguistics and translation and for the evaluation of legislation. Our resource is based on the classified compilation of Swiss federal legislation. All texts contained in the classified compilation exist in German, French and Italian, some of them are also available in Romansh and English. Our resource is currently being exploited (a) as a testing environment for developing methods of automated style checking for legislative drafts, (b) as the basis of a statistical multilingual word concordance, and (c) for the empirical evaluation of legislation. The paper describes the domain- and language-specific procedures that we have implemented to provide the automatic annotations needed for these applications.

P2 - Crowdsourcing

Wednesday, May 28, 11:35

Chairperson: **Alain Couillault** Poster Session

A Study on Expert Sourcing Enterprise Question Collection and Classification

Yuan Luo, Thomas Boucher, Tolga Oral, David Osofsky and Sara Weber

Large enterprises, such as IBM, accumulate petabytes of free-text data within their organizations. To mine this big data, a

critical ability is to enable meaningful question answering beyond keywords search. In this paper, we present a study on the characteristics and classification of IBM sales questions. The characteristics are analyzed both semantically and syntactically, from where a question classification guideline evolves. We adopted an enterprise level expert sourcing approach to gather questions, annotate questions based on the guideline and manage the quality of annotations via enhanced inter-annotator agreement analysis. We developed a question feature extraction system and experimented with rule-based, statistical and hybrid question classifiers. We share our annotated corpus of questions and report our experimental results. Statistical classifiers separately based on n-grams and hand-crafted rule features give reasonable macro-f1 scores at 61.7% and 63.1% respectively. Rule based classifier gives a macro-f1 at 77.1%. The hybrid classifier with n-gram and rule features using a second guess model further improves the macro-f1 to 83.9%.

Can the Crowd be Controlled?: A Case Study on Crowd Sourcing and Automatic Validation of Completed Tasks based on User Modeling

Balamurali A.R

Annotation is an essential step in the development cycle of many Natural Language Processing (NLP) systems. Lately, crowd-sourcing has been employed to facilitate large scale annotation at a reduced cost. Unfortunately, verifying the quality of the submitted annotations is a daunting task. Existing approaches address this problem either through sampling or redundancy. However, these approaches do have a cost associated with it. Based on the observation that a crowd-sourcing worker returns to do a task that he has done previously, a novel framework for automatic validation of crowd-sourced task is proposed in this paper. A case study based on sentiment analysis is presented to elucidate the framework and its feasibility. The result suggests that validation of the crowd-sourced task can be automated to a certain extent.

When Transliteration Met Crowdsourcing : An Empirical Study of Transliteration via Crowdsourcing using Efficient, Non-redundant and Fair Quality Control

Mitesh M. Khapra, Ananthakrishnan Ramanathan, Anoop Kunchukuttan, Karthik Visweswariah and Pushpak Bhattacharyya

Sufficient parallel transliteration pairs are needed for training state of the art transliteration engines. Given the cost involved, it is often infeasible to collect such data using experts. Crowdsourcing could be a cheaper alternative, provided that a good quality control (QC) mechanism can be devised for this task. Most QC mechanisms employed in crowdsourcing are aggressive (unfair to workers) and expensive (unfair to requesters). In contrast, we propose a low-cost QC mechanism which is fair to both workers and requesters. At the heart of our approach, lies a rule-based Transliteration Equivalence approach which takes as input a list of vowels in the two languages and a mapping of the consonants in the two languages. We empirically show that our approach outperforms other popular QC mechanisms (*viz.*, consensus and sampling) on two vital parameters : (i) fairness to requesters (lower cost per correct transliteration) and (ii) fairness to workers (lower rate of rejecting correct answers). Further, as an extrinsic evaluation we use the standard NEWS 2010 test set and show that such quality controlled crowdsourced data compares well to expert data when used for training a transliteration engine.

Design and Development of an Online Computational Framework to Facilitate Language Comprehension Research on Indian Languages

Manjira Sinha, Tirthankar Dasgupta and Anupam Basu

In this paper we have developed an open-source online computational framework that can be used by different research groups to conduct reading researches on Indian language texts. The framework can be used to develop a large annotated Indian language text comprehension data from different user-based experiments. The novelty in this framework lies in the fact that it brings different empirical data-collection techniques for text comprehension under one roof. The framework has been customized specifically to address language particularities for Indian languages. It will also offer many types of automatic analysis on the data at different levels such as full text, sentence and word level. To address the subjectivity of text difficulty perception, the framework allows to capture user background against multiple factors. The assimilated data can be automatically cross referenced against varying strata of readers.

Collaboration in the Production of a Massively Multilingual Lexicon

Martin Benjamin

This paper discusses the multiple approaches to collaboration that the Kamusi Project is employing in the creation of a massively multilingual lexical resource. The project's data structure enables the inclusion of large amounts of rich data within each sense-specific entry, with transitive concept-based links across languages. Data collection involves mining existing data sets, language experts using an online editing system, crowdsourcing, and games with a purpose. The paper discusses the benefits and drawbacks of each of these elements, and the steps the project is taking to account for those. Special attention is paid to guiding crowd members with targeted questions that produce results in

a specific format. Collaboration is seen as an essential method for generating large amounts of linguistic data, as well as for validating the data so it can be considered trustworthy.

A SICK Cure for the Evaluation of Compositional Distributional Semantic Models

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi and Roberto Zamparelli

Shared and internationally recognized benchmarks are fundamental for the development of any computational system. We aim to help the research community working on compositional distributional semantic models (CDSMs) by providing SICK (Sentences Involving Compositional Knowldedge), a large size English benchmark tailored for them. SICK consists of about 10,000 English sentence pairs that include many examples of the lexical, syntactic and semantic phenomena that CDSMs are expected to account for, but do not require dealing with other aspects of existing sentential data sets (idiomatic multiword expressions, named entities, telegraphic language) that are not within the scope of CDSMs. By means of crowdsourcing techniques, each pair was annotated for two crucial semantic tasks: relatedness in meaning (with a 5-point rating scale as gold score) and entailment relation between the two elements (with three possible gold labels: entailment, contradiction, and neutral). The SICK data set was used in SemEval-2014 Task 1, and it freely available for research purposes.

Can Crowdsourcing be used for Effective Annotation of Arabic?

Wajdi Zaghouani and Kais Dukes

Crowdsourcing has been used recently as an alternative to traditional costly annotation by many natural language processing groups. In this paper, we explore the use of Amazon Mechanical Turk (AMT) in order to assess the feasibility of using AMT workers (also known as Turkers) to perform linguistic annotation of Arabic. We used a gold standard data set taken from the Quran corpus project annotated with part-of-speech and morphological information. An Arabic language qualification test was used to filter out potential non-qualified participants. Two experiments were performed, a part-of-speech tagging task in where the annotators were asked to choose a correct word-category from a multiple choice list and case ending identification task. The results obtained so far showed that annotating Arabic grammatical case is harder than POS tagging, and crowdsourcing for Arabic linguistic annotation requiring expert annotators could be not as effective as other crowdsourcing experiments requiring less expertise and qualifications.

Crowdsourcing as a Preprocessing for Complex Semantic Annotation Tasks

Héctor Martínez Alonso and Lauren Romeo

This article outlines a methodology that uses crowdsourcing to reduce the workload of experts for complex semantic tasks. We split turker-annotated datasets into a high-agreement block, which is not modified, and a low-agreement block, which is re-annotated by experts. The resulting annotations have higher observed agreement. We identify different biases in the annotation for both turkers and experts.

Online Experiments with the Percy Software Framework - Experiences and some Early Results

Christoph Draxler

In early 2012 the online perception experiment software Percy was deployed on a production server at our lab. Since then, 38 experiments have been made publicly available, with a total of 3078 experiment sessions. In the course of time, the software has been continuously updated and extended to adapt to changing user requirements. Web-based editors for the structure and layout of the experiments have been developed. This paper describes the system architecture, presents usage statistics, discusses typical characteristics of online experiments, and gives an outlook on ongoing work. webapp.phonetik.uni-muenchen.de/WebExperiment lists all currently active experiments.

A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic

Ryan Cotterell and Chris Callison-Burch

This paper presents a multi-dialect, multi-genre, human annotated corpus of dialectal Arabic. We collected utterances in five Arabic dialects: Levantine, Gulf, Egyptian, Iraqi and Maghrebi. We scraped newspaper websites for user commentary and Twitter for two distinct types of dialectal content. To the best of the authors' knowledge, this work is the most diverse corpus of dialectal Arabic in both the source of the content and the number of dialects. Every utterance in the corpus was human annotated on Amazon's Mechanical Turk; this stands in contrast to Al-Sabbagh and Girju (2012) where only a small subset was human annotated in order to train a classifier to automatically annotate the remainder of the corpus. We provide a discussion of the methodology used for the annotation in addition to the performance of the individual workers. We extend the Arabic dialect identification task to the Iraqi and Maghrebi dialects and

improve the results of Zaidan and Callison-Burch (2011a) on Levantine, Gulf and Egyptian.

P3 - Dialogue

Wednesday, May 28, 11:35

Chairperson: **Dan Cristea** Poster Session

First Insight into Quality-Adaptive Dialogue

Stefan Ultes, Hüseyin Dikme and Wolfgang Minker

While Spoken Dialogue Systems have gained in importance in recent years, most systems applied in the real world are still static and error-prone. To overcome this, the user is put into the focus of dialogue management. Hence, an approach for adapting the course of the dialogue to Interaction Quality, an objective variant of user satisfaction, is presented in this work. In general, rendering the dialogue adaptive to user satisfaction enables the dialogue system to improve the course of the dialogue and to handle problematic situations better. In this contribution, we present a pilot study of quality-adaptive dialogue. By selecting the confirmation strategy based on the current IQ value, the course of the dialogue is adapted in order to improve the overall user experience. In a user experiment comparing three different confirmation strategies in a train booking domain, the adaptive strategy performs successful and is among the two best rated strategies based on the overall user experience.

The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues

Volha Petukhova, Martin Gropp, Dietrich Klakow, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz, Steffen Liersch and Anna Schmidt

This paper describes the data collection and annotation carried out within the DBOX project (Eureka project, number E! 7152). This project aims to develop interactive games based on spoken natural language human-computer dialogues, in 3 European languages: English, German and French. We collect the DBOX data continuously. We first start with human-human Wizard of Oz experiments to collect human-human data in order to model natural human dialogue behaviour, for better understanding of phenomena of human interactions and predicting interlocutors actions, and then replace the human Wizard by an increasingly advanced dialogue system, using evaluation data for system improvement. The designed dialogue system relies on a Question-Answering (QA) approach, but showing truly interactive gaming behaviour, e.g., by providing feedback, managing turns and contact, producing social signals and acts, e.g., encouraging vs. downplaying, polite vs. rude, positive vs. negative attitude towards players or their actions, etc. The DBOX dialogue corpus has required substantial investment. We expect it to have a great impact on the rest of the project. The DBOX project consortium will continue to maintain the corpus and to take an interest in its growth, e.g., expand to other languages. The resulting corpus will be publicly released.

Modeling and Evaluating Dialog Success in the LAST MINUTE Corpus

Dietmar Rösner, Rafael Friesen, Stephan Günther and Rico Andrich

The LAST MINUTE corpus comprises records and transcripts of naturalistic problem solving dialogs between N = 130 subjects and a companion system simulated in a Wizard of Oz experiment. Our goal is to detect dialog situations where subjects might break up the dialog with the system which might happen when the subject is unsuccessful. We present a dialog act-based representation of the dialog courses in the problem solving phase of the experiment and propose and evaluate measures for dialog success or failure derived from this representation. This dialog act representation refines our previous coarse measure as it enables the correct classification of many dialog sequences that were ambiguous before. The dialog act representation is useful for the identification of different subject groups and the exploration of interesting dialog courses in the corpus. We find young females to be most successful in the challenging last part of the problem solving phase and young subjects to have the initiative in the dialog more often than the elderly.

NASTIA: Negotiating Appointment Setting Interface

Layla El Asri, Rémi Lemonnier, Romain Laroche, Olivier Pietquin and Hatim Khouzaimi

This paper describes a French Spoken Dialogue System (SDS) named NASTIA (Negotiating Appointment SeTting InterfAce). Appointment scheduling is a hybrid task halfway between slot-filling and negotiation. NASTIA implements three different negotiation strategies. These strategies were tested on 1734 dialogues with 385 users who interacted at most 5 times with the SDS and gave a rating on a scale of 1 to 10 for each dialogue. Previous appointment scheduling systems were evaluated with the same experimental protocol. NASTIA is different from these systems in that it can adapt its strategy during the dialogue. The

highest system task completion rate with these systems was 81% whereas NASTIA had an 88% average and its best performing strategy even reached 92%. This strategy also significantly outperformed previous systems in terms of overall user rating with an average of 8.28 against 7.40. The experiment also enabled highlighting global recommendations for building spoken dialogue systems.

DINASTI: Dialogues with a Negotiating Appointment Setting Interface

Layla El Asri, Romain Laroche and Olivier Pietquin

This paper describes the DINASTI (DIalogues with a Negotiating Appointment SeTting Interface) corpus, which is composed of 1734 dialogues with the French spoken dialogue system NASTIA (Negotiating Appointment SeTting InterfAce). NASTIA is a reinforcement learning-based system. The DINASTI corpus was collected while the system was following a uniform policy. Each entry of the corpus is a system-user exchange annotated with 120 automatically computable features.The corpus contains a total of 21587 entries, with 385 testers. Each tester performed at most five scenario-based interactions with NASTIA. The dialogues last an average of 10.82 dialogue turns, with 4.45 reinforcement learning decisions. The testers filled an evaluation questionnaire after each dialogue. The questionnaire includes three questions to measure task completion. In addition, it comprises 7 Likert-scaled items evaluating several aspects of the interaction, a numerical overall evaluation on a scale of 1 to 10, and a free text entry. Answers to this questionnaire are provided with DINASTI. This corpus is meant for research on reinforcement learning modelling for dialogue management.

El-WOZ: a Client-Server Wizard-of-Oz Interface

Thomas Pellegrini, Vahid Hedayati and Angela Costa

In this paper, we present a speech recording interface developed in the context of a project on automatic speech recognition for elderly native speakers of European Portuguese. In order to collect spontaneous speech in a situation of interaction with a machine, this interface was designed as a Wizard-of-Oz (WOZ) plateform. In this setup, users interact with a fake automated dialog system controled by a human wizard. It was implemented as a client-server application and the subjects interact with a talking head. The human wizard chooses pre-defined questions or sentences in a graphical user interface, which are then synthesized and spoken aloud by the avatar on the client side. A small spontaneous speech corpus was collected in a daily center. Eight speakers between 75 and 90 years old were recorded. They appreciated the interface and felt at ease with the avatar. Manual orthographic transcriptions were created for the total of about 45 minutes of speech.

P4 - Phonetic Databases and Prosody

Wednesday, May 28, 11:35

Chairperson: **Philippe Martin** Poster Session

Tools for Arabic Natural Language Processing: a case study in qalqalah prosody

Claire Brierley, Majdi Sawalha and Eric Atwell

In this paper, we focus on the prosodic effect of qalqalah or "vibration" applied to a subset of Arabic consonants under certain constraints during correct Qur'anic recitation or taǧwīd, using our Boundary-Annotated Qur'an dataset of 77430 words (Brierley et al 2012; Sawalha et al 2014). These qalqalah events are rule-governed and are signified orthographically in the Arabic script. Hence they can be given abstract definition in the form of regular expressions and thus located and collected automatically. High frequency qalqalah content words are also found to be statistically significant discriminators or keywords when comparing Meccan and Medinan chapters in the Qur'an using a state-of-the-art Visual Analytics toolkit: Semantic Pathways. Thus we hypothesise that qalqalah prosody is one way of highlighting salient items in the text. Finally, we implement Arabic transcription technology (Brierley et al under review; Sawalha et al forthcoming) to create a qalqalah pronunciation guide where each word is transcribed phonetically in IPA and mapped to its chapter-verse ID. This is funded research under the EPSRC "Working Together" theme.

A Benchmark Database of Phonetic Alignments in Historical Linguistics and Dialectology

Johann-Mattis List and Jelena Prokić

In the last two decades, alignment analyses have become an important technique in quantitative historical linguistics and dialectology. Phonetic alignment plays a crucial role in the identification of regular sound correspondences and deeper genealogical relations between and within languages and language families. Surprisingly, up to today, there are no easily accessible benchmark data sets for phonetic alignment analyses. Here we present a publicly available database of manually edited phonetic alignments which can serve as a platform for testing and improving the performance of automatic alignment algorithms. The database consists of a great variety of alignments drawn from a large number of different sources. The data is arranged in a such way that typical problems encountered in phonetic

alignment analyses (metathesis, diversity of phonetic sequences) are represented and can be directly tested.

Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea and Atanas Tchobanov

The main objective of the Rhapsodie project (ANR Rhapsodie 07 Corp-030-01) was to define rich, explicit, and reproducible schemes for the annotation of prosody and syntax in different genres (± spontaneous, ± planned, face-to-face interviews vs. broadcast, etc.), in order to study the prosody/syntax/discourse interface in spoken French, and their roles in the segmentation of speech into discourse units (Lacheret, Kahane, & Pietrandrea forthcoming). We here describe the deliverable, a syntactic and prosodic treebank of spoken French, composed of 57 short samples of spoken French (5 minutes long on average, amounting to 3 hours of speech and 33000 words), orthographically and phonetically transcribed. The transcriptions and the annotations are all aligned on the speech signal: phonemes, syllables, words, speakers, overlaps. This resource is freely available at www.projet-rhapsodie.fr. The sound samples (wav/mp3), the acoustic analysis (original F0 curve manually corrected and automatic stylized F0, pitch format), the orthographic transcriptions (txt), the microsyntactic annotations (tabular format), the macrosyntactic annotations (txt, tabular format), the prosodic annotations (xml, textgrid, tabular format), and the metadata (xml and html) can be freely downloaded under the terms of the Creative Commons licence Attribution - Noncommercial - Share Alike 3.0 France. The metadata are encoded in the IMDI-CMFI format and can be parsed on line.

C-PhonoGenre: a 7-hour Corpus of 7 Speaking Styles in French: Relations between Situational Features and Prosodic Properties

Jean-Philippe Goldman, Tea Prsir and Antoine Auchlin

Phonogenres, or speaking styles, are typified acoustic images associated to types of language activities, causing prosodic and phonostylistic variations. This communication presents a large speech corpus (7 hours) in French, extending a previous work by Goldman et al. (2011a), Simon et al. (2010), with a greater number and complementary repertoire of considered phonogenres. The corpus is available with segmentation at phonetic, syllabic and word levels, as well as manual annotation. Segmentations and annotations were achieved semi-automatically, through a set of Praat implemented tools, and manual steps. The phonogenres are also described with a reduced set of situational dimensions as in Lucci (1983) and Koch & Oesterreicher's (2001). A preliminary acoustic study, joining rhythmical comparative measurements (Dellwo 2010) to Goldman et al.'s (2007a) ProsoReport, reports acoustic differences between phonogenres.

A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition

Abir Masmoudi, Mariem Ellouze Khmekhem, Yannick Estève, Lamia Hadrich Belguith and Nizar Habash

In this paper we describe an effort to create a corpus and phonetic dictionary for Tunisian Arabic Automatic Speech Recognition (ASR). The corpus, named TARIC (Tunisian Arabic Railway Interaction Corpus) has a collection of audio recordings and transcriptions from dialogues in the Tunisian Railway Transport Network. The phonetic (or pronunciation) dictionary is an important ASR component that serves as an intermediary between acoustic models and language models in ASR systems. The method proposed in this paper, to automatically generate a phonetic dictionary, is rule-based. For that reason, we define a set of pronunciation rules and a lexicon of exceptions. To determine the performance of our phonetic rules, we chose to evaluate our pronunciation dictionary on two types of corpora. The word error rate of word grapheme-to-phoneme mapping is around 9%.

Towards Automatic Transformation Between Different Transcription Conventions: Prediction of Intonation Markers from Linguistic and Acoustic Features

Yuichi Ishimoto, Tomoyuki Tsuchiya, Hanae Koiso and Yasuharu Den

Because of the tremendous effort required for recording and transcription, large-scale spoken language corpora have been hardly developed in Japanese, with a notable exception of the Corpus of Spontaneous Japanese (CSJ). Various research groups have individually developed conversation corpora in Japanese, but these corpora are transcribed by different conventions and have few annotations in common, and some of them lack fundamental annotations, which are prerequisites for conversation research. To solve this situation by sharing existing conversation corpora that cover diverse styles and settings, we have tried to automatically transform a transcription made by one convention into that made by another convention. Using a conversation corpus transcribed in both the Conversation-Analysis-style (CA-style) and CSJ-style, we analyzed the correspondence between CA's 'intonation markers' and CSJ's 'tone labels,' and constructed a statistical model that converts tone labels into intonation markers with reference to linguistic and acoustic features of the speech. The result showed that there is considerable variance in intonation

marking even between trained transcribers. The model predicted with 85% accuracy the presence of the intonation markers, and classified the types of the markers with 72% accuracy.

RSS-TOBI - a Prosodically Enhanced Romanian Speech Corpus

Tiberiu Boroş, Adriana Stan, Oliver Watts and Stefan Daniel Dumitrescu

This paper introduces a recent development of a Romanian Speech corpus to include prosodic annotations of the speech data in the form of ToBI labels. We describe the methodology of determining the required pitch patterns that are common for the Romanian language, annotate the speech resource, and then provide a comparison of two text-to-speech synthesis systems to establish the benefits of using this type of information to our speech resource. The result is a publicly available speech dataset which can be used to further develop speech synthesis systems or to automatically learn the prediction of ToBI labels from text in Romanian language.

Segmentation Evaluation Metrics, a Comparison Grounded on Prosodic and Discourse Units

Klim Peshkov and Laurent Prévot

Knowledge on evaluation metrics and best practices of using them have improved fast in the recent years Fort et al. (2012). However, the advances concern mostly evaluation of classification related tasks. Segmentation tasks have received less attention. Nevertheless, there are crucial in a large number of linguistic studies. A range of metrics is available (F-score on boundaries, F-score on units, WindowDiff ((WD), Boundary Similarity (BS) but it is still relatively difficult to interpret these metrics on various linguistic segmentation tasks, such as prosodic and discourse segmentation. In this paper, we consider real segmented datasets (introduced in Peshkov et al. (2012)) as references which we deteriorate in different ways (random addition of boundaries, random removal boundaries, near-miss errors introduction). This provide us with various measures on controlled datasets and with an interesting benchmark for various linguistic segmentation tasks.

A Cross-language Corpus for Studying the Phonetics and Phonology of Prominence

Bistra Andreeva, William Barry and Jacques Koreman

The present article describes a corpus which was collected for the cross-language comparison of prominence. In the data analysis, the acoustic-phonetic properties of words spoken with two different levels of accentuation (de-accented and nuclear accented in non-contrastive narrow-focus) are examined in question-answer elicited sentences and iterative imitations (on the syllable 'da') produced by Bulgarian, Russian, French, German and Norwegian speakers (3 male and 3 female per language). Normalized parameter values allow a comparison of the properties employed in differentiating the two levels of accentuation. Across the five languages there are systematic differences in the degree to which duration, f0, intensity and spectral vowel definition change with changing prominence under different focus conditions. The link with phonological differences between the languages is discussed.

Using a Machine Learning Model to Assess the Complexity of Stress Systems

Liviu Dinu, Alina Maria Ciobanu, Ioana Chitoran and Vlad Niculae

We address the task of stress prediction as a sequence tagging problem. We present sequential models with averaged perceptron training for learning primary stress in Romanian words. We use character n-grams and syllable n-grams as features and we account for the consonant-vowel structure of the words. We show in this paper that Romanian stress is predictable, though not deterministic, by using data-driven machine learning techniques.

GlobalPhone: Pronunciation Dictionaries in 20 Languages

Tanja Schultz and Tim Schlippe

This paper describes the advances in the multilingual text and speech database GlobalPhone, a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in 20 languages. GlobalPhone was designed to be uniform across languages with respect to the amount of data, speech quality, the collection scenario, the transcription and phone set conventions. With more than 400 hours of transcribed audio data from more than 2000 native speakers GlobalPhone supplies an excellent basis for research in the areas of multilingual speech recognition, rapid deployment of speech processing systems to yet unsupported languages, language identification tasks, speaker recognition in multiple languages, multilingual speech synthesis, as well as monolingual speech recognition in a large variety of languages. Very recently the GlobalPhone pronunciation dictionaries have been made available for research and commercial purposes by the European Language Resources Association (ELRA).

P5 - Speech Resources

Wednesday, May 28, 11:35

Chairperson: **Martine Adda-Decker** Poster Session

New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease

Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva and Elmar Nöth

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder after Alzheimer's, affecting about 1% of the people older than 65 and about 89% of the people with PD develop different speech disorders. Different researchers are currently working in the analysis of speech of people with PD, including the study of different dimensions in speech such as phonation, articulation,prosody and intelligibility. The study of phonation and articulation has been addressed mainly considering sustained vowels; however, the analysis of prosody and intelligibility requires the inclusion of words, sentences and monologue. In this paper we present a new database with speech recordings of 50 patients with PD and their respective healthy controls, matched by age and gender. All of the participants are Spanish native speakers and the recordings were collected following a protocol that considers both technical requirements and several recommendations given by experts in linguistics, phoniatry and neurology. This corpus includes tasks such as sustained phonations of the vowels, diadochokinetic evaluation, 45 words, 10 sentences, a reading text and a monologue. The paper also includes results of the characterization of the Spanish vowels considering different measures used in other works to characterize different speech impairments.

An Effortless Way To Create Large-Scale Datasets For Famous Speakers

François Salmon and Félicien Vallet

The creation of large-scale multimedia datasets has become a scientific matter in itself. Indeed, the fully-manual annotation of hundreds or thousands of hours of video and/or audio turns out to be practically infeasible. In this paper, we propose an extremly handy approach to automatically construct a database of famous speakers from TV broadcast news material.We then run a user experiment with a correctly designed tool that demonstrates that very reliable results can be obtained with this method. In particular, a thorough error analysis demonstrates the value of the approach and provides hints for the improvement of the quality of the dataset.

German Alcohol Language Corpus - the Question of Dialect

Florian Schiel and Thomas Kisler

Speech uttered under the influence of alcohol is known to deviate from the speech of the same person when sober. This is an important feature in forensic investigations and could also be used to detect intoxication in the automotive environment. Aside from acoustic-phonetic features and speech content which have already been studied by others in this contribution we address the question whether speakers use dialectal variation or dialect words more frequently when intoxicated than when sober. We analyzed 300,000 recorded word tokens in read and spontaneous speech uttered by 162 female and male speakers within the German Alcohol Language Corpus. We found that contrary to our expectations the frequency of dialectal forms decreases significantly when speakers are under the influence. We explain this effect with a compensatory over-shoot mechanism: speakers are aware of their intoxication and that they are being monitored. In forensic analysis of speech this 'awareness factor' must be taken into account.

Vulnerability in Acquisition, Language Impairments in Dutch: Creating a VALID Data Archive

Jetske Klatter, Roeland van Hout, Henk van den Heuvel, Paula Fikkert, Anne Baker, Jan de Jong, Frank Wijnen, Eric Sanders and Paul Trilsbeek

The VALID Data Archive is an open multimedia data archive (under construction) with data from speakers suffering from language impairments. We report on a pilot project in the CLARIN-NL framework in which five data resources were curated. For all data sets concerned, written informed consent from the participants or their caretakers has been obtained. All materials were anonymized. The audio files were converted into wav (linear PCM) files and the transcriptions into CHAT or ELAN format. Research data that consisted of test, SPSS and Excel files were documented and converted into CSV files. All data sets obtained appropriate CMDI metadata files. A new CMDI metadata profile for this type of data resources was established and care was taken that ISOcat metadata categories were used to optimize interoperability.

After curation all data are deposited at the Max Planck Institute for Psycholinguistics Nijmegen where persistent identifiers are linked to all resources. The content of the transcriptions in CHAT and plain text format can be searched with the TROVA search engine.

The Nijmegen Corpus of Casual Czech

Mirjam Ernestus, Lucie Kočková-Amortová and Petr Pollak

This article introduces a new speech corpus, the Nijmegen Corpus of Casual Czech (NCCCz), which contains more than 30 hours of high-quality recordings of casual conversations in Common Czech, among ten groups of three male and ten groups of three female friends. All speakers were native speakers of Czech, raised in Prague or in the region of Central Bohemia, and were between 19 and 26 years old. Every group of speakers consisted of one confederate, who was instructed to keep the conversations lively, and two speakers naive to the purposes of the recordings. The naive speakers were engaged in conversations for approximately 90 minutes, while the confederate joined them for approximately the last 72 minutes. The corpus was orthographically annotated by experienced transcribers and this orthographic transcription was aligned with the speech signal. In addition, the conversations were videotaped. This corpus can form the basis for all types of research on casual conversations in Czech, including phonetic research and research on how to improve automatic speech recognition. The corpus will be freely available.

CIEMPIESS: A New Open-Sourced Mexican Spanish Radio Corpus

Carlos Daniel Hernandez Mena and Abel Herrera Camacho

Corpus de Investigación en Español de México del Posgrado de Ingeniería Eléctrica y Servicio Social" (CIEMPIESS) is a new open-sourced corpus extracted from Spanish spoken FM podcasts in the dialect of the center of Mexico. The CIEMPIESS corpus was designed to be used in the field of automatic speech recongnition (ASR) and it is provided with two different kind of pronouncing dictionaries, one of them containing the phonemes of Mexican Spanish and the other containing this same phonemes plus allophones. Corpus annotation took into account the tonic vowel of every word and the four different sounds that letter "x" presents in the Spanish language. CIEMPIESS corpus is also provided with two different language models extracted from electronic newsletters, one of them takes into account the tonic vowels but not the other one. Both the dictionaries and the language models allow users to experiment different scenarios for the recognition task in order to adequate the corpus to their needs.

Mapping Diatopic and Diachronic Variation in Spoken Czech: the Ortofon and Dialekt Corpora

Marie Kopřivová, Hana Goláňová, Petra Klimešová and David Lukeš

ORTOFON and DIALEKT are two corpora of spoken Czech (recordings + transcripts) which are currently being built at the Institute of the Czech National Corpus. The first one (ORTOFON) continues the tradition of the CNC's ORAL series of spoken corpora by focusing on collecting recordings of unscripted informal spoken interactions ("prototypically spoken texts"), but also provides new features, most notably an annotation scheme with multiple tiers per speaker, including orthographic and phonetic transcripts and allowing for a more precise treatment of overlapping speech. Rich speaker- and situation-related metadata are also collected for possible use as factors in sociolinguistic analyses. One of the stated goals is to make the data in the corpus balanced with respect to a subset of these. The second project, DIALEKT, consists in annotating (in a way partially compatible with the ORTOFON corpus) and providing electronic access to historical (1960s–80s) dialect recordings, mainly of a monological nature, from all over the Czech Republic. The goal is to integrate both corpora into one map-based browsing interface, allowing an intuitive and informative spatial visualization of query results or dialect feature maps, confrontation with isoglosses previously established through the effort of dialectologists etc.

The Research and Teaching Corpus of Spoken German – FOLK

Thomas Schmidt

FOLK is the "Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)" (eng.: research and teaching corpus of spoken German). The project has set itself the aim of building a corpus of German conversations which a) covers a broad range of interaction types in private, institutional and public settings, b) is sufficiently large and diverse and of sufficient quality to support different qualitative and quantitative research approaches, c) is transcribed, annotated and made accessible according to current technological standards, and d) is available to the scientific community on a sound legal basis and without unnecessary restrictions of usage. This paper gives an overview of the corpus design, the strategies for acquisition of a diverse range of interaction data, and the

corpus construction workflow from recording via transcription an annotation to dissemination.

Free Acoustic and Language Models for Large Vocabulary Continuous Speech Recognition in Swedish

Niklas Vanhainen and Giampiero Salvi

This paper presents results for large vocabulary continuous speech recognition (LVCSR) in Swedish. We trained acoustic models on the public domain NST Swedish corpus and made them freely available to the community. The training procedure corresponds to the reference recogniser (RefRec) developed for the SpeechDat databases during the COST249 action. We describe the modifications we made to the procedure in order to train on the NST database, and the language models we created based on the N-gram data available at the Norwegian Language Council. Our tests include medium vocabulary isolated word recognition and LVCSR. Because no previous results are available for LVCSR in Swedish, we use as baseline the performance of the SpeechDat models on the same tasks. We also compare our best results to the ones obtained in similar conditions on resource rich languages such as American English. We tested the acoustic models with HTK and Julius and plan to make them available in CMU Sphinx format as well in the near future. We believe that the free availability of these resources will boost research in speech and language technology in Swedish, even in research groups that do not have resources to develop ASR systems.

O5 - Linked Data (Special Session)

Wednesday, May 28, 14:45

Chairperson: **Asuncion Gomez-Perez** Oral Session

Metadata as Linked Open Data: Mapping Disparate XML Metadata Registries into one RDF/OWL Registry

Marta Villegas, Maite Melero and Núria Bel

The proliferation of different metadata schemas and models pose serious problems of interoperability. Maintaining isolated repositories with overlapping data is costly in terms of time and effort. In this paper, we describe how we have achieved a Linked Open Data version of metadata descriptions coming from heterogeneous sources, originally encoded in XML. The resulting model is much simpler than the original XSD schema and avoids problems typical of XML syntax, such as semantic ambiguity and order constraint. Moreover, the open world assumption of RDF/OWL allows to naturally integrate objects from different schemas and to add further extensions, facilitating merging of different models as well as linking to external data. Apart from the advantages in terms of interoperability and maintainability, the merged repository enables end-users to query multiple sources using a unified schema and is able to present them with implicit knowledge derived from the linked data. The approach we present here is easily scalable to any number of sources and schemas.

Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano and Roberto Navigli

Recent years have witnessed a surge in the amount of semantic information published on the Web. Indeed, the Web of Data, a subset of the Semantic Web, has been increasing steadily in both volume and variety, transforming the Web into a 'global database' in which resources are linked across sites. Linguistic fields – in a broad sense – have not been left behind, and we observe a similar trend with the growth of linguistic data collections on the so-called 'Linguistic Linked Open Data (LLOD) cloud'. While both Semantic Web and Natural Language Processing communities can obviously take advantage of this growing and distributed linguistic knowledge base, they are today faced with a new challenge, i.e., that of facilitating multilingual access to the Web of data. In this paper we present the publication of BabelNet 2.0, a wide-coverage multilingual encyclopedic dictionary and ontology, as Linked Data. The conversion made use of lemon, a lexicon model for ontologies particularly well-suited for this enterprise. The result is an interlinked multilingual (lexical) resource which can not only be accessed on the LOD, but also be used to enrich existing datasets with linguistic information, or to support the process of mapping datasets across languages.

Enabling Language Resources to Expose Translations as Linked Data on the Web

Jorge Gracia, Elena Montiel-Ponsoda, Daniel Vila-Suero and Guadalupe Aguado-de-Cea

Language resources, such as multilingual lexica and multilingual electronic dictionaries, contain collections of lexical entries in several languages. Having access to the corresponding explicit or implicit translation relations between such entries might be of great interest for many NLP-based applications. By using Semantic Web-based techniques, translations can be available on the Web to be consumed by other (semantic enabled) resources in a direct manner, not relying on application-specific formats. To that end, in this paper we propose a model for representing translations as linked data, as an extension of the lemon model. Our translation module represents some core information associated to term translations and does not commit to specific

views or translation theories. As a proof of concept, we have extracted the translations of the terms contained in Terminesp, a multilingual terminological database, and represented them as linked data. We have made them accessible on the Web both for humans (via a Web interface) and software agents (with a SPARQL endpoint).

A SKOS-based Schema for TEI encoded Dictionaries at ICLTT

Thierry Declerck, Karlheinz Mörth and Eveline Wandl-Vogt

At our institutes we are working with quite some dictionaries and lexical resources in the field of less-resourced language data, like dialects and historical languages. We are aiming at publishing those lexical data in the Linked Open Data framework in order to link them with available data sets for highly-resourced languages and elevating them thus to the same "digital dignity" the mainstream languages have gained. In this paper we concentrate on two TEI encoded variants of the Arabic language and propose a mapping of this TEI encoded data onto SKOS, showing how the lexical entries of the two dialectal dictionaries can be linked to other language resources available in the Linked Open Data cloud.

O6 - Audiovisual

Wednesday, May 28, 14:45

Chairperson: **Gilles Adda** Oral Session

TVD: A Reproducible and Multiply Aligned TV Series Dataset

Anindya Roy, Camille Guinaudeau, Herve Bredin and Claude Barras

We introduce a new dataset built around two TV series from different genres, The Big Bang Theory, a situation comedy and Game of Thrones, a fantasy drama. The dataset has multiple tracks extracted from diverse sources, including dialogue (manual and automatic transcripts, multilingual subtitles), crowd-sourced textual descriptions (brief episode summaries, longer episode outlines) and various metadata (speakers, shots, scenes). The paper describes the dataset and provide tools to reproduce it for research purposes provided one has legally acquired the DVD set of the series. Tools are also provided to temporally align a major subset of dialogue and description tracks, in order to combine complementary information present in these tracks for enhanced accessibility. For alignment, we consider tracks as comparable corpora and first apply an existing algorithm for aligning such corpora based on dynamic time warping and TFIDF-based similarity scores. We improve this baseline algorithm using contextual information, WordNet-based word similarity and scene location information. We report the performance of these algorithms on a manually aligned subset of the data. To highlight the interest of the database, we report a use case involving rich speech retrieval and propose other uses.

Image Annotation with ISO-Space: Distinguishing Content from Structure

James Pustejovsky and Zachary Yocum

Natural language descriptions of visual media present interesting problems for linguistic annotation of spatial information. This paper explores the use of ISO-Space, an annotation specification to capturing spatial information, for encoding spatial relations mentioned in descriptions of images. Especially, we focus on the distinction between references to representational content and structural components of images, and the utility of such a distinction within a compositional semantics. We also discuss how such a structure-content distinction within the linguistic annotation can be leveraged to compute further inferences about spatial configurations depicted by images with verbal captions. We construct a composition table to relate content-based relations to structure-based relations in the image, as expressed in the captions. While still preliminary, our initial results suggest that a weak composition table is both sound and informative for deriving new spatial relations.

SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling

Arantza del Pozo, Carlo Aliprandi, Aitor Álvarez, Carlos Mendes, Joao P. Neto, Sérgio Paulo, Nicola Piccinini and Matteo Raffaelli

This paper describes the data collection, annotation and sharing activities carried out within the FP7 EU-funded SAVAS project. The project aims to collect, share and reuse audiovisual language resources from broadcasters and subtitling companies to develop large vocabulary continuous speech recognisers in specific domains and new languages, with the purpose of solving the automated subtitling needs of the media industry.

Phoneme Similarity Matrices to Improve Long Audio Alignment for Automatic Subtitling

Pablo Ruiz, Aitor Álvarez and Haritz Arzelus

Long audio alignment systems for Spanish and English are presented, within an automatic subtitling application. Language-specific phone decoders automatically recognize audio contents at phoneme level. At the same time, language-dependent grapheme-to-phoneme modules perform a transcription of the script for

the audio. A dynamic programming algorithm (Hirschberg's algorithm) finds matches between the phonemes automatically recognized by the phone decoder and the phonemes in the script's transcription. Alignment accuracy is evaluated when scoring alignment operations with a baseline binary matrix, and when scoring alignment operations with several continuous-score matrices, based on phoneme similarity as assessed through comparing multivalued phonological features. Alignment accuracy results are reported at phoneme, word and subtitle level. Alignment accuracy when using the continuous scoring matrices based on phonological similarity was clearly higher than when using the baseline binary matrix.

KALAKA-3: a Database for the Recognition of Spoken European Languages on YouTube Audios

Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez and German Bordel

This paper describes the main features of KALAKA-3, a speech database specifically designed for the development and evaluation of language recognition systems. The database provides TV broadcast speech for training, and audio data extracted from YouTube videos for tuning and testing. The database was created to support the Albayzin 2012 Language Recognition Evaluation, which featured two language recognition tasks, both dealing with European languages. The first one involved six target languages (Basque, Catalan, English, Galician, Portuguese and Spanish) for which there was plenty of training data, whereas the second one involved four target languages (French, German, Greek and Italian) for which no training data was provided. Two separate sets of YouTube audio files were provided to test the performance of language recognition systems on both tasks. To allow open-set tests, these datasets included speech in 11 additional (Out-Of-Set) European languages. The paper also presents a summary of the results attained in the evaluation, along with the performance of state-of-the-art systems on the four evaluation tracks defined on the database, which demonstrates the extreme difficulty of some of them. As far as we know, this is the first database specifically designed to benchmark spoken language recognition technology on YouTube audios.

O7 - Processing of Social Media

Wednesday, May 28, 14:45

Chairperson: **Paul Rayson** Oral Session

Named Entity Recognition on Turkish Tweets

Dilek Kucuk, Guillaume Jacquet and Ralf Steinberger

Various recent studies show that the performance of named entity recognition (NER) systems developed for well-formed text types drops significantly when applied to tweets. The only existing study for the highly inflected agglutinative language Turkish reports a drop in F-Measure from 91% to 19% when ported from news articles to tweets. In this study, we present a new named entity-annotated tweet corpus and a detailed analysis of the various tweet-specific linguistic phenomena. We perform comparative NER experiments with a rule-based multilingual NER system adapted to Turkish on three corpora: a news corpus, our new tweet corpus, and another tweet corpus. Based on the analysis and the experimentation results, we suggest system features required to improve NER results for social media like Twitter.

Comprehensive Annotation of Multiword Expressions in a Social Web Corpus

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad and Noah A. Smith

Multiword expressions (MWEs) are quite frequent in languages such as English, but their diversity, the scarcity of individual MWE types, and contextual ambiguity have presented obstacles to corpus-based studies and NLP systems addressing them as a class. Here we advocate for a comprehensive annotation approach: proceeding sentence by sentence, our annotators manually group tokens into MWEs according to guidelines that cover a broad range of multiword phenomena. Under this scheme, we have fully annotated an English web corpus for multiword expressions, including those containing gaps.

Re-using an Argument Corpus to Aid in the Curation of Social Media Collections

Clare Llewellyn, Claire Grover, Jon Oberlander and Ewan Klein

This work investigates how automated methods can be used to classify social media text into argumentation types. In particular it is shown how supervised machine learning was used to annotate a Twitter dataset (London Riots) with argumentation classes. An investigation of issues arising from a natural inconsistency within social media data found that machine learning algorithms tend to over fit to the data because Twitter contains a lot of repetition in the form of retweets. It is also noted that when learning argumentation classes we must be aware that the classes will most likely be of very different sizes and this must be kept in mind when analysing the results. Encouraging results were found in adapting a model from one domain of Twitter data (London Riots) to another (OR2012). When adapting a model to another dataset

the most useful feature was punctuation. It is probable that the nature of punctuation in Twitter language, the very specific use in links, indicates argumentation class.

#mygoal: Finding Motivations on Twitter

Marc Tomlinson, David Bracewell, Wayne Krug and David Hinote

Our everyday language reflects our psychological and cognitive state and effects the states of other individuals. In this contribution we look at the intersection between motivational state and language. We create a set of hashtags, which are annotated for the degree to which they are used by individuals to mark-up language that is indicative of a collection of factors that interact with an individual's motivational state. We look for tags that reflect a goal mention, reward, or a perception of control. Finally, we present results for a language-model-based classifier which is able to predict the presence of one of these factors in a tweet with between 69% and 80% accuracy on a balanced testing set. Our approach suggests that hashtags can be used to understand, not just the language of topics, but the deeper psychological and social meaning of a tweet.

A Framework for Public Health Surveillance

Andrew Yates, Jon Parker, Nazli Goharian and Ophir Frieder

With the rapid growth of social media, there is increasing potential to augment traditional public health surveillance methods with data from social media. We describe a framework for performing public health surveillance on Twitter data. Our framework, which is publicly available, consists of three components that work together to detect health-related trends in social media: a concept extraction component for identifying health-related concepts, a concept aggregation component for identifying how the extracted health-related concepts relate to each other, and a trend detection component for determining when the aggregated health-related concepts are trending. We describe the architecture of the framework and several components that have been implemented in the framework, identify other components that could be used with the framework, and evaluate our framework on approximately 1.5 years of tweets. While it is difficult to determine how accurately a Twitter trend reflects a trend in the real world, we discuss the differences in trends detected by several different methods and compare flu trends detected by our framework to data from Google Flu Trends.

O8 - Acquisition

Wednesday, May 28, 14:45

Chairperson: **Xavier Tannier** Oral Session

Bootstrapping Term Extractors for Multiple Languages

Ahmet Aker, Monica Paramita, Emma Barker and Robert Gaizauskas

Terminology extraction resources are needed for a wide range of human language technology applications, including knowledge management, information extraction, semantic search, cross-language information retrieval and automatic and assisted translation. We create a low cost method for creating terminology extraction resources for 21 non-English EU languages. Using parallel corpora and a projection method, we create a General POS Tagger for these languages. We also investigate the use of EuroVoc terms and Wikipedia corpus to automatically create term grammar for each language. Our results show that these automatically generated resources can assist term extraction process with similar performance to manually generated resources. All resources resulted in this experiment are freely available for download.

Evaluation of Automatic Hypernym Extraction from Technical Corpora in English and Dutch

Els Lefever, Marjan van de Kauter and Véronique Hoste

In this research, we evaluate different approaches for the automatic extraction of hypernym relations from English and Dutch technical text. The detected hypernym relations should enable us to semantically structure automatically obtained term lists from domain- and user-specific data. We investigated three different hypernymy extraction approaches for Dutch and English: a lexico-syntactic pattern-based approach, a distributional model and a morpho-syntactic method. To test the performance of the different approaches on domain-specific data, we collected and manually annotated English and Dutch data from two technical domains, viz. the dredging and financial domain. The experimental results show that especially the morpho-syntactic approach obtains good results for automatic hypernym extraction from technical and domain-specific texts.

Resources for the Detection of Conventionalized Metaphors in Four Languages

Lori Levin, Teruko Mitamura, Brian MacWhinney, Davida Fromm, Jaime Carbonell, Weston Feely, Robert Frederking, Anatole Gershman and Carlos Ramirez

This paper describes a suite of tools for extracting conventionalized metaphors in English, Spanish, Farsi, and

Russian. The method depends on three significant resources for each language: a corpus of conventionalized metaphors, a table of conventionalized conceptual metaphors (CCM table), and a set of extraction rules. Conventionalized metaphors are things like "escape from poverty" and "burden of taxation". For each metaphor, the CCM table contains the metaphorical source domain word (such as "escape") the target domain word (such as "poverty") and the grammatical construction in which they can be found. The extraction rules operate on the output of a dependency parser and identify the grammatical configurations (such as a verb with a prepositional phrase complement) that are likely to contain conventional metaphors. We present results on detection rates for conventional metaphors and analysis of the similarity and differences of source domains for conventional metaphors in the four languages.

A Language-independent and fully Unsupervised Approach to Lexicon Induction and Part-of-Speech Tagging for Closely Related Languages

Yves Scherrer and Benoît Sagot

In this paper, we describe our generic approach for transferring part-of-speech annotations from a resourced language towards an etymologically closely related non-resourced language, without using any bilingual (i.e., parallel) data. We first induce a translation lexicon from monolingual corpora, based on cognate detection followed by cross-lingual contextual similarity. Second, POS information is transferred from the resourced language along translation pairs to the non-resourced language and used for tagging the corpus. We evaluate our methods on three language families, consisting of five Romance languages, three Germanic languages and five Slavic languages. We obtain tagging accuracies of up to 91.6%.

Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality

Stefan Bott and Sabine Schulte im Walde

In the work presented here we assess the degree of compositionality of German Particle Verbs with a Distributional Semantics Model which only relies on word window information and has no access to syntactic information as such. Our method only takes the lexical distributional distance between the Particle Verb to its Base Verb as a predictor for compositionality. We show that the ranking of distributional similarity correlates significantly with the ranking of human judgements on semantic compositionality for a series of Particle Verbs and the Base Verbs they are derived from. We also investigate the influence of further linguistic factors, such as the ambiguity and the overall frequency of the verbs and a syntactically separate occurrences of verbs and particles that causes difficulties for the correct lemmatization of Particle Verbs. We analyse in how far these factors may influence the success with which the compositionality of the Particle Verbs may be predicted.

P6 - Endangered Languages

Wednesday, May 28, 14:45

Chairperson: **Laurette Pretorius** Poster Session

Open-domain Interaction and Online Content in the Sami Language

Kristiina Jokinen

This paper presents data collection and collaborative community events organised within the project Digital Natives on the North Sami language. The project is one of the collaboration initiatives on endangered Finno-Ugric languages, supported by the larger framework between the Academy of Finland and the Hungarian Academy of Sciences. The goal of the project is to improve digital visibility and viability of the targeted Finno-Ugric languages, as well as to develop language technology tools and resources in order to assist automatic language processing and experimenting with multilingual interactive applications.

The Gulf of Guinea Creole Corpora

Tjerk Hagemeijer, Michel Généreux, Iris Hendrickx, Amália Mendes, Abigail Tiny and Armando Zamora

We present the process of building linguistic corpora of the Portuguese-related Gulf of Guinea creoles, a cluster of four historically related languages: Santome, Angolar, Principense and Fa d'Ambô. We faced the typical difficulties of languages lacking an official status, such as lack of standard spelling, language variation, lack of basic language instruments, and small data sets, which comprise data from the late 19th century to the present. In order to tackle these problems, the compiled written and transcribed spoken data collected during field work trips were adapted to a normalized spelling that was applied to the four languages. For the corpus compilation we followed corpus linguistics standards. We recorded meta data for each file and added morphosyntactic information based on a part-of-speech tag set that was designed to deal with the specificities of these languages. The corpora of three of the four creoles are already available and searchable via an online web interface.

Languagesindanger.eu - Including Multimedia Language Resources to disseminate Knowledge and Create Educational Material on less-Resourced Languages

Dagmar Jung, Katarzyna Klessa, Zsuzsa Duray, Beatrix Oszkó, Mária Sipos, Sándor Szeverényi, Zsuzsa Várnai, Trilsbeek Paul and Tamás Váradi

The present paper describes the development of the languagesindanger.eu interactive website as an example of including multimedia language resources to disseminate knowledge and create educational material on less-resourced languages. The website is a product of INNET (Innovative networking in infrastructure for endangered languages), European FP7 project. Its main functions can be summarized as related to the three following areas: (1) raising students' awareness of language endangerment and arouse their interest in linguistic diversity, language maintenance and language documentation; (2) informing both students and teachers about these topics and show ways how they can enlarge their knowledge further with a special emphasis on information about language archives; (3) helping teachers include these topics into their classes. The website has been localized into five language versions with the intention to be accessible to both scientific and non-scientific communities such as (primarily) secondary school teachers and students, beginning university students of linguistics, journalists, the interested public, and also members of speech communities who speak minority languages.

Casa De La Lhéngua: a Set of Language Resources and Natural Language Processing Tools for Mirandese

José Pedro Ferreira, Cristiano Chesi, Daan Baldewijns, Fernando Miguel Pinto, Margarita Correia, Daniela Braga, Hyongsil Cho, Amadeu Ferreira and Miguel Dias

This paper describes the efforts for the construction of Language Resources and NLP tools for Mirandese, a minority language spoken in North-eastern Portugal, now available on a community-led portal, Casa de la Lhéngua. The resources were developed in the context of a collaborative citizenship project led by Microsoft, in the context of the creation of the first TTS system for Mirandese. Development efforts encompassed the compilation of a corpus with over 1M tokens, the construction of a GTP system, syllable-division, inflection and a Part-of-Speech (POS) tagger modules, leading to the creation of an inflected lexicon of about 200.000 entries with phonetic transcription, detailed POS tagging, syllable division, and stress mark-up. Alongside these tasks, which were made easier through the adaptation and reuse of existing tools for closely related languages, a casting for voice talents among the speaking community was conducted and the first speech database for speech synthesis was recorded for Mirandese. These resources were combined to fulfil the requirements of a well-tested statistical parameter synthesis model, leading to an intelligible voice font. These language resources are available freely at Casa de la Lhéngua, aiming at promoting the development of real-life applications and fostering linguistic research on Mirandese.

A Finite-State Morphological Analyzer for a Lakota HPSG Grammar

Christian Curtis

This paper reports on the design and implementation of a morphophonological analyzer for Lakota, a member of the Siouan language family. The initial motivation for this work was to support development of a precision implemented grammar for Lakota on the basis of the LinGO Grammar Matrix. A finite-state transducer (FST) was developed to adapt Lakota's complex verbal morphology into a form directly usable as input to the Grammar Matrix-derived grammar. As the FST formalism can be applied in both directions, this approach also supports generative output of correct surface forms from the implemented grammar. This article describes the approach used to model Lakota verbal morphology using finite-state methods. It also discusses the results of developing a lexicon from existing text and evaluating its application to related but novel text. The analyzer presented here, along with its companion precision grammar, explores an approach that may have application in enabling machine translation for endangered and under-resourced languages.

P7 - Evaluation Methodologies

Wednesday, May 28, 14:45

Chairperson: **Violeta Seretan** Poster Session

Extrinsic Corpus Evaluation with a Collocation Dictionary Task

Adam Kilgarriff, Pavel Rychlý, Milos Jakubicek, Vojtěch Kovář, Vít Baisa and Lucia Kocincová

The NLP researcher or application-builder often wonders "what corpus should I use, or should I build one of my own? If I build one of my own, how will I know if I have done a good job?" Currently there is very little help available for them. They are in need of a framework for evaluating corpora. We develop such a framework, in relation to corpora which aim for good coverage of "general language". The task we set is automatic creation of a publication-quality collocations dictionary. For a sample

of 100 headwords of Czech and 100 of English, we identify a gold standard dataset of (ideally) all the collocations that should appear for these headwords in such a dictionary. The datasets are being made available alongside this paper. We then use them to determine precision and recall for a range of corpora, with a range of parameters.

Evaluating the Effects of Interactivity in a Post-Editing Workbench

Nancy Underwood, Bartolomé Mesa-Lao, Mercedes García Martínez, Michael Carl, Vicent Alabau, Jesús González-Rubio, Luis A. Leiva, Germán Sanchis-Trilles, Daniel Ortíz-Martínez and Francisco Casacuberta

This paper describes the field trial and subsequent evaluation of a post-editing workbench which is currently under development in the EU-funded CasMaCat project. Based on user evaluations of the initial prototype of the workbench, this second prototype of the workbench includes a number of interactive features designed to improve productivity and user satisfaction. Using CasMaCat's own facilities for logging keystrokes and eye tracking, data were collected from nine post-editors in a professional setting. These data were then used to investigate the effects of the interactive features on productivity, quality, user satisfaction and cognitive load as reflected in the post-editors' gaze activity. These quantitative results are combined with the qualitative results derived from user questionnaires and interviews conducted with all the participants.

Bridging the Gap between Speech Technology and Natural Language Processing: An Evaluation Toolbox for Term Discovery Systems

Bogdan Ludusan, Maarten Versteegh, Aren Jansen, Guillaume Gravier, Xuan-Nga Cao, Mark Johnson and Emmanuel Dupoux

The unsupervised discovery of linguistic terms from either continuous phoneme transcriptions or from raw speech has seen an increasing interest in the past years both from a theoretical and a practical standpoint. Yet, there exists no common accepted evaluation method for the systems performing term discovery. Here, we propose such an evaluation toolbox, drawing ideas from both speech technology and natural language processing. We first transform the speech-based output into a symbolic representation and compute five types of evaluation metrics on this representation: the quality of acoustic matching, the quality of the clusters found, and the quality of the alignment with real words (type, token, and boundary scores). We tested our approach on two term discovery systems taking speech as input, and one using symbolic input. The latter was run using both the gold transcription and a transcription obtained from an automatic speech recognizer, in order to simulate the case when only imperfect symbolic information is available. The results obtained are analysed through the use of the proposed evaluation metrics and the implications of these metrics are discussed.

Introducing a Framework for the Evaluation of Music Detection Tools

Paula Lopez-Otero, Laura Docio-Fernandez and Carmen Garcia-Mateo

The huge amount of multimedia information available nowadays makes its manual processing prohibitive, requiring tools for automatic labelling of these contents. This paper describes a framework for assessing a music detection tool; this framework consists of a database, composed of several hours of radio recordings that include different types of radio programmes, and a set of evaluation measures for evaluating the performance of a music detection tool in detail. A tool for automatically detecting music in audio streams, with application to music information retrieval tasks, is presented as well. The aim of this tool is to discard the audio excerpts that do not contain music in order to avoid their unnecessary processing. This tool applies fingerprinting to different acoustic features extracted from the audio signal in order to remove perceptual irrelevancies, and a support vector machine is trained for classifying these fingerprints in classes music and no-music. The validity of this tool is assessed in the proposed evaluation framework.

Measuring Readability of Polish Texts: Baseline Experiments

Bartosz Broda, Bartłomiej Nitoń, Włodzimierz Gruszczyński and Maciej Ogrodniczuk

Measuring readability of a text is the first sensible step to its simplification. In this paper we present an overview of the most common approaches to automatic measuring of readability. Of the described ones, we implemented and evaluated: Gunning FOG index, Flesch-based Pisarek method. We also present two other approaches. The first one is based on measuring distributional lexical similarity of a target text and comparing it to reference texts. In the second one, we propose a novel method for automation of Taylor test – which, in its base form, requires performing a large amount of surveys. The automation of Taylor test is performed using a technique called statistical language modelling. We have developed a free on-line web-based system and constructed plugins for the most common text editors, namely Microsoft Word and OpenOffice.org. Inner workings of the system are described in detail. Finally, extensive evaluations are performed for Polish – a Slavic, highly inflected language. We

show that Pisarek's method is highly correlated to Gunning FOG Index, even if different in form, and that both the similarity-based approach and automated Taylor test achieve high accuracy. Merits of using either of them are discussed.

Fuzzy V-Measure - An Evaluation Method for Cluster Analyses of Ambiguous Data

Jason Utt, Sylvia Springorum, Maximilian Köper and Sabine Schulte im Walde

This paper discusses an extension of the V-measure (Rosenberg and Hirschberg, 2007), an entropy-based cluster evaluation metric. While the original work focused on evaluating hard clusterings, we introduce the Fuzzy V-measure which can be used on data that is inherently ambiguous. We perform multiple analyses varying the sizes and ambiguity rates and show that while entropy-based measures in general tend to suffer when ambiguity increases, a measure with desirable properties can be derived from these in a straightforward manner.

Finding a Tradeoff between Accuracy and Rater's Workload in Grading Clustered Short Answers

Andrea Horbach, Alexis Palmer and Magdalena Wolska

n this paper we investigate the potential of answer clustering for semi-automatic scoring of short answer questions for German as a foreign language. We use surface features like word and character n-grams to cluster answers to listening comprehension exercises per question and simulate having human graders only label one answer per cluster and then propagating this label to all other members of the cluster. We investigate various ways to select this single item to be labeled and find that choosing the item closest to the centroid of a cluster leads to improved (simulated) grading accuracy over random item selection. Averaged over all questions, we can reduce a teacher's workload to labeling only 40% of all different answers for a question, while still maintaining a grading accuracy of more than 85%.

Improving Evaluation of English-Czech MT through Paraphrasing

Petra Barancikova, Rudolf Rosa and Ales Tamchyna

In this paper, we present a method of improving the accuracy of machine translation evaluation of Czech sentences. Given a reference sentence, our algorithm transforms it by targeted paraphrasing into a new synthetic reference sentence that is closer in wording to the machine translation output, but at the same time preserves the meaning of the original reference sentence. Grammatical correctness of the new reference sentence is provided by applying Depfix on newly created paraphrases. Depfix is a system for post-editing English-to-Czech machine translation outputs. We adjusted it to fix the errors in paraphrased sentences. Due to a noisy source of our paraphrases, we experiment with adding word alignment. However, the alignment reduces the number of paraphrases found and the best results were achieved by a simple greedy method with only one-word paraphrases thanks to their intensive filtering. BLEU scores computed using these new reference sentences show significantly higher correlation with human judgment than scores computed on the original reference sentences.

On the Reliability and Inter-Annotator Agreement of Human Semantic MT Evaluation via HMEANT

Chi-kiu Lo and Dekai Wu

We present analyses showing that HMEANT is a reliable, accurate and fine-grained semantic frame-based human MT evaluation metric with high inter-annotator agreement (IAA) and correlation with human adequacy judgments, despite only requiring a minimal training of about 15 minutes for lay annotators. Previous work shows that the IAA on the semantic role labeling (SRL) subtask within HMEANT is over 70%. In this paper we focus on (1) the IAA on the semantic role alignment task and (2) the overall IAA of HMEANT. Our results show that the IAA on the alignment task of HMEANT is over 90% when humans align SRL output from the same SRL annotator, which shows that the instructions on the alignment task are sufficiently precise, although the overall IAA where humans align SRL output from different SRL annotators falls to only 61% due to the pipeline effect on the disagreement in the two annotation task. We show that instead of manually aligning the semantic roles using an automatic algorithm not only helps maintaining the overall IAA of HMEANT at 70%, but also provides a finer-grained assessment on the phrasal similarity of the semantic role fillers. This suggests that HMEANT equipped with automatic alignment is reliable and accurate for humans to evaluate MT adequacy while achieving higher correlation with human adequacy judgments than HTER.

P8 - Language Resource Infrastructures

Wednesday, May 28, 14:45

Chairperson: **Georg Rehm** Poster Session

The Evolving Infrastructure for Language Resources and the Role for Data Scientists

Nelleke Oostdijk and Henk van den Heuvel

In the context of ongoing developments as regards the creation of a sustainable, interoperable language resource infrastructure and spreading ideas of the need for open access, not only of research

publications but also of the underlying data, various issues present themselves which require that different stakeholders reconsider their positions. In the present paper we relate the experiences from the CLARIN-NL data curation service (DCS) over the two years that it has been operational, and the future role we envisage for expertise centres like the DCS in the evolving infrastructure.

Using TEI, CMDI and ISOcat in CLARIN-DK

Dorte Haltrup Hansen, Lene Offersgaard and Sussi Olsen

This paper presents the challenges and issues encountered in the conversion of TEI header metadata into the CMDI format. The work is carried out in the Danish research infrastructure, CLARIN-DK, in order to enable the exchange of language resources nationally as well as internationally, in particular with other partners of CLARIN ERIC. The paper describes the task of converting an existing TEI specification applied to all the text resources deposited in DK-CLARIN. During the task we have tried to reuse and share CMDI profiles and components in the CLARIN Component Registry, as well as linking the CMDI components and elements to the relevant data categories in the ISOcat Data Category Registry. The conversion of the existing metadata into the CMDI format turned out not to be a trivial task and the experience and insights gained from this work have resulted in a proposal for a work flow for future use. We also present a core TEI header metadata set.

ROOTS: a Toolkit for Easy, Fast and Consistent Processing of Large Sequential Annotated Data Collections

Jonathan Chevelu, Gwénolé Lecorvé and Damien Lolive

The development of new methods for given speech and natural language processing tasks usually consists in annotating large corpora of data before applying machine learning techniques to train models or to extract information. Beyond scientific aspects, creating and managing such annotated data sets is a recurrent problem. While using human annotators is obviously expensive in time and money, relying on automatic annotation processes is not a simple solution neither. Typically, the high diversity of annotation tools and of data formats, as well as the lack of efficient middleware to interface them all together, make such processes very complex and painful to design. To circumvent this problem, this paper presents the toolkit ROOTS, a freshly released open source toolkit (http://roots-toolkit.gforge.inria.fr) for easy, fast and consistent management of heterogeneously annotated data. ROOTS is designed to efficiently handle massive complex sequential data and to allow quick and light prototyping, as this is often required for research purposes. To illustrate these properties, three sample applications are presented in the field of speech and language processing, though ROOTS can more generally be easily extended to other application domains.

Sharing Cultural Heritage: the Clavius on the Web Project

Matteo Abrate, Angelo Mario del Grosso, Emiliano Giovannetti, Angelica Lo Duca, Damiana Luzzi, Lorenzo Mancini, Andrea Marchetti, Irene Pedretti and Silvia Piccini

In the last few years the amount of manuscripts digitized and made available on the Web has been constantly increasing. However, there is still a considarable lack of results concerning both the explicitation of their content and the tools developed to make it available. The objective of the Clavius on the Web project is to develop a Web platform exposing a selection of Christophorus Clavius letters along with three different levels of analysis: linguistic, lexical and semantic. The multilayered annotation of the corpus involves a XML-TEI encoding followed by a tokenization step where each token is univocally identified through a CTS urn notation and then associated to a part-of-speech and a lemma. The text is lexically and semantically annotated on the basis of a lexicon and a domain ontology, the former structuring the most relevant terms occurring in the text and the latter representing the domain entities of interest (e.g. people, places, etc.). Moreover, each entity is connected to linked and non linked resources, including DBpedia and VIAF. Finally, the results of the three layers of analysis are gathered and shown through interactive visualization and storytelling techniques. A demo version of the integrated architecture was developed.

'interHist' - an Interactive Visual Interface for Corpus Exploration

Verena Lyding, Lionel Nicolas and Egon Stemle

In this article, we present interHist, a compact visualization for the interactive exploration of results to complex corpus queries. Integrated with a search interface to the PAISA corpus of Italian web texts, interHist aims at facilitating the exploration of large results sets to linguistic corpus searches. This objective is approached by providing an interactive visual overview of the data, which supports the user-steered navigation by means of interactive filtering. It allows to dynamically switch between an overview on the data and a detailed view on results in their immediate textual context, thus helping to detect and inspect relevant hits more efficiently. We provide background information on corpus linguistics and related work on visualizations for language and linguistic data. We introduce the architecture of interHist, by detailing the data structure it relies on, describing the visualization design and providing technical details of the

implementation and its integration with the corpus querying environment. Finally, we illustrate its usage by presenting a use case for the analysis of the composition of Italian noun phrases.

P9 - Machine Translation

Wednesday, May 28, 14:45

Chairperson: **Jordi Atserias** Poster Session

Constructing a Chinese–Japanese Parallel Corpus from Wikipedia

Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi

Parallel corpora are crucial for statistical machine translation (SMT). However, they are quite scarce for most language pairs, such as Chinese–Japanese. As comparable corpora are far more available, many studies have been conducted to automatically construct parallel corpora from comparable corpora. This paper presents a robust parallel sentence extraction system for constructing a Chinese–Japanese parallel corpus from Wikipedia. The system is inspired by previous studies that mainly consist of a parallel sentence candidate filter and a binary classifier for parallel sentence identification. We improve the system by using the common Chinese characters for filtering and two novel feature sets for classification. Experiments show that our system performs significantly better than the previous studies for both accuracy in parallel sentence extraction and SMT performance. Using the system, we construct a Chinese–Japanese parallel corpus with more than 126k highly accurate parallel sentences from Wikipedia. The constructed parallel corpus is freely available at http://orchid.kuee.kyoto-u.ac.jp/ chu/resource/wiki_ zh_ ja.tgz.

An Iterative Approach for Mining Parallel Sentences in a Comparable Corpus

Lise Rebout and Phillippe Langlais

We describe an approach for mining parallel sentences in a collection of documents in two languages. While several approaches have been proposed for doing so, our proposal differs in several respects. First, we use a document level classifier in order to focus on potentially fruitful document pairs, an understudied approach. We show that mining less, but more parallel documents can lead to better gains in machine translation. Second, we compare different strategies for post-processing the output of a classifier trained to recognize parallel sentences. Last, we report a simple bootstrapping experiment which shows that promising sentence pairs extracted in a first stage can help to mine new sentence pairs in a second stage. We applied our approach on the English-French Wikipedia. Gains of a statistical machine translation (SMT) engine are analyzed along different test sets.

Large SMT data-sets extracted from Wikipedia

Dan Tufiş

The article presents experiments on mining Wikipedia for extracting SMT useful sentence pairs in three language pairs. Each extracted sentence pair is associated with a cross-lingual lexical similarity score based on which, several evaluations have been conducted to estimate the similarity thresholds which allow the extraction of the most useful data for training three-language pairs SMT systems. The experiments showed that for a similarity score higher than 0.7 all sentence pairs in the three language pairs were fully parallel. However, including in the training sets less parallel sentence pairs (that is with a lower similarity score) showed significant improvements in the translation quality (BLEU-based evaluations). The optimized SMT systems were evaluated on unseen test-sets also extracted from Wikipedia. As one of the main goals of our work was to help Wikipedia contributors to translate (with as little post editing as possible) new articles from major languages into less resourced languages and vice-versa, we call this type of translation experiments "in-genre" translation. As in the case of "in-domain" translation, our evaluations showed that using only "in-genre" training data for translating same genre new texts is better than mixing the training data with "out-of-genre" (even) parallel texts.

Production of Phrase Tables in 11 European Languages using an Improved Sub-sentential Aligner

Juan Luo and Yves Lepage

This paper is a partial report of an on-going Kakenhi project which aims to improve sub-sentential alignment and release multilingual syntactic patterns for statistical and example-based machine translation. Here we focus on improving a sub-sentential aligner which is an instance of the association approach. Phrase table is not only an essential component in the machine translation systems but also an important resource for research and usage in other domains. As part of this project, all phrase tables produced in the experiments will also be made freely available.

Collection of a Simultaneous Translation Corpus for Comparative Analysis

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda and Satoshi Nakamura

This paper describes the collection of an English-Japanese/Japanese-English simultaneous interpretation corpus. There are two main features of the corpus. The first is that

professional simultaneous interpreters with different amounts of experience cooperated with the collection. By comparing data from simultaneous interpretation of each interpreter, it is possible to compare better interpretations to those that are not as good. The second is that for part of our corpus there are already translation data available. This makes it possible to compare translation data with simultaneous interpretation data. We recorded the interpretations of lectures and news, and created time-aligned transcriptions. A total of 387k words of transcribed data were collected. The corpus will be helpful to analyze differences in interpretations styles and to construct simultaneous interpretation systems.

English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling

Sharid Loaiciga, Thomas Meyer and Andrei Popescu-Belis

This paper presents a method for verb phrase (VP) alignment in an English-French parallel corpus and its use for improving statistical machine translation (SMT) of verb tenses. The method starts from automatic word alignment performed with GIZA++, and relies on a POS tagger and a parser, in combination with several heuristics, in order to identify non-contiguous components of VPs, and to label the aligned VPs with their tense and voice on each side. This procedure is applied to the Europarl corpus, leading to the creation of a smaller, high-precision parallel corpus with about 320,000 pairs of finite VPs, which is made publicly available. This resource is used to train a tense predictor for translation from English into French, based on a large number of surface features. Three MT systems are compared: (1) a baseline phrase-based SMT; (2) a tense-aware SMT system using the above predictions within a factored translation model; and (3) a system using oracle predictions from the aligned VPs. For several tenses, such as the French "imparfait", the tense-aware SMT system improves significantly over the baseline and is closer to the oracle system.

Two-Step Machine Translation with Lattices

Bushra Jawaid and Ondrej Bojar

The idea of two-step machine translation was introduced to divide the complexity of the search space into two independent steps: (1) lexical translation and reordering, and (2) conjugation and declination in the target language. In this paper, we extend the two-step machine translation structure by replacing state-of-the-art phrase-based machine translation with the hierarchical machine translation in the 1st step. We further extend the fixed string-based input format of the 2nd step with word lattices (Dyer et al., 2008); this provides the 2nd step with the opportunity to choose among a sample of possible reorderings instead of relying on the single best one as produced by the 1st step.

P10 - Metadata

Wednesday, May 28, 14:45

Chairperson: **Victoria Arranz** Poster Session

The CMD Cloud

Matej Durco and Menzo Windhouwer

The CLARIN Component Metadata Infrastructure (CMDI) established means for flexible resource descriptions for the domain of language resources with sound provisions for semantic interoperability weaved deeply into the meta model and the infrastructure. Based on this solid grounding, the infrastructure accommodates a growing collection of metadata records. In this paper, we give a short overview of the current status in the CMD data domain on the schema and instance level and harness the installed mechanisms for semantic interoperability to explore the similarity relations between individual profiles/schemas. We propose a method to use the semantic links shared among the profiles to generate/compile a similarity graph. This information is further rendered in an interactive graph viewer: the SMC Browser. The resulting interactive graph offers an intuitive view on the complex interrelations of the discussed dataset revealing clusters of more similar profiles. This information is useful both for metadata modellers, for metadata curation tasks as well as for general audience seeking for a 'big picture' of the complex CMD data domain.

The eIdentity Text ExplorationWorkbench

Fritz Kliche, Andre Blessing, Dr. Ulrich Heid and Jonathan Sonntag

We work on tools to explore text contents and metadata of newspaper articles as provided by news archives. Our tool components are being integrated into an "Exploration Workbench" for Digital Humanities researchers. Next to the conversion of different data formats and character encodings, a prominent feature of our design is its "Wizard" function for corpus building: Researchers import raw data and define patterns to extract text contents and metadata. The Workbench also comprises different tools for data cleaning. These include filtering of off-topic articles, duplicates and near-duplicates, corrupted and empty articles. We currently work on ca. 860.000 newspaper articles from different media archives, provided in different data formats. We index the data with state-of-the-art systems to allow for large scale information retrieval. We extract metadata on publishing dates, author names, newspaper sections, etc., and split articles into

segments such as headlines, subtitles, paragraphs, etc. After cleaning the data and compiling a thematically homogeneous corpus, the sample can be used for quantitative analyses which are not affected by noise. Users can retrieve sets of articles on different topics, issues or otherwise defined research questions ("subcorpora") and investigate quantitatively their media attention on the timeline ("Issue Cycles").

Visualization of Language Relations and Families: MultiTree

Damir Cavar and Malgorzata Cavar

MultiTree is an NFS-funded project collecting scholarly hypotheses about language relationships, and visualizing them on a web site in the form of trees or graphs. Two open online interfaces allow scholars, students, and the general public an easy access to search for language information or comparisons of competing hypotheses. One objective of the project was to facilitate research in historical linguistics. MultiTree has evolved to a much more powerful tool, it is not just a simple repository of scholarly information. In this paper we present the MultiTree interfaces and the impact of the project beyond the field of historical linguistics, including, among others, the use of standardized ISO language codes, and creating an interconnected database of language and dialect names, codes, publications, and authors. Further, we offer the dissemination of linguistic findings world-wide to both scholars and the general public, thus boosting the collaboration and accelerating the scientific exchange. We discuss also the ways MultiTree will develop beyond the time of the duration of the funding.

P11 - MultiWord Expressions and Terms

Wednesday, May 28, 14:45

Chairperson: **Valeria Quochi** Poster Session

Linked Open Data and Web Corpus Data for noun compound bracketing

Pierre André Ménard and Caroline Barriere

This research provides a comparison of a linked open data resource (DBpedia) and web corpus data resources (Google Web Ngrams and Google Books Ngrams) for noun compound bracketing. Large corpus statistical analysis has often been used for noun compound bracketing, and our goal is to introduce a linked open data (LOD) resource for such task. We show its particularities and its performance on the task. Results obtained on resources tested individually are promising, showing a potential for DBpedia to be included in future hybrid systems.

4FX: Light Verb Constructions in a Multilingual Parallel Corpus

Anita Rácz, István Nagy T. and Veronika Vincze

In this paper, we describe 4FX, a quadrilingual (English-Spanish-German-Hungarian) parallel corpus annotated for light verb constructions. We present the annotation process, and report statistical data on the frequency of LVCs in each language. We also offer inter-annotator agreement rates and we highlight some interesting facts and tendencies on the basis of comparing multilingual data from the four corpora. According to the frequency of LVC categories and the calculated Kendall's coefficient for the four corpora, we found that Spanish and German are very similar to each other, Hungarian is also similar to both, but German differs from all these three. The qualitative and quantitative data analysis might prove useful in theoretical linguistic research for all the four languages. Moreover, the corpus will be an excellent testbed for the development and evaluation of machine learning-based methods aiming at extracting or identifying light verb constructions in these four languages.

Identifying Idioms in Chinese Translations

Wan Yu Ho, Christine Kng, Shan Wang and Francis Bond

Optimally, a translated text should preserve information while maintaining the writing style of the original. When this is not possible, as is often the case with figurative speech, a common practice is to simplify and make explicit the implications. However, in our investigations of translations from English to another language, English-to-Chinese texts were often found to include idiomatic expressions (usually in the form of Chengyu) where there were originally no idiomatic, metaphorical, or even figurative expressions. We have created an initial small lexicon of Chengyu, with which we can use to find all occurrences of Chengyu in a given corpus, and will continue to expand the database. By examining the rates and patterns of occurrence across four genres in the NTU Multilingual Corpus, a resource may be created to aid machine translation or, going further, predict Chinese translational trends in any given genre.

Narrowing the Gap Between Termbases and Corpora in Commercial Environments

Kara Warburton

Terminological resources offer potential to support applications beyond translation, such as controlled authoring and indexing, which are increasingly of interest to commercial enterprises. The ad-hoc semasiological approach adopted by commercial

terminographers diverges considerably from methodologies prescribed by conventional theory. The notion of termhood in such production-oriented environments is driven by pragmatic criteria such as frequency and repurposability of the terminological unit. A high degree of correspondence between the commercial corpus and the termbase is desired. Research carried out at the City University of Hong Kong using four IT companies as case studies revealed a large gap between corpora and termbases. Problems in selecting terms and in encoding them properly in termbases account for a significant portion of this gap. A rigorous corpus-based approach to term selection would significantly reduce this gap and improve the effectiveness of commercial termbases. In particular, single-word terms (keywords) identified by comparison to a reference corpus offer great potential for identifying important multi-word terms in this context. We conclude that terminography for production purposes should be more corpus-based than is currently the norm.

Identification of Multiword Expressions in the brWaC

Rodrigo Boos, Kassius Prestes and Aline Villavicencio

Although corpus size is a well known factor that affects the performance of many NLP tasks, for many languages large freely available corpora are still scarce. In this paper we describe one effort to build a very large corpus for Brazilian Portuguese, the brWaC, generated following the Web as Corpus kool yinitiative. To indirectly assess the quality of the resulting corpus we examined the impact of corpus origin in a specific task, the identification of Multiword Expressions with association measures, against a standard corpus. Focusing on nominal compounds, the expressions obtained from each corpus are of comparable quality and indicate that corpus origin has no impact on this task.

Collocation or Free Combination? – Applying Machine Translation Techniques to identify collocations in Japanese

Lis Pereira, Elga Strafella and Yuji Matsumoto

This work presents an initial investigation on how to distinguish collocations from free combinations. The assumption is that, while free combinations can be literally translated, the overall meaning of collocations is different from the sum of the translation of its parts. Based on that, we verify whether a machine translation system can help us perform such distinction. Results show that it improves the precision compared with standard methods of collocation identification through statistical association measures.

Building a Crisis Management Term Resource for Social Media: The Case of Floods and Protests

Irina Temnikova, Andrea Varga and Dogan Biyikli

Extracting information from social media is being currently exploited for a variety of tasks, including the recognition of emergency events in Twitter. This is done in order to supply Crisis Management agencies with additional crisis information. The existing approaches, however, mostly rely on geographic location and hashtags/keywords, obtained via a manual Twitter search. As we expect that Twitter crisis terminology would differ from existing crisis glossaries, we start collecting a specialized terminological resource to support this task. The aim of this resource is to contain sets of crisis-related Twitter terms which are the same for different instances of the same type of event. This article presents a preliminary investigation of the nature of terms used in four events of two crisis types, tests manual and automatic ways to collect these terms and comes up with an initial collection of terms for these two types of events. As contributions, a novel annotation schema is presented, along with important insights into the differences in annotations between different specialists, descriptive term statistics, and performance results of existing automatic terminology recognition approaches for this task.

P12 - Treebanks

Wednesday, May 28, 14:45

Chairperson: **Beatrice Daille** Poster Session

Towards Building a Kashmiri Treebank: Setting up the Annotation Pipeline

Riyaz Ahmad Bhat, Shahid Musjtaq Bhat and Dipti Misra Sharma

Kashmiri is a resource poor language with very less computational and language resources available for its text processing. As the main contribution of this paper, we present an initial version of the Kashmiri Dependency Treebank. The treebank consists of 1,000 sentences (17,462 tokens), annotated with part-of-speech (POS), chunk and dependency information. The treebank has been manually annotated using the Paninian Computational Grammar (PCG) formalism (Begum et al., 2008; Bharati et al., 2009). This version of Kashmiri treebank is an extension of its earlier verion of 500 sentences (Bhat, 2012), a pilot experiment aimed at defining the annotation guidelines on a small subset of Kashmiri corpora. In this paper, we have refined the guidelines with some significant changes and have carried out inter-annotator

agreement studies to ascertain its quality. We also present a dependency parsing pipeline, consisting of a tokenizer, a stemmer, a POS tagger, a chunker and an inter-chunk dependency parser. It, therefore, constitutes the first freely available, open source dependency parser of Kashmiri, setting the initial baseline for Kashmiri dependency parsing.

A Japanese Word Dependency Corpus

Shinsuke Mori, Hideki Ogura and Tetsuro Sasada

In this paper, we present a corpus annotated with dependency relationships in Japanese. It contains about 30 thousand sentences in various domains. Six domains in Balanced Corpus of Contemporary Written Japanese have part-of-speech and pronunciation annotation as well. Dictionary example sentences have pronunciation annotation and cover basic vocabulary in Japanese with English sentence equivalent. Economic newspaper articles also have pronunciation annotation and the topics are similar to those of Penn Treebank. Invention disclosures do not have other annotation, but it has a clear application, machine translation. The unit of our corpus is word like other languages contrary to existing Japanese corpora whose unit is phrase called bunsetsu. Each sentence is manually segmented into words. We first present the specification of our corpus. Then we give a detailed explanation about our standard of word dependency. We also report some preliminary results of an MST-based dependency parser on our corpus.

A Compact Interactive Visualization of Dependency Treebank Query Results

Chris Culy, Marco Passarotti and Ulla König-Cardanobile

One of the challenges of corpus querying is making sense of the results of a query, especially when a large number of results and linguistically annotated data are concerned. While the most widespread tools for querying syntactically annotated corpora tend to focus on single occurrences, one aspect that is not fully exploited yet in this area is that language is a complex system whose units are connected to each other at both microscopic (the single occurrences) and macroscopic level (the whole system itself). Assuming that language is a system, we describe a tool (using the DoubleTreeJS visualization) to visualize the results of querying dependency treebanks by forming a node from a single item type, and building a network in which the heads and the dependents of the central node are respectively the left and the right vertices of the tree, which are connected to the central node by dependency relations. One case study is presented, consisting in the exploitation of DoubleTreeJS for supporting one assumption in theoretical linguistics with evidence provided by the data of a dependency treebank of Medieval Latin.

Thomas Aquinas in the TüNDRA: Integrating the Index Thomisticus Treebank into CLARIN-D

Scott Martens and Marco Passarotti

This paper describes the integration of the Index Thomisticus Treebank (IT-TB) into the web-based treebank search and visualization application TueNDRA (Tuebingen aNnotated Data Retrieval & Analysis). TueNDRA was originally designed to provide access via the Internet to constituency treebanks and to tools for searching and visualizing them, as well as tabulating statistics about their contents. TueNDRA has now been extended to also provide full support for dependency treebanks with non-projective dependencies, in order to integrate the IT-TB and future treebanks with similar properties. These treebanks are queried using an adapted form of the TIGERSearch query language, which can search both hierarchical and sequential information in treebanks in a single query. As a web application, making the IT-TB accessible via TueNDRA makes the treebank and the tools to use of it available to a large community without having to distribute software and show users how to install it.

Boosting the Creation of a Treebank

Blanca Arias, Núria Bel, Mercè Lorente, Montserrat Marimón, Alba Milà, Jorge Vivaldi, Muntsa Padró, Marina Fomicheva and Imanol Larrea

In this paper we present the results of an ongoing experiment of bootstrapping a Treebank for Catalan by using a Dependency Parser trained with Spanish sentences. In order to save time and cost, our approach was to profit from the typological similarities between Catalan and Spanish to create a first Catalan data set quickly by automatically: (i) annotating with a de-lexicalized Spanish parser, (ii) manually correcting the parses, and (iii) using the Catalan corrected sentences to train a Catalan parser. The results showed that the number of parsed sentences required to train a Catalan parser is about 1000 that were achieved in 4 months, with 2 annotators.

The IULA Spanish LSP Treebank

Montserrat Marimón, Núria Bel, Beatriz Fisas, Blanca Arias, Silvia Vázquez, Jorge Vivaldi, Carlos Morell and Mercè Lorente

This paper presents the IULA Spanish LSP Treebank, a dependency treebank of over 41,000 sentences of different domains (Law, Economy, Computing Science, Environment, and Medicine), developed in the framework of the European project

METANET4U. Dependency annotations in the treebank were automatically derived from manually selected parses produced by an HPSG-grammar by a deterministic conversion algorithm that used the identifiers of grammar rules to identify the heads, the dependents, and some dependency types that were directly transferred onto the dependency structure (e.g., subject, specifier, and modifier), and the identifiers of the lexical entries to identify the argument-related dependency functions (e.g. direct object, indirect object, and oblique complement). The treebank is accessible with a browser that provides concordance-based search functions and delivers the results in two formats: (i) a column-based format, in the style of CoNLL-2006 shared task, and (ii) a dependency graph, where dependency relations are noted by an oriented arrow which goes from the dependent node to the head node. The IULA Spanish LSP Treebank is the first technical corpus of Spanish annotated at surface syntactic level following the dependency grammar theory. The treebank has been made publicly and freely available from the META-SHARE platform with a Creative Commons CC-by licence.

The Norwegian Dependency Treebank

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen and Janne Bondi Johannessen

The Norwegian Dependency Treebank is a new syntactic treebank for Norwegian Bokmål and Nynorsk with manual syntactic and morphological annotation, developed at the National Library of Norway in collaboration with the University of Oslo. It is the first publically available treebank for Norwegian. This paper presents the core principles behind the syntactic annotation and how these principles were employed in certain specific cases. We then present the selection of texts and distribution between genres, as well as the annotation process and an evaluation of the inter-annotator agreement. Finally, we present the first results of data-driven dependency parsing of Norwegian, contrasting four state-of-the-art dependency parsers trained on the treebank. The consistency and the parsability of this treebank is shown to be comparable to other large treebank initiatives.

A Persian Treebank with Stanford Typed Dependencies

Mojgan Seraji, Carina Jahani, Beáta Megyesi and Joakim Nivre

We present the Uppsala Persian Dependency Treebank (UPDT) with a syntactic annotation scheme based on Stanford Typed Dependencies. The treebank consists of 6,000 sentences and 151,671 tokens with an average sentence length of 25 words. The data is from different genres, including newspaper articles and fiction, as well as technical descriptions and texts about culture and art, taken from the open source Uppsala Persian Corpus (UPC). The syntactic annotation scheme is extended for Persian to include all syntactic relations that could not be covered by the primary scheme developed for English. In addition, we present open source tools for automatic analysis of Persian containing a text normalizer, a sentence segmenter and tokenizer, a part-of-speech tagger, and a parser. The treebank and the parser have been developed simultaneously in a bootstrapping procedure. The result of a parsing experiment shows an overall labeled attachment score of 82.05% and an unlabeled attachment score of 85.29%. The treebank is freely available as an open source resource.

Converting an HPSG-based Treebank into its Parallel Dependency-based Treebank

Masood Ghayoomi and Jonas Kuhn

A treebank is an important language resource for supervised statistical parsers. The parser induces the grammatical properties of a language from this language resource and uses the model to parse unseen data automatically. Since developing such a resource is very time-consuming and tedious, one can take advantage of already extant resources by adapting them to a particular application. This reduces the amount of human effort required to develop a new language resource. In this paper, we introduce an algorithm to convert an HPSG-based treebank into its parallel dependency-based treebank. With this converter, we can automatically create a new language resource from an existing treebank developed based on a grammar formalism. Our proposed algorithm is able to create both projective and non-projective dependency trees.

O9 - Sentiment Analysis and Social Media (1)

Wednesday, May 28, 16:45

Chairperson: **Stelios Piperidis** Oral Session

On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter

Hassan Saif, Miriam Fernandez, Yulan He and Harith Alani

Sentiment classification over Twitter is usually affected by the noisy nature (abbreviations, irregular forms) of tweets data. A popular procedure to reduce the noise of textual data is to remove stopwords by using pre-compiled stopword lists or more sophisticated methods for dynamic stopword identification. However, the effectiveness of removing stopwords in the context of Twitter sentiment classification has been debated in the last few years. In this paper we investigate whether removing stopwords helps or hampers the effectiveness of Twitter sentiment

classification methods. To this end, we apply six different stopword identification methods to Twitter data from six different datasets and observe how removing stopwords affects two well-known supervised sentiment classification methods. We assess the impact of removing stopwords by observing fluctuations on the level of data sparsity, the size of the classifier's feature space and its classification performance. Our results show that using pre-compiled lists of stopwords negatively impacts the performance of Twitter sentiment classification approaches. On the other hand, the dynamic generation of stopword lists, by removing those infrequent terms appearing only once in the corpus, appears to be the optimal method to maintaining a high classification performance while reducing the data sparsity and shrinking the feature space.

Investigating the Image of Entities in Social Media: Dataset Design and First Results

Julien Velcin, Young-Min Kim, Caroline Brun, Jean-Yves Dormagen, Eric SanJuan, Leila Khouas, Anne Peradotto, Stéphane Bonnevay, Claude Roux, Julien Boyadjian, Alejandro Molina and Marie Neihouser

The objective of this paper is to describe the design of a dataset that deals with the image (i.e., representation, web reputation) of various entities populating the Internet: politicians, celebrities, companies, brands etc. Our main contribution is to build and provide an original annotated French dataset. This dataset consists of 11527 manually annotated tweets expressing the opinion on specific facets (e.g., ethic, communication, economic project) describing two French policitians over time. We believe that other researchers might benefit from this experience, since designing and implementing such a dataset has proven quite an interesting challenge. This design comprises different processes such as data selection, formal definition and instantiation of an image. We have set up a full open-source annotation platform. In addition to the dataset design, we present the first results that we obtained by applying clustering methods to the annotated dataset in order to extract the entity images.

Benchmarking Twitter Sentiment Analysis Tools

Ahmed Abbasi, Ammar Hassan and Milan Dhar

Twitter has become one of the quintessential social media platforms for user-generated content. Researchers and industry practitioners are increasingly interested in Twitter sentiments. Consequently, an array of commercial and freely available Twitter sentiment analysis tools have emerged, though it remains unclear how well these tools really work. This study presents the findings of a detailed benchmark analysis of Twitter sentiment analysis tools, incorporating 20 tools applied to 5 different test beds. In addition to presenting detailed performance evaluation results, a thorough error analysis is used to highlight the most prevalent challenges facing Twitter sentiment analysis tools. The results have important implications for various stakeholder groups, including social media analytics researchers, NLP developers, and industry managers and practitioners using social media sentiments as input for decision-making.

Recognising Suicidal Messages in Dutch Social Media

Bart Desmet and Véronique Hoste

Early detection of suicidal thoughts is an important part of effective suicide prevention. Such thoughts may be expressed online, especially by young people. This paper presents on-going work on the automatic recognition of suicidal messages in social media. We present experiments for automatically detecting relevant messages (with suicide-related content), and those containing suicide threats. A sample of 1357 texts was annotated in a corpus of 2674 blog posts and forum messages from Netlog, indicating relevance, origin, severity of suicide threat and risks as well as protective factors. For the classification experiments, Naive Bayes, SVM and KNN algorithms are combined with shallow features, i.e. bag-of-words of word, lemma and character ngrams, and post length. The best relevance classification is achieved by using SVM with post length, lemma and character ngrams, resulting in an F-score of 85.6% (78.7% precision and 93.8% recall). For the second task (threat detection), a cascaded setup which first filters out irrelevant messages with SVM and then predicts the severity with KNN, performs best: 59.2% F-score (69.5% precision and 51.6% recall).

O10 - Conversational (1)

Wednesday, May 28, 16:45

Chairperson: **Nick Campbell** Oral Session

Automatic Detection of Other-Repetition Occurrences: Application to French Conversational Speech

Brigitte Bigi, Roxane Bertrand and Mathilde Guardiola

This paper investigates the discursive phenomenon called other-repetitions (OR), particularly in the context of spontaneous French dialogues. It focuses on their automatic detection and characterization. A method is proposed to retrieve automatically OR: this detection is based on rules that are applied on the lexical material only. This automatic detection process has been used to label other-repetitions on 8 dialogues of CID - Corpus of Interactional Data. Evaluations performed on one speaker are

good with a F1-measure of 0.85. Retrieved OR occurrences are then statistically described: number of words, distance, etc.

ANCOR_ Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol and Jeanne Villaneau

This article presents ANCOR_ Centre, a French coreference corpus, available under the Creative Commons Licence. With a size of around 500,000 words, the corpus is large enough to serve the needs of data-driven approaches in NLP and represents one of the largest coreference resources currently available. The corpus focuses exclusively on spoken language, it aims at representing a certain variety of spoken genders. ANCOR_ Centre includes anaphora as well as coreference relations which involve nominal and pronominal mentions. The paper describes into details the annotation scheme and the reliability measures computed on the resource.

Phone Boundary Annotation in Conversational Speech

Yi-Fen Liu, Shu-Chuan Tseng and J.-S Roger Jang

Phone-aligned spoken corpora are indispensable language resources for quantitative linguistic analyses and automatic speech systems. However, producing this type of data resources is not an easy task due to high costs of time and man power as well as difficulties of applying valid annotation criteria and achieving reliable inter-labeler's consistency. Among different types of spoken corpora, conversational speech that is often filled with extreme reduction and varying pronunciation variants is particularly challenging. By adopting a combined verification procedure, we obtained reasonably good annotation results. Preliminary phone boundaries that were automatically generated by a phone aligner were provided to human labelers for verifying. Instead of making use of the visualization of acoustic cues, the labelers should solely rely on their perceptual judgments to locate a position that best separates two adjacent phones. Impressionistic judgments in cases of reduction and segment deletion were helpful and necessary, as they balanced subtle nuance caused by differences in perception.

Automatically Enriching Spoken Corpora with Syntactic Information for Linguistic Studies

Alexis Nasr, Frédéric Béchet , Benoit Favre, Thierry Bazillon, Jose Deulofeu and Andre Valli

Syntactic parsing of speech transcriptions faces the problem of the presence of disfluencies that break the syntactic structure of the utterances. We propose in this paper two solutions to this problem. The first one relies on a disfluencies predictor that detects disfluencies and removes them prior to parsing. The second one integrates the disfluencies in the syntactic structure of the utterances and train a disfluencies aware parser.

O11 - Collaborative Resources (1)

Wednesday, May 28, 16:45

Chairperson: **Iryna Gurevych** Oral Session

Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines

Marta Sabou, Kalina Bontcheva, Leon Derczynski and Arno Scharl

Crowdsourcing is an emerging collaborative approach that can be used for the acquisition of annotated corpora and a wide range of other linguistic resources. Although the use of this approach is intensifying in all its key genres (paid-for crowdsourcing, games with a purpose, volunteering-based approaches), the community still lacks a set of best-practice guidelines similar to the annotation best practices for traditional, expert-based corpus acquisition. In this paper we focus on the use of crowdsourcing methods for corpus acquisition and propose a set of best practice guidelines based in our own experiences in this area and an overview of related literature. We also introduce GATE Crowd, a plugin of the GATE platform that relies on these guidelines and offers tool support for using crowdsourcing in a more principled and efficient manner.

Towards an Environment for the Production and the Validation of Lexical Semantic Resources

Mikaël Morardo and Eric de La Clergerie

We present the components of a processing chain for the creation, visualization, and validation of lexical resources (formed of terms and relations between terms). The core of the chain is a component for building lexical networks relying on Harris' distributional hypothesis applied on the syntactic dependencies produced by the French parser FRMG on large corpora. Another important aspect concerns the use of an online interface for the visualization and collaborative validation of the resulting resources.

Towards an Encyclopedia of Compositional Semantics: Documenting the Interface of the English Resource Grammar

Dan Flickinger, Emily M. Bender and Stephan Oepen

We motivate and describe the design and development of an emerging encyclopedia of compositional semantics, pursuing

three objectives. We first seek to compile a comprehensive catalogue of interoperable semantic analyses, i.e., a precise characterization of meaning representations for a broad range of common semantic phenomena. Second, we operationalize the discovery of semantic phenomena and their definition in terms of what we call their semantic fingerprint, a formal account of the building blocks of meaning representation involved and their configuration. Third, we ground our work in a carefully constructed semantic test suite of minimal exemplars for each phenomenon, along with a 'target' fingerprint that enables automated regression testing. We work towards these objectives by codifying and documenting the body of knowledge that has been constructed in a long-term collaborative effort, the development of the LinGO English Resource Grammar. Documentation of its semantic interface is a prerequisite to use by non-experts of the grammar and the analyses it produces, but this effort also advances our own understanding of relevant interactions among phenomena, as well as of areas for future work in the grammar.

Mapping CPA Patterns onto OntoNotes Senses

Octavian Popescu, Martha Palmer and Patrick Hanks

In this paper we present an alignment experiment between patterns of verb use discovered by Corpus Pattern Analysis (CPA; Hanks 2004, 2008, 2012) and verb senses in OntoNotes (ON; Hovy et al. 2006, Weischedel et al. 2011). We present a probabilistic approach for mapping one resource into the other. Firstly we introduce a basic model, based on conditional probabilities, which determines for any given sentence the best CPA pattern match. On the basis of this model, we propose a joint source channel model (JSCM) that computes the probability of compatibility of semantic types between a verb phrase and a pattern, irrespective of whether the verb phrase is a norm or an exploitation. We evaluate the accuracy of the proposed mapping using cluster similarity metrics based on entropy.

O12 - Semantics (1)

Wednesday, May 28, 16:45

Chairperson: **Eva Hajičová** Oral Session

T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini and Octavian Popescu

The goal of this paper is to introduce T-PAS, a resource of typed predicate argument structures for Italian, acquired from corpora by manual clustering of distributional information about Italian verbs, to be used for linguistic analysis and semantic processing tasks. T-PAS is the first resource for Italian in which semantic selection properties and sense-in-context distinctions of verbs are characterized fully on empirical ground. In the paper, we first describe the process of pattern acquisition and corpus annotation (section 2) and its ongoing evaluation (section 3). We then demonstrate the benefits of pattern tagging for NLP purposes (section 4), and discuss current effort to improve the annotation of the corpus (section 5). We conclude by reporting on ongoing experiments using semiautomatic techniques for extending coverage (section 6).

The N2 Corpus: a Semantically Annotated Collection of Islamist Extremist Stories

Mark Finlayson, Jeffry Halverson and Steven Corman

We describe the N2 (Narrative Networks) Corpus, a new language resource. The corpus is unique in three important ways. First, every text in the corpus is a story, which is in contrast to other language resources that may contain stories or story-like texts, but are not specifically curated to contain only stories. Second, the unifying theme of the corpus is material relevant to Islamist Extremists, having been produced by or often referenced by them. Third, every text in the corpus has been annotated for 14 layers of syntax and semantics, including: referring expressions and co-reference; events, time expressions, and temporal relationships; semantic roles; and word senses. In cases where analyzers were not available to do high-quality automatic annotations, layers were manually double-annotated and adjudicated by trained annotators. The corpus comprises 100 texts and 42,480 words. Most of the texts were originally in Arabic but all are provided in English translation. We explain the motivation for constructing the corpus, the process for selecting the texts, the detailed contents of the corpus itself, the rationale behind the choice of annotation layers, and the annotation procedure.

Predicate Matrix: Extending SemLink through WordNet mappings

Maddalen Lopez de Lacalle, Egoitz Laparra and German Rigau

This paper presents the Predicate Matrix v1.1, a new lexical resource resulting from the integration of multiple sources of predicate information including FrameNet, VerbNet, PropBank and WordNet. We start from the basis of SemLink. Then, we use advanced graph-based algorithms to further extend the mapping coverage of SemLink. Second, we also exploit the current content of SemLink to infer new role mappings among the different predicate schemas. As a result, we have obtained a new version of

the Predicate Matrix which largely extends the current coverage of SemLink and the previous version of the Predicate Matrix.

A Unified Annotation Scheme for the Semantic/Pragmatic Components of Definiteness

Archna Bhatia, Mandy Simons, Lori Levin, Yulia Tsvetkov, Chris Dyer and Jordan Bender

We present a definiteness annotation scheme that captures the semantic, pragmatic, and discourse information, which we call communicative functions, associated with linguistic descriptions such as "a story about my speech", "the story", "every time I give it", "this slideshow". A survey of the literature suggests that definiteness does not express a single communicative function but is a grammaticalization of many such functions, for example, identifiability, familiarity, uniqueness, specificity. Our annotation scheme unifies ideas from previous research on definiteness while attempting to remove redundancy and make it easily annotatable. This annotation scheme encodes the communicative functions of definiteness rather than the grammatical forms of definiteness. We assume that the communicative functions are largely maintained across languages while the grammaticalization of this information may vary. One of the final goals is to use our semantically annotated corpora to discover how definiteness is grammaticalized in different languages. We release our annotated corpora for English and Hindi, and sample annotations for Hebrew and Russian, together with an annotation manual.

P13 - Discourse Annotation, Representation and Processing

Wednesday, May 28, 16:45

Chairperson: **Ann Bies** Poster Session

A Model for Processing Illocutionary Structures and Argumentation in Debates

Kasia Budzynska, Mathilde Janier, Chris Reed, Patrick Ssaint-dizier, Manfred Stede and Olena Yakorska

In this paper, we briefly present the objectives of Inference Anchoring Theory (IAT) and the formal structure which is proposed for dialogues. Then, we introduce our development corpus, and a computational model designed for the identification of discourse minimal units in the context of argumentation and the illocutionary force associated with each unit. We show the categories of resources which are needed and how they can be reused in different contexts.

Potsdam Commentary Corpus 2.0: Annotation for Discourse Research

Manfred Stede and Arne Neumann

We present a revised and extended version of the Potsdam Commentary Corpus, a collection of 175 German newspaper commentaries (op-ed pieces) that has been annotated with syntax trees and three layers of discourse-level information: nominal coreference,connectives and their arguments (similar to the PDTB, Prasad et al. 2008), and trees reflecting discourse structure according to Rhetorical Structure Theory (Mann/Thompson 1988). Connectives have been annotated with the help of a semi-automatic tool, Conano (Stede/Heintze 2004), which identifies most connectives and suggests arguments based on their syntactic category. The other layers have been created manually with dedicated annotation tools. The corpus is made available on the one hand as a set of original XML files produced with the annotation tools, based on identical tokenization. On the other hand, it is distributed together with the open-source linguistic database ANNIS3 (Chiarcos et al. 2008; Zeldes et al. 2009), which provides multi-layer search functionality and layer-specific visualization modules. This allows for comfortable qualitative evaluation of the correlations between annotation layers.

Verbs of Saying with a Textual Connecting Function in the Prague Discourse Treebank

Magdalena Rysova

The paper tries to contribute to the general discussion on discourse connectives, concretely to the question whether it is meaningful to distinguish two separate groups of connectives – i.e. "classical" connectives limited to few predefined classes like conjunctions or adverbs (e.g. "but") vs. alternative lexicalizations of connectives (i.e. unrestricted expressions and phrases like "the reason is", "he added", "the condition was" etc.). In this respect, the paper focuses on one group of these broader connectives in Czech – the selected verbs of saying "doplnit/doplňovat" ("to complement"), "upřesnit/upřesňovat" ("to specify"), "dodat/dodávat" ("to add"), "pokračovat" ("to continue") – and analyses their occurrence and function in texts from the Prague Discourse Treebank. The paper demonstrates that these verbs of saying have a special place within the other connectives, as they contain two items – e.g. "he added" means "and he said" so the verb "to add" contains an information about the relation to the previous context ("and") plus the verb of saying ("to say"). This information led us to a more general observation, i.e. discourse connectives in broader sense do not necessarily connect two pieces of a text but some of them carry the second argument right in their semantics, which "classical" connectives can never do.

Building a Corpus of Manually Revised Texts from Discourse Perspective

Ryu Iida and Takenobu Tokunaga

This paper presents building a corpus of manually revised texts which includes both before and after-revision information. In

order to create such a corpus, we propose a procedure for revising a text from a discourse perspective, consisting of dividing a text to discourse units, organising and reordering groups of discourse units and finally modifying referring and connective expressions, each of which imposes limits on freedom of revision. Following the procedure, six revisers who have enough experience in either teaching Japanese or scoring Japanese essays revised 120 Japanese essays written by Japanese native speakers. Comparing the original and revised texts, we found some specific manual revisions frequently occurred between the original and revised texts, e.g. 'thesis' statements were frequently placed at the beginning of a text. We also evaluate text coherence using the original and revised texts on the task of pairwise information ordering, identifying a more coherent text. The experimental results using two text coherence models demonstrated that the two models did not outperform the random baseline.

The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank

Lanjun Zhou, Binyang Li, Zhongyu Wei and Kam-Fai Wong

The lack of open discourse corpus for Chinese brings limitations for many natural language processing tasks. In this work, we present the first open discourse treebank for Chinese, namely, the Discourse Treebank for Chinese (DTBC). At the current stage, we annotated explicit intra-sentence discourse connectives, their corresponding arguments and senses for all 890 documents of the Chinese Treebank 5. We started by analysing the characteristics of discourse annotation for Chinese, adapted the annotation scheme of Penn Discourse Treebank 2 (PDTB2) to Chinese language while maintaining the compatibility as far as possible. We made adjustments to 3 essential aspects according to the previous study of Chinese linguistics. They are sense hierarchy, argument scope and semantics of arguments. Agreement study showed that our annotation scheme could achieve highly reliable results.

Computational Narratology: Extracting Tense Clusters from Narrative Texts

Thomas Bögel, Jannik Strötgen and Michael Gertz

Computational Narratology is an emerging field within the Digital Humanities. In this paper, we tackle the problem of extracting temporal information as a basis for event extraction and ordering, as well as further investigations of complex phenomena in narrative texts. While most existing systems focus on news texts and extract explicit temporal information exclusively, we show that this approach is not feasible for narratives. Based on tense information of verbs, we define temporal clusters as an annotation task and validate the annotation schema by showing that the task can be performed with high inter-annotator agreement. To alleviate and reduce the manual annotation effort, we propose a rule-based approach to robustly extract temporal clusters using a multi-layered and dynamic NLP pipeline that combines off-the-shelf components in a heuristic setting. Comparing our results against human judgements, our system is capable of predicting the tense of verbs and sentences with very high reliability: for the most prevalent tense in our corpus, more than 95% of all verbs are annotated correctly.

Can Numerical Expressions Be Simpler? Implementation and Demonstration of a Numerical Simplification System for Spanish

Susana Bautista and Horacio Saggion

Information in newspapers is often showed in the form of numerical expressions which present comprehension problems for many people, including people with disabilities, illiteracy or lack of access to advanced technology. The purpose of this paper is to motivate, describe, and demonstrate a rule-based lexical component that simplifies numerical expressions in Spanish texts. We propose an approach that makes news articles more accessible to certain readers by rewriting difficult numerical expressions in a simpler way. We will showcase the numerical simplification system with a live demo based on the execution of our components over different texts, and which will consider both successful and unsuccessful simplification cases.

Cross-Linguistic Annotation of Narrativity for English/French Verb Tense Disambiguation

Cristina Grisot and Thomas Meyer

This paper presents manual and automatic annotation experiments for a pragmatic verb tense feature (narrativity) in English/French parallel corpora. The feature is considered to play an important role for translating English Simple Past tense into French, where three different tenses are available. Whether the French Passé Composé, Passé Simple or Imparfait should be used is highly dependent on a longer-range context, in which either narrative events ordered in time or mere non-narrative state of affairs in the past are described. This longer-range context is usually not available to current machine translation (MT) systems, that are trained on parallel corpora. Annotating narrativity prior to translation is therefore likely to help current MT systems. Our experiments show that narrativity can be reliably identified with kappa-values of up to 0.91 in manual annotation and with F1 scores of up to 0.72 in automatic annotation.

P14 - Grammar and Syntax

Wednesday, May 28, 16:45

Chairperson: **Cristina Bosco** Poster Session

A Database for Measuring Linguistic Information Content

Richard Sproat, Bruno Cartoni, HyunJeong Choe, David Huynh, Linne Ha, Ravindran Rajakumar and Evelyn Wenzel-Grondie

Which languages convey the most information in a given amount of space? This is a question often asked of linguists, especially by engineers who often have some information theoretic measure of "information" in mind, but rarely define exactly how they would measure that information. The question is, in fact remarkably hard to answer, and many linguists consider it unanswerable. But it is a question that seems as if it ought to have an answer. If one had a database of close translations between a set of typologically diverse languages, with detailed marking of morphosyntactic and morphosemantic features, one could hope to quantify the differences between how these different languages convey information. Since no appropriate database exists we decided to construct one. The purpose of this paper is to present our work on the database, along with some preliminary results. We plan to release the dataset once complete.

Valency and Word Order in Czech – A Corpus Probe

Katerina Rysova and Jiří Mírovský

We present a part of broader research on word order aiming at finding factors influencing word order in Czech (i.e. in an inflectional language) and their intensity. The main aim of the paper is to test a hypothesis that obligatory adverbials (in terms of the valency) follow the non-obligatory (i.e. optional) ones in the surface word order. The determined hypothesis was tested by creating a list of features for the decision trees algorithm and by searching in data of the Prague Dependency Treebank using the search tool PML Tree Query. Apart from the valency, our experiment also evaluates importance of several other features, such as argument length and deep syntactic function. Neither of the used methods has proved the given hypothesis but according to the results, there are several other features that influence word order of contextually non-bound free modifiers of a verb in Czech, namely position of the sentence in the text, form and length of the verb modifiers (the whole subtrees), and the semantic dependency relation (functor) of the modifiers.

Mörkum Njálu. An Annotated Corpus to Analyse and Explain Grammatical Divergences Between 14th-century Manuscripts of Njál's Saga

Ludger Zeevaert

The work of the research project "Variance of Njáls saga" at the Árni Magnússon Institute for Icelandic Studies in Reykjavík relies mainly on an annotated XML-corpus of manuscripts of Brennu-Njáls saga or 'The Story of Burnt Njál', an Icelandic prose narrative from the end of the 13th century. One part of the project is devoted to linguistic variation in the earliest transmission of the text in parchment manuscripts and fragments from the 14th century. The article gives a short overview over the design of the corpus that has to serve quite different purposes from palaeographic over stemmatological to literary research. It focuses on features important for the analysis of certain linguistic variables and the challenge lying in their implementation in a corpus consisting of close transcriptions of medieval manuscripts and gives examples for the use of the corpus for linguistic research in the frame of the project that mainly consists of the analysis of different grammatical/syntactic constructions that are often referred to in connection with stylistic research (narrative inversion, historical present tense, indirect-speech constructions).

GenitivDB – a Corpus-Generated Database for German Genitive Classification

Roman Schneider

We present a novel NLP resource for the explanation of linguistic phenomena, built and evaluated exploring very large annotated language corpora. For the compilation, we use the German Reference Corpus (DeReKo) with more than 5 billion word forms, which is the largest linguistic resource worldwide for the study of contemporary written German. The result is a comprehensive database of German genitive formations, enriched with a broad range of intra- und extralinguistic metadata. It can be used for the notoriously controversial classification and prediction of genitive endings (short endings, long endings, zero-marker). We also evaluate the main factors influencing the use of specific endings. To get a general idea about a factor's influences and its side effects, we calculate chi-square-tests and visualize the residuals with an association plot. The results are evaluated against a gold standard by implementing tree-based machine learning algorithms. For the statistical analysis, we applied the supervised LMT Logistic Model Trees algorithm, using the WEKA software. We intend to use this gold standard to evaluate GenitivDB, as well as to explore methodologies for a predictive genitive model.

Building a Reference Lexicon for Countability in English

Tibor Kiss, Francis Jeffry Pelletier and Tobias Stadtfeld

The present paper describes the construction of a resource to determine the lexical preference class of a large number of English noun-senses ($\approx$ \$14,000) with respect to the distinction between mass and count interpretations. In constructing the lexicon, we have employed a questionnaire-based approach based on existing resources such as the Open ANC (http://www.anc.org) and WordNet. The questionnaire requires annotators to answer six questions about a noun-sense pair. Depending on the answers, a given noun-sense pair can be assigned to fine-grained noun classes, spanning the area between count and mass. The reference lexicon contains almost 14,000 noun-sense pairs. An initial data set of 1,000 has been annotated together by four native speakers, while the remaining 12,800 noun-sense pairs have been annotated in parallel by two annotators each. We can confirm the general feasibility of the approach by reporting satisfactory values between 0.694 and 0.755 in inter-annotator agreement using Krippendorff's α.

P15 - Lexicons

Wednesday, May 28, 16:45

Chairperson: **Amália Mendes** Poster Session

Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing

Ismail El Maarouf, Jane Bradbury, Vít Baisa and Patrick Hanks

This paper reports the results of Natural Language Processing (NLP) experiments in semantic parsing, based on a new semantic resource, the Pattern Dictionary of English Verbs (PDEV) (Hanks, 2013). This work is set in the DVC (Disambiguating Verbs by Collocation) project , a project in Corpus Lexicography aimed at expanding PDEV to a large scale. This project springs from a long-term collaboration of lexicographers with computer scientists which has given rise to the design and maintenance of specific, adapted, and user-friendly editing and exploration tools. Particular attention is drawn on the use of NLP deep semantic methods to help in data processing. Possible contributions of NLP include pattern disambiguation, the focus of this article. The present article explains how PDEV differs from other lexical resources and describes its structure in detail. It also presents new classification experiments on a subset of 25 verbs. The SVM model obtained a micro-average F1 score of 0.81.

GLÀFF, a Large Versatile French Lexicon

Nabil Hathout, Franck Sajous and Basilio Calderone

This paper introduces GLAFF, a large-scale versatile French lexicon extracted from Wiktionary, the collaborative online dictionary. GLAFF contains, for each entry, inflectional features and phonemic transcriptions. It distinguishes itself from the other available French lexicons by its size, its potential for constant updating and its copylefted license. We explain how we have built GLAFF and compare it to other known resources in terms of coverage and quality of the phonemic transcriptions. We show that its size and quality are strong assets that could allow GLAFF to become a reference lexicon for French NLP and linguistics. Moreover, other derived lexicons can easily be based on GLAFF to satisfy specific needs of various fields such as psycholinguistics.

Bilingual Dictionary Construction with Transliteration Filtering

John Richardson, Toshiaki Nakazawa and Sadao Kurohashi

In this paper we present a bilingual transliteration lexicon of 170K Japanese-English technical terms in the scientific domain. Translation pairs are extracted by filtering a large list of transliteration candidates generated automatically from a phrase table trained on parallel corpora. Filtering uses a novel transliteration similarity measure based on a discriminative phrase-based machine translation approach. We demonstrate that the extracted dictionary is accurate and of high recall (F1 score 0.8). Our lexicon contains not only single words but also multi-word expressions, and is freely available. Our experiments focus on Katakana-English lexicon construction, however it would be possible to apply the proposed methods to transliteration extraction for a variety of language pairs.

Bootstrapping Open-Source English-Bulgarian Computational Dictionary

Krasimir Angelov

We present an open-source English-Bulgarian dictionary which is a unification and consolidation of existing and freely available resources for the two languages. The new resource can be used as either a pair of two monolingual morphological lexicons, or as a bidirectional translation dictionary between the languages. The structure of the resource is compatible with the existing synchronous English-Bulgarian grammar in Grammatical Framework (GF). This makes it possible to immediately plug it

in as a component in a grammar-based translation system that is currently under development in the same framework. This also meant that we had to enrich the dictionary with additional syntactic and semantic information that was missing in the original resources.

MotàMot Project: Conversion of a French-Khmer Published Dictionary for Building a Multilingual Lexical System

Mathieu Mangeot

Economic issues related to the information processing techniques are very important. The development of such technologies is a major asset for developing countries like Cambodia and Laos, and emerging ones like Vietnam, Malaysia and Thailand. The MotAMot project aims to computerize an under-resourced language: Khmer, spoken mainly in Cambodia. The main goal of the project is the development of a multilingual lexical system targeted for Khmer. The macrostructure is a pivot one with each word sense of each language linked to a pivot axi. The microstructure comes from a simplification of the explanatory and combinatory dictionary. The lexical system has been initialized with data coming mainly from the conversion of the French-Khmer bilingual dictionary of Denis Richer from Word to XML format. The French part was completed with pronunciation and parts-of-speech coming from the FeM French-english-Malay dictionary. The Khmer headwords noted in IPA in the Richer dictionary were converted to Khmer writing with OpenFST, a finite state transducer tool. The resulting resource is available online for lookup, editing, download and remote programming via a REST API on a Jibiki platform.

RELISH LMF: Unlocking the Full Power of the Lexical Markup Framework

Menzo Windhouwer, Justin Petro and Shakila Shayan

The Lexical Markup Framework (ISO 24613:2008) provides a core class diagram and various extensions as the basis for constructing lexical resources. Unfortunately the informative Document Type Definition provided by the standard and other available LMF serializations lack support for many of the powerful features of the model. This paper describes RELISH LMF, which unlocks the full power of the LMF model by providing a set of extensible modern schema modules. As use cases RELISH LL LMF and support by LEXUS, an online lexicon tool, are described.

Building a Dataset of Multilingual Cognates for the Romanian Lexicon

Liviu Dinu and Alina Maria Ciobanu

Identifying cognates is an interesting task with applications in numerous research areas, such as historical and comparative linguistics, language acquisition, cross-lingual information retrieval, readability and machine translation. We propose a dictionary-based approach to identifying cognates based on etymology and etymons. We account for relationships between languages and we extract etymology-related information from electronic dictionaries. We employ the dataset of cognates that we obtain as a gold standard for evaluating to which extent orthographic methods can be used to detect cognate pairs. The question that arises is whether they are able to discriminate between cognates and non-cognates, given the orthographic changes undergone by foreign words when entering new languages. We investigate some orthographic approaches widely used in this research area and some original metrics as well. We run our experiments on the Romanian lexicon, but the method we propose is adaptable to any language, as far as resources are available.

LexTec - a Rich Language Resource for Technical Domains in Portuguese

Palmira Marrafa, Raquel Amaro and Sara Mendes

The growing amount of available information and the importance given to the access to technical information enhance the potential role of NLP applications in enabling users to deal with information for a variety of knowledge domains. In this process, language resources are crucial. This paper presents Lextec, a rich computational language resource for technical vocabulary in Portuguese. Encoding a representative set of terms for ten different technical domains, this concept-based relational language resource combines a wide range of linguistic information by integrating each entry in a domain-specific wordnet and associating it with a precise definition for each lexicalization in the technical domain at stake, illustrative texts and information for translation into English.

P16 - Morphology

Wednesday, May 28, 16:45

Chairperson: **Benoît Sagot** Poster Session

Amazigh Verb Conjugator

Fadoua Ataa Allah and Siham Boulaknadel

With the aim of preserving the Amazigh heritage from being threatened with disappearance, it seems suitable to provide Amazigh with required resources to confront the stakes of access to the domain of New Information and Communication Technologies (ICT). In this context and in the perspective to build linguistic resources and natural language processing tools for this language, we have undertaken to develop an online

conjugating tool that generates the inflectional forms of the Amazigh verbs. This tool is based on novel linguistically motivated morphological rules describing the verbal paradigm for all the Moroccan Amazigh varieties. Furthermore, it is based on the notion of morphological tree structure and uses transformational rules which are attached to the leaf nodes. Each rule may have numerous mutually exclusive clauses, where each part of a clause is a regular expression pattern that is matched against the radical pattern. This tool is an interactive conjugator that provides exhaustive coverage of linguistically accurate conjugation paradigms for over 3584 Armazigh verbs. It has been made simple and easy to use and designed from the ground up to be a highly effective learning aid that stimulates a desire to learn.

The Development of Dutch and Afrikaans Language Resources for Compound Boundary Analysis.

Menno van Zaanen, Gerhard van Huyssteen, Suzanne Aussems, Chris Emmery and Roald Eiselen

In most languages, new words can be created through the process of compounding, which combines two or more words into a new lexical unit. Whereas in languages such as English the components that make up a compound are separated by a space, in languages such as Finnish, German, Afrikaans and Dutch these components are concatenated into one word. Compounding is very productive and leads to practical problems in developing machine translators and spelling checkers, as newly formed compounds cannot be found in existing lexicons. The Automatic Compound Processing (AuCoPro) project deals with the analysis of compounds in two closely-related languages, Afrikaans and Dutch. In this paper, we present the development and evaluation of two datasets, one for each language, that contain compound words with annotated compound boundaries. Such datasets can be used to train classifiers to identify the compound components in novel compounds. We describe the process of annotation and provide an overview of the annotation guidelines as well as global properties of the datasets. The inter-rater agreements between the annotators are considered highly reliable. Furthermore, we show the usability of these datasets by building an initial automatic compound boundary detection system, which assigns compound boundaries with approximately 90% accuracy.

Zmorge: A German Morphological Lexicon Extracted from Wiktionary

Rico Sennrich and Beat Kunz

We describe a method to automatically extract a German lexicon from Wiktionary that is compatible with the finite-state morphological grammar SMOR. The main advantage of the resulting lexicon over existing lexica for SMOR is that it is open and permissively licensed. A recall-oriented evaluation shows that a morphological analyser built with our lexicon has comparable coverage compared to existing lexica, and continues to improve as Wiktionary grows. We also describe modifications to the SMOR grammar that result in a more conventional lemmatisation of words.

A New Form of Humor – Mapping Constraint-based Computational Morphologies to a Finite-State Representation

Attila Novák

MorphoLogic's Humor morphological analyzer engine has been used for the development of several high-quality computational morphologies, among them ones for complex agglutinative languages. However, Humor's closed source licensing scheme has been an obstacle to making these resources widely available. Moreover, there are other limitations of the rule-based Humor engine: lack of support for morphological guessing and for the integration of frequency information or other weighting of the models. These problems were solved by converting the databases to a finite-state representation that allows for morphological guessing and the addition of weights. Moreover, it has open-source implementations.

Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus

Veronika Vincze, Viktor Varga, Katalin Ilona Simkó, János Zsibrita, Ágoston Nagy, Richárd Farkas and János Csirik

The Szeged Corpus is the largest manually annotated database containing the possible morphological analyses and lemmas for each word form. In this work, we present its latest version, Szeged Corpus 2.5, in which the new harmonized morphological coding system of Hungarian has been employed and, on the other hand, the majority of misspelled words have been corrected and tagged with the proper morphological code. New morphological codes are introduced for participles, causative / modal / frequentative verbs, adverbial pronouns and punctuation marks, moreover, the distinction between common and proper nouns is eliminated. We also report some statistical data on the frequency of the new morphological codes. The new version of the corpus made it possible to train magyarlanc, a data-driven POS-tagger of Hungarian on a dataset with the new harmonized codes. According to the results, magyarlanc is able to achieve a state-of-the-art accuracy score on the 2.5 version as well.

A Set of Open Source Tools for Turkish Natural Language Processing

Çağrı Çöltekin

This paper introduces a set of freely available, open-source tools for Turkish that are built around TRmorph, a morphological analyzer introduced earlier in Coltekin (2010). The article first provides an update on the analyzer, which includes a complete rewrite using a different finite-state description language and tool set as well as major tagset changes to comply better with the state-of-the-art computational processing of Turkish and the user requests received so far. Besides these major changes to the analyzer, this paper introduces tools for morphological segmentation, stemming and lemmatization, guessing unknown words, grapheme to phoneme conversion, hyphenation and a morphological disambiguation.

Word-Formation Network for Czech

Magda Sevcikova and Zdeněk Žabokrtský

In the present paper, we describe the development of the lexical network DeriNet, which captures core word-formation relations on the set of around 266 thousand Czech lexemes. The network is currently limited to derivational relations because derivation is the most frequent and most productive word-formation process in Czech. This limitation is reflected in the architecture of the network: each lexeme is allowed to be linked up with just a single base word; composition as well as combined processes (composition with derivation) are thus not included. After a brief summarization of theoretical descriptions of Czech derivation and the state of the art of NLP approaches to Czech derivation, we discuss the linguistic background of the network and introduce the formal structure of the network and the semi-automatic annotation procedure. The network was initialized with a set of lexemes whose existence was supported by corpus evidence. Derivational links were created using three sources of information: links delivered by a tool for morphological analysis, links based on an automatically discovered set of derivation rules, and on a grammar-based set of rules. Finally, we propose some research topics which could profit from the existence of such lexical network.

MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan Roth

In this paper, we present MADAMIRA, a system for morphological analysis and disambiguation of Arabic that combines some of the best aspects of two previously commonly used systems for Arabic processing, MADA (Habash and Rambow, 2005; Habash et al., 2009; Habash et al., 2013) and AMIRA (Diab et al., 2007). MADAMIRA improves upon the two systems with a more streamlined Java implementation that is more robust, portable, extensible, and is faster than its ancestors by more than an order of magnitude. We also discuss an online demo (see http://nlp.ldeo.columbia.edu/madamira/) that highlights these aspects.

Computer-aided Morphology Expansion for Old Swedish

Yvonne Adesam, Malin Ahlberg, Peter Andersson, Gerlof Bouma, Markus Forsberg and Mans Hulden

In this paper we describe and evaluate a tool for paradigm induction and lexicon extraction that has been applied to Old Swedish. The tool is semi-supervised and uses a small seed lexicon and unannotated corpora to derive full inflection tables for input lemmata. In the work presented here, the tool has been modified to deal with the rich spelling variation found in Old Swedish texts. We also present some initial experiments, which are the first steps towards creating a large-scale morphology for Old Swedish.

Morfeusz Reloaded

Marcin Woliński

The paper presents recent developments in Morfeusz – a morphological analyser for Polish. The program, being already a fundamental resource for processing Polish, has been reimplemented with some important changes in the tagset, some new options, added information on proper names, and ability to perform simple prefix derivation. The present version of Morfeusz (including its dictionaries) is made available under the very liberal 2-clause BSD license. The program can be downloaded from http://sgjp.pl/morfeusz/.

P17 - WordNet

Wednesday, May 28, 16:45

Chairperson: **Francis Bond** Poster Session

Automatic Creation of WordNets from Parallel Corpora

Antoni Oliver and Salvador Climent

In this paper we present the evaluation results for the creation of WordNets for five languages (Spanish, French, German, Italian and Portuguese) using an approach based on parallel corpora. We have used three very large parallel corpora for our experiments:

DGT-TM, EMEA and ECB. The English part of each corpus is semantically tagged using Freeling and UKB. After this step, the process of WordNet creation is converted into a word alignment problem, where we want to alignWordNet synsets in the English part of the corpus with lemmata on the target language part of the corpus. The word alignment algorithm used in these experiments is a simple most frequent translation algorithm implemented into the WN-Toolkit. The obtained precision values are quite satisfactory, but the overall number of extracted synset-variant pairs is too low, leading into very poor recall values. In the conclusions, the use of more advanced word alignment algorithms, such as Giza++, Fast Align or Berkeley aligner is suggested.

Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources

Spandana Gella, Carlo Strapparava and Vivi Nastase

In this paper we present the mapping between WordNet domains and WordNet topics, and the emergent Wikipedia categories. This mapping leads to a coarse alignment between WordNet and Wikipedia, useful for producing domain-specific and multilingual corpora. Multilinguality is achieved through the cross-language links between Wikipedia categories. Research in word-sense disambiguation has shown that within a specific domain, relevant words have restricted senses. The multilingual, and comparable, domain-specific corpora we produce have the potential to enhance research in word-sense disambiguation and terminology extraction in different languages, which could enhance the performance of various NLP tasks.

Adapting VerbNet to French using Existing Resources

Quentin Pradet, Laurence Danlos and Gaël de Chalendar

VerbNet is an English lexical resource for verbs that has proven useful for English NLP due to its high coverage and coherent classification. Such a resource doesn't exist for other languages, despite some (mostly automatic and unsupervised) attempts. We show how to semi-automatically adapt VerbNet using existing resources designed for different purposes. This study focuses on French and uses two French resources: a semantic lexicon (Les Verbes Français) and a syntactic lexicon (Lexique-Grammaire).

Bootstrapping an Italian VerbNet: data-driven analysis of verb alternations

GianLuca Lebani, Veronica Viola and Alessandro Lenci

The goal of this paper is to propose a classification of the syntactic alternations admitted by the most frequent Italian verbs. The data-driven two-steps procedure exploited and the structure of the identified classes of alternations are presented in depth and discussed. Even if this classification has been developed with a practical application in mind, namely the semi-automatic building of a VerbNet-like lexicon for Italian verbs, partly following the methodology proposed in the context of the VerbNet project, its availability may have a positive impact on several related research topics and Natural Language Processing tasks

Dense Components in the Structure of WordNet

Ahti Lohk, Kaarel Allik, Heili Orav and Leo Võhandu

This paper introduces a test-pattern named a dense component for checking inconsistencies in the hierarchical structure of a wordnet. Dense component (viewed as substructure) points out the cases of regular polysemy in the context of multiple inheritance. Definition of the regular polysemy is redefined – instead of lexical units there are used lexical concepts (synsets). All dense components are evaluated by expert lexicographer. Based on this experiment we give an overview of the inconsistencies which the test-pattern helps to detect. Special attention is turned to all different kind of corrections made by lexicographer. Authors of this paper find that the greatest benefit of the use of dense components is helping to detect if the regular polysemy is justified or not. In-depth analysis has been performed for Estonian Wordnet Version 66. Some comparative figures are also given for the Estonian Wordnet (EstWN) Version 67 and Princeton WordNet (PrWN) Version 3.1. Analysing hierarchies only hypernym-relations are used.

The Making of Ancient Greek WordNet

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo del Gratta, Monica Monachini and Gregory Crane

This paper describes the process of creation and review of a new lexico-semantic resource for the classical studies: AncientGreekWordNet. The candidate sets of synonyms (synsets) are extracted from Greek-English dictionaries, on the assumption that Greek words translated by the same English word or phrase have a high probability of being synonyms or at least semantically closely related. The process of validation and the web interface developed to edit and query the resource are described in detail. The lexical coverage of Ancient Greek WordNet is illustrated and the accuracy is evaluated. Finally, scenarios for exploiting the resource are discussed.

Etymological WordNet: Tracing the History of Words

Gerard de Melo

Research on the history of words has led to remarkable insights about language and also about the history of human civilization

more generally. This paper presents the Etymological Wordnet, the first database that aims at making word origin information available as a large, machine-readable network of words in many languages. The information in this resource is obtained from Wiktionary. Extracting a network of etymological information from Wiktionary requires significant effort, as much of the etymological information is only given in prose. We rely on custom pattern matching techniques and mine a large network with over 500,000 word origin links as well as over 2 million derivational/compositional links.

O13 - Sentiment Analysis (1)

Wednesday, May 28, 18:10

Chairperson: **Paul Buitelaar** Oral Session

Generating Polarity Lexicons with WordNet Propagation in 5 Languages

Isa Maks, Ruben Izquierdo, Francesca Frontini, Rodrigo Agerri, Piek Vossen and Andoni Azpeitia

In this paper we focus on the creation of general-purpose (as opposed to domain-specific) polarity lexicons in five languages: French, Italian, Dutch, English and Spanish using WordNet propagation. WordNet propagation is a commonly used method to generate these lexicons as it gives high coverage of general purpose language and the semantically rich WordNets where concepts are organised in synonym , antonym and hyperonym/hyponym structures seem to be well suited to the identification of positive and negative words. However, WordNets of different languages may vary in many ways such as the way they are compiled, the number of synsets, number of synonyms and number of semantic relations they include. In this study we investigate whether this variability translates into differences of performance when these WordNets are used for polarity propagation. Although many variants of the propagation method are developed for English, little is known about how they perform with WordNets of other languages. We implemented a propagation algorithm and designed a method to obtain seed lists similar with respect to quality and size, for each of the five languages. We evaluated the results against gold standards also developed according to a common method in order to achieve as less variance as possible between the different languages.

SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis

Muhammad Abdul-Mageed and Mona Diab

The computational treatment of subjectivity and sentiment in natural language is usually significantly improved by applying features exploiting lexical resources where entries are tagged with semantic orientation (e.g., positive, negative values). In spite of the fair amount of work on Arabic sentiment analysis over the past few years (e.g., (Abbasi et al., 2008; Abdul-Mageed et al., 2014; Abdul-Mageed et al., 2012; Abdul-Mageed and Diab, 2012a; Abdul-Mageed and Diab, 2012b; Abdul-Mageed et al., 2011a; Abdul-Mageed and Diab, 2011)), the language remains under-resourced as to these polarity repositories compared to the English language. In this paper, we report efforts to build and present SANA, a large-scale, multi-genre, multi-dialect multi-lingual lexicon for the subjectivity and sentiment analysis of the Arabic language and dialects.

On the Importance of Text Analysis for Stock Price Prediction

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney and Dan Jurafsky

We investigate the importance of text analysis for stock price prediction. In particular, we introduce a system that forecasts companies' stock price changes (UP, DOWN, STAY) in response to financial events reported in 8-K documents. Our results indicate that using text boosts prediction accuracy over 10% (relative) over a strong baseline that incorporates many financially-rooted features. This impact is most important in the short term (i.e., the next day after the financial event) but persists for up to five days.

O14 - Paralinguistics

Wednesday, May 28, 18:10

Chairperson: **Sophie Rosset** Oral Session

An Analysis of Older Users' Interactions with Spoken Dialogue Systems

Jamie Bost and Johanna Moore

This study explores communication differences between older and younger users with a task-oriented spoken dialogue system. Previous analyses of the MATCH corpus show that older users have significantly longer dialogues than younger users and that they are less satisfied with the system. Open questions remain regarding the relationship between information recall and cognitive abilities. This study documents a length annotation scheme designed to explore causes of additional length in the dialogues and the relationships between length, cognitive abilities, user satisfaction, and information recall. Results show that primary causes of older users' additional length include using polite vocabulary, providing additional information relevant to the task, and using full sentences to respond to the system. Regression models were built to predict length from cognitive

abilities and user satisfaction from length. Overall, users with higher cognitive ability scores had shorter dialogues than users with lower cognitive ability scores, and users with shorter dialogues were more satisfied with the system than users with longer dialogues. Dialogue length and cognitive abilities were significantly correlated with information recall. Overall, older users tended to use a human-to-human communication style with the system, whereas younger users tended to adopt a factual interaction style.

Alert!... Calm Down, There is Nothing to Worry About. Warning and Soothing Speech Synthesis

Milan Rusko, Sakhia Darjaa, Marian Trnka, Marian Ritomsky and Robert Sabo

Presence of appropriate acoustic cues of affective features in the synthesized speech can be a prerequisite for the proper evaluation of the semantic content by the message recipient. In the recent work the authors have focused on the research of expressive speech synthesis capable of generating naturally sounding synthetic speech at various levels of arousal. The synthesizer should be able to produce speech in Slovak in different styles from extremely urgent warnings, insisting messages, alerts, through comments, and neutral style speech to soothing messages and very calm speech. A three-step method was used for recording both - the high-activation and low-activation expressive speech databases. The acoustic properties of the obtained databases are discussed. Several synthesizers with different levels of arousal were designed using these databases and their outputs are compared to the original voice of the voice talent. A possible ambiguity of acoustic cues is pointed out and the relevance of the semantic meaning of the sentences both in the sentence set for the speech database recording and in the set for subjective synthesizer testing is discussed.

Prosodic, Syntactic, Semantic Guidelines for Topic Structures Across Domains and Corpora

Ana Isabel Mata, Helena Moniz, Telmo Móia, Anabela Gonçalves, Fátima Silva, Fernando Batista, Inês Duarte, Fátima Oliveira and Isabel Falé

This paper presents the annotation guidelines applied to naturally occurring speech, aiming at an integrated account of contrast and parallel structures in European Portuguese. These guidelines were defined to allow for the empirical study of interactions among intonation and syntax-discourse patterns in selected sets of different corpora (monologues and dialogues, by adults and teenagers). In this paper we focus on the multilayer annotation process of left periphery structures by using a small sample of highly spontaneous speech in which the distinct types of topic structures are displayed. The analysis of this sample provides fundamental training and testing material for further application in a wider range of domains and corpora. The annotation process comprises the following time-linked levels (manual and automatic): phone, syllable and word level transcriptions (including co-articulation effects); tonal events and break levels; part-of-speech tagging; syntactic-discourse patterns (construction type; construction position; syntactic function; discourse function), and disfluency events as well. Speech corpora with such a multi-level annotation are a valuable resource to look into grammar module relations in language use from an integrated viewpoint. Such viewpoint is innovative in our language, and has not been often assumed by studies for other languages.

O15 - Multiword Expressions

Wednesday, May 28, 18:10

Chairperson: **Aline Villavicencio** Oral Session

How to Tell a Schneemann from a Milchmann: An Annotation Scheme for Compound-Internal Relations

Corina Dima, Verena Henrich, Erhard Hinrichs and Christina Hoppermann

This paper presents a language-independent annotation scheme for the semantic relations that link the constituents of noun-noun compounds, such as Schneemann 'snow man' or Milchmann 'milk man'. The annotation scheme is hybrid in the sense that it assigns each compound a two-place label consisting of a semantic property and a prepositional paraphrase. The resulting inventory combines the insights of previous annotation schemes that rely exclusively on either semantic properties or prepositions, thus avoiding the known weaknesses that result from using only one of the two label types. The proposed annotation scheme has been used to annotate a set of 5112 German noun-noun compounds. A release of the dataset is currently being prepared and will be made available via the CLARIN Center Tübingen. In addition to the presentation of the hybrid annotation scheme, the paper also reports on an inter-annotator agreement study that has resulted in a substantial agreement among annotators.

Semi-Compositional Method for Synonym Extraction of Multi-Word Terms

Béatrice Daille and Amir Hazem

Automatic synonyms and semantically related word extraction is a challenging task, useful in many NLP applications such as question answering, search query expansion, text summarization, etc. While different studies addressed the task of word synonym

extraction, only a few investigations tackled the problem of acquiring synonyms of multi-word terms (MWT) from specialized corpora. To extract pairs of synonyms of multi-word terms, we propose in this paper an unsupervised semi-compositional method that makes use of distributional semantics and exploit the compositional property shared by most MWT. We show that our method outperforms significantly the state-of-the-art.

Multiword Expressions in Machine Translation

Valia Kordoni and Iliana Simova

This work describes an experimental evaluation of the significance of phrasal verb treatment for obtaining better quality statistical machine translation (SMT) results. The importance of the detection and special treatment of phrasal verbs is measured in the context of SMT, where the word-for-word translation of these units often produces incoherent results. Two ways of integrating phrasal verb information in a phrase-based SMT system are presented. Automatic and manual evaluations of the results reveal improvements in the translation quality in both experiments.

O16 - Spelling Normalisation

Wednesday, May 28, 18:10

Chairperson: **Hrafn Loftsson** Oral Session

A Database of Freely Written Texts of German School Students for the Purpose of Automatic Spelling Error Classification

Kay Berkling, Johanna Fay, Masood Ghayoomi, Katrin Hein, Rémi Lavalley, Ludwig Linhuber and Sebastian Stüker

The spelling competence of school students is best measured on freely written texts, instead of pre-determined, dictated texts. Since the analysis of the error categories in these kinds of texts is very labor intensive and costly, we are working on an automatic systems to perform this task. The modules of the systems are derived from techniques from the area of natural language processing, and are learning systems that need large amounts of training data. To obtain the data necessary for training and evaluating the resulting system, we conducted data collection of freely written, German texts by school children. 1,730 students from grade 1 through 8 participated in this data collection. The data was transcribed electronically and annotated with their corrected version. This resulted in a total of 14,563 sentences that can now be used for research regarding spelling diagnostics. Additional meta-data was collected regarding writers' language biography, teaching methodology, age, gender, and school year. In order to do a detailed manual annotation of the categories of the spelling errors committed by the students we developed a tool specifically tailored to the task.

Towards Shared Datasets for Normalization Research

Orphee de Clercq, Sarah Schulz, Bart Desmet and Véronique Hoste

In this paper we present a Dutch and English dataset that can serve as a gold standard for evaluating text normalization approaches. With the combination of text messages, message board posts and tweets, these datasets represent a variety of user generated content. All data was manually normalized to their standard form using newly-developed guidelines. We perform automatic lexical normalization experiments on these datasets using statistical machine translation techniques. We focus on both the word and character level and find that we can improve the BLEU score with ca. 20% for both languages. In order for this user generated content data to be released publicly to the research community some issues first need to be resolved. These are discussed in closer detail by focussing on the current legislation and by investigating previous similar data collection projects. With this discussion we hope to shed some light on various difficulties researchers are facing when trying to share social media data.

Synergy of Nederlab and @PhilosTEI: Diachronic and Multilingual Text-induced Corpus Clean-up

Martin Reynaert

In two concurrent projects in the Netherlands we are further developing TICCL or Text-Induced Corpus Clean-up. In project Nederlab TICCL is set to work on diachronic Dutch text. To this end it has been equipped with the largest diachronic lexicon and a historical name list developed at the Institute for Dutch Lexicology or INL. In project @PhilosTEI TICCL will be set to work on a fair range of European languages. We present a new implementation in C++ of the system which has been tailored to be easily adaptable to different languages. We further revisit prior work on diachronic Portuguese in which it was compared to VARD2 which had been manually adapted to Portuguese. This tested the new mechanisms for ranking correction candidates we have devised. We then move to evaluating the new TICCL port on a very large corpus of Dutch books known as EDBO, digitized by the Dutch National Library. The results show that TICCL scales to the largest corpus sizes and performs excellently raising the quality of the Gold Standard EDBO book by about 20% to 95% word accuracy. Simultaneous unsupervised post-correction of 10,000 digitized books is now a real option.

P18 - Corpora and Annotation

Wednesday, May 28, 18:10 - 19:30

Chairperson: **Steve Cassidy** Poster Session

Translation Errors from English to Portuguese: an Annotated Corpus

Angela Costa, Tiago Luís and Luísa Coheur

Analysing the translation errors is a task that can help us finding and describing translation problems in greater detail, but can also suggest where the automatic engines should be improved. Having these aims in mind we have created a corpus composed of 150 sentences, 50 from the TAP magazine, 50 from a TED talk and the other 50 from the from the TREC collection of factoid questions. We have automatically translated these sentences from English into Portuguese using Google Translate and Moses. After we have analysed the errors and created the error annotation taxonomy, the corpus was annotated by a linguist native speaker of Portuguese. Although Google's overall performance was better in the translation task (we have also calculated the BLUE and NIST scores), there are some error types that Moses was better at coping with, specially discourse level errors.

CoRoLa – The Reference Corpus of Contemporary Romanian Language

Verginica Barbu Mititelu, Elena Irimia and Dan Tufiş

We present the project of creating CoRoLa, a reference corpus of contemporary Romanian (from 1945 onwards). In the international context, the project finds its place among the initiatives of gathering huge collections of texts, of pre-processing and annotating them at several levels, and also of documenting them with metadata (CMDI). Our project is a joined effort of two institutes of the Romanian Academy. We foresee a corpus of more than 500 million word forms, covering all functional styles of the language. Although the vast majority of texts will be in written form, we target about 300 hours of oral texts, too, obligatorily with associated transcripts. Most of the texts will be from books, while the rest will be harvested from newspapers, booklets, technical reports, etc. The pre-processing includes cleaning the data and harmonising the diacritics, sentence splitting and tokenization. Annotation will be done at a morphological level in a first stage, followed by lemmatization, with the possibility of adding syntactic, semantic and discourse annotation in a later stage. A core of CoRoLa is described in the article. The target users of our corpus will be researchers in linguistics and language processing, teachers of Romanian, students.

A Multidialectal Parallel Corpus of Arabic

Houda Bouamor, Nizar Habash and Kemal Oflazer

The daily spoken variety of Arabic is often termed the colloquial or dialect form of Arabic. There are many Arabic dialects across the Arab World and within other Arabic speaking communities. These dialects vary widely from region to region and to a lesser extent from city to city in each region. The dialects are not standardized, they are not taught, and they do not have official status. However they are the primary vehicles of communication (face-to-face and recently, online) and have a large presence in the arts as well. In this paper, we present the first multidialectal Arabic parallel corpus, a collection of 2,000 sentences in Standard Arabic, Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic, in addition to English. Such parallel data does not exist naturally, which makes this corpus a very valuable resource that has many potential applications such as Arabic dialect identification and machine translation.

YouDACC: the Youtube Dialectal Arabic Comment Corpus

Ahmed Salama, Houda Bouamor, Behrang Mohit and Kemal Oflazer

This paper presents YOUDACC, an automatically annotated large-scale multi-dialectal Arabic corpus collected from user comments on Youtube videos. Our corpus covers different groups of dialects: Egyptian (EG), Gulf (GU), Iraqi (IQ), Maghrebi (MG) and Levantine (LV). We perform an empirical analysis on the crawled corpus and demonstrate that our location-based proposed method is effective for the task of dialect labeling.

Comparing Two Acquisition Systems for Automatically Building an English-Croatian Parallel Corpus from Multilingual Websites

Miquel Esplà-Gomis, Filip Klubička, Nikola Ljubešić, Sergio Ortiz-Rojas, Vassilis Papavassiliou and Prokopis Prokopidis

In this paper we compare two tools for automatically harvesting bitexts from multilingual websites: bitextor and ILSP-FC. We used both tools for crawling 21 multilingual websites from the tourism domain to build a domain-specific English–Croatian parallel corpus. Different settings were tried for both tools and 10,662 unique document pairs were obtained. A sample of about 10% of them was manually examined and the success rate was computed on the collection of pairs of documents detected by each setting. We compare the performance of the settings and

the amount of different corpora detected by each setting. In addition, we describe the resource obtained, both by the settings and through the human evaluation, which has been released as a high-quality parallel corpus.

Towards an Integration of Syntactic and Temporal Annotations in Estonian

Siim Orasmaa

We investigate the question how manually created syntactic annotations can be used to analyse and improve consistency in manually created temporal annotations. Our work introduces an annotation project for Estonian, where temporal annotations in TimeML framework were manually added to a corpus containing gold standard morphological and dependency syntactic annotations. In the first part of our work, we evaluate the consistency of manual temporal annotations, focusing on event annotations. We use syntactic annotations to distinguish different event annotation models, and we observe highest inter-annotator agreements on models representing ”prototypical events” (event verbs and events being part of the syntactic predicate of clause). In the second part of our work, we investigate how to improve consistency between syntactic and temporal annotations. We test on whether syntactic annotations can be used to validate temporal annotations: to find missing or partial annotations. Although the initial results indicate that such validation is promising, we also note that a better bridging between temporal (semantic) and syntactic annotations is needed for a complete automatic validation.

Annotation of Specialized Corpora using a Comprehensive Entity and Relation Scheme

Louise Deleger, Anne-Laure Ligozat, Cyril Grouin, Pierre Zweigenbaum and Aurelie Neveol

Annotated corpora are essential resources for many applications in Natural Language Processing. They provide insight on the linguistic and semantic characteristics of the genre and domain covered, and can be used for the training and evaluation of automatic tools. In the biomedical domain, annotated corpora of English texts have become available for several genres and subfields. However, very few similar resources are available for languages other than English. In this paper we present an effort to produce a high-quality corpus of clinical documents in French, annotated with a comprehensive scheme of entities and relations. We present the annotation scheme as well as the results of a pilot annotation study covering 35 clinical documents in a variety of subfields and genres. We show that high inter-annotator agreement can be achieved using a complex annotation scheme.

Developing Politeness Annotated Corpus of Hindi Blogs

Ritesh Kumar

In this paper I discuss the creation and annotation of a corpus of Hindi blogs. The corpus consists of a total of over 479,000 blog posts and blog comments. It is annotated with the information about the politeness level of each blog post and blog comment. The annotation is carried out using four levels of politeness – neutral, appropriate, polite and impolite. For the annotation, three classifiers – were trained and tested maximum entropy (MaxEnt), Support Vector Machines (SVM) and C4.5 - using around 30,000 manually annotated texts. Among these, C4.5 gave the best accuracy. It achieved an accuracy of around 78% which is within 2% of the human accuracy during annotation. Consequently this classifier is used to annotate the rest of the corpus

The MERLIN corpus: Learner Language and the CEFR

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová and Chiara Vettori

The MERLIN corpus is a written learner corpus for Czech, German,and Italian that has been designed to illustrate the Common European Framework of Reference for Languages (CEFR) with authentic learner data. The corpus contains 2,290 learner texts produced in standardized language certifications covering CEFR levels A1-C1. The MERLIN annotation scheme includes a wide range of language characteristics that enable research into the empirical foundations of the CEFR scales and provide language teachers, test developers, and Second Language Acquisition researchers with concrete examples of learner performance and progress across multiple proficiency levels. For computational linguistics, it provide a range of authentic learner data for three target languages, supporting a broadening of the scope of research in areas such as automatic proficiency classification or native language identification. The annotated corpus and related information will be freely available as a corpus resource and through a freely accessible, didactically-oriented online platform.

DysList: An Annotated Resource of Dyslexic Errors

Luz Rello, Ricardo Baeza-Yates and Joaquim Llisterri

We introduce a language resource for Spanish, DysList, composed of a list of unique errors extracted from a collection of texts written by people with dyslexia. Each of the errors was annotated with a set of characteristics as well as visual and phonetic features. To

the best of our knowledge this is the largest resource of this kind, especially given the difficulty of finding texts written by people with dyslexia

Criteria for Identifying and Annotating Caused Motion Constructions in Corpus Data

Jena D. Hwang, Annie Zaenen and Martha Palmer

While natural language processing performance has been improved through the recognition that there is a relationship between the semantics of the verb and the syntactic context in which the verb is realized, sentences where the verb does not conform to the expected syntax-semantic patterning behavior remain problematic. For example, in the sentence "The crowed laughed the clown off the stage", a verb of non-verbal communication laugh is used in a caused motion construction and gains a motion entailment that is atypical given its inherent lexical semantics. This paper focuses on our efforts at defining the semantic types and varieties of caused motion constructions (CMCs) through an iterative annotation process and establishing annotation guidelines based on these criteria to aid in the production of a consistent and reliable annotation. The annotation will serve as training and test data for classifiers for CMCs, and the CMC definitions developed throughout this study will be used in extending VerbNet to handle representations of sentences in which a verb is used in a syntactic context that is atypical for its lexical semantics.

The American Local News Corpus

Ann Irvine, Joshua Langfus and Chris Callison-Burch

We present the American Local News Corpus (ALNC), containing over 4 billion words of text from 2,652 online newspapers in the United States. Each article in the corpus is associated with a timestamp, state, and city. All 50 U.S. states and 1,924 cities are represented. We detail our method for taking daily snapshots of thousands of local and national newspapers and present two example corpus analyses. The first explores how different sports are talked about over time and geography. The second compares per capita murder rates with news coverage of murders across the 50 states. The ALNC is about the same size as the Gigaword corpus and is growing continuously. Version 1.0 is available for research use.

P19 - Document Classification, Text Categorisation

Wednesday, May 28, 18:10 - 19:30

Chairperson: **Karën Fort** Poster Session

A LDA-based Topic Classification Approach from highly Imperfect Automatic Transcriptions

Mohamed Morchid, Richard Dufour and Georges Linares

Although the current transcription systems could achieve high recognition performance, they still have a lot of difficulties to transcribe speech in very noisy environments. The transcription quality has a direct impact on classification tasks using text features. In this paper, we propose to identify themes of telephone conversation services with the classical Term Frequency-Inverse Document Frequency using Gini purity criteria (TF-IDF-Gini) method and with a Latent Dirichlet Allocation (LDA) approach. These approaches are coupled with a Support Vector Machine (SVM) classification to resolve theme identification problem. Results show the effectiveness of the proposed LDA-based method compared to the classical TF-IDF-Gini approach in the context of highly imperfect automatic transcriptions. Finally, we discuss the impact of discriminative and non-discriminative words extracted by both methods in terms of transcription accuracy.

How to Use Less Features and Reach Better Performance in Author Gender Identification

Juan Soler and Leo Wanner

Over the last years, author profiling in general and author gender identification in particular have become a popular research area due to their potential attractive applications that range from forensic investigations to online marketing studies. However, nearly all state-of-the-art works in the area still very much depend on the datasets they were trained and tested on, since they heavily draw on content features, mostly a large number of recurrent words or combinations of words extracted from the training sets. We show that using a small number of features that mainly depend on the structure of the texts we can outperform other approaches that depend mainly on the content of the texts and that use a huge number of features in the process of identifying if the author of a text is a man or a woman. Our system has been tested against a dataset constructed for our work as well as against two datasets that were previously used in other papers.

Genres in the Prague Discourse Treebank

Lucie Poláková, Pavlína Jínová and Jiří Mírovský

We present the project of classification of Prague Discourse Treebank documents (Czech journalistic texts) for their genres. Our main interest lies in opening the possibility to observe how text coherence is realized in different types (in the genre sense) of language data and, in the future, in exploring the ways of using genres as a feature for multi-sentence-level language technologies. In the paper, we first describe the motivation and the concept of the genre annotation, and briefly introduce the Prague Discourse Treebank. Then, we elaborate on the process of manual annotation of genres in the treebank, from the annotators' manual work to post-annotation checks and to the inter-annotator agreement measurements. The annotated genres are subsequently analyzed together with discourse relations (already annotated in the treebank) – we present distributions of the annotated genres and results of studying distinctions of distributions of discourse relations across the individual genres.

Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers

Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, Ekaterina Lapshinova-Koltunski, Noam Ordan and Elke Teich

We present a methodology to analyze the linguistic evolution of scientific registers with data mining techniques, comparing the insights gained from shallow vs. linguistic features. The focus is on selected scientific disciplines at the boundaries to computer science (computational linguistics, bioinformatics, digital construction, microelectronics). The data basis is the English Scientific Text Corpus (SCITEX) which covers a time range of roughly thirty years (1970/80s to early 2000s) (Degaetano-Ortlieb et al., 2013; Teich and Fankhauser, 2010). In particular, we investigate the diversification of scientific registers over time. Our theoretical basis is Systemic Functional Linguistics (SFL) and its specific incarnation of register theory (Halliday and Hasan, 1985). In terms of methods, we combine corpus-based methods of feature extraction and data mining techniques.

Detecting Document Structure in a Very Large Corpus of UK Financial Reports

Mahmoud El-Haj, Paul Rayson, Steve Young and Martin Walker

In this paper we present the evaluation of our automatic methods for detecting and extracting document structure in annual financial reports. The work presented is part of the Corporate Financial Information Environment (CFIE) project in which we are using Natural Language Processing (NLP) techniques to study the causes and consequences of corporate disclosure and financial reporting outcomes. We aim to uncover the determinants of financial reporting quality and the factors that influence the quality of information disclosed to investors beyond the financial statements. The CFIE consists of the supply of information by firms to investors, and the mediating influences of information intermediaries on the timing, relevance and reliability of information available to investors. It is important to compare and contrast specific elements or sections of each annual financial report across our entire corpus rather than working at the full document level. We show that the values of some metrics e.g. readability will vary across sections, thus improving on previous research research based on full texts.

Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing

Noushin Rezapour Asheghi, Serge Sharoff and Katja Markert

Research in Natural Language Processing often relies on a large collection of manually annotated documents. However, currently there is no reliable genre-annotated corpus of web pages to be employed in Automatic Genre Identification (AGI). In AGI, documents are classified based on their genres rather than their topics or subjects. The major shortcoming of available web genre collections is their relatively low inter-coder agreement. Reliability of annotated data is an essential factor for reliability of the research result. In this paper, we present the first web genre corpus which is reliably annotated. We developed precise and consistent annotation guidelines which consist of well-defined and well-recognized categories. For annotating the corpus, we used crowd-sourcing which is a novel approach in genre annotation. We computed the overall as well as the individual categories' chance-corrected inter-annotator agreement. The results show that the corpus has been annotated reliably.

Locating Requests among Open Source Software Communication Messages

Ioannis Korkontzelos and Sophia Ananiadou

As a first step towards assessing the quality of support offered online for Open Source Software (OSS), we address the task of locating requests, i.e., messages that raise an issue to be addressed by the OSS community, as opposed to any other message. We present a corpus of online communication messages randomly sampled from newsgroups and bug trackers, manually annotated as requests or non-requests. We identify several linguistically shallow, content-based heuristics that correlate with the classification and investigate the extent to which they can serve

as independent classification criteria. Then, we train machine-learning classifiers on these heuristics. We experiment with a wide range of settings, such as different learners, excluding some heuristics and adding unigram features of various parts-of-speech and frequency. We conclude that some heuristics can perform well, while their accuracy can be improved further using machine learning, at the cost of obtaining manual annotations.

Sockpuppet Detection in Wikipedia: A Corpus of Real-World Deceptive Writing for Linking Identities

Thamar Solorio, Ragib Hasan and Mainul Mizan

This paper describes a corpus of sockpuppet cases from Wikipedia. A sockpuppet is an online user account created with a fake identity for the purpose of covering abusive behavior and/or subverting the editing regulation process. We used a semi-automated method for crawling and curating a dataset of real sockpuppet investigation cases. To the best of our knowledge, this is the first corpus available on real-world deceptive writing. We describe the process for crawling the data and some preliminary results that can be used as baseline for benchmarking research. The dataset has been released under a Creative Commons license from our project website (http://docsig.cis.uab.edu/tools-and-datasets/).

P20 - FrameNet

Wednesday, May 28, 18:10 - 19:30

Chairperson: **Alessandro Lenci** Poster Session

Reusing Swedish FrameNet for Training Semantic Roles

Ildikó Pilán and Elena Volodina

In this article we present the first experiences of reusing the Swedish FrameNet (SweFN) as a resource for training semantic roles. We give an account of the procedure we used to adapt SweFN to the needs of students of Linguistics in the form of an automatically generated exercise. During this adaptation, the mapping of the fine-grained distinction of roles from SweFN into learner-friendlier coarse-grained roles presented a major challenge. Besides discussing the details of this mapping, we describe the resulting multiple-choice exercise and its graphical user interface. The exercise was made available through Lärka, an online platform for students of Linguistics and learners of Swedish as a second language. We outline also aspects underlying the selection of the incorrect answer options which include semantic as well as frequency-based criteria. Finally, we present our own observations and initial user feedback about the applicability of such a resource in the pedagogical domain. Students' answers indicated an overall positive experience, the majority found the exercise useful for learning semantic roles.

Discovering Frames in Specialized Domains

Marie-Claude L' Homme, Benoît Robichaud and Carlos Subirats Rüggeberg

This paper proposes a method for discovering semantic frames (Fillmore, 1982, 1985; Fillmore et al., 2003) in specialized domains. It is assumed that frames are especially relevant for capturing the lexical structure in specialized domains and that they complement structures such as ontologies that appear better suited to represent specific relationships between entities. The method we devised is based on existing lexical entries recorded in a specialized database related to the field of the environment (erode, impact, melt, recycling, warming). The frames and the data encoded in FrameNet are used as a reference. Selected information was extracted automatically from the database on the environment (and, when possible, compared to FrameNet), and presented to a linguist who analyzed this information to discover potential frames. Several different frames were discovered with this method. About half of them correspond to frames already described in FrameNet; some new frames were also defined and part of these might be specific to the field of the environment.

Developing a French FrameNet: Methodology and First Results

Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël de Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot and Laure Vieu

The Asfalda project aims to develop a French corpus with frame-based semantic annotations and automatic tools for shallow semantic analysis. We present the first part of the project: focusing on a set of notional domains, we delimited a subset of English frames, adapted them to French data when necessary, and developed the corresponding French lexicon. We believe that working domain by domain helped us to enforce the coherence of the resulting resource, and also has the advantage that, though the number of frames is limited (around a hundred), we obtain full coverage within a given domain.

P21 - Semantics

Wednesday, May 28, 18:10 - 19:30

Chairperson: **Peter Anick** Poster Session

Corpus-based Computation of Reverse Associations

Reinhard Rapp

According to psychological learning theory an important principle governing language acquisition is co-occurrence. For example, when we perceive language, our brain seems to unconsciously analyze and store the co-occurrence patterns of the words. And during language production, these co-occurrence patterns are reproduced. The applicability of this principle is particularly obvious in the case of word associations. There is evidence that the associative responses people typically come up with upon presentation of a stimulus word are often words which frequently co-occur with it. It is thus possible to predict a response by looking at co-occurrence data. The work presented here is along these lines. However, it differs from most previous work in that it investigates the direction from the response to the stimulus rather than vice-versa, and that it also deals with the case when several responses are known. Our results indicate that it is possible to predict a stimulus word from its responses, and that it helps if several responses are given.

First Approach toward Semantic Role Labeling for Basque

Haritz Salaberri, Olatz Arregi and Beñat Zapirain

In this paper, we present the first Semantic Role Labeling system developed for Basque. The system is implemented using machine learning techniques and trained with the Reference Corpus for the Processing of Basque (EPEC). In our experiments the classifier that offers the best results is based on Support Vector Machines. Our system achieves 84.30 F1 score in identifying the PropBank semantic role for a given constituent and 82.90 F1 score in identifying the VerbNet role. Our study establishes a baseline for Basque SRL. Although there are no directly comparable systems for English we can state that the results we have achieved are quite good. In addition, we have performed a Leave-One-Out feature selection procedure in order to establish which features are the worthiest regarding argument classification. This will help smooth the way for future stages of Basque SRL and will help draw some of the guidelines of our research.

Constructing a Corpus of Japanese Predicate Phrases for Synonym/Antonym Relations

Tomoko Izumi, Tomohide Shibata, Hisako Asano, Yoshihiro Matsuo and Sadao Kurohashi

We construct a large corpus of Japanese predicate phrases for synonym-antonym relations. The corpus consists of 7,278 pairs of predicates such as "receive-permission (ACC)" vs. "obtain-permission (ACC)", in which each predicate pair is accompanied by a noun phrase and case information. The relations are categorized as synonyms, entailment, antonyms, or unrelated. Antonyms are further categorized into three different classes depending on their aspect of oppositeness. Using the data as a training corpus, we conduct the supervised binary classification of synonymous predicates based on linguistically-motivated features. Combining features that are characteristic of synonymous predicates with those that are characteristic of antonymous predicates, we succeed in automatically identifying synonymous predicates at the high F-score of 0.92, a 0.4 improvement over the baseline method of using the Japanese WordNet. The results of an experiment confirm that the quality of the corpus is high enough to achieve automatic classification. To the best of our knowledge, this is the first and the largest publicly available corpus of Japanese predicate phrases for synonym-antonym relations.

Distributed Distributional Similarities of Google Books over the Centuries

Martin Riedl, Richard Steuer and Chris Biemann

This paper introduces a distributional thesaurus and sense clusters computed on the complete Google Syntactic N-grams, which is extracted from Google Books, a very large corpus of digitized books published between 1520 and 2008. We show that a thesaurus computed on such a large text basis leads to much better results than using smaller corpora like Wikipedia. We also provide distributional thesauri for equal-sized time slices of the corpus. While distributional thesauri can be used as lexical resources in NLP tasks, comparing word similarities over time can unveil sense change of terms across different decades or centuries, and can serve as a resource for diachronic lexicography. Thesauri and clusters are available for download.

Lexical Substitution Dataset for German

Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler and Iryna Gurevych

This article describes a lexical substitution dataset for German. The whole dataset contains 2,040 sentences from the German Wikipedia, with one target word in each sentence. There are

51 target nouns, 51 adjectives, and 51 verbs randomly selected from 3 frequency groups based on the lemma frequency list of the German WaCKy corpus. 200 sentences have been annotated by 4 professional annotators and the remaining sentences by 1 professional annotator and 5 additional annotators who have been recruited via crowdsourcing. The resulting dataset can be used to evaluate not only lexical substitution systems, but also different sense inventories and word sense disambiguation systems.

Buy One Get One Free: Distant Annotation of Chinese Tense, Event Type and Modality

Nianwen Xue and Yuchen Zhang

We describe a "distant annotation" method where we mark up the semantic tense, event type, and modality of Chinese events via a word-aligned parallel corpus. We first map Chinese verbs to their English counterparts via word alignment, and then annotate the resulting English text spans with coarse-grained categories for semantic tense, event type, and modality that we believe apply to both English and Chinese. Because English has richer morpho-syntactic indicators for semantic tense, event type and modality than Chinese, our intuition is that this distant annotation approach will yield more consistent annotation than if we annotate the Chinese side directly. We report experimental results that show stable annotation agreement statistics and that event type and modality have significant influence on tense prediction. We also report the size of the annotated corpus that we have obtained, and how different domains impact annotation consistency.

Latent Semantic Analysis Models on Wikipedia and TASA

Dan Stefanescu, Rajendra Banjade and Vasile Rus

This paper introduces a collection of freely available Latent Semantic Analysis models built on the entire English Wikipedia and the TASA corpus. The models differ not only on their source, Wikipedia versus TASA, but also on the linguistic items they focus on: all words, content-words, nouns-verbs, and main concepts. Generating such models from large datasets (e.g. Wikipedia), that can provide a large coverage for the actual vocabulary in use, is computationally challenging, which is the reason why large LSA models are rarely available. Our experiments show that for the task of word-to-word similarity, the scores assigned by these models are strongly correlated with human judgment, outperforming many other frequently used measures, and comparable to the state of the art.

Annotation of Computer Science Papers for Semantic Relation Extrac-tion

Yuka Tateisi, Yo Shidahara, Yusuke Miyao and Akiko Aizawa

We designed a new annotation scheme for formalising relation structures in research papers, through the investigation of computer science papers. The annotation scheme is based on the hypothesis that identifying the role of entities and events that are described in a paper is useful for intelligent information retrieval in academic literature, and the role can be determined by the relationship between the author and the described entities or events, and relationships among them. Using the scheme, we have annotated research abstracts from the IPSJ Journal published in Japanese by the Information Processing Society of Japan. On the basis of the annotated corpus, we have developed a prototype information extraction system which has the facility to classify sentences according to the relationship between entities mentioned, to help find the role of the entity in which the searcher is interested.

Automatic Extraction of Synonyms for German Particle Verbs from Parallel Data with Distributional Similarity as a Re-Ranking Feature

Moritz Wittmann, Marion Weller and Sabine Schulte im Walde

We present a method for the extraction of synonyms for German particle verbs based on a word-aligned German-English parallel corpus: by translating the particle verb to a pivot, which is then translated back, a set of synonym candidates can be extracted and ranked according to the respective translation probabilities. In order to deal with separated particle verbs, we apply re-ordering rules to the German part of the data. In our evaluation against a gold standard, we compare different pre-processing strategies (lemmatized vs. inflected forms) and introduce language model scores of synonym candidates in the context of the input particle verb as well as distributional similarity as additional re-ranking criteria. Our evaluation shows that distributional similarity as a re-ranking feature is more robust than language model scores and leads to an improved ranking of the synonym candidates. In addition to evaluating against a gold standard, we also present a small-scale manual evaluation.

DBpedia Domains: Augmenting DBpedia with Domain Information

Gregor Titze, Volha Bryl, Cäcilia Zirn and Simone Paolo Ponzetto

We present an approach for augmenting DBpedia, a very large ontology lying at the heart of the Linked Open Data (LOD) cloud, with domain information. Our approach uses the thematic labels provided for DBpedia entities by Wikipedia categories, and groups them based on a kernel based k-means clustering algorithm. Experiments on gold-standard data show that our approach provides a first solution to the automatic annotation of

DBpedia entities with domain labels, thus providing the largest LOD domain-annotated ontology to date.

Classifying Inconsistencies in DBpedia Language Specific Chapters

Elena Cabrio, Serena Villata and Fabien Gandon

This paper proposes a methodology to identify and classify the semantic relations holding among the possible different answers obtained for a certain query on DBpedia language specific chapters. The goal is to reconcile information provided by language specific DBpedia chapters to obtain a consistent results set. Starting from the identified semantic relations between two pieces of information, we further classify them as positive or negative, and we exploit bipolar abstract argumentation to represent the result set as a unique graph, where using argumentation semantics we are able to detect the (possible multiple) consistent sets of elements of the query result. We experimented with the proposed methodology over a sample of triples extracted from 10 DBpedia ontology properties. We define the LingRel ontology to represent how the extracted information from different chapters is related to each other, and we map the properties of the LingRel ontology to the properties of the SIOC-Argumentation ontology to built argumentation graphs. The result is a pilot resource that can be profitably used both to train and to evaluate NLP applications querying linked data in detecting the semantic relations among the extracted values, in order to output consistent information sets.

P22 - Speech Resources

Wednesday, May 28, 18:10 - 19:30

Chairperson: **Giuseppe Riccardi** Poster Session

The Database for Spoken German – DGD2

Thomas Schmidt

The Database for Spoken German (Datenbank für Gesprochenes Deutsch, DGD2, http://dgd.ids-mannheim.de) is the central platform for publishing and disseminating spoken language corpora from the Archive of Spoken German (Archiv für Gesprochenes Deutsch, AGD, http://agd.ids-mannheim.de) at the Institute for the German Language in Mannheim. The corpora contained in the DGD2 come from a variety of sources, some of them in-house projects, some of them external projects. Most of the corpora were originally intended either for research into the (dialectal) variation of German or for studies in conversation analysis and related fields. The AGD has taken over the task of permanently archiving these resources and making them available for reuse to the research community. To date, the DGD2 offers access to 19 different corpora, totalling around 9000 speech events, 2500 hours of audio recordings or 8 million transcribed words. This paper gives an overview of the data made available via the DGD2, of the technical basis for its implementation, and of the most important functionalities it offers. The paper concludes with information about the users of the database and future plans for its development.

The EASR Corpora of European Portuguese, French, Hungarian and Polish Elderly Speech

Annika Hämäläinen, Jairo Avelar, Silvia Rodrigues, Miguel Sales Dias, Artur Kolesiński, Tibor Fegyó, Géza Németh, Petra Csobánka, Karine Lan and David Hewson

Currently available speech recognisers do not usually work well with elderly speech. This is because several characteristics of speech (e.g. fundamental frequency, jitter, shimmer and harmonic noise ratio) change with age and because the acoustic models used by speech recognisers are typically trained with speech collected from younger adults only. To develop speech-driven applications capable of successfully recognising elderly speech, this type of speech data is needed for training acoustic models from scratch or for adapting acoustic models trained with younger adults' speech. However, the availability of suitable elderly speech corpora is still very limited. This paper describes an ongoing project to design, collect, transcribe and annotate large elderly speech corpora for four European languages: Portuguese, French, Hungarian and Polish. The Portuguese, French and Polish corpora contain read speech only, whereas the Hungarian corpus also contains spontaneous command and control type of speech. Depending on the language in question, the corpora contain 76 to 205 hours of speech collected from 328 to 986 speakers aged 60 and over. The final corpora will come with manually verified orthographic transcriptions, as well as annotations for filled pauses, noises and damaged words.

GRASS: the Graz corpus of Read And Spontaneous Speech

Barbara Schuppler, Martin Hagmueller, Juan A. Morales-Cordovilla and Hannes Pessentheiner

This paper provides a description of the preparation, the speakers, the recordings, and the creation of the orthographic transcriptions of the first large scale speech database for Austrian German. It contains approximately 1900 minutes of (read and spontaneous) speech produced by 38 speakers. The corpus consists of three components. First, the Conversation Speech (CS) component contains free conversations of one hour length between friends, colleagues, couples, or family members. Second, the Commands Component (CC) contains

commands and keywords which were either read or elicited by pictures. Third, the Read Speech (RS) component contains phonetically balanced sentences and digits. The speech of all components has been recorded at super-wideband quality in a soundproof recording-studio with head-mounted microphones, large-diaphragm microphones, a laryngograph, and with a video camera. The orthographic transcriptions, which have been created and subsequently corrected manually, contain approximately 290 000 word tokens from 15 000 different word types.

Design and Development of an RDB Version of the Corpus of Spontaneous Japanese

Hanae Koiso, Yasuharu Den, Ken'ya Nishikawa and Kikuo Maekawa

In this paper, we describe the design and development of a new version of the Corpus of Spontaneous Japanese (CSJ), which is a large-scale spoken corpus released in 2004. CSJ contains various annotations that are represented in XML format (CSJ-XML). CSJ-XML, however, is very complicated and suffers from some problems. To overcome this problem, we have developed and released, in 2013, a relational database version of CSJ (CSJ-RDB). CSJ-RDB is based on an extension of the segment and link-based annotation scheme, which we adapted to handle multi-channel and multi-modal streams. Because this scheme adopts a stand-off framework, CSJ-RDB can represent three hierarchical structures at the same time: inter-pausal-unit-top, clause-top, and intonational-phrase-top. CSJ-RDB consists of five different types of tables: segment, unaligned-segment, link, relation, and meta-information tables. The database was automatically constructed from annotation files extracted from CSJ-XML by using general-purpose corpus construction tools. CSJ-RDB enables us to easily and efficiently conduct complex searches required for corpus-based studies of spoken language.

Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process

Camille Fauth, Anne Bonneau, Frank Zimmerer, Juergen Trouvain, Bistra Andreeva, Vincent Colotte, Dominique Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, Odile Mella and Bernd Möbius

We present the design of a corpus of native and non-native speech for the language pair French-German, with a special emphasis on phonetic and prosodic aspects. To our knowledge there is no suitable corpus, in terms of size and coverage, currently available for the target language pair. To select the target L1-L2 interference phenomena we prepare a small preliminary corpus (corpus1), which is analyzed for coverage and cross-checked jointly by French and German experts. Based on this analysis, target phenomena on the phonetic and phonological level are selected on the basis of the expected degree of deviation from the native performance and the frequency of occurrence. 14 speakers performed both L2 (either French or German) and L1 material (either German or French). This allowed us to test, recordings duration, recordings material, the performance of our automatic aligner software. Then, we built corpus2 taking into account what we learned about corpus1. The aims are the same but we adapted speech material to avoid too long recording sessions. 100 speakers will be recorded. The corpus (corpus1 and corpus2) will be prepared as a searchable database, available for the scientific community after completion of the project.

Semi-Automatic Annotation of the UCU Accents Speech Corpus

Rosemary Orr, Marijn Huijbregts, Roeland van Beek, Lisa Teunissen, Kate Backhouse and David van Leeuwen

Annotation and labeling of speech tasks in large multitask speech corpora is a necessary part of preparing a corpus for distribution. We address three approaches to annotation and labeling: manual, semi automatic and automatic procedures for labeling the UCU Accent Project speech data, a multilingual multitask longitudinal speech corpus. Accuracy and minimal time investment are the priorities in assessing the efficacy of each procedure. While manual labeling based on aural and visual input should produce the most accurate results, this approach is error-prone because of its repetitive nature. A semi automatic event detection system requiring manual rejection of false alarms and location and labeling of misses provided the best results. A fully automatic system could not be applied to entire speech recordings because of the variety of tasks and genres. However, it could be used to annotate separate sentences within a specific task. Acoustic confidence measures can correctly detect sentences that do not match the text with an EER of 3.3%

A Corpus of European Portuguese Child and Child-directed Speech

Ana Lúcia Santos, Michel Généreux, Aida Cardoso, Celina Agostinho and Silvana Abalada

We present a corpus of child and child-directed speech of European Portuguese. This corpus results from the expansion of an already existing database (Santos, 2006). It includes around 52 hours of child-adult interaction and now contains 27,595 child utterances and 70,736 adult utterances. The corpus was transcribed according to the CHILDES system (Child Language Data Exchange System) and using the CLAN software (MacWhinney, 2000). The corpus itself represents a valuable

resource for the study of lexical, syntax and discourse acquisition. In this paper, we also show how we used an existing part-of-speech tagger trained on written material (Généreux, Hendrickx & Mendes, 2012) to automatically lemmatize and tag child and child-directed speech and generate a line with part-of-speech information compatible with the CLAN interface. We show that a POS-tagger trained on the analysis of written language can be exploited for the treatment of spoken material with minimal effort, with only a small number of written rules assisting the statistical model.

The SSPNet-Mobile Corpus: Social Signal Processing Over Mobile Phones.

Anna Polychroniou, Hugues Salamin and Alessandro Vinciarelli

This article presents the SSPNet-Mobile Corpus, a collection of 60 mobile phone calls between unacquainted individuals (120 subjects). The corpus is designed to support research on non-verbal behavior and it has been manually annotated into conversational topics and behavioral events (laughter, fillers, back-channel, etc.). Furthermore, the corpus includes, for each subject, psychometric questionnaires measuring personality, conflict attitude and interpersonal attraction. Besides presenting the main characteristics of the corpus (scenario, subjects, experimental protocol, sensing approach, psychometric measurements), the paper reviews the main results obtained so far using the data.

Annotation Pro + TGA: Automation of Speech Timing Analysis

Katarzyna Klessa and Dafydd Gibbon

This paper reports on two tools for the automatic statistical analysis of selected properties of speech timing on the basis of speech annotation files. The tools, one online (TGA, Time Group Analyser) and one offline (Annotation Pro+TGA), are intended to support the rapid analysis of speech timing data without the need to create specific scripts or spreadsheet functions for this purpose. The software calculates, inter alia, mean, median, rPVI, nPVI, slope and intercept functions within interpausal groups, provides visualisations of timing patterns, as well as correlations between these, and parses interpausal groups into hierarchies based on duration relations. Although many studies, especially in speech technology, use computational means, enquiries have shown that a large number of phoneticians and phonetics students do not have script creation skills and therefore use traditional copy+spreadsheet techniques, which are slow, preclude the analysis of large data sets, and are prone to inconsistencies. The present tools have been tested in a number of studies on English, Mandarin and Polish, and are introduced here with reference to results from these studies.

The Munich Biovoice Corpus: Effects of Physical Exercising, Heart Rate, and Skin Conductance on Human Speech Production

Björn Schuller, Felix Friedmann and Florian Eyben

We introduce a spoken language resource for the analysis of impact that physical exercising has on human speech production. In particular, the database provides heart rate and skin conductance measurement information alongside the audio recordings. It contains recordings from 19 subjects in a relaxed state and after exercising. The audio material includes breathing, sustained vowels, and read text. Further, we describe pre-extracted audio-features from our openSMILE feature extractor together with baseline performances for the recognition of high and low heart rate using these features. The baseline results clearly show the feasibility of automatic estimation of heart rate from the human voice, in particular from sustained vowels. Both regression - in order to predict the exact heart rate value - and a binary classification setting for high and low heart rate classes are investigated. Finally, we give tendencies on feature group relevance in the named contexts of heart rate estimation and skin conductivity estimation.

Keynote Speech 1

Thursday, May 29, 9:00

Chairperson: **Khalid Choukri**

Language Technology for Commerce, the eBay Way

Hassan Sawaf

Machine Translation and Human Language Technology plays a key role in expanding the eBay user experience to other countries. But eBay has to use MT very differently from most other companies, so a range of challenges arise. Challenges include the amount, complexity, and type of data, and also the expectations on speed, and the notion what “good” translation is. Hassan will present an overview of eBay’s work in research on language resource management, computational linguistics, machine learning for language technology, machine translation and evaluation.

O17 - Infrastructures for LRs

Thursday, May 29, 9:45

Chairperson: **Nancy Ide** Oral Session

ELRA's Consolidated Services for the HLT Community

Victoria Arranz, Khalid Choukri, Valérie Mapelli and Hélène Mazo

This paper emphasises on ELRA's contribution to the HLT field thanks to the consolidation of its services since LREC 2012. Among the most recent contributions is the establishment of the International Standard Language Resource Number (ISLRN), with the creation and exploitation of an associated web portal to enable the procurement of unique identifiers for Language Resources. Interoperability, consolidation and synchronization remain also a strong focus in ELRA's cataloguing work, in particular with ELRA's involvement in the META-SHARE project, whose platform is to become ELRA's next instrument of sharing LRs. Since last LREC, ELRA has continued its action to offer free LRs to the research community. Cooperation is another watchword within ELRA's activities on multiple aspects: 1) at the legal level, ELRA is supporting the EC in identifying the gaps to be fulfilled to reach harmonized copyright regulations for the HLT community in Europe; 2) at the production level, ELRA is participating in several international projects, in the field of LR production and evaluation of technologies; 3) at the communication level, ELRA has organised the NLP12 meeting with the aim of boosting co-operation and strengthening the bridges between various communities.

The Strategic Impact of META-NET on the Regional, National and International Level

Georg Rehm, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen Garcia-Mateo, Josef van Genabith, Jan Hajic, Inma Hernaez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Linden, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asuncion Moreno, Jan Odijk, Maciej Ogrodniczuk, Piotr Pezik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Michael Rosner, Bolette Pedersen, Inguna Skadina, Koenraad de Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiljevs, Kadri Vider and Jolanta Zabarskaite

This article provides an overview of the dissemination work carried out in META-NET from 2010 until early 2014; we describe its impact on the regional, national and international level, mainly with regard to politics and the situation of funding for LT topics. This paper documents the initiative's work throughout Europe in order to boost progress and innovation in our field.

The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars

Erhard Hinrichs and Steven Krauwer

CLARIN is the short name for the Common Language Resources and Technology Infrastructure, which aims at providing easy and sustainable access for scholars in the humanities and social sciences to digital language data and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of where they are located. CLARIN is in the process of building a networked federation of European data repositories, service centers and centers of expertise, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centers will be interoperable so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. Interoperability of language resources and tools in the federation of CLARIN Centers is ensured by adherence to TEI and ISO standards for text encoding, by the use of persistent identifiers, and by the observance of common protocols. The purpose of the present paper is to give an overview of language resources, tools, and services that CLARIN presently offers.

META-SHARE: One Year After

Stelios Piperidis, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini and Christian Girardi

This paper presents META-SHARE (www.meta-share.eu), an open language resource infrastructure, and its usage since its Europe-wide deployment in early 2013. META-SHARE is a network of repositories that store language resources (data, tools and processing services) documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access. META-SHARE was developed by META-NET (www.meta-net.eu) and aims to serve as an important component of a language technology marketplace for researchers, developers, professionals and industrial players, catering for the full development cycle of language technology, from research through to innovative products and services. The observed usage in its initial steps, the steadily increasing number of network nodes, resources, users, queries, views and downloads are all

encouraging and considered as supportive of the choices made so far. In tandem, take-up activities like direct linking and processing of datasets by language processing services as well as metadata transformation to RDF are expected to open new avenues for data and resources linking and boost the organic growth of the infrastructure while facilitating language technology deployment by much wider research communities and industrial sectors.

New Directions for Language Resource Development and Distribution

Christopher Cieri, Denise DiPersio, Mark Liberman, Andrea Mazzucchi, Stephanie Strassel and Jonathan Wright

Despite the growth in the number of linguistic data centers around the world, their accomplishments and expansions and the advances they have help enable, the language resources that exist are a small fraction of those required to meet the goals of Human Language Technologies (HLT) for the world's languages and the promises they offer: broad access to knowledge, direct communication across language boundaries and engagement in a global community. Using the Linguistic Data Consortium as a focus case, this paper sketches the progress of data centers, summarizes recent activities and then turns to several issues that have received inadequate attention and proposes some new approaches to their resolution.

O18 - Speech Resources Annotation

Thursday, May 29, 9:45

Chairperson: **Satoshi Nakamura** Oral Session

Designing the Latvian Speech Recognition Corpus

Mārcis Pinnis, Ilze Auzina and Kārlis Goba

In this paper the authors present the first Latvian speech corpus designed specifically for speech recognition purposes. The paper outlines the decisions made in the corpus designing process through analysis of related work on speech corpora creation for different languages. The authors provide also guidelines that were used for the creation of the Latvian speech recognition corpus. The corpus creation guidelines are fairly general for them to be re-used by other researchers when working on different language speech recognition corpora. The corpus consists of two parts – an orthographically annotated corpus containing 100 hours of orthographically transcribed audio data and a phonetically annotated corpus containing 4 hours of phonetically transcribed audio data. Metadata files in XML format provide additional details about the speakers, noise levels, speech styles, etc. The speech recognition corpus is phonetically balanced and phonetically rich and the paper describes also the methodology how the phonetical balancedness has been assessed.

A Corpus of Spontaneous Speech in Lectures: The KIT Lecture Corpus for Spoken Language Processing and Translation

Eunah Cho, Sarah Fünfer, Sebastian Stüker and Alex Waibel

With the increasing number of applications handling spontaneous speech, the needs to process spoken languages become stronger. Speech disfluency is one of the most challenging tasks to deal with in automatic speech processing. As most applications are trained with well-formed, written texts, many issues arise when processing spontaneous speech due to its distinctive characteristics. Therefore, more data with annotated speech disfluencies will help the adaptation of natural language processing applications, such as machine translation systems. In order to support this, we have annotated speech disfluencies in German lectures at KIT. In this paper we describe how we annotated the disfluencies in the data and provide detailed statistics on the size of the corpus and the speakers. Moreover, machine translation performance on a source text including disfluencies is compared to the results of the translation of a source text without different sorts of disfluencies or no disfluencies at all.

The Pragmatic Annotation of a Corpus of Academic Lectures

Sian Alsop and Hilary Nesi

This paper will describe a process of 'pragmatic annotation' (c.f. Simpson-Vlach and Leicher 2006) which systematically identifies pragmatic meaning in spoken text. The annotation of stretches of text that perform particular pragmatic functions allows conclusions to be drawn across data sets at a different level than that of the individual lexical item, or structural content. The annotation of linguistic features, which cannot be identified by purely objective means, is distinguished here from structural mark-up of speaker identity, turns, pauses etc. The features annotated are 'explaining', 'housekeeping', 'humour', 'storytelling' and 'summarising'. Twenty-two subcategories are attributed to these elements. Data is from the Engineering Lecture Corpus (ELC), which includes 76 English-medium engineering lectures from the UK, New Zealand and Malaysia. The annotation allows us to compare differences in the use of these discourse features across cultural subcorpora. Results show that cultural context does impact on the linguistic realisation of commonly occurring discourse features in engineering lectures.

HESITA(te) in Portuguese

Sara Candeias, Dirce Celorico, Jorge Proença, Arlindo Veiga, Carla Lopes and Fernando Perdigão

Hesitations, so-called disfluencies, are a characteristic of spontaneous speech, playing a primary role in its structure, reflecting aspects of the language production and the management of inter-communication. In this paper we intend to present a database of hesitations in European Portuguese speech - HESITA - as a relevant base of work to study a variety of speech phenomena. Patterns of hesitations, hesitation distribution according to speaking style, and phonetic properties of the fillers are some of the characteristics we extrapolated from the HESITA database. This database also represents an important resource for improvement in synthetic speech naturalness as well as in robust acoustic modelling for automatic speech recognition. The HESITA database is the output of a project in the speech-processing field for European Portuguese held by an interdisciplinary group in intimate articulation between engineering tools and experience and the linguistic approach.

VOCE Corpus: Ecologically Collected Speech Annotated with Physiological and Psychological Stress Assessments

Ana Aguiar, Mariana Kaiseler, Hugo Meinedo, Pedro Almeida, Mariana Cunha and Jorge Silva

Public speaking is a widely requested professional skill, and at the same time an activity that causes one of the most common adult phobias (Miller and Stone, 2009). It is also known that the study of stress under laboratory conditions, as it is most commonly done, may provide only limited ecological validity (Wilhelm and Grossman, 2010). Previously, we introduced an inter-disciplinary methodology to enable collecting a large amount of recordings under consistent conditions (Aguiar et al., 2013). This paper introduces the VOCE corpus of speech annotated with stress indicators under naturalistic public speaking (PS) settings, and makes it available at http://paginas.fe.up.pt/ voce/articles.html. The novelty of this corpus is that the recordings are carried out in objectively stressful PS situations, as recommended in (Zanstra and Johnston, 2011). The current database contains a total of 38 recordings, 13 of which contain full psychologic and physiologic annotation. We show that the collected recordings validate the assumptions of the methodology, namely that participants experience stress during the PS events. We describe the various metrics that can be used for physiologic and psychologic annotation, and we characterise the sample collected so far, providing evidence that demographics do not affect the relevant psychologic or physiologic annotation. The collection activities are on-going, and we expect to increase the number of complete recordings in the corpus to 30 by June 2014.

O19 - Summarisation

Thursday, May 29, 9:45

Chairperson: **Horacio Saggion** Oral Session

The Impact of Cohesion Errors in Extraction-based Summaries

Evelina Rennes and Arne Jonsson

We present results from an eye tracking study of automatic text summarization. Automatic text summarization is a growing field due to the modern world's Internet based society, but to automatically create perfect summaries is challenging. One problem is that extraction based summaries often have cohesion errors. By the usage of an eye tracking camera, we have studied the nature of four different types of cohesion errors occurring in extraction based summaries. A total of 23 participants read and rated four different texts and marked the most difficult areas of each text. Statistical analysis of the data revealed that absent cohesion or context and broken anaphoric reference (pronouns) caused some disturbance in reading, but that the impact is restricted to the effort to read rather than the comprehension of the text. However, erroneous anaphoric references (pronouns) were not always detected by the participants which poses a problem for automatic text summarizers. The study also revealed other potential disturbing factors.

Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline

Matthew Shardlow

Lexical simplification is the task of automatically reducing the complexity of a text by identifying difficult words and replacing them with simpler alternatives. Whilst this is a valuable application of natural language generation, rudimentary lexical simplification systems suffer from a high error rate which often results in nonsensical, non-simple text. This paper seeks to characterise and quantify the errors which occur in a typical baseline lexical simplification system. We expose 6 distinct categories of error and propose a classification scheme for these. We also quantify these errors for a moderate size corpus, showing the magnitude of each error type. We find that for 183 identified simplification instances, only 19 (10.38%) result in a valid simplification, with the rest causing errors of varying gravity.

LQVSumm: A Corpus of Linguistic Quality Violations in Multi-Document Summarization

Annemarie Friedrich, Marina Valeeva and Alexis Palmer

We present LQVSumm, a corpus of about 2000 automatically created extractive multi-document summaries from the TAC 2011

shared task on Guided Summarization, which we annotated with several types of linguistic quality violations. Examples for such violations include pronouns that lack antecedents or ungrammatical clauses. We give details on the annotation scheme and show that inter-annotator agreement is good given the open-ended nature of the task. The annotated summaries have previously been scored for Readability on a numeric scale by human annotators in the context of the TAC challenge; we show that the number of instances of violations of linguistic quality of a summary correlates with these intuitively assigned numeric scores. On a system-level, the average number of violations marked in a system's summaries achieves higher correlation with the Readability scores than current supervised state-of-the-art methods for assigning a single readability score to a summary. It is our hope that our corpus facilitates the development of methods that not only judge the linguistic quality of automatically generated summaries as a whole, but which also allow for detecting, labeling, and fixing particular violations in a text.

Summarizing News Clusters on the Basis of Thematic Chains

Natalia Loukachevitch and Aleksey Alekseev

In this paper we consider a method for extraction of sets of semantically similar language expressions representing different partici-pants of the text story – thematic chains. The method is based on the structural organization of news clusters and exploits comparison of various contexts of words. The word contexts are used as a basis for extracting multiword expressions and constructing thematic chains. The main difference of thematic chains in comparison with lexical chains is the basic principle of their construction: thematic chains are intended to model different participants (concrete or abstract) of the situation described in the analyzed texts, what means that elements of the same thematic chain cannot often co-occur in the same sentences of the texts under consideration. We evaluate our method on the multi-document summarization task

A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin and Ani Nenkova

In the period since 2004, many novel sophisticated approaches for generic multi-document summarization have been developed. Intuitive simple approaches have also been shown to perform unexpectedly well for the task. Yet it is practically impossible to compare the existing approaches directly, because systems have been evaluated on different datasets, with different evaluation measures, against different sets of comparison systems. Here we present a corpus of summaries produced by several state-of-the-art extractive summarization systems or by popular baseline systems. The inputs come from the 2004 DUC evaluation, the latest year in which generic summarization was addressed in a shared task. We use the same settings for ROUGE automatic evaluation to compare the systems directly and analyze the statistical significance of the differences in performance. We show that in terms of average scores the state-of-the-art systems appear similar but that in fact they produce very different summaries. Our corpus will facilitate future research on generic summarization and motivates the need for development of more sensitive evaluation measures and for approaches to system combination in summarization.

O20 - Grammar, Lexicon and Morphology

Thursday, May 29, 9:45

Chairperson: **Lori Levin** Oral Session

The Interplay Between Lexical and Syntactic Resources in Incremental Parsebanking

Victoria Rosén, Petter Haugereid, Martha Thunes, Gyri S. Losnegaard and Helge Dyvik

Automatic syntactic analysis of a corpus requires detailed lexical and morphological information that cannot always be harvested from traditional dictionaries. In building the INESS Norwegian treebank, it is often the case that necessary lexical information is missing in the morphology or lexicon. The approach used to build the treebank is incremental parsebanking; a corpus is parsed with an existing grammar, and the analyses are efficiently disambiguated by annotators. When the intended analysis is unavailable after parsing, the reason is often that necessary information is not available in the lexicon. INESS has therefore implemented a text preprocessing interface where annotators can enter unrecognized words before parsing. This may concern words that are unknown to the morphology and/or lexicon, and also words that are known, but for which important information is missing. When this information is added, either during text preprocessing or during disambiguation, the result is that after reparsing the intended analysis can be chosen and stored in the treebank. The lexical information added to the lexicon in this way may be of great interest both to lexicographers and to other language technology efforts, and the enriched lexical resource being developed will be made available at the end of the project.

An Efficient Language Independent Toolkit for Complete Morphological Disambiguation

László Laki and György Orosz

In this paper a Moses SMT toolkit-based language-independent complete morphological annotation tool is presented called

HuLaPos2. Our system performs PoS tagging and lemmatization simultaneously. Amongst others, the algorithm used is able to handle phrases instead of unigrams, and can perform the tagging in a not strictly left-to-right order. With utilizing these gains, our system outperforms the HMM-based ones. In order to handle the unknown words, a suffix-tree based guesser was integrated into HuLaPos2. To demonstrate the performance of our system it was compared with several systems in different languages and PoS tag sets. In general, it can be concluded that the quality of HuLaPos2 is comparable with the state-of-the-art systems, and in the case of PoS tagging it outperformed many available systems.

Language Resource Addition: Dictionary or Corpus?

Shinsuke Mori and Graham Neubig

In this paper, we investigate the relative effect of two strategies of language resource additions to the word segmentation problem and part-of-speech tagging problem in Japanese. The first strategy is adding entries to the dictionary and the second is adding annotated sentences to the training corpus. The experimental results showed that the annotated sentence addition to the training corpus is better than the entries addition to the dictionary. And the annotated sentence addition is efficient especially when we add new words with contexts of three real occurrences as partially annotated sentences. According to this knowledge, we executed annotation on the invention disclosure texts and observed word segmentation accuracy.

Utilizing Constituent Structure for Compound Analysis

Kristín Bjarnadóttir and Jón Daðason

Compounding is extremely productive in Icelandic and multi-word compounds are common. The likelihood of finding previously unseen compounds in texts is thus very high, which makes out-of-vocabulary words a problem in the use of NLP tools. The tool de-scribed in this paper splits Icelandic compounds and shows their binary constituent structure. The probability of a constituent in an unknown (or unanalysed) compound forming a combined constituent with either of its neighbours is estimated, with the use of data on the constituent structure of over 240 thousand compounds from the Database of Modern Icelandic Inflection, and word frequencies from Íslenskur orðasjóður, a corpus of approx. 550 million words. Thus, the structure of an unknown compound is derived by com-parison with compounds with partially the same constituents and similar structure in the training data. The granularity of the split re-turned by the decompounder is important in tasks such as semantic analysis or machine translation, where a flat (non-structured) se-quence of constituents is insufficient.

Word Semantic Similarity for Morphologically Rich Languages

Kalliopi Zervanou, Elias Iosif and Alexandros Potamianos

In this work, we investigate the role of morphology on the performance of semantic similarity for morphologically rich languages, such as German and Greek. The challenge in processing languages with richer morphology than English, lies in reducing estimation error while addressing the semantic distortion introduced by a stemmer or a lemmatiser. For this purpose, we propose a methodology for selective stemming, based on a semantic distortion metric. The proposed algorithm is tested on the task of similarity estimation between words using two types of corpus-based similarity metrics: co-occurrence-based and context-based. The performance on morphologically rich languages is boosted by stemming with the context-based metric, unlike English, where the best results are obtained by the co-occurrence-based metric. A key finding is that the estimation error reduction is different when a word is used as a feature, rather than when it is used as a target word.

P23 - Collaborative Resource Construction

Thursday, May 29, 9:45

Chairperson: **Christian Chiarcos** Poster Session

Digital Library 2.0: Source of Knowledge and Research Collaboration Platform

Włodzimierz Gruszczyński and Maciej Ogrodniczuk

Digital libraries are frequently treated just as a new method of storage of digitized artifacts, with all consequences of transferring long-established ways of dealing with physical objects into the digital world. Such attitude improves availability, but often neglects other opportunities offered by global and immediate access, virtuality and linking – as easy as never before. The article presents the idea of transforming a conventional digital library into knowledge source and research collaboration platform, facilitating content augmentation, interpretation and co-operation of geographically distributed researchers representing different academic fields. This concept has been verified by the process of extending descriptions stored in thematic Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries with extended item-associated information provided by historians, philologists, librarians and computer scientists. It resulted in associating the customary fixed metadata and digitized content with historical comments, mini-dictionaries of foreign interjections or explanation of less-known background details.

Exploiting networks in Law

Livio Robaldo, Guido Boella, Luigi di Caro and Andrea Violato

In this paper we first introduce the working context related to the understanding of an heterogeneous network of references contained in the Italian regulatory framework. We then present an extended analysis of a large network of laws, providing several types of analytical evaluation that can be used within a legal management system for understanding the data through summarization, visualization, and browsing. In the legal domain, yet several tasks are strictly supervised by humans, with strong consumption of time and energy that would dramatically drop with the help of automatic or semi-automatic supporting tools. We overview different techniques and methodologies explaining how they can be helpful in actual scenarios.

Guampa: a Toolkit for Collaborative Translation

Alex Rudnick, Taylor Skidmore, Alberto Samaniego and Michael Gasser

Here we present Guampa, a new software package for online collaborative translation. This system grows out of our discussions with Guarani-language activists and educators in Paraguay, and attempts to address problems faced by machine translation researchers and by members of any community speaking an under-represented language. Guampa enables volunteers and students to work together to translate documents into heritage languages, both to make more materials available in those languages, and also to generate bitext suitable for training machine translation systems. While many approaches to crowdsourcing bitext corpora focus on Mechanical Turk and temporarily engaging anonymous workers, Guampa is intended to foster an online community in which discussions can take place, language learners can practice their translation skills, and complete documents can be translated. This approach is appropriate for the Spanish-Guarani language pair as there are many speakers of both languages, and Guarani has a dedicated activist community. Our goal is to make it easy for anyone to set up their own instance of Guampa and populate it with documents – such as automatically imported Wikipedia articles – to be translated for their particular language pair. Guampa is freely available and relatively easy to use.

The Halliday Centre Tagger: An Online Platform for Semi-automatic Text Annotation and Analysis

Billy T.M. Wong, Ian C. Chow, Jonathan J. Webster and Hengbin Yan

This paper reports the latest development of The Halliday Centre Tagger (the Tagger), an online platform provided with semi-automatic features to facilitate text annotation and analysis. The Tagger is featured for its web-based architecture with all functionalities and file storage space provided online, and a theory-neutral design where users can define their own labels for annotating various kinds of linguistic information. The Tagger is currently optimized for text annotation of Systemic Functional Grammar (SFG), providing by default a pre-defined set of SFG grammatical features, and the function of automatic identification of process types for English verbs. Apart from annotation, the Tagger also offers the features of visualization and summarization to aid text analysis. The visualization feature combines and illustrates multi-dimensional layers of annotation in a unified way of presentation, while the summarization feature categorizes annotated entries according to different SFG systems, i.e., transitivity, theme, logical-semantic relations, etc. Such features help users identify grammatical patterns in an annotated text.

Modeling, Managing, Exposing, and Linking Ontologies with a Wiki-based Tool

Mauro Dragoni, Alessio Bosca, Matteo Casu and Andi Rexha

In the last decade, the need of having effective and useful tools for the creation and the management of linguistic resources significantly increased. One of the main reasons is the necessity of building linguistic resources (LRs) that, besides the goal of expressing effectively the domain that users want to model, may be exploited in several ways. In this paper we present a wiki-based collaborative tool for modeling ontologies, and more in general any kind of linguistic resources, called MoKi. This tool has been customized in the context of an EU-funded project for addressing three important aspects of LRs modeling: (i) the exposure of the created LRs, (ii) for providing features for linking the created resources to external ones, and (iii) for producing multilingual LRs in a safe manner.

Propa-L: a Semantic Filtering Service from a Lexical Network Created using Games With A Purpose

Mathieu Lafourcade and Karën Fort

This article presents Propa-L, a freely accessible Web service that allows to semantically filter a lexical network. The language resources behind the service are dynamic and created through Games With A Purpose. We show an example of application of this service: the generation of a list of keywords for parental filtering on the Web, but many others can be envisaged. Moreover, the propagation algorithm we present here can be applied to any lexical network, in any language.

Open Philology at the University of Leipzig

Frederik Baumgardt, Giuseppe Celano, Gregory R. Crane, Stella Dee, Maryam Foradi, Emily Franzini, Greta Franzini, Monica Lent, Maria Moritz and Simona Stoyanova

The Open Philology Project at the University of Leipzig aspires to re-assert the value of philology in its broadest sense. Philology signifies the widest possible use of the linguistic record to enable a deep understanding of the complete lived experience of humanity. Pragmatically, we focus on Greek and Latin because (1) substantial collections and services are already available within these languages, (2) substantial user communities exist (c. 35,000 unique users a month at the Perseus Digital Library), and (3) a European-based project is better positioned to process extensive cultural heritage materials in these languages rather than in Chinese or Sanskrit. The Open Philology Project has been designed with the hope that it can contribute to any historical language that survives within the human record. It includes three tasks: (1) the creation of an open, extensible, repurposable collection of machine-readable linguistic sources; (2) the development of dynamic textbooks that use annotated corpora to customize the vocabulary and grammar of texts that learners want to read, and at the same time engage students in collaboratively producing new annotated data; (3) the establishment of new workflows for, and forms of, publication, from individual annotations with argumentation to traditional publications with integrated machine-actionable data.

LexTerm Manager: Design for an Integrated Lexicography and Terminology System

Joshua Elliot, Logan Kearsley, Jason Housley and Alan Melby

We present a design for a multi-modal database system for lexical information that can be accessed in either lexicographical or terminological views. The use of a single merged data model makes it easy to transfer common information between termbases and dictionaries, thus facilitating information sharing and re-use. Our combined model is based on the LMF and TMF metamodels for lexicographical and terminological databases and is compatible with both, thus allowing for the import of information from existing dictionaries and termbases, which may be transferred to the complementary view and re-exported. We also present a new Linguistic Configuration Model, analogous to a TBX XCS file, which can be used to specify multiple language-specific schemata for validating and understanding lexical information in a single database. Linguistic configurations are mutable and can be refined and evolved over time as understanding of documentary needs improves. The system is designed with a client-server architecture using the HTTP protocol, allowing for the independent implementation of multiple clients for specific use cases and easy deployment over the web.

RESTful Annotation and Efficient Collaboration

Jonathan Wright

As linguistic collection and annotation scale up and collaboration across sites increases, novel technologies are necessary to support projects. Recent events at LDC, namely the move to a web-based infrastructure, the formation of the Software Group, and our involvement in the NSF LAPPS Grid project, have converged on concerns of efficient collaboration. The underlying design of the Web, typically referred to as RESTful principles, is crucial for collaborative annotation, providing data and processing services, and participating in the Linked Data movement. This paper outlines recommendations that will facilitate such collaboration.

P24 - Corpora and Annotation

Thursday, May 29, 9:45

Chairperson: **Maria Gavrilidou** Poster Session

Collecting Natural SMS and Chat Conversations in Multiple Languages: The BOLT Phase 2 Corpus

Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan and Ann Sawyer

The DARPA BOLT Program develops systems capable of allowing English speakers to retrieve and understand information from informal foreign language genres. Phase 2 of the program required large volumes of naturally occurring informal text (SMS) and chat messages from individual users in multiple languages to support evaluation of machine translation systems. We describe the design and implementation of a robust collection system capable of capturing both live and archived SMS and chat conversations from willing participants. We also discuss the challenges recruitment at a time when potential participants have acute and growing concerns about their personal privacy in the realm of digital communication, and we outline the techniques adopted to confront those challenges. Finally, we review the properties of the resulting BOLT Phase 2 Corpus, which comprises over 6.5 million words of naturally-occurring chat and SMS in English, Chinese and Egyptian Arabic.

The Slovak Categorized News Corpus

Daniel Hladek, Jan Stas and Jozef Juhar

The presented corpus aims to be the first attempt to create a representative sample of the contemporary Slovak language from

various domains with easy searching and automated processing. This first version of the corpus contains words and automatic morphological and named entity annotations and transcriptions of abbreviations and numerals. Integral part of the proposed paper is a word boundary and sentence boundary detection algorithm that utilizes characteristic features of the language.

TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation

Matus Pleva and Jozef Juhar

This article presents an overview of the existing acoustical corpuses suitable for broadcast news automatic transcription task in the Slovak language. The TUKE-BNews-SK database created in our department was built to support the application development for automatic broadcast news processing and spontaneous speech recognition of the Slovak language. The audio corpus is composed of 479 Slovak TV broadcast news shows from public Slovak television called STV1 or "Jednotka" containing 265 hours of material and 186 hours of clean transcribed speech (4 hours subset extracted for testing purposes). The recordings were manually transcribed using Transcriber tool modified for Slovak annotators and automatic Slovak spell checking. The corpus design, acquisition, annotation scheme and pronunciation transcription is described together with corpus statistics and tools used. Finally the evaluation procedure using automatic speech recognition is presented on the broadcast news and parliamentary speeches test sets.

Sublanguage Corpus Analysis Toolkit: a Tool for Assessing the Representativeness and Sublanguage Characteristics of Corpora

Irina Temnikova, William A. Baumgartner Jr., Negacy D. Hailu, Ivelina Nikolova, Tony McEnery, Adam Kilgarriff, Galia Angelova and K. Bretonnel Cohen

Sublanguages are varieties of language that form "subsets" of the general language, typically exhibiting particular types of lexical, semantic, and other restrictions and deviance. SubCAT, the Sublanguage Corpus Analysis Toolkit, assesses the representativeness and closure properties of corpora to analyze the extent to which they are either sublanguages, or representative samples of the general language. The current version of SubCAT contains scripts and applications for assessing lexical closure, morphological closure, sentence type closure, over-represented words, and syntactic deviance. Its operation is illustrated with three case studies concerning scientific journal articles, patents, and clinical records. Materials from two language families are analyzed—English (Germanic), and Bulgarian (Slavic). The software is available at sublanguage.sourceforge.net under a liberal Open Source license.

The Hungarian Gigaword Corpus

Csaba Oravecz, Tamás Váradi and Bálint Sass

The paper reports on the development of the Hungarian Gigaword Corpus (HGC), an extended new edition of the Hungarian National Corpus, with upgraded and redesigned linguistic annotation and an increased size of 1.5 billion tokens. Issues concerning the standard steps of corpus collection and preparation are discussed with special emphasis on linguistic analysis and annotation due to Hungarian having some challenging characteristics with respect to computational processing. As the HGC is designed to serve as a resource for a wide range of linguistic research as well as for the interested public, a number of issues had to be resolved which were raised by trying to find a balance between the above two application areas. The following main objectives have been defined for the development of the HGC, focusing on the pivotal concept of increase in: - size: extending the corpus to minimum 1 billion words, - quality: using new technology for development and analysis, - coverage and representativity: taking new samples of language use and including further variants (transcribed spoken language data and user generated content (social media) from the internet in particular).

The SETimes.HR Linguistically Annotated Corpus of Croatian

Željko Agić and Nikola Ljubešić

We present SETimes.HR – the first linguistically annotated corpus of Croatian that is freely available for all purposes. The corpus is built on top of the SETimes parallel corpus of nine Southeast European languages and English. It is manually annotated for lemmas, morphosyntactic tags, named entities and dependency syntax. We couple the corpus with domain-sensitive test sets for Croatian and Serbian to support direct model transfer evaluation between these closely related languages. We build and evaluate statistical models for lemmatization, morphosyntactic tagging, named entity recognition and dependency parsing on top of SETimes.HR and the test sets, providing the state of the art in all the tasks. We make all resources presented in the paper freely available under a very permissive licensing scheme.

caWaC - a Web Corpus of Catalan and its Application to Language Modeling and Machine Translation

Nikola Ljubešić and Antonio Toral

In this paper we present the construction process of a web corpus of Catalan built from the content of the .cat top-level domain. For

collecting and processing data we use the Brno pipeline with the spiderling crawler and its accompanying tools. To the best of our knowledge the corpus represents the largest existing corpus of Catalan containing 687 million words, which is a significant increase given that until now the biggest corpus of Catalan, CuCWeb, counts 166 million words. We evaluate the resulting resource on the tasks of language modeling and statistical machine translation (SMT) by calculating LM perplexity and incorporating the LM in the SMT pipeline. We compare language models trained on different subsets of the resource with those trained on the Catalan Wikipedia and the target side of the parallel data used to train the SMT system.

ACTIV-ES: a Comparable, Cross-Dialect Corpus of "everyday" Spanish from Argentina, Mexico, and Spain

Jerid Francom, Mans Hulden and Adam Ussishkin

Corpus resources for Spanish have proved invaluable for a number of applications in a wide variety of fields. However, a majority of resources are based on formal, written language and/or are not built to model language variation between varieties of the Spanish language, despite the fact that most language in ‘everyday’ use is informal/ dialogue-based and shows rich regional variation. This paper outlines the development and evaluation of the ACTIV-ES corpus, a first-step to produce a comparable, cross-dialect corpus representative of the ‘everyday’ language of various regions of the Spanish-speaking world.

Language Editing Dataset of Academic Texts

Vidas Daudaravicius

We describe the VTeX Language Editing Dataset of Academic Texts (LEDAT), a dataset of text extracts from scientific papers that were edited by professional native English language editors at VTeX. The goal of the LEDAT is to provide a large data resource for the development of language evaluation and grammar error correction systems for the scientific community. We describe the data collection and the compilation process of the LEDAT. The new dataset can be used in many NLP studies and applications where deeper knowledge of the academic language and language editing is required. The dataset can be used also as a knowledge base of English academic language to support many writers of scientific papers.

Annotating the Focus of Negation in Japanese Text

Suguru Matsuyoshi, Ryo Otsuki and Fumiyo Fukumoto

This paper proposes an annotation scheme for the focus of negation in Japanese text. Negation has its scope and the focus within the scope. The scope of negation is the part of the sentence that is negated; the focus is the part of the scope that is most prominently or explicitly negated. In natural language processing, correct interpretation of negated statements requires precise detection of the focus of negation in the statements. As a foundation for developing a negation focus detector for Japanese, we have annotated textdata of "Rakuten Travel: User review data" and the newspaper subcorpus of the "Balanced Corpus of Contemporary Written Japanese" with labels proposed in our annotation scheme. We report 1,327 negation cues and the foci in the corpora, and present classification of these foci based on syntactic types and semantic types. We also propose a system for detecting the focus of negation in Japanese using 16 heuristic rules and report the performance of the system.

A Corpus of Participant Roles in Contentious Discussions

Siddharth Jain, Archna Bhatia, Angelique Rein and Eduard Hovy

The expansion of social roles is, nowadays, a fact due to the ability of users to interact, discuss, exchange ideas and opinions, and form social networks though social media. Users in online social environment play a variety of social roles. The concept of "social role" has long been used in social science describe the intersection of behavioural, meaningful, and structural attributes that emerge regularly in particular settings. In this paper, we present a new corpus for social roles in online contentious discussions. We explore various behavioural attributes such as stubbornness, sensibility, influence, and ignorance to create a model of social roles to distinguish among various social roles participants assume in such setup. We annotate discussions drawn from two different sets of corpora in order to ensure that our model of social roles and their signals hold up in general. We discuss the various criteria for deciding values for each behavioural attributes which define the roles.

P25 - Machine Translation

Thursday, May 29, 9:45

Chairperson: **Holger Schwenk** Poster Session

CFT13: a Resource for Research into the Post-editing Process

Michael Carl, Mercedes Martínez García and Bartolomé Mesa-Lao

This paper describes the most recent dataset that has been added to the CRITT Translation Process Research Database (TPR-DB). Under the name CFT13, this new study contains user activity data (UAD) in the form of key-logging and eye-tracking collected during the second CasMaCat field trial in June

2013. The CFT13 is a publicly available resource featuring a number of simple and compound process and product units suited to investigate human-computer interaction while post-editing machine translation outputs.

Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech

Nianwen Xue, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova and Xiuhong Zhang

Abstract Meaning Representations (AMRs) are rooted, directional and labeled graphs that abstract away from morpho-syntactic idiosyncrasies such as word category (verbs and nouns), word order, and function words (determiners, some prepositions). Because these syntactic idiosyncrasies account for many of the cross-lingual differences, it would be interesting to see if this representation can serve, e.g., as a useful, minimally divergent transfer layer in machine translation. To answer this question, we have translated 100 English sentences that have existing AMRs into Chinese and Czech to create AMRs for them. A cross-linguistic comparison of English to Chinese and Czech AMRs reveals both cases where the AMRs for the language pairs align well structurally and cases of linguistic divergence. We found that the level of compatibility of AMR between English and Chinese is higher than between English and Czech. We believe this kind of comparison is beneficial to further refining the annotation standards for each of the three languages and will lead to more compatible annotation guidelines between the languages.

On Complex Word Alignment Configurations

Miriam Kaeshammer and Anika Westburg

Resources of manual word alignments contain configurations that are beyond the alignment capacity of current translation models, hence the term complex alignment configuration. They have been the matter of some debate in the machine translation community, as they call for more powerful translation models that come with further complications. In this work we investigate instances of complex alignment configurations in data sets of four different language pairs to shed more light on the nature and cause of those configurations. For the English-German alignments from Padó and Lapata (2006), for instance, we find that only a small fraction of the complex configurations are due to real annotation errors. While a third of the complex configurations in this data set could be simplified when annotating according to a different style guide, the remaining ones are phenomena that one would like to be able to generate during translation. Those instances are mainly caused by the different word order of English and German. Our findings thus motivate further research in the area of translation beyond phrase-based and context-free translation modeling.

Shata-Anuvadak: Tackling Multiway Translation of Indian Languages

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah and Pushpak Bhattacharyya

We present a compendium of 110 Statistical Machine Translation systems built from parallel corpora of 11 Indian languages belonging to both Indo-Aryan and Dravidian families. We analyze the relationship between translation accuracy and the language families involved. We feel that insights obtained from this analysis will provide guidelines for creating machine translation systems of specific Indian language pairs. We build phrase based systems and some extensions. Across multiple languages, we show improvements on the baseline phrase based systems using these extensions: (1) source side reordering for English-Indian language translation, and (2) transliteration of untranslated words for Indian language-Indian language translation. These enhancements harness shared characteristics of Indian languages. To stimulate similar innovation widely in the NLP community, we have made the trained models for these language pairs publicly available.

Automatic Annotation of Machine Translation Datasets with Binary Quality Judgements

Marco Turchi and Matteo Negri

The automatic estimation of machine translation (MT) output quality is an active research area due to its many potential applications (e.g. aiding human translation and post-editing, re-ranking MT hypotheses, MT system combination). Current approaches to the task rely on supervised learning methods for which high-quality labelled data is fundamental. In this framework, quality estimation (QE) has been mainly addressed as a regression problem where models trained on (source, target) sentence pairs annotated with continuous scores (in the [0-1] interval) are used to assign quality scores (in the same interval) to unseen data. Such definition of the problem assumes that continuous scores are informative and easily interpretable by different users. These assumptions, however, conflict with the subjectivity inherent to human translation and evaluation. On one side, the subjectivity of human judgements adds noise and biases to annotations based on scaled values. This problem reduces the usability of the resulting datasets, especially in application scenarios where a sharp distinction between "good" and "bad" translations is needed. On the other side, continuous scores are not always sufficient to decide whether a translation is actually acceptable or not. To overcome these issues, we present an

automatic method for the annotation of (source, target) pairs with binary judgements that reflect an empirical, and easily interpretable notion of quality. The method is applied to annotate with binary judgements three QE datasets for different language combinations. The three datasets are combined in a single resource, called BinQE, which can be freely downloaded from http://hlt.fbk.eu/technologies/binqe.

A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation

Violeta Seretan, Pierrette Bouillon and Johanna Gerlach

The user-generated content represents an increasing share of the information available today. To make this type of content instantly accessible in another language, the ACCEPT project focuses on developing pre-editing technologies for correcting the source text in order to increase its translatability. Linguistically-informed pre-editing rules have been developed for English and French for the two domains considered by the project, namely, the technical domain and the healthcare domain. In this paper, we present the evaluation experiments carried out to assess the impact of the proposed pre-editing rules on translation quality. Results from a large-scale evaluation campaign show that pre-editing helps indeed attain a better translation quality for a high proportion of the data, the difference with the number of cases where the adverse effect is observed being statistically significant. The ACCEPT pre-editing technology is freely available online and can be used in any Web-based environment to enhance the translatability of user-generated content so that it reaches a broader audience.

Rule-based Reordering Space in Statistical Machine Translation

Nicolas Pécheux, Alexander Allauzen and François Yvon

In Statistical Machine Translation (SMT), the constraints on word reorderings have a great impact on the set of potential translations that are explored. Notwithstanding computationnal issues, the reordering space of a SMT system needs to be designed with great care: if a larger search space is likely to yield better translations, it may also lead to more decoding errors, because of the added ambiguity and the interaction with the pruning strategy. In this paper, we study this trade-off using a state-of-the art translation system, where all reorderings are represented in a word lattice prior to decoding. This allows us to directly explore and compare different reordering spaces. We study in detail a rule-based preordering system, varying the length or number of rules, the tagset used, as well as contrasting with oracle settings and purely combinatorial subsets of permutations. We focus on two language pairs: English-French, a close language pair and English-German, known to be a more challenging reordering pair.

Hindi to English Machine Translation: Using Effective Selection in Multi-Model SMT

Kunal Sachdeva, Rishabh Srivastava, Sambhav Jain and Dipti Sharma

Recent studies in machine translation support the fact that multi-model systems perform better than the individual models. In this paper, we describe a Hindi to English statistical machine translation system and improve over the baseline using multiple translation models. We have considered phrase based as well as hierarchical models and enhanced over both these baselines using a regression model. The system is trained over textual as well as syntactic features extracted from source and target of the aforementioned translations. Our system shows significant improvement over the baseline systems for both automatic as well as human evaluations. The proposed methodology is quite generic and easily be extended to other language pairs as well.

P26 - Parallel Corpora

Thursday, May 29, 9:45

Chairperson: **Dan Tufiş** Poster Session

Benchmarking of English-Hindi Parallel Corpora

Jayendra Rakesh Yeka, Prasanth Kolachina and Dipti Misra Sharma

In this paper we present several parallel corpora for English↔Hindi and talk about their natures and domains. We also discuss briefly a few previous attempts in MT for translation from English to Hindi. The lack of uniformly annotated data makes it difficult to compare these attempts and precisely analyze their strengths and shortcomings. With this in mind, we propose a standard pipeline to provide uniform linguistic annotations to these resources using state-of-art NLP technologies. We conclude the paper by presenting evaluation scores of different statistical MT systems on the corpora detailed in this paper for English→Hindi and present the proposed plans for future work. We hope that both these annotated parallel corpora resources and MT systems will serve as benchmarks for future approaches to MT in English →Hindi. This was and remains the main motivation for the attempts detailed in this paper.

Transliteration and Alignment of Parallel Texts from Cyrillic to Latin

Petic Mircea and Daniela Gîfu

This article describes a methodology of recovering and preservation of old Romanian texts and problems related to

their recognition. Our focus is to create a gold corpus for Romanian language (the novella Sania), for both alphabets used in Transnistria – Cyrillic and Latin. The resource is available for similar researches. This technology is based on transliteration and semiautomatic alignment of parallel texts at the level of letter/lexem/multiwords. We have analysed every text segment present in this corpus and discovered other conventions of writing at the level of transliteration, academic norms and editorial interventions. These conventions allowed us to elaborate and implement some new heuristics that make a correct automatic transliteration process. Sometimes the words of Latin script are modified in Cyrillic script from semantic reasons (for instance, editor's interpretation). Semantic transliteration is seen as a good practice in introducing multiwords from Cyrillic to Latin. Not only does it preserve how a multiwords sound in the source script, but also enables the translator to modify in the original text (here, choosing the most common sense of an expression). Such a technology could be of interest to lexicographers, but also to specialists in computational linguistics to improve the actual transliteration standards.

Exploiting Catenae in a Parallel Treebank Alignment

Manuela Sanguinetti, Cristina Bosco and Loredana Cupi

This paper aims to introduce the issues related to the syntactic alignment of a dependency-based multilingual parallel treebank, ParTUT. Our approach to the task starts from a lexical mapping and then attempts to expand it using dependency relations. In developing the system, however, we realized that the only dependency relations between the individual nodes were not sufficient to overcome some translation divergences, or shifts, especially in the absence of a direct lexical mapping and a different syntactic realization. For this purpose, we explored the use of a novel syntactic notion introduced in dependency theoretical framework, i.e. that of catena (Latin for "chain"), which is intended as a group of words that are continuous with respect to dominance. In relation to the task of aligning parallel dependency structures, catenae can be used to explain and identify those cases of one-to-many or many-to-many correspondences, typical of several translation shifts, that cannot be detected by means of direct word-based mappings or bare syntactic relations. The paper presented here describes the overall structure of the alignment system as it has been currently designed, how catenae are extracted from the parallel resource, and their potential relevance to the completion of tree alignment in ParTUT sentences.

SwissAdmin: a Multilingual Tagged Parallel Corpus of Press Releases

Yves Scherrer, Luka Nerima, Lorenza Russo, Maria Ivanova and Eric Wehrli

SwissAdmin is a new multilingual corpus of press releases from the Swiss Federal Administration, available in German, French, Italian and English. We provide SwissAdmin in three versions: (i) plain texts of approximately 6 to 8 million words per language; (ii) sentence-aligned bilingual texts for each language pair; (iii) a part-of-speech-tagged version consisting of annotations in both the Universal tagset and the richer Fips tagset, along with grammatical functions, verb valencies and collocations. The SwissAdmin corpus is freely available at www.latl.unige.ch/swissadmin.

UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira and Lu Yi

Parallel corpus is a valuable resource for cross-language information retrieval and data-driven natural language processing systems, especially for Statistical Machine Translation (SMT). However, most existing parallel corpora to Chinese are subject to in-house use, while others are domain specific and limited in size. To a certain degree, this limits the SMT research. This paper describes the acquisition of a large scale and high quality parallel corpora for English and Chinese. The corpora constructed in this paper contain about 15 million English-Chinese (E-C) parallel sentences, and more than 2 million training data and 5,000 testing sentences are made publicly available. Different from previous work, the corpus is designed to embrace eight different domains. Some of them are further categorized into different topics. The corpus will be released to the research community, which is available at the NLP2CT website.

Quality Estimation for Synthetic Parallel Data Generation

Raphael Rubino, Antonio Toral, Nikola Ljubešić and Gema Ramírez-Sánchez

This paper presents a novel approach for parallel data generation using machine translation and quality estimation. Our study focuses on pivot-based machine translation from English to Croatian through Slovene. We generate an English–Croatian version of the Europarl parallel corpus based on the English–Slovene Europarl corpus and the Apertium rule-based translation system for Slovene–Croatian. These experiments are to be considered as a first step towards the generation of reliable synthetic parallel data for under-resourced languages.

We first collect small amounts of aligned parallel data for the Slovene–Croatian language pair in order to build a quality estimation system for sentence-level Translation Edit Rate (TER) estimation. We then infer TER scores on automatically translated Slovene to Croatian sentences and use the best translations to build an English–Croatian statistical MT system. We show significant improvement in terms of automatic metrics obtained on two test sets using our approach compared to a random selection of synthetic parallel data.

Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus

Raivis Skadinš, Jörg Tiedemann, Roberts Rozis and Daiga Deksne

The European Union is a great source of high quality documents with translations into several languages. Parallel corpora from its publications are frequently used in various tasks, machine translation in particular. A source that has not systematically been explored yet is the EU Bookshop – an online service and archive of publications from various European institutions. The service contains a large body of publications in the 24 official of the EU. This paper describes our efforts in collecting those publications and converting them to a format that is useful for natural language processing in particular statistical machine translation. We report our procedure of crawling the website and various pre-processing steps that were necessary to clean up the data after the conversion from the original PDF files. Furthermore, we demonstrate the use of this dataset in training SMT models for English, French, German, Spanish, and Latvian.

The AMARA Corpus: Building Parallel Language Resources for the Educational Domain

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad and Stephan Vogel

This paper presents the AMARA corpus of on-line educational content: a new parallel corpus of educational video subtitles, multilingually aligned for 20 languages, i.e. 20 monolingual corpora and 190 parallel corpora. This corpus includes both resource-rich languages such as English and Arabic, and resource-poor languages such as Hindi and Thai. In this paper, we describe the gathering, validation, and preprocessing of a large collection of parallel, community-generated subtitles. Furthermore, we describe the methodology used to prepare the data for Machine Translation tasks. Additionally, we provide a document-level, jointly aligned development and test sets for 14 language pairs, designed for tuning and testing Machine Translation systems. We provide baseline results for these tasks, and highlight some of the challenges we face when building machine translation systems for educational content.

Incorporating Alternate Translations into English Translation Treebank

Ann Bies, Justin Mott, Seth Kulick, Jennifer Garland and Colin Warner

New annotation guidelines and new processing methods were developed to accommodate English treebank annotation of a parallel English/Chinese corpus of web data that includes alternate English translations (one fluent, one literal) of expressions that are idiomatic in the Chinese source. In previous machine translation programs, alternate translations of idiomatic expressions had been present in untreebanked data only, but due to the high frequency of such expressions in informal genres such as discussion forums, machine translation system developers requested that alternatives be added to the treebanked data as well. In consultation with machine translation researchers, we chose a pragmatic approach of syntactically annotating only the fluent translation, while retaining the alternate literal translation as a segregated node in the tree. Since the literal translation alternates are often incompatible with English syntax, this approach allows us to create fluent trees without losing information. This resource is expected to support machine translation efforts, and the flexibility provided by the alternate translations is an enhancement to the treebank for this purpose.

Dual Subtitles as Parallel Corpora

Shikun Zhang, Wang Ling and Chris Dyer

In this paper, we leverage the existence of dual subtitles as a source of parallel data. Dual subtitles present viewers with two languages simultaneously, and are generally aligned in the segment level, which removes the need to automatically perform this alignment. This is desirable as extracted parallel data does not contain alignment errors present in previous work that aligns different subtitle files for the same movie. We present a simple heuristic to detect and extract dual subtitles and show that more than 20 million sentence pairs can be extracted for the Mandarin-English language pair. We also show that extracting data from this source can be a viable solution for improving Machine Translation systems in the domain of subtitles.

Aligning Parallel Texts with InterText

Pavel Vondřička

InterText is a flexible manager and editor for alignment of parallel texts aimed both at individual and collaborative creation of parallel corpora of any size or translational memories. It is available in

two versions: as a multi-user server application with a web-based interface and as a native desktop application for personal use. Both versions are able to cooperate with each other. InterText can process plain text or custom XML documents, deploy existing automatic aligners and provide a comfortable interface for manual post-alignment correction of both the alignment and the text contents and segmentation of the documents. One language version may be aligned with several other versions (using stand-off alignment) and the application ensures consistency between them. The server version supports different user levels and privileges and it can also track changes made to the texts for easier supervision. It also allows for batch import, alignment and export and can be connected to other tools and scripts for better integration in a more complex project workflow.

P27 - Sign Language

Thursday, May 29, 9:45

Chairperson: **Thomas Hanke** Poster Session

Expanding N-gram Analytics in ELAN and a Case Study for Sign Synthesis

Rosalee Wolfe, John McDonald, Larwan Berke and Marie Stumbo

Corpus analysis is a powerful tool for signed language synthesis. A new extension to ELAN offers expanded n-gram analysis tools including improved search capabilities and an extensive library of statistical measures of association for n-grams. Uncovering and exploring coarticulatory timing effects via corpus analysis requires n-gram analysis to discover the most frequently occurring bigrams. This paper presents an overview of the new tools and a case study in American Sign Language synthesis that exploits these capabilities for computing more natural timing in generated sentences. The new extension provides a time-saving convenience for language researchers using ELAN.

SLMotion - an Extensible Sign Language Oriented Video Analysis Tool

Matti Karppa, Ville Viitaniemi, Marcos Luzardo, Jorma Laaksonen and Tommi Jantunen

We present a software toolkit called SLMotion which provides a framework for automatic and semiautomatic analysis, feature extraction and annotation of individual sign language videos, and which can easily be adapted to batch processing of entire sign language corpora. The program follows a modular design, and exposes a Numpy-compatible Python application programming interface that makes it easy and convenient to extend its functionality through scripting. The program includes support for exporting the annotations in ELAN format. The program is released as free software, and is available for GNU/Linux and MacOS platforms.

S-pot - a Benchmark in Spotting Signs Within Continuous Signing

Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa and Jorma Laaksonen

In this paper we present S-pot, a benchmark setting for evaluating the performance of automatic spotting of signs in continuous sign language videos. The benchmark includes 5539 video files of Finnish Sign Language, ground truth sign spotting results, a tool for assessing the spottings against the ground truth, and a repository for storing information on the results. In addition we will make our sign detection system and results made with it publicly available as a baseline for comparison and further developments.

A Colloquial Corpus of Japanese Sign Language: Linguistic Resources for Observing Sign Language Conversations

Mayumi Bono, Kouhei Kikuchi, Paul Cibulka and Yutaka Osugi

We began building a corpus of Japanese Sign Language (JSL) in April 2011. The purpose of this project was to increase awareness of sign language as a distinctive language in Japan. This corpus is beneficial not only to linguistic research but also to hearing-impaired and deaf individuals, as it helps them to recognize and respect their linguistic differences and communication styles. This is the first large-scale JSL corpus developed for both academic and public use. We collected data in three ways: interviews (for introductory purposes only), dialogues, and lexical elicitation. In this paper, we focus particularly on data collected during a dialogue to discuss the application of conversation analysis (CA) to signed dialogues and signed conversations. Our annotation scheme was designed not only to elucidate theoretical issues related to grammar and linguistics but also to clarify pragmatic and interactional phenomena related to the use of JSL.

Exploring Factors that Contribute to Successful Fingerspelling Comprehension

Leah Geer and Jonathan Keane

Using a novel approach, we examine which cues in a fingerspelling stream, namely holds or transitions, allow for more successful comprehension by students learning American Sign Language (ASL). Sixteen university-level ASL students participated in this study. They were shown video clips of a native signer fingerspelling common English words. Clips were modified

in the following ways: all were slowed down to half speed, one-third of the clips were modified to black out the transition portion of the fingerspelling stream, and one-third modified to have holds blacked out. The remaining third of clips were free of blacked out portions, which we used to establish a baseline of comprehension. Research by Wilcox (1992), among others, suggested that transitions provide more rich information, and thus items with the holds blacked out should be easier to comprehend than items with the transitions blacked out. This was not found to be the case here. Students achieved higher comprehension scores when hold information was provided. Data from this project can be used to design training tools to help students become more proficient at fingerspelling comprehension, a skill with which most students struggle.

Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt and Hermann Ney

This paper introduces the RWTH-PHOENIX-Weather 2014, a video-based, large vocabulary, German sign language corpus which has been extended over the last two years, tripling the size of the original corpus. The corpus contains weather forecasts simultaneously interpreted into sign language which were recorded from German public TV and manually annotated using glosses on the sentence level and semi-automatically transcribed spoken German extracted from the videos using the open-source speech recognition system RASR. Spatial annotations of the signers' hands as well as shape and orientation annotations of the dominant hand have been added for more than 40k respectively 10k video frames creating one of the largest corpora allowing for quantitative evaluation of object tracking algorithms. Further, over 2k signs have been annotated using the SignWriting annotation system, focusing on the shape, orientation, movement as well as spatial contacts of both hands. Finally, extended recognition and translation setups are defined, and baseline results are presented.

The Use of a FileMaker Pro Database in Evaluating Sign Language Notation Systems

Julie Hochgesang

In this paper, FileMaker Pro has been used to create a database in order to evaluate sign language notation systems used for representing hand configurations. The database cited in this paper focuses on child acquisition data, particularly the dataset of one child and one adult productions of the same American Sign Language (ASL) signs produced in a two-year span. The hand configurations in selected signs have been coded using Stokoe notation (Stokoe, Casterline & Croneberg, 1965), the Hamburg Notation System or HamNoSys (Prillwitz et al, 1989), the revised Prosodic Model Handshape Coding system or PM (Eccarius & Brentari, 2008) and Sign Language Phonetic Annotation or SLPA, a notation system that has grown from the Movement-Hold Model (Johnson & Liddell, 2010, 2011a, 2011b, 2012). Data was pulled from ELAN transcripts, organized and notated in a FileMaker Pro database created to investigate the representativeness of each system. Representativeness refers to the ability of the notation system to represent the hand configurations in the dataset. This paper briefly describes the design of the FileMaker Pro database intended to provide both quantitative and qualitative information in order to allow the sign language researcher to examine the representativeness of sign language notation systems.

A New Framework for Sign Language Recognition based on 3D Handshape Identification and Linguistic Modeling

Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle and Dimitris Metaxas

Current approaches to sign recognition by computer generally have at least some of the following limitations: they rely on laboratory conditions for sign production, are limited to a small vocabulary, rely on 2D modeling (and therefore cannot deal with occlusions and off-plane rotations), and/or achieve limited success. Here we propose a new framework that (1) provides a new tracking method less dependent than others on laboratory conditions and able to deal with variations in background and skin regions (such as the face, forearms, or other hands); (2) allows for identification of 3D hand configurations that are linguistically important in American Sign Language (ASL); and (3) incorporates statistical information reflecting linguistic constraints in sign production. For purposes of large-scale computer-based sign language recognition from video, the ability to distinguish hand configurations accurately is critical. Our current method estimates the 3D hand configuration to distinguish among 77 hand configurations linguistically relevant for ASL. Constraining the problem in this way makes recognition of 3D hand configuration more tractable and provides the information specifically needed for sign recognition. Further improvements are obtained by incorporation of statistical information about linguistic dependencies among handshapes within a sign derived from an annotated corpus of almost 10,000 sign tokens.

O21 - Collaborative Resources (2)

Thursday, May 29, 11:45

Chairperson: **Thierry Declerck** Oral Session

Annotating Arguments: The NOMAD Collaborative Annotation Tool

Georgios Petasis

The huge amount of the available information in the Web creates the need for effective information extraction systems that are able to produce metadata that satisfy user's information needs. The development of such systems, in the majority of cases, depends on the availability of an appropriately annotated corpus in order to learn or evaluate extraction models. The production of such corpora can be significantly facilitated by annotation tools, which provide user-friendly facilities and enable annotators to annotate documents according to a predefined annotation schema. However, the construction of annotation tools that operate in a distributed environment is a challenging task: the majority of these tools are implemented as Web applications, having to cope with the capabilities offered by browsers. This paper describes the NOMAD collaborative annotation tool, which implements an alternative architecture: it remains a desktop application, fully exploiting the advantages of desktop applications, but provides collaborative annotation through the use of a centralised server for storing both the documents and their metadata, and instance messaging protocols for communicating events among all annotators. The annotation tool is implemented as a component of the Ellogon language engineering platform, exploiting its extensive annotation engine, its cross-platform abilities and its linguistic processing components, if such a need arises. Finally, the NOMAD annotation tool is distributed with an open source license, as part of the Ellogon platform.

Pivot-based Multilingual Dictionary Building using Wiktionary

Judit Ács

We describe a method for expanding existing dictionaries in several languages by discovering previously non-existent links between translations. We call this method triangulation and we present and compare several variations of it. We assess precision manually, and recall by comparing the extracted dictionaries with independently obtained basic vocabulary sets.We featurize the translation candidates and train a maximum entropy classifier to identify correct translations in the noisy data.

Terminology Resources and Terminology Work Benefit from Cloud Services

Tatiana Gornostay and Andrejs Vasiljevs

This paper presents the concept of the innovative platform TaaS "Terminology as a Service". TaaS brings the benefits of cloud services to the user, in order to foster the creation of terminology resources and to maintain their up-to-datedness by integrating automated data extraction and user-supported clean-up of raw terminological data and sharing user-validated terminology. The platform is based on cutting-edge technologies, provides single-access-point terminology services, and facilitates the establishment of emerging trends beyond conventional praxis and static models in terminology work. A cloud-based, user-oriented, collaborative, portable, interoperable, and multilingual platform offers such terminology services as terminology project creation and sharing, data collection for translation lookup, user document upload and management, terminology extraction customisation and execution, raw terminological data management, validated terminological data export and reuse, and other terminology services.

Crowdsourcing for the Identification of Event Nominals: an Experiment

Rachele Sprugnoli and Alessandro Lenci

This paper presents the design and results of a crowdsourcing experiment on the recognition of Italian event nominals. The aim of the experiment was to assess the feasibility of crowdsourcing methods for a complex semantic task such as distinguishing the eventive interpretation of polysemous nominals taking into consideration various types of syntagmatic cues. Details on the theoretical background and on the experiment set up are provided together with the final results in terms of accuracy and inter-annotator agreement. These results are compared with the ones obtained by expert annotators on the same task. The low values in accuracy and Fleiss' kappa of the crowdsourcing experiment demonstrate that crowdsourcing is not always optimal for complex linguistic tasks. On the other hand, the use of non-expert contributors allows to understand what are the most ambiguous patterns of polysemy and the most useful syntagmatic cues to be used to identify the eventive reading of nominals.

O22 - Conversational (2)

Thursday, May 29, 11:45

Chairperson: **Dafydd Gibbon** Oral Session

Combining Elicited Imitation and Fluency Features for Oral Proficiency Measurement

Deryle Lonsdale and Carl Christensen

The automatic grading of oral language tests has been the subject of much research in recent years. Several obstacles lie in the way of achieving this goal. Recent work suggests a testing technique called elicited imitation (EI) that can serve to accurately approximate global oral proficiency. This testing methodology, however, does not incorporate some fundamental aspects of language, such as fluency. Other work has suggested another testing technique, simulated speech (SS), as a supplement or an alternative to EI that can provide automated fluency metrics. In this work, we investigate a combination of fluency features extracted from SS tests and EI test scores as a means to more accurately predict oral language proficiency. Using machine learning and statistical modeling, we identify which features automatically extracted from SS tests best predicted hand-scored SS test results, and demonstrate the benefit of adding EI scores to these models. Results indicate that the combination of EI and fluency features do indeed more effectively predict hand-scored SS test scores. We finally discuss implications of this work for future automated oral testing scenarios.

On the Use of a Fuzzy Classifier to Speed Up the Sp ToBI Labeling of the Glissando Spanish Corpus

David Escudero, Aguilar-Cuevas Lourdes, González-Ferreras César, Gutiérrez-González Yurena and Valentín Cardeñoso-Payo

In this paper, we present the application of a novel automatic prosodic labeling methodology for speeding up the manual labeling of the Glissando corpus (Spanish read news items). The methodology is based on the use of soft classification techniques. The output of the automatic system consists on a set of label candidates per word. The number of predicted candidates depends on the degree of certainty assigned by the classifier to each of the predictions. The manual transcriber checks the sets of predictions to select the correct one. We describe the fundamentals of the fuzzy classification tool and its training with a corpus labeled with Sp TOBI labels. Results show a clear coherence between the most confused labels in the output of the automatic classifier and the most confused labels detected in inter-transcriber consistency tests. More importantly, in a preliminary test, the real time ratio of the labeling process was 1:66 when the template of predictions is used and 1:80 when it is not.

The RATS Collection: Supporting HLT Research with Degraded Audio Data

David Graff, Kevin Walker, Stephanie Strassel, Xiaoyi Ma, Karen Jones and Ann Sawyer

The DARPA RATS program was established to foster development of language technology systems that can perform well on speaker-to-speaker communications over radio channels that evince a wide range in the type and extent of signal variability and acoustic degradation. Creating suitable corpora to address this need poses an equally wide range of challenges for the collection, annotation and quality assessment of relevant data. This paper describes the LDC's multi-year effort to build the RATS data collection, summarizes the content and properties of the resulting corpora, and discusses the novel problems and approaches involved in ensuring that the data would satisfy its intended use, to provide speech recordings and annotations for training and evaluating HLT systems that perform 4 specific tasks on difficult radio channels: Speech Activity Detection (SAD), Language Identification (LID), Speaker Identification (SID) and Keyword Spotting (KWS).

Eliciting and Annotating Uncertainty in Spoken Language

Heather Pon-Barry, Stuart Shieber and Nicholas Longenbaugh

A major challenge in the field of automatic recognition of emotion and affect in speech is the subjective nature of affect labels. The most common approach to acquiring affect labels is to ask a panel of listeners to rate a corpus of spoken utterances along one or more dimensions of interest. For applications ranging from educational technology to voice search to dictation, a speaker's level of certainty is a primary dimension of interest. In such applications, we would like to know the speaker's actual level of certainty, but past research has only revealed listeners' perception of the speaker's level of certainty. In this paper, we present a method for eliciting spoken utterances using stimuli that we design such that they have a quantitative, crowdsourced legibility score. While we cannot control a speaker's actual internal level of certainty, the use of these stimuli provides a better estimate of internal certainty compared to existing speech corpora. The Harvard Uncertainty Speech Corpus, containing speech data, certainty annotations, and prosodic features, is made available to the research community.

O23 - Text Mining

Thursday, May 29, 11:45

Chairperson: **Lucia Specia** Oral Session

The Meta-knowledge of Causality in Biomedical Scientific Discourse

Claudiu Mihăilă and Sophia Ananiadou

Causality lies at the heart of biomedical knowledge, being involved in diagnosis, pathology or systems biology. Thus, automatic causality recognition can greatly reduce the human workload by suggesting possible causal connections and aiding in the curation of pathway models. For this, we rely on corpora that are annotated with classified, structured representations of important facts and findings contained within text. However, it is impossible to correctly interpret these annotations without additional information, e.g., classification of an event as fact, hypothesis, experimental result or analysis of results, confidence of authors about the validity of their analyses etc. In this study, we analyse and automatically detect this type of information, collectively termed meta-knowledge (MK), in the context of existing discourse causality annotations. Our effort proves the feasibility of identifying such pieces of information, without which the understanding of causal relations is limited.

Co-clustering of Bilingual Datasets as a Mean for Assisting the Construction of Thematic Bilingual Comparable Corpora

Guiyao Ke and Pierre-Francois Marteau

We address in this paper the assisted construction of bilingual thematic comparable corpora by means of co-clustering bilingual documents collected from raw sources such as the Web. The proposed approach is based on a quantitative comparability measure and a co-clustering approach which allow to mix similarity measures existing in each of the two linguistic spaces with a "thematic" comparability measure that defines a mapping between these two spaces. With the improvement of the co-clustering (k-medoids) performance we get, we use a comparability threshold and a manual verification to ensure the good and robust alignment of co-clusters (co-medoids). Finally, from any available raw corpus, we enrich the aligned clusters in order to provide "thematic" comparable corpora of good quality and controlled size. On a case study that exploit raw web data, we show that this approach scales reasonably well and is quite suited for the construction of thematic comparable corpora of good quality.

NewsReader: Recording History from Daily News Streams

Piek Vossen, German Rigau, Luciano Serafini, Pim Stouten, Francis Irving and Willem van Hage

The European project NewsReader develops technology to process daily news streams in 4 languages, extracting what happened, when, where and who was involved. NewsReader does not just read a single newspaper but massive amounts of news coming from thousands of sources. It compares the results across sources to complement information and determine where they disagree. Furthermore, it merges news of today with previous news, creating a long-term history rather than separate events. The result is stored in a KnowledgeStore, that cumulates information over time, producing an extremely large knowledge graph that is visualized using new techniques to provide more comprehensive access. We present the first version of the system and the results of processing first batches of data.

Identification of Technology Terms in Patents

Peter Anick, Marc Verhagen and James Pustejovsky

Natural language analysis of patents holds promise for the development of tools designed to assist analysts in the monitoring of emerging technologies. One component of such tools is the identification of technology terms. We describe an approach to the discovery of technology terms using supervised machine learning and evaluate its performance on subsets of patents in three languages: English, German, and Chinese.

O24 - Document Classification

Thursday, May 29, 11:45

Chairperson: **Robert Frederking** Oral Session

Cross-Language Authorship Attribution

Dasha Bogdanova and Angeliki Lazaridou

This paper presents a novel task of cross-language authorship attribution (CLAA), an extension of authorship attribution task to multilingual settings: given data labelled with authors in language X, the objective is to determine the author of a document written in language Y , where X is different from Y . We propose a number of cross-language stylometric features for the task of CLAA, such as those based on sentiment and emotional markers. We also explore an approach based on machine translation (MT) with both lexical and cross-language features. We experimentally show that MT could be used as a starting point to CLAA, since it allows good

attribution accuracy to be achieved. The cross-language features provide acceptable accuracy while using jointly with MT, though do not outperform lexical features.

Learning from Domain Complexity

Robert Remus and Dominique Ziegelmayer

Sentiment analysis is genre and domain dependent, i.e. the same method performs differently when applied to text that originates from different genres and domains. Intuitively, this is due to different language use in different genres and domains. We measure such differences in a sentiment analysis gold standard dataset that contains texts from 1 genre and 10 domains. Differences in language use are quantified using certain language statistics, viz. domain complexity measures. We investigate 4 domain complexity measures: percentage of rare words, word richness, relative entropy and corpus homogeneity. We relate domain complexity measurements to performance of a standard machine learning-based classifier and find strong correlations. We show that we can accurately estimate its performance based on domain complexity using linear regression models fitted using robust loss functions. Moreover, we illustrate how domain complexity may guide us in model selection, viz. in deciding what word n-gram order to employ in a discriminative model and whether to employ aggressive or conservative word n-gram feature selection.

Using Word Familiarities and Word Associations to Measure Corpus Representativeness

Reinhard Rapp

The definition of corpus representativeness used here assumes that a representative corpus should reflect as well as possible the average language use a native speaker encounters in everyday life over a longer period of time. As it is not practical to observe people's language input over years, we suggest to utilize two types of experimental data capturing two forms of human intuitions: Word familiarity norms and word association norms. If it is true that human language acquisition is corpus-based, such data should reflect people's perceived language input. Assuming so, we compute a representativeness score for a corpus by extracting word frequency and word association statistics from it and by comparing these statistics to the human data. The higher the similarity, the more representative the corpus should be for the language environments of the test persons. We present results for five different corpora and for truncated versions thereof. The results confirm the expectation that corpus size and corpus balance are crucial aspects for corpus representativeness.

A Modular System for Rule-based Text Categorisation

Marco del Tredici and Malvina Nissim

We introduce a modular rule-based approach to text categorisation which is more flexible and less time consuming to build than a standard rule-based system because it works with a hierarchical structure and allows for re-usability of rules. When compared to currently more wide-spread machine learning models on a case study, our modular system shows competitive results, and it has the advantage of reducing manual effort over time, since only fewer rules must be written when moving to a (partially) new domain, while annotation of training data is always required in the same amount.

P28 - Information Extraction

Thursday, May 29, 11:45

Chairperson: **Diana Maynard** Poster Session

Extracting News Web Page Creation Time with DCTFinder

Xavier Tannier

Web pages do not offer reliable metadata concerning their creation date and time. However, getting the document creation time is a necessary step for allowing to apply temporal normalization systems to web pages. In this paper, we present DCTFinder, a system that parses a web page and extracts from its content the title and the creation date of this web page. DCTFinder combines heuristic title detection, supervised learning with Conditional Random Fields (CRFs) for document date extraction, and rule-based creation time recognition. Using such a system allows further deep and efficient temporal analysis of web pages. Evaluation on three corpora of English and French web pages indicates that the tool can extract document creation times with reasonably high accuracy (between 87 and 92%). DCTFinder is made freely available on http://sourceforge.net/projects/dctfinder/, as well as all resources (vocabulary and annotated documents) built for training and evaluating the system in English and French, and the English trained model itself.

Information Extraction from German Patient Records via Hybrid Parsing and Relation Extraction Strategies

Hans-Ulrich Krieger, Christian Spurk, Hans Uszkoreit, Feiyu Xu, Yi Zhang, Frank Müller and Thomas Tolxdorff

In this paper, we report on first attempts and findings to analyzing German patient records, using a hybrid parsing architecture and a

combination of two relation extraction strategies. On a practical level, we are interested in the extraction of concepts and relations among those concepts, a necessary cornerstone for building medical information systems. The parsing pipeline consists of a morphological analyzer, a robust chunk parser adapted to Latin phrases used in medical diagnosis, a repair rule stage, and a probabilistic context-free parser that respects the output from the chunker. The relation extraction stage is a combination of two systems: SProUT, a shallow processor which uses hand-written rules to discover relation instances from local text units and DARE which extracts relation instances from complete sentences, using rules that are learned in a bootstrapping process, starting with semantic seeds. Two small experiments have been carried out for the parsing pipeline and the relation extraction stage.

Media Monitoring and Information Extraction for the Highly Inflected Agglutinative Language Hungarian

Júlia Pajzs, Ralf Steinberger, Maud Ehrmann, Mohamed Ebrahim, Leonida Della Rocca, Stefano Bucci, Eszter Simon and Tamás Váradi

The Europe Media Monitor (EMM) is a fully-automatic system that analyses written online news by gathering articles in over 70 languages and by applying text analysis software for currently 21 languages, without using linguistic tools such as parsers, part-of-speech taggers or morphological analysers. In this paper, we describe the effort of adding to EMM Hungarian text mining tools for news gathering; document categorisation; named entity recognition and classification for persons, organisations and locations; name lemmatisation; quotation recognition; and cross-lingual linking of related news clusters. The major challenge of dealing with the Hungarian language is its high degree of inflection and agglutination. We present several experiments where we apply linguistically light-weight methods to deal with inflection and we propose a method to overcome the challenges. We also present detailed frequency lists of Hungarian person and location name suffixes, as found in real-life news texts. This empirical data can be used to draw further conclusions and to improve existing Named Entity Recognition software. Within EMM, the solutions described here will also be applied to other morphologically complex languages such as those of the Slavic language family. The media monitoring and analysis system EMM is freely accessible online via the web page http://emm.newsbrief.eu/overview.html.

Creating a Gold Standard Corpus for the Extraction of Chemistry-Disease Relations from Patent Texts

Antje Schlaf, Claudia Bobach and Matthias Irmer

This paper describes the creation of a gold standard for chemistry-disease relations in patent texts. We start with an automated annotation of named entities of the domains chemistry (e.g. "propranolol") and diseases (e.g. "hypertension") as well as of related domains like methods and substances. After that, domain-relevant relations between these entities, e.g. "propranolol treats hypertension", have been manually annotated. The corpus is intended to be suitable for developing and evaluating relation extraction methods. In addition, we present two reasoning methods of high precision for automatically extending the set of extracted relations. Chain reasoning provides a method to infer and integrate additional, indirectly expressed relations occurring in relation chains. Enumeration reasoning exploits the frequent occurrence of enumerations in patents and automatically derives additional relations. These two methods are applicable both for verifying and extending the manually annotated data as well as for potential improvements of automatic relation extraction.

T2K^2: a System for Automatically Extracting and Organizing Knowledge from Texts

Felice Dell'Orletta, Giulia Venturi, Andrea Cimino and Simonetta Montemagni

In this paper, we present T2K^2, a suite of tools for automatically extracting domain–specific knowledge from collections of Italian and English texts. T2K^2 (Text–To–Knowledge v2) relies on a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine learning which are dynamically integrated to provide an accurate and incremental representation of the content of vast repositories of unstructured documents. Extracted knowledge ranges from domain–specific entities and named entities to the relations connecting them and can be used for indexing document collections with respect to different information types. T2K^2 also includes "linguistic profiling" functionalities aimed at supporting the user in constructing the acquisition corpus, e.g. in selecting texts belonging to the same genre or characterized by the same degree of specialization or in monitoring the "added value" of newly inserted documents. T2K^2 is a web application which can be accessed from any browser through a personal account which has been tested in a wide range of domains.

Freepal: A Large Collection of Deep Lexico-Syntactic Patterns for Relation Extraction

Johannes Kirschnick, Alan Akbik and Holmer Hemsen

The increasing availability and maturity of both scalable computing architectures and deep syntactic parsers is opening up

new possibilities for Relation Extraction (RE) on large corpora of natural language text. In this paper, we present Freepal, a resource designed to assist with the creation of relation extractors for more than 5,000 relations defined in the Freebase knowledge base (KB). The resource consists of over 10 million distinct lexico-syntactic patterns extracted from dependency trees, each of which is assigned to one or more Freebase relations with different confidence strengths. We generate the resource by executing a large-scale distant supervision approach on the ClueWeb09 corpus to extract and parse over 260 million sentences labeled with Freebase entities and relations. We make Freepal freely available to the research community, and present a web demonstrator to the dataset, accessible from free-pal.appspot.com.

Ranking Job Offers for Candidates: Learning Hidden Knowledge from Big Data

Marc Poch, Núria Bel, Sergio Espeja and Felipe Navio

This paper presents a system for suggesting a ranked list of appropriate vacancy descriptions to job seekers in a job board web site. In particular our work has explored the use of supervised classifiers with the objective of learning implicit relations which cannot be found with similarity or pattern based search methods that rely only on explicit information. Skills, names of professions and degrees, among other examples, are expressed in different languages, showing high variation and the use of ad-hoc resources to trace the relations is very costly. This implicit information is unveiled when a candidate applies for a job and therefore it is information that can be used for learning a model to predict new cases. The results of our experiments, which combine different clustering, classification and ranking methods, show the validity of the approach.

Hot Topics and Schisms in NLP: Community and Trend Analysis with Saffron on ACL and LREC Proceedings

Paul Buitelaar, Georgeta Bordea and Barry Coughlan

In this paper we present a comparative analysis of two series of conferences in the field of Computational Linguistics, the LREC conference and the ACL conference. Conference proceedings were analysed using Saffron by performing term extraction and topical hierarchy construction with the goal of analysing topic trends and research communities. The system aims to provide insight into a research community and to guide publication and participation strategies, especially of novice researchers.

Textual Emigration Analysis (TEA)

Andre Blessing and Jonas Kuhn

We present a web-based application which is called TEA (Textual Emigration Analysis) as a showcase that applies textual analysis for the humanities. The TEA tool is used to transform raw text input into a graphical display of emigration source and target countries (under a global or an individual perspective). It provides emigration-related frequency information, and gives access to individual textual sources, which can be downloaded by the user. Our application is built on top of the CLARIN infrastructure which targets researchers of the humanities. In our scenario, we focus on historians, literary scientists, and other social scientists that are interested in the semantic interpretation of text. Our application processes a large set of documents to extract information about people who emigrated. The current implementation integrates two data sets: A data set from the Global Migrant Origin Database, which does not need additional processing, and a data set which was extracted from the German Wikipedia edition. The TEA tool can be accessed by using the following URL: http://clarin01.ims.uni-stuttgart.de/geovis/showcase.html

P29 - Lexicons

Thursday, May 29, 11:45

Chairperson: **Nianwen Xue** Poster Session

WordNet–Wikipedia–Wiktionary: Construction of a Three-way Alignment

Tristan Miller and Iryna Gurevych

The coverage and quality of conceptual information contained in lexical semantic resources is crucial for many tasks in natural language processing. Automatic alignment of complementary resources is one way of improving this coverage and quality; however, past attempts have always been between pairs of specific resources. In this paper we establish some set-theoretic conventions for describing concepts and their alignments, and use them to describe a method for automatically constructing n-way alignments from arbitrary pairwise alignments. We apply this technique to the production of a three-way alignment from previously published WordNet-Wikipedia and WordNet-Wiktionary alignments. We then present a quantitative and informal qualitative analysis of the aligned resource. The three-way alignment was found to have greater coverage, an enriched sense representation, and coarser sense granularity than both the original resources and their pairwise alignments, though this came at the cost of accuracy. An evaluation of the induced word sense clusters in a word sense disambiguation task showed that they were no better than random clusters of equivalent granularity. However, use of the alignments to enrich a sense inventory with

additional sense glosses did significantly improve the performance of a baseline knowledge-based WSD algorithm.

xLiD-Lexica: Cross-lingual Linked Data Lexica

Lei Zhang, Michael Färber and Achim Rettinger

In this paper, we introduce our cross-lingual linked data lexica, called xLiD-Lexica, which are constructed by exploiting the multilingual Wikipedia and linked data resources from Linked Open Data (LOD). We provide the cross-lingual groundings of linked data resources from LOD as RDF data, which can be easily integrated into the LOD data sources. In addition, we build a SPARQL endpoint over our xLiD-Lexica to allow users to easily access them using SPARQL query language. Multilingual and cross-lingual information access can be facilitated by the availability of such lexica, e.g., allowing for an easy mapping of natural language expressions in different languages to linked data resources from LOD. Many tasks in natural language processing, such as natural language generation, cross-lingual entity linking, text annotation and question answering, can benefit from our xLiD-Lexica.

Turkish Resources for Visual Word Recognition

Begum Erten, Cem Bozsahin and Deniz Zeyrek

We report two tools to conduct psycholinguistic experiments on Turkish words. KelimetriK allows experimenters to choose words based on desired orthographic scores of word frequency, bigram and trigram frequency, ON, OLD20, ATL and subset/superset similarity. Turkish version of Wuggy generates pseudowords from one or more template words using an efficient method. The syllabified version of the words are used as the input, which are decomposed into their sub-syllabic components. The bigram frequency chains are constructed by the entire words' onset, nucleus and coda patterns. Lexical statistics of stems and their syllabification are compiled by us from BOUN corpus of 490 million words. Use of these tools in some experiments is shown.

Computer-Aided Quality Assurance of an Icelandic Pronunciation Dictionary

Martin Jansche

We propose a model-driven method for ensuring the quality of pronunciation dictionaries. The key ingredient is computing an alignment between letter strings and phoneme strings, a standard technique in pronunciation modeling. The novel aspect of our method is the use of informative, parametric alignment models which are refined iteratively as they are tested against the data. We discuss the use of alignment failures as a signal for detecting and correcting problematic dictionary entries. We illustrate this method using an existing pronunciation dictionary for Icelandic. Our method is completely general and has been applied in the construction of pronunciation dictionaries for commercially deployed speech recognition systems in several languages.

Bring vs. MTRoget: Evaluating Automatic Thesaurus Translation

Lars Borin, Jens Allwood and Gerard de Melo

Evaluation of automatic language-independent methods for language technology resource creation is difficult, and confounded by a largely unknown quantity, viz. to what extent typological differences among languages are significant for results achieved for one language or language pair to be applicable across languages generally. In the work presented here, as a simplifying assumption, language-independence is taken as axiomatic within certain specified bounds. We evaluate the automatic translation of Roget's "Thesaurus" from English into Swedish using an independently compiled Roget-style Swedish thesaurus, S.C. Bring's "Swedish vocabulary arranged into conceptual classes" (1930). Our expectation is that this explicit evaluation of one of the thesaureses created in the MTRoget project will provide a good estimate of the quality of the other thesauruses created using similar methods.

Bilingual Dictionary Induction as an Optimization Problem

Wushouer Mairidan, Toru Ishida, Donghui Lin and Katsutoshi Hirayama

Bilingual dictionaries are vital in many areas of natural language processing, but such resources are rarely available for lower-density language pairs, especially for those that are closely related. Pivot-based induction consists of using a third language to bridge a language pair. As an approach to create new dictionaries, it can generate wrong translations due to polysemy and ambiguous words. In this paper we propose a constraint approach to pivot-based dictionary induction for the case of two closely related languages. In order to take into account the word senses, we use an approach based on semantic distances, in which possibly missing translations are considered, and instance of induction is encoded as an optimization problem to generate new dictionary. Evaluations show that the proposal achieves 83.7% accuracy and approximately 70.5% recall, thus outperforming the baseline pivot-based method.

Enriching the "Senso Comune" Platform with Automatically Acquired Data

Tommaso Caselli, Laure Vieu, Carlo Strapparava and Guido Vetere

This paper reports on research activities on automatic methods for the enrichment of the Senso Comune platform. At this stage

of development, we will report on two tasks, namely word sense alignment with MultiWordNet and automatic acquisition of Verb Shallow Frames from sense annotated data in the MultiSemCor corpus. The results obtained are satisfying. We achieved a final F-measure of 0.64 for noun sense alignment and a F-measure of 0.47 for verb sense alignment, and an accuracy of 68% on the acquisition of Verb Shallow Frames.

MUHIT: A Multilingual Harmonized Dictionary

Sameh Alansary

This paper discusses a trial to build a multilingual harmonized dictionary that contains more than 40 languages, with special reference to Arabic which represents about 20% of the whole size of the dictionary. This dictionary is called MUHIT which is an interactive multilingual dictionary application. It is a web application that makes it easily accessible to all users. MUHIT is developed within the Universal Networking Language (UNL) framework by the UNDL Foundation, in cooperation with Bibliotheca Alexandrina (BA). This application targets to serve specialists and non-specialists. It provides users with full linguistic description to each lexical item. This free application is useful to many NLP tasks such as multilingual translation and cross-language synonym search. This dictionary is built depending on WordNet and corpus based approaches, in a specially designed linguistic environment called UNLariam that is developed by the UNLD foundation. This dictionary is the first launched application by the UNLD foundation.

Language Resources for French in the Biomedical Domain

Aurelie Neveol, Julien Grosjean, Stéfan Darmoni and Pierre Zweigenbaum

The biomedical domain offers a wealth of linguistic resources for Natural Language Processing, including terminologies and corpora. While many of these resources are prominently available for English, other languages including French benefit from substantial coverage thanks to the contribution of an active community over the past decades. However, access to terminological resources in languages other than English may not be as straight-forward as access to their English counterparts. Herein, we review the extent of resource coverage for French and give pointers to access French-language resources. We also discuss the sources and methods for making additional material available for French.

Gold-standard for Topic-specific Sentiment Analysis of Economic Texts

Pyry Takala, Pekka Malo, Ankur Sinha and Oskar Ahlgren

Public opinion, as measured by media sentiment, can be an important indicator in the financial and economic context. These are domains where traditional sentiment estimation techniques often struggle, and existing annotated sentiment text collections are of less use. Though considerable progress has been made in analyzing sentiments at sentence-level, performing topic-dependent sentiment analysis is still a relatively uncharted territory. The computation of topic-specific sentiments has commonly relied on naive aggregation methods without much consideration to the relevance of the sentences to the given topic. Clearly, the use of such methods leads to a substantial increase in noise-to-signal ratio. To foster development of methods for measuring topic-specific sentiments in documents, we have collected and annotated a corpus of financial news that have been sampled from Thomson Reuters newswire. In this paper, we describe the annotation process and evaluate the quality of the dataset using a number of inter-annotator agreement metrics. The annotations of 297 documents and over 9000 sentences can be used for research purposes when developing methods for detecting topic-wise sentiment in financial text.

P30 - Large Projects and Infrastructural Issues

Thursday, May 29, 11:45

Chairperson: **Yohei Murakami** Poster Session

A Decade of HLT Agency Activities in the Low Countries: from Resource Maintenance (BLARK) to Service Offerings (BLAISE)

Peter Spyns and Remco van Veenendaal

In this paper we report on the Flemish-Dutch Agency for Human Language Technologies (HLT Agency or TST-Centrale in Dutch) in the Low Countries. We present its activities in its first decade of existence. The main goal of the HLT Agency is to ensure the sustainability of linguistic resources for Dutch. 10 years after its inception, the HLT Agency faces new challenges and opportunities. An important contextual factor is the rise of the infrastructure networks and proliferation of resource centres. We summarise some lessons learnt and we propose as future work to define and build for Dutch (which by extension can apply to any national language) a set of Basic LAnguage Infrastructure SErvices (BLAISE). As a conclusion, we state that the HLT Agency, also by its peculiar institutional status, has fulfilled and still is fulfilling an important role in maintaining Dutch as a digitally fully fledged functional language.

CLARA: A New Generation of Researchers in Common Language Resources and Their Applications

Koenraad de Smedt, Erhard Hinrichs, Detmar Meurers, Inguna Skadina, Bolette Pedersen, Costanza Navarretta, Núria Bel, Krister Linden, Marketa Lopatkova, Jan Hajic, Gisle Andersen and Przemyslaw Lenkiewicz

CLARA (Common Language Resources and Their Applications) is a Marie Curie Initial Training Network which ran from 2009 until 2014 with the aim of providing researcher training in crucial areas related to language resources and infrastructure. The scope of the project was broad and included infrastructure design, lexical semantic modeling, domain modeling, multimedia and multimodal communication, applications, and parsing technologies and grammar models. An international consortium of 9 partners and 12 associate partners employed researchers in 19 new positions and organized a training program consisting of 10 thematic courses and summer/winter schools. The project has resulted in new theoretical insights as well as new resources and tools. Most importantly, the project has trained a new generation of researchers who can perform advanced research and development in language resources and technologies.

Encompassing a Spectrum of LT Users in the CLARIN-DK Infrastructure

Lina Henriksen, Dorte Haltrup Hansen, Bente Maegaard, Bolette Sandford Pedersen and Claus Povlsen

CLARIN-DK is a platform with language resources constituting the Danish part of the European infrastructure CLARIN ERIC. Unlike some other language based infrastructures CLARIN-DK is not solely a repository for upload and storage of data, but also a platform of web services permitting the user to process data in various ways. This involves considerable complications in relation to workflow requirements. The CLARIN-DK interface must guide the user to perform the necessary steps of a workflow; even when the user is inexperienced and perhaps has an unclear conception of the requested results. This paper describes a user driven approach to creating a user interface specification for CLARIN-DK. We indicate how different user profiles determined different crucial interface design options. We also describe some use cases established in order to give illustrative examples of how the platform may facilitate research.

Legal Aspects of Text Mining

Maarten Truyens and Patrick van Eecke

Unlike data mining, text mining has received only limited attention in legal circles. Nevertheless, interesting legal stumbling blocks exist, both with respect to the data collection and data sharing phases, due to the strict rules of copyright and database law. Conflicts are particularly likely when content is extracted from commercial databases, and when texts that have a minimal level of creativity are stored in a permanent way. In all circumstances, even with non-commercial research, license agreements and website terms of use can impose further restrictions. Accordingly, only for some delineated areas (very old texts for which copyright expired, legal statutes, texts in the public domain) strong legal certainty can be obtained without case-by-case assessments. As a result, while prior permission is certainly not required in all cases, many researchers tend to err on the side of caution, and seek permission from publishers, institutions and individual authors before including texts in their corpora, although this process can be difficult and very time-consuming. In the United States, the legal assessment is very different, due to the open-ended nature and flexibility offered by the "fair use" doctrine.

CLARIN-NL: Major Results

Jan Odijk

In this paper I provide a high level overview of the major results of CLARIN-NL so far. I will show that CLARIN-NL is starting to provide the data, facilities and services in the CLARIN infrastructure to carry out humanities research supported by large amounts of data and tools. These services have easy interfaces and easy search options (no technical background needed). Still some training is required, to understand both the possibilities and the limitations of the data and the tools. Actual use of the facilities leads to suggestions for improvements and to suggestions for new functionality. All researchers are therefore invited to start using the elements in the CLARIN infrastructure offered by CLARIN-NL. Though I will show that a lot has been achieved in the CLARIN-NL project, I will also provide a long list of functionality and interoperability cases that have not been dealt with in CLARIN-NL and must remain for future work.

An Innovative World Language Centre : Challenges for the Use of Language Technology

Auður Hauksdóttir

The Vigdis International Centre of Multilingualism and Intercultural Understanding at the University of Iceland will work under the auspices of UNESCO. The main objective of the Centre is to promote linguistic diversity and to raise awareness of the importance of multilingualism. The focus will be on research on translations, foreign language learning, language policy and language planning. The centre will also serve as a platform for promoting collaborative activities on languages and cultures, in

particular, organizing exhibitions and other events aimed at both the academic community and the general public. The Centre will work in close collaboration with the national and international research community. The Centre aims to create state-of-the-art infrastructure, using Language Technology resources in research and academic studies, in particular in translations and language learning (Computer-Assisted Language Learning). In addition, the centre will provide scholars with a means to conduct corpus-based research for synchronic investigations and for comparative studies. The Centre will also function as a repository for language data corpora. Facilities will be provided so that these corpora can be used by the research community on site and online. Computer technology resources will also be exploited in creating tools and exhibitions for the general audience.

Facing the Identification Problem in Language-Related Scientific Data Analysis

Joseph Mariani, Christopher Cieri, Gil Francopoulo, Patrick Paroubek and Marine Delaborde

This paper describes the problems that must be addressed when studying large amounts of data over time which require entity normalization applied not to the usual genres of news or political speech, but to the genre of academic discourse about language resources, technologies and sciences. It reports on the normalization processes that had to be applied to produce data usable for computing statistics in three past studies on the LRE Map, the ISCA Archive and the LDC Bibliography. It shows the need for human expertise during normalization and the necessity to adapt the work to the study objectives. It investigates possible improvements for reducing the workload necessary to produce comparable results. Through this paper, we show the necessity to define and agree on international persistent and unique identifiers.

Taalportaal: an Online Grammar of Dutch and Frisian

Frank Landsbergen, Carole Tiberius and Roderik Dernison

In this paper, we present the Taalportaal project. Taalportaal will create an online portal containing an exhaustive and fully searchable electronic reference of Dutch and Frisian phonology, morphology and syntax. Its content will be in English. The main aim of the project is to serve the scientific community by organizing, integrating and completing the grammatical knowledge of both languages, and to make this data accessible in an innovative way. The project is carried out by a consortium of four universities and research institutions. Content is generated in two ways: (1) by a group of authors who, starting from existing grammatical resources, write text directly in XML, and (2) by integrating the full Syntax of Dutch into the portal, after an automatic conversion from Word to XML. We discuss the project's workflow, content creation and management, the actual web application, and the way in which we plan to enrich the portal's content, such as by crosslinking between topics and linking to external resources.

P31 - Opinion Mining and Reviews Analysis

Thursday, May 29, 11:45

Chairperson: **Manfred Stede** Poster Session

The USAGE Review Corpus for Fine Grained Multi Lingual Opinion Analysis

Roman Klinger and Philipp Cimiano

Opinion mining has received wide attention in recent years. Models for this task are typically trained or evaluated with a manually annotated dataset. However, fine-grained annotation of sentiments including information about aspects and their evaluation is very labour-intensive. The data available so far is limited. Contributing to this situation, this paper describes the Bielefeld University Sentiment Analysis Corpus for German and English (USAGE), which we offer freely to the community and which contains the annotation of product reviews from Amazon with both aspects and subjective phrases. It provides information on segments in the text which denote an aspect or a subjective evaluative phrase which refers to the aspect. Relations and coreferences are explicitly annotated. This dataset contains 622 English and 611 German reviews, allowing to investigate how to port sentiment analysis systems across languages and domains. We describe the methodology how the corpus was created and provide statistics including inter-annotator agreement. We further provide figures for a baseline system and results for German and English as well as in a cross-domain setting. The results are encouraging in that they show that aspects and phrases can be extracted robustly without the need of tuning to a particular type of products.

PACE Corpus: a Multilingual Corpus of Polarity-Annotated Textual Data from the Domains Automotive and CEllphone

Christian Haenig, Andreas Niekler and Carsten Wuensch

In this paper, we describe a publicly available multilingual evaluation corpus for phrase-level Sentiment Analysis that can be used to evaluate real world applications in an industrial context. This corpus contains data from English and German Internet forums (1000 posts each) focusing on the automotive domain.

The major topic of the corpus is connecting and using cellphones to/in cars. The presented corpus contains different types of annotations: objects (e.g. my car, my new cellphone), features (e.g. address book, sound quality) and phrase-level polarities (e.g. the best possible automobile, big problem). Each of the posts has been annotated by at least four different annotators – these annotations are retained in their original form. The reliability of the annotations is evaluated by inter-annotator agreement scores. Besides the corpus data and format, we provide comprehensive corpus statistics. This corpus is one of the first lexical resources focusing on real world applications that analyze the voice of the customer which is crucial for various industrial use cases.

Adapting Freely Available Resources to Build an Opinion Mining Pipeline in Portuguese

Patrik Lambert and Carlos Rodriguez-Penagos

We present a complete UIMA-based pipeline for sentiment analysis in Portuguese news using freely available resources and a minimal set of manually annotated training data. We obtained good precision on binary classification but concluded that news feed is a challenging environment to detect the extent of opinionated text.

The NewSoMe Corpus: a Unifying Opinion Annotation Framework Across Genres and in Multiple Languages

Roser Saurí, Judith Domingo and Toni Badia

We present the NewSoMe (News and Social Media) Corpus, a set of subcorpora with annotations on opinion expressions across genres (news reports, blogs, product reviews and tweets) and covering multiple languages (English, Spanish, Catalan and Portuguese). NewSoMe is the result of an effort to increase the opinion corpus resources available in languages other than English, and to build a unifying annotation framework for analyzing opinion in different genres, including controlled text, such as news reports, as well as different types of user generated contents (UGC). Given the broad design of the resource, most of the annotation effort were carried out resorting to crowdsourcing platforms: Amazon Mechanical Turk and CrowdFlower. This created an excellent opportunity to research on the feasibility of crowdsourcing methods for annotating big amounts of text in different languages.

The Dangerous Myth of the Star System

André Bittar, Dini Luca, Sigrid Maurel and Mathieu Ruhlmann

In recent years we have observed two parallel trends in computational linguistics research and e-commerce development. On the research side, there has been an increasing interest in algorithms and approaches that are able to capture the polarity of opinions expressed by users on products, institutions and services. On the other hand, almost all big e-commerce and aggregator sites are by now providing users the possibility of writing comments and expressing their appreciation with a numeric score (usually represented as a number of stars). This generates the impression that the work carried out in the research community is made partially useless (at least for economic exploitation) by an evolution in web practices. In this paper we describe an experiment on a large corpus which shows that the score judgments provided by users are often conflicting with the text contained in the opinion, and to such a point that a rule-based opinion mining system can be demonstrated to perform better than the users themselves in ranking their opinions.

A Corpus of Comparisons in Product Reviews

Wiltrud Kessler and Jonas Kuhn

Sentiment analysis (or opinion mining) deals with the task of determining the polarity of an opinionated document or sentence. Users often express sentiment about one product by comparing it to a different product. In this work, we present a corpus of comparison sentences from English camera reviews. For our purposes we define a comparison to be any statement about the similarity or difference of two entities. For each sentence we have annotated detailed information about the comparisons it contains: The comparative predicate that expresses the comparison, the type of the comparison, the two entities that are being compared, and the aspect they are compared in. The results of our agreement study show that the decision whether a sentence contains a comparison is difficult to make even for trained human annotators. Once that decision is made, we can achieve consistent results for the very detailed annotations. In total, we have annotated 2108 comparisons in 1707 sentences from camera reviews which makes our corpus the largest resource currently available. The corpus and the annotation guidelines are publicly available on our website.

P32 - Social Media Processing

Thursday, May 29, 11:45

Chairperson: **Fei Xia** Poster Session

Finding Romanized Arabic Dialect in Code-Mixed Tweets

Clare Voss, Stephen Tratz, Jamal Laoudi and Douglas Briesch

Recent computational work on Arabic dialect identification has focused primarily on building and annotating corpora written in

Arabic script. Arabic dialects however also appear written in Roman script, especially in social media. This paper describes our recent work developing tweet corpora and a token-level classifier that identifies a Romanized Arabic dialect and distinguishes it from French and English in tweets. We focus on Moroccan Darija, one of several spoken vernaculars in the family of Maghrebi Arabic dialects. Even given noisy, code-mixed tweets,the classifier achieved token-level recall of 93.2% on Romanized Arabic dialect, 83.2% on English, and 90.1% on French. The classifier, now integrated into our tweet conversation annotation tool (Tratz et al. 2013), has semi-automated the construction of a Romanized Arabic-dialect lexicon. Two datasets, a full list of Moroccan Darija surface token forms and a table of lexical entries derived from this list with spelling variants, as extracted from our tweet corpus collection, will be made available in the LRE MAP.

Hashtag Occurrences, Layout and Translation: A Corpus-driven Analysis of Tweets Published by the Canadian Government

Fabrizio Gotti, Phillippe Langlais and Atefeh Farzindar

We present an aligned bilingual corpus of 8758 tweet pairs in French and English, derived from Canadian government agencies. Hashtags appear in a tweet's prologue, announcing its topic, or in the tweet's text in lieu of traditional words, or in an epilogue. Hashtags are words prefixed with a pound sign in 80% of the cases. The rest is mostly multiword hashtags, for which we describe a segmentation algorithm. A manual analysis of the bilingual alignment of 5000 hashtags shows that 5% (French) to 18% (English) of them don't have a counterpart in their containing tweet's translation. This analysis shows that 80% of multiword hashtags are correctly translated by humans, and that the mistranslation of the rest may be due to incomplete translation directives regarding social media. We show how these resources and their analysis can guide the design of a machine translation pipeline, and its evaluation. A baseline system implementing a tweet-specific tokenizer yields promising results. The system is improved by translating epilogues, prologues, and text separately. We attempt to feed the SMT engine with the original hashtag and some alternatives ("dehashed" version or a segmented version of multiword hashtags), but translation quality improves at the cost of hashtag recall.

Clustering Tweets using Wikipedia Concepts

Guoyu Tang, Yunqing Xia, Weizhi Wang, Raymond Lau and Fang Zheng

Two challenging issues are notable in tweet clustering. Firstly, the sparse data problem is serious since no tweet can be longer than 140 characters. Secondly, synonymy and polysemy are rather common because users intend to present a unique meaning with a great number of manners in tweets. Enlightened by the recent research which indicates Wikipedia is promising in representing text, we exploit Wikipedia concepts in representing tweets with concept vectors. We address the polysemy issue with a Bayesian model, and the synonymy issue by exploiting the Wikipedia redirections. To further alleviate the sparse data problem, we further make use of three types of out-links in Wikipedia. Evaluation on a twitter dataset shows that the concept model outperforms the traditional VSM model in tweet clustering.

An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis

Eshrag Refaee and Verena Rieser

We present a newly collected data set of 8,868 gold-standard annotated Arabic feeds. The corpus is manually labelled for subjectivity and sentiment analysis (SSA) (= 0:816). In addition, the corpus is annotated with a variety of motivated feature-sets that have previously shown positive impact on performance. The paper highlights issues posed by twitter as a genre, such as mixture of language varieties and topic-shifts. Our next step is to extend the current corpus, using online semi-supervised learning. A first sub-corpus will be released via the ELRA repository as part of this submission.

TweetNorm_ es: an Annotated Corpus for Spanish Microtext Normalization

Iñaki Alegria, Nora Aranberri, Pere Comas, Victor Fresno, Pablo Gamallo, Lluís Padró, Iñaki San Vicente, Jordi Turmo and Arkaitz Zubiaga

In this paper we introduce TweetNorm_ es, an annotated corpus of tweets in Spanish language, which we make publicly available under the terms of the CC-BY license. This corpus is intended for development and testing of microtext normalization systems. It was created for Tweet-Norm, a tweet normalization workshop and shared task, and is the result of a joint annotation effort from different research groups. In this paper we describe the methodology defined to build the corpus as well as the guidelines followed in the annotation process. We also present a brief overview of the Tweet-Norm shared task, as the first evaluation environment where the corpus was used.

TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages

Nikola Ljubešić, Darja Fišer and Tomaž Erjavec

This paper presents TweetCaT, an open-source Python tool for building Twitter corpora that was designed for smaller languages.

Using the Twitter search API and a set of seed terms, the tool identifies users tweeting in the language of interest together with their friends and followers. By running the tool for 235 days we tested it on the task of collecting two monitor corpora, one for Croatian and Serbian and the other for Slovene, thus also creating new and valuable resources for these languages. A post-processing step on the collected corpus is also described, which filters out users that tweet predominantly in a foreign language thus further cleans the collected corpora. Finally, an experiment on discriminating between Croatian and Serbian Twitter users is reported.

A German Twitter Snapshot

Tatjana Scheffler

We present a new corpus of German tweets. Due to the relatively small number of German messages on Twitter, it is possible to collect a virtually complete snapshot of German twitter messages over a period of time. In this paper, we present our collection method which produced a 24 million tweet corpus, representing a large majority of all German tweets sent in April, 2013. Further, we analyze this representative data set and characterize the German twitterverse. While German Twitter data is similar to other Twitter data in terms of its temporal distribution, German Twitter users are much more reluctant to share geolocation information with their tweets. Finally, the corpus collection method allows for a study of discourse phenomena in the Twitter data, structured into discussion threads.

P33 - Treebanks

Thursday, May 29, 11:45

Chairperson: **Montserrat Marimón** Poster Session

The Procedure of Lexico-Semantic Annotation of Składnica Treebank

Elżbieta Hajnicz

In this paper, the procedure of lexico-semantic annotation of Składnica Treebank using Polish WordNet is presented. Other semantically annotated corpora, in particular treebanks, are outlined first. Resources involved in annotation as well as a tool called Semantikon used for it are described. The main part of the paper is the analysis of the applied procedure. It consists of the basic and correction phases. During basic phase all nouns, verbs and adjectives are annotated with wordnet senses. The annotation is performed independently by two linguists. During the correction phase, conflicts are resolved by the linguist supervising the process. Multi-word units obtain special tags, synonyms and hypernyms are used for senses absent in Polish WordNet. Additionally, each sentence receives its general assessment. Finally, some statistics of the results of annotation are given, including inter-annotator agreement. The final resource is represented in XML files preserving the structure of Składnica.

Deep Syntax Annotation of the Sequoia French Treebank

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah and Eric de la Clergerie

We define a deep syntactic representation scheme for French, which abstracts away from surface syntactic variation and diathesis alternations, and describe the annotation of deep syntactic representations on top of the surface dependency trees of the Sequoia corpus. The resulting deep-annotated corpus, named deep-sequoia, is freely available, and hopefully useful for corpus linguistics studies and for training deep analyzers to prepare semantic analysis.

Projection-based Annotation of a Polish Dependency Treebank

Alina Wróblewska and Adam Przepiórkowski

This paper presents an approach of automatic annotation of sentences with dependency structures. The approach builds on the idea of cross-lingual dependency projection. The presented method of acquiring dependency trees involves a weighting factor in the processes of projecting source dependency relations to target sentences and inducing well-formed target dependency trees from sets of projected dependency relations. Using a parallel corpus, source trees are transferred onto equivalent target sentences via an extended set of alignment links. Projected arcs are initially weighted according to the certainty of word alignment links. Then, arc weights are recalculated using a method based on the EM selection algorithm. Maximum spanning trees selected from EM-scored digraphs and labelled with appropriate grammatical functions constitute a target dependency treebank. Extrinsic evaluation shows that parsers trained on such a treebank may perform comparably to parsers trained on a manually developed treebank.

Croatian Dependency Treebank 2.0: New Annotation Guidelines for Improved Parsing

Željko Agić, Daša Berović, Danijela Merkler and Marko Tadić

We present a new version of the Croatian Dependency Treebank. It constitutes a slight departure from the previously closely observed Prague Dependency Treebank syntactic layer annotation

guidelines as we introduce a new subset of syntactic tags on top of the existing tagset. These new tags are used in explicit annotation of subordinate clauses via subordinate conjunctions. Introducing the new annotation to Croatian Dependency Treebank, we also modify head attachment rules addressing subordinate conjunctions and subordinate clause predicates. In an experiment with data-driven dependency parsing, we show that implementing these new annotation guidelines leeds to a statistically significant improvement in parsing accuracy. We also observe a substantial improvement in inter-annotator agreement, facilitating more consistent annotation in further treebank development.

Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie

Rachel Bawden, Marie-Amélie Botalla, Kim Gerdes and Sylvain Kahane

This article presents the methods, results, and precision of the syntactic annotation process of the Rhapsodie Treebank of spoken French. The Rhapsodie Treebank is an 33,000 word corpus annotated for prosody and syntax, licensed in its entirety under Creative Commons. The syntactic annotation contains two levels: a macro-syntactic level, containing a segmentation into illocutionary units (including discourse markers, parentheses ...) and a micro-syntactic level including dependency relations and various paradigmatic structures, called pile constructions, the latter being particularly frequent and diverse in spoken language. The micro-syntactic annotation process, presented in this paper, includes a semi-automatic preparation of the transcription, the application of a syntactic dependency parser, transcoding of the parsing results to the Rhapsodie annotation scheme, manual correction by multiple annotators followed by a validation process, and finally the application of coherence rules that check common errors. The good inter-annotator agreement scores are presented and analyzed in greater detail. The article also includes the list of functions used in the dependency annotation and for the distinction of various pile constructions and presents the ideas underlying these choices.

Because Size Does Matter: The Hamburg Dependency Treebank

Kilian A. Foth, Arne Köhn, Niels Beuck and Wolfgang Menzel

We present the Hamburg Dependency Treebank (HDT), which to our knowledge is the largest dependency treebank currently available. It consists of genuine dependency annotations, i. e. they have not been transformed from phrase structures. We explore characteristics of the treebank and compare it against others. To exemplify the benefit of large dependency treebanks, we evaluate different parsers on the HDT. In addition, a set of tools will be described which help working with and searching in the treebank.

HamleDT 2.0: Thirty Dependency Treebanks Stanfordized

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman and Zdeněk Žabokrtský

We present HamleDT 2.0 (HArmonized Multi-LanguagE Dependency Treebank). HamleDT 2.0 is a collection of 30 existing treebanks harmonized into a common annotation style, the Prague Dependencies, and further transformed into Stanford Dependencies, a treebank annotation style that became popular in recent years. We use the newest basic Universal Stanford Dependencies, without added language-specific subtypes. We describe both of the annotation styles, including adjustments that were necessary to make, and provide details about the conversion process. We also discuss the differences between the two styles, evaluating their advantages and disadvantages, and note the effects of the differences on the conversion. We regard the stanfordization as generally successful, although we admit several shortcomings, especially in the distinction between direct and indirect objects, that have to be addressed in future. We release part of HamleDT 2.0 freely; we are not allowed to redistribute the whole dataset, but we do provide the conversion pipeline.

Bidirectionnal Converter Between Syntactic Annotations: from French Treebank Dependencies to PASSAGE Annotations, and back

Munshi Asadullah, Patrick Paroubek and Anne Vilnat

We present here part of a bidirectional converter between the French Tree-bank Dependency (FTB - DEP) annotations into the PASSAGE format. FTB - DEP is the representation used by several freely available parsers and the PASSAGE annotation was used to hand-annotate a relatively large sized corpus, used as gold-standard in the PASSAGE evaluation campaigns. Our converter will give the means to evaluate these parsers on the PASSAGE corpus. We shall illustrate the mapping of important syntactic phenomena using the corpus made of the examples of the FTB - DEP annotation guidelines, which we have hand-annotated with PASSAGE annotations and used to compute quantitative performance measures on the FTB - DEP guidelines.n this paper we will briefly introduce the two annotation formats. Then, we detail the two converters, and the rules which have been written. The last part will detail the results we obtained on the phenomenon we mostly study, the passive forms. We evaluate the converters by a double conversion, from PASSAGE to CoN LL

and back to PASSAGE. We will detailed in this paper the linguistic phenomenon we detail here, the passive form.

Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash and Ramy Eskander

This paper describes the parallel development of an Egyptian Arabic Treebank and a morphological analyzer for Egyptian Arabic (CALIMA). By the very nature of Egyptian Arabic, the data collected is informal, for example Discussion Forum text, which we use for the treebank discussed here. In addition, Egyptian Arabic, like other Arabic dialects, is sufficiently different from Modern Standard Arabic (MSA) that tools and techniques developed for MSA cannot be simply transferred over to work on Egyptian Arabic work. In particular, a morphological analyzer for Egyptian Arabic is needed to mediate between the written text and the segmented, vocalized form used for the syntactic trees. This led to the necessity of a feedback loop between the treebank team and the analyzer team, as improvements in each area were fed to the other. Therefore, by necessity, there needed to be close cooperation between the annotation team and the tool development team, which was to their mutual benefit. Collaboration on this type of challenge, where tools and resources are limited, proved to be remarkably synergistic and opens the way to further fruitful work on Arabic dialects.

Icelandic Invited Talk

Thursday, May 29, 13:10

Chairperson: **Eiríkur Rögnvaldsson**

Icelandic Quirks: Testing Linguistic Theories and Language Technology

Thórhallur Eythórsson

Linguists working on Icelandic have brought to the fore a number of important empirical facts that at the time of their initial discussion in the theoretical literature were believed to be crosslinguistically very rare, even unattested. Among such "quirks" are the following syntactic phenomena:

- Oblique ("quirky") subjects (Andrews 1976, Thráinsson 1979)
- Stylistic Fronting (Maling 1980)
- Long Distance Reflexivization (Thráinsson 1979)
- Object Shift of full NPs (Holmberg 1986)
- The Transitive Expletive Construction (Ottósson 1989, Jonas Bobaljik 1993)
- The New Passive (New Impersonal)(Maling Sigurjónsdóttir 2001, Eythórsson 2008)

These phenomena provided a testing ground for various theoretical models because they contradicted conventional views on the nature of grammatical categories and syntactic structure; some even went as far as claiming that Icelandic is "not a natural language". This pessimistic view was authoritatively examined and dismissed by Thráinsson (1996). The present paper takes the issue one step further, by showing how the discovery of various linguistic structures of Icelandic has led to the recognition of similar facts in other (Germanic, Indo-European and even unrelated) languages, where they had previously gone unnoticed, or had at least not been problematized in terms of linguistic theory. For example, the insight that syntactic subjects can have a morphological case other than nominative was not generally acknowledged until after the oblique subject hypothesis had been proposed for Icelandic. As a consequence, earlier theories on the relation between case and grammatical function had to be revised. Thus, numerous descriptive facts from Icelandic have advanced theoretical linguistics, in that any model of natural language must take them into account. In addition to their synchronic status, the syntactic phenomena listed above raise questions about the historical development of such "quirks". On the one hand, Icelandic is known to be a "conservative" language that has preserved many archaic features; on the other hand, despite its relative stability, numerous innovations are known have taken place in Icelandic, including a number of syntactic changes. Fortunately, we are now in a position to be able to map, at least to a certain degree, the diachrony of Icelandic syntax from the earliest attested documents in the 12th century AD until the present day. This is in particular due to the existence of the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al. 2011), which is currently being put to use in work on Icelandic diachronic syntax. Among other things, this research tool is invaluable in distinguishing between archaisms and innovations in Icelandic syntax. A further corpus, Greinir skáldskapar ("Analyzer of Poetry") (Karlsson et al. 2012), is particularly useful for the analysis of the syntax of the earliest poetic texts of Icelandic. In conclusion, the above "quirks" present a challenge both to Linguistic Theory and Language Technology. This paper illustrates, by means of selected examples, how this challenge has been successfully met and how advances in linguistic research proceed in a constant interplay between description and theorizing.

O25 - Machine Translation and Evaluation (2)

Thursday, May 29, 14:55

Chairperson: **Toru Ishida** Oral Session

The taraXÜ Corpus of Human-Annotated Machine Translations

Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Maja Popović, Cindy Tscherwinka, David Vilar and Hans Uszkoreit

Human translators are the key to evaluating machine translation (MT) quality and also to addressing the so far unanswered question when and how to use MT in professional translation workflows. This paper describes the corpus developed as a result of a detailed large scale human evaluation consisting of three tightly connected tasks: ranking, error classification and post-editing.

Measuring the Impact of Spelling Errors on the Quality of Machine Translation

Irina Galinskaya, Valentin Gusev, Elena Mescheryakova and Mariya Shmatova

In this paper we show how different types of spelling errors influence the quality of machine translation. We also propose a method to evaluate the impact of spelling errors correction on translation quality without expensive manual work of providing reference translations.

A Quality-based Active Sample Selection Strategy for Statistical Machine Translation

Varvara Logacheva and Lucia Specia

This paper presents a new active learning technique for machine translation based on quality estimation of automatically translated sentences. It uses an error-driven strategy, i.e., it assumes that the more errors an automatically translated sentence contains, the more informative it is for the translation system. Our approach is based on a quality estimation technique which involves a wider range of features of the source text, automatic translation, and machine translation system compared to previous work. In addition, we enhance the machine translation system training data with post-edited machine translations of the sentences selected, instead of simulating this using previously created reference translations. We found that re-training systems with additional post-edited data yields higher quality translations regardless of the selection strategy used. We relate this to the fact that post-editions tend to be closer to source sentences as compared to references, making the rule extraction process more reliable.

A Comparison of MT Errors and ESL Errors

Homa B. Hashemi and Rebecca Hwa

Generating fluent and grammatical sentences is a major goal for both Machine Translation (MT) and second-language Grammar Error Correction (GEC), but there have not been a lot of cross-fertilization between the two research communities. Arguably, an automatic translate-to-English system might be seen as an English as a Second Language (ESL) writer whose native language is the source language. This paper investigates whether research findings from the GEC community may help with characterizing MT error analysis. We describe a method for the automatic classification of MT errors according to English as a Second Language (ESL) error categories and conduct a large comparison experiment that includes both high-performing and low-performing translate-to-English MT systems for several source languages. Comparing the distribution of MT error types for all the systems suggests that MT systems have fairly similar distributions regardless of their source languages, and the high-performing MT systems have error distributions that are more similar to those of the low-performing MT systems than to those of ESL learners with the same L1.

VERTa: Facing a Multilingual Experience of a Linguistically-based MT Evaluation

Elisabet Comelles, Jordi Atserias, Victoria Arranz, Irene Castellon and Jordi Sesé

There are several MT metrics used to evaluate translation into Spanish, although most of them use partial or little linguistic information. In this paper we present the multilingual capability of VERTa, an automatic MT metric that combines linguistic information at lexical, morphological, syntactic and semantic level. In the experiments conducted we aim at identifying those linguistic features that prove the most effective to evaluate adequacy in Spanish segments. This linguistic information is tested both as independent modules (to observe what each type of feature provides) and in a combinatory fastion (where different kinds of information interact with each other). This allows us to extract the optimal combination. In addition we compare these linguistic features to those used in previous versions of VERTa aimed at evaluating adequacy for English segments. Finally, experiments show that VERTa can be easily adapted to other languages than English and that its collaborative approach correlates better with human judgements on adequacy than other well-known metrics.

O26 - Computer Aided Language Learning

Thursday, May 29, 14:55

Chairperson: **Justus Roux** Oral Session

ASR-based CALL Systems and Learner Speech Data: New Resources and Opportunities for Research and Development in Second Language Learning

Catia Cucchiarini, Steve Bodnar, Bart Penning de Vries, Roeland van Hout and Helmer Strik

In this paper we describe the language resources developed within the project "Feedback and the Acquisition of Syntax in Oral Proficiency" (FASOP), which is aimed at investigating the effectiveness of various forms of practice and feedback on the acquisition of syntax in second language (L2) oral proficiency, as well as their interplay with learner characteristics such as education level, learner motivation and confidence. For this purpose, use is made of a Computer Assisted Language Learning (CALL) system that employs Automatic Speech Recognition (ASR) technology to allow spoken interaction and to create an experimental environment that guarantees as much control over the language learning setting as possible. The focus of the present paper is on the resources that are being produced in FASOP. In line with the theme of this conference, we present the different types of resources developed within this project and the way in which these could be used to pursue innovative research in second language acquisition and to develop and improve ASR-based language learning applications.

The Dutch LESLLA Corpus

Eric Sanders, Ineke van de Craats and Vanja de Lint

This paper describes the Dutch LESLLA data and its curation. LESLLA stands for Low-Educated Second Language and Literacy Acquisition. The data was collected for research in this field and would have been disappeared if it were not saved. Within the CLARIN project Data Curation Service the data was made into a spoken language resource and made available to other researchers.

Student Achievement and French Sentence Repetition Test Scores

Deryle Lonsdale and Benjamin Millard

Sentence repetition (SR) tests are one way of probing a language learner's oral proficiency. Test-takers listen to a set of carefully engineered sentences of varying complexity one-by-one, and then try to repeat them back as exactly as possible. In this paper we explore how well an SR test that we have developed for French corresponds with the test-taker's achievement levels, represented by proficiency interview scores and by college class enrollment. We describe how we developed our SR test items using various language resources, and present pertinent facts about the test administration. The responses were scored by humans and also by a specially designed automatic speech recognition (ASR) engine; we sketch both scoring approaches. Results are evaluated in several ways: correlations between human and ASR scores, item response analysis to quantify the relative difficulty of the items, and criterion-referenced analysis setting thresholds of consistency across proficiency levels. We discuss several observations and conclusions prompted by the analyses, and suggestions for future work.

Using a Serious Game to Collect a Child Learner Speech Corpus

Claudia Baur, Manny Rayner and Nikos Tsourakis

We present an English-L2 child learner speech corpus, produced by 14 year old Swiss German-L1 students in their third year of learning English, which is currently in the process of being collected. The collection method uses a web-enabled multimodal language game implemented using the CALL-SLT platform, in which subjects hold prompted conversations with an animated agent. Prompts consist of a short animated English-language video clip together with a German-language piece of text indicating the semantic content of the requested response. Grammar-based speech understanding is used to decide whether responses are accepted or rejected, and dialogue flow is controlled using a simple XML-based scripting language; the scripts are written to allow multiple dialogue paths, the choice being made randomly. The system is gamified using a score-and-badge framework with four levels of badges. We describe the application, the data collection and annotation procedures, and the initial tranche of data. The full corpus, when complete, should contain at least 5,000 annotated utterances.

Sprinter: Language Technologies for Interactive and Multimedia Language Learning

Renlong Ai, Marcela Charfuelan, Walter Kasper, Tina Klüwer, Hans Uszkoreit, Feiyu Xu, Sandra Gasber and Philip Gienandt

Modern language learning courses are no longer exclusively based on books or face-to-face lectures. More and more lessons make use of multimedia and personalized learning methods. Many of these are based on e-learning solutions. Learning via the Internet provides 7/24 services that require sizeable human resources. Therefore we witness a growing economic pressure to employ computer-assisted methods for improving language learning in

quality, efficiency and scalability. In this paper, we will address three applications of language technologies for language learning: 1) Methods and strategies for pronunciation training in second language learning, e.g., multimodal feedback via visualization of sound features, speech verification and prosody transplantation; 2) Dialogue-based language learning games; 3) Application of parsing and generation technologies to the automatic generation of paraphrases for the semi-automatic production of learning material.

O27 - Information Extraction (1)

Thursday, May 29, 14:55

Chairperson: **Sophia Ananiadou** Oral Session

Automatic Semantic Relation Extraction from Portuguese Texts

Leonardo Sameshima Taba and Helena Caseli

Nowadays we are facing a growing demand for semantic knowledge in computational applications, particularly in Natural Language Processing (NLP). However, there aren't sufficient human resources to produce that knowledge at the same rate of its demand. Considering the Portuguese language, which has few resources in the semantic area, the situation is even more alarming. Aiming to solve that problem, this work investigates how some semantic relations can be automatically extracted from Portuguese texts. The two main approaches investigated here are based on (i) textual patterns and (ii) machine learning algorithms. Thus, this work investigates how and to which extent these two approaches can be applied to the automatic extraction of seven binary semantic relations (is-a, part-of, location-of, effect-of, property-of, made-of and used-for) in Portuguese texts. The results indicate that machine learning, in particular Support Vector Machines, is a promising technique for the task, although textual patterns presented better results for the used-for relation.

Estimation of Speaking Style in Speech Corpora Focusing on Speech Transcriptionss

Raymond Shen and Hideaki Kikuchi

Recent developments in computer technology have allowed the construction and widespread application of large-scale speech corpora. To foster ease of data retrieval for people interested in utilising these speech corpora, we attempt to characterise speaking style across some of them. In this paper, we first introduce the 3 scales of speaking style proposed by Eskenazi in 1993. We then use morphological features extracted from speech transcriptions that have proven effective in style discrimination and author identification in the field of natural language processing to construct an estimation model of speaking style. More specifically, we randomly choose transcriptions from various speech corpora as text stimuli with which to conduct a rating experiment on speaking style perception; then, using the features extracted from those stimuli and the rating results, we construct an estimation model of speaking style by a multi-regression analysis. After the cross validation (leave-1-out), the results show that among the 3 scales of speaking style, the ratings of 2 scales can be estimated with high accuracies, which prove the effectiveness of our method in the estimation of speaking style.

Annotating Clinical Events in Text Snippets for Phenotype Detection

Prescott Klassen, Fei Xia, Lucy Vanderwende and Meliha Yetisgen

Early detection and treatment of diseases that onset after a patient is admitted to a hospital, such as pneumonia, is critical to improving and reducing costs in healthcare. NLP systems that analyze the narrative data embedded in clinical artifacts such as x-ray reports can help support early detection. In this paper, we consider the importance of identifying the change of state for events - in particular, clinical events that measure and compare the multiple states of a patient's health across time. We propose a schema for event annotation comprised of five fields <location, attribute, value, change-of-state, reference> and create preliminary annotation guidelines for annotators to apply the schema. We then train annotators, measure their performance, and finalize our guidelines. With the complete guidelines, we then annotate a corpus of snippets extracted from chest x-ray reports in order to integrate the annotations as a new source of features for classification tasks.

Annotating Inter-Sentence Temporal Relations in Clinical Notes

Jennifer D'Souza and Vincent Ng

Owing in part to the surge of interest in temporal relation extraction, a number of datasets manually annotated with temporal relations between event-event pairs and event-time pairs have been produced recently. However, it is not uncommon to find missing annotations in these manually annotated datasets. Many researchers attributed this problem to "annotator fatigue". While some of these missing relations can be recovered automatically, many of them cannot. Our goals in this paper are to (1) manually annotate certain types of missing links that cannot be automatically recovered in the i2b2 Clinical Temporal Relations Challenge Corpus, one of the recently released evaluation corpora for temporal relation extraction; and (2) empirically determine the usefulness of these additional annotations. We will make

our annotations publicly available, in hopes of enabling a more accurate evaluation of temporal relation extraction systems.

Characterizing and Predicting Bursty Events: the Buzz Case Study on Twitter

Mohamed Morchid, Georges Linares and Richard Dufour

The prediction of bursty events on the Internet is a challenging task. Difficulties are due to the diversity of information sources, the size of the Internet, dynamics of popularity, user behaviors... On the other hand, Twitter is a structured and limited space. In this paper, we present a new method for predicting bursty events using content-related indices. Prediction is performed by a neural network that combines three features in order to predict the number of retweets of a tweet on the Twitter platform. The indices are related to popularity, expressivity and singularity. Popularity index is based on the analysis of RSS streams. Expressivity uses a dictionary that contains words annotated in terms of expressivity load. Singularity represents outlying topic association estimated via a Latent Dirichlet Allocation (LDA) model. Experiments demonstrate the effectiveness of the proposal with a 72% F-measure prediction score for the tweets that have been forwarded at least 60 times.

O28 - Lexicon

Thursday, May 29, 14:55

Chairperson: **Núria Bel** Oral Session

Building Domain Specific Bilingual Dictionaries

Lucas Hilgert, Lucelene Lopes, Artur Freitas, Renata Vieira, Denise Hogetop and Aline Vanin

This paper proposes a method to build bilingual dictionaries for specific domains defined by a parallel corpora. The proposed method is based on an original method that is not domain specific. Both the original and the proposed methods are constructed with previously available natural language processing tools. Therefore, this paper contribution resides in the choice and parametrization of the chosen tools. To illustrate the proposed method benefits we conduct an experiment over technical manuals in English and Portuguese. The results of our proposed method were analyzed by human specialists and our results indicates significant increases in precision for unigrams and muli-grams. Numerically, the precision increase is as big as 15% according to our evaluation.

DeLex, a Freely-available, Large-scale and Linguistically Grounded Morphological Lexicon for German

Benoît Sagot

We introduce DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German developed within the Alexina framework. We extracted lexical information from the German wiktionary and developed a morphological inflection grammar for German, based on a linguistically sound model of inflectional morphology. Although the developement of DeLex involved some manual work, we show that is represents a good tradeoff between development cost, lexical coverage and resource accuracy.

Walenty: Towards a Comprehensive Valence Dictionary of Polish

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski and Marek Świdziński

This paper presents Walenty, a comprehensive valence dictionary of Polish, with a number of novel features, as compared to other such dictionaries. The notion of argument is based on the coordination test and takes into consideration the possibility of diverse morphosyntactic realisations. Some aspects of the internal structure of phraseological (idiomatic) arguments are handled explicitly. While the current version of the dictionary concentrates on syntax, it already contains some semantic features, including semantically defined arguments, such as locative, temporal or manner, as well as control and raising, and work on extending it with semantic roles and selectional preferences is in progress. Although Walenty is still being intensively developed, it is already by far the largest Polish valence dictionary, with around 8600 verbal lemmata and almost 39 000 valence schemata. The dictionary is publicly available on the Creative Commons BY SA licence and may be downloaded from http://zil.ipipan.waw.pl/Walenty.

A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon

Marion Baranes and Benoît Sagot

In this paper, we describe and evaluate an unsupervised method for acquiring pairs of lexical entries belonging to the same morphological family, i.e., derivationally related words, starting from a purely inflectional lexicon. Our approach relies on transformation rules that relate lexical entries with the one another, and which are automatically extracted from the inflected lexicon based on surface form analogies and on part-of-speech information. It is generic enough to be applied to any language with a mainly concatenative derivational morphology. Results were obtained and evaluated on English, French, German and Spanish. Precision results are satisfying, and our French results favorably compare with another resource, although its construction relied on manually developed lexicographic

information whereas our approach only requires an inflectional lexicon.

Automatic Expansion of the MRC Psycholinguistic Database Imageability Ratings

Ting Liu, Kit Cho, G. Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz, Ignacio Cases and Ching-Sheng Lin

Recent studies in metaphor extraction across several languages (Broadwell et al., 2013; Strzalkowski et al., 2013) have shown that word imageability ratings are highly correlated with the presence of metaphors in text. Information about imageability of words can be obtained from the MRC Psycholinguistic Database (MRCPD) for English words and Léxico Informatizado del Español Programa (LEXESP) for Spanish words, which is a collection of human ratings obtained in a series of controlled surveys. Unfortunately, word imageability ratings were collected for only a limited number of words: 9,240 words in English, 6,233 in Spanish; and are unavailable at all in the other two languages studied: Russian and Farsi. The present study describes an automated method for expanding the MRCPD by conferring imageability ratings over the synonyms and hyponyms of existing MRCPD words, as identified in Wordnet. The result is an expanded MRCPD+ database with imagea-bility scores for more than 100,000 words. The appropriateness of this expansion process is assessed by examining the structural coherence of the expanded set and by validating the expanded lexicon against human judgment. Finally, the performance of the metaphor extraction system is shown to improve significantly with the expanded database. This paper describes the process for English MRCPD+ and the resulting lexical resource. The process is analogous for other languages.

P34 - Corpora and Annotation

Thursday, May 29, 14:55

Chairperson: **Zygmunt Vetulani** Poster Session

A Conventional Orthography for Tunisian Arabic

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith and Nizar Habash

Tunisian Arabic is a dialect of the Arabic language spoken in Tunisia. Tunisian Arabic is an under-resourced language. It has neither a standard orthography nor large collections of written text and dictionaries. Actually, there is no strict separation between Modern Standard Arabic, the official language of the government, media and education, and Tunisian Arabic; the two exist on a continuum dominated by mixed forms. In this paper, we present a conventional orthography for Tunisian Arabic, following a previous effort on developing a conventional orthography for Dialectal Arabic (or CODA) demonstrated for Egyptian Arabic. We explain the design principles of CODA and provide a detailed description of its guidelines as applied to Tunisian Arabic.

Large Scale Arabic Error Annotation: Guidelines and Framework

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani and Kemal Oflazer

We present annotation guidelines and a web-based annotation framework developed as part of an effort to create a manually annotated Arabic corpus of errors and corrections for various text types. Such a corpus will be invaluable for developing Arabic error correction tools, both for training models and as a gold standard for evaluating error correction algorithms. We summarize the guidelines we created. We also describe issues encountered during the training of the annotators, as well as problems that are specific to the Arabic language that arose during the annotation process. Finally, we present the annotation tool that was developed as part of this project, the annotation pipeline, and the quality of the resulting annotations.

Flow Graph Corpus from Recipe Texts

Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata and Tetsuro Sasada

In this paper, we present our attempt at annotating procedural texts with a flow graph as a representation of understanding. The domain we focus on is cooking recipe. The flow graphs are directed acyclic graphs with a special root node corresponding to the final dish. The vertex labels are recipe named entities, such as foods, tools, cooking actions, etc. The arc labels denote relationships among them. We converted 266 Japanese recipe texts into flow graphs manually. 200 recipes are randomly selected from a web site and 66 are of the same dish. We detail the annotation framework and report some statistics on our corpus. The most typical usage of our corpus may be automatic conversion from texts to flow graphs which can be seen as an entire understanding of procedural texts. With our corpus, one can also try word segmentation, named entity recognition, predicate-argument structure analysis, and coreference resolution.

Recent Developments in DeReKo

Marc Kupietz and Harald Lüngen

This paper gives an overview of recent developments in the German Reference Corpus DeReKo in terms of growth,

maximising relevant corpus strata, metadata, legal issues, and its current and future research interface. Due to the recent acquisition of new licenses, DeReKo has grown by a factor of four in the first half of 2014, mostly in the area of newspaper text, and presently contains over 24 billion word tokens. Other strata, like fictional texts, web corpora, in particular CMC texts, and spoken but conceptually written texts have also increased significantly. We report on the newly acquired corpora that led to the major increase, on the principles and strategies behind our corpus acquisition activities, and on our solutions for the emerging legal, organisational, and technical challenges.

Why Chinese Web-as-Corpus is Wacky? Or: How Big Data is Killing Chinese Corpus Linguistics

Shu-Kai Hsieh

This paper aims to examine and evaluate the current development of using Web-as-Corpus (WaC) paradigm in Chinese corpus linguistics. I will argue that the unstable notion of wordhood in Chinese and the resulting diverse ideas of implementing word segmentation systems have posed great challenges for those who are keen on building web-scaled corpus data. Two lexical measures are proposed to illustrate the issues and methodological discussions are provided.

Extending HeidelTime for Temporal Expressions Referring to Historic Dates

Jannik Strötgen, Thomas Bögel, Julian Zell, Ayser Armiti, Tran van Canh and Michael Gertz

Research on temporal tagging has achieved a lot of attention during the last years. However, most of the work focuses on processing news-style documents. Thus, references to historic dates are often not well handled by temporal taggers although they frequently occur in narrative-style documents about history, e.g., in many Wikipedia articles. In this paper, we present the AncientTimes corpus containing documents about different historic time periods in eight languages, in which we manually annotated temporal expressions. Based on this corpus, we explain the challenges of temporal tagging documents about history. Furthermore, we use the corpus to extend our multilingual, cross-domain temporal tagger HeidelTime to extract and normalize temporal expressions referring to historic dates, and to demonstrate HeidelTime's new capabilities. Both, the AncientTimes corpus as well as the new HeidelTime version are made publicly available.

A 500-Million Word POS-Tagged Icelandic Corpus

Thomas Eckart, Erla Hallsteinsdóttir, Sigrún Helgadóttir, Uwe Quasthoff and Dirk Goldhahn

The new POS-tagged Icelandic corpus of the Leipzig Corpora Collection is an extensive resource for the analysis of the Icelandic language. As it contains a large share of all Web documents hosted under the .is top-level domain, it is especially valuable for investigations on modern Icelandic and non-standard language varieties. The corpus is accessible via a dedicated web portal and large shares are available for download. Focus of this paper will be the description of the tagging process and evaluation of statistical properties like word form frequencies and part of speech tag distributions. The latter will be in particular compared with values from the Icelandic Frequency Dictionary (IFD) Corpus.

Building the Sense-Tagged Multilingual Parallel Corpus

Shan Wang and Francis Bond

Sense-annotated parallel corpora play a crucial role in natural language processing. This paper introduces our progress in creating such a corpus for Asian languages using English as a pivot, which is the first such corpus for these languages. Two sets of tools have been developed for sequential and targeted tagging, which are also easy to set up for any new language in addition to those we are annotating. This paper also briefly presents the general guidelines for doing this project. The current results of monolingual sense-tagging and multilingual linking are illustrated, which indicate the differences among genres and language pairs. All the tools, guidelines and the manually annotated corpus will be freely available at compling.ntu.edu.sg/ntumc.

A Hindi-English Code-Switching Corpus

Anik Dey and Pascale Fung

The aim of this paper is to investigate the rules and constraints of code-switching (CS) in Hindi-English mixed language data. In this paper, we'll discuss how we collected the mixed language corpus. This corpus is primarily made up of student interview speech. The speech was manually transcribed and verified by bilingual speakers of Hindi and English. The code-switching cases in the corpus are discussed and the reasons for code-switching are explained.

KoKo: an L1 Learner Corpus for German

Andrea Abel, Aivars Glaznieks, Lionel Nicolas and Egon Stemle

We introduce the KoKo corpus, a collection of German L1 learner texts annotated with learner errors, along with the methods and tools used in its construction and evaluation. The corpus contains both texts and corresponding survey information from 1,319 pupils and amounts to around 716,000 tokens. The evaluation of the performed transcriptions and annotations shows an accuracy

of orthographic error annotations of approximately 80% as well as high accuracies of transcriptions (>99%), automatic tokenisation (>99%), sentence splitting (>96%) and POS-tagging (>94%). The KoKo corpus will be published at the end of 2014. It will be the first accessible linguistically annotated German L1 learner corpus and a valuable source for research on L1 learner language as well as for teachers of German as L1, in particular with regards to writing skills.

On Paraphrase Identification Corpora

Vasile Rus, Rajendra Banjade and Mihai Lintean

We analyze in this paper a number of data sets proposed over the last decade or so for the task of paraphrase identification. The goal of the analysis is to identify the advantages as well as shortcomings of the previously proposed data sets. Based on the analysis, we then make recommendations about how to improve the process of creating and using such data sets for evaluating in the future approaches to the task of paraphrase identification or the more general task of semantic similarity. The recommendations are meant to improve our understanding of what a paraphrase is, offer a more fair ground for comparing approaches, increase the diversity of actual linguistic phenomena that future data sets will cover, and offer ways to improve our understanding of the contributions of various modules or approaches proposed for solving the task of paraphrase identification or similar tasks.

Construction and Annotation of a French Folkstale Corpus

Anne Garcia-Fernandez, Anne-Laure Ligozat and Anne Vilnat

In this paper, we present the digitization and annotation of a tales corpus - which is to our knowledge the only French tales corpus available and classified according to the Aarne&Thompson classification - composed of historical texts (with old French parts). We first studied whether the pre-processing tools, namely OCR and PoS-tagging, have good enough accuracies to allow automatic analysis. We also manually annotated this corpus according to several types of information which could prove useful for future work: character references, episodes, and motifs. The contributions are the creation of an corpus of French tales from classical anthropology material, which will be made available to the community; the evaluation of OCR and NLP tools on this corpus; and the annotation with anthropological information.

Statistical Analysis of Multilingual Text Corpus and Development of Language Models

Shyam Sundar Agrawal, Abhimanue, Shweta Bansal and Minakshi Mahajan

This paper presents two studies, first a statistical analysis for three languages i.e. Hindi, Punjabi and Nepali and the other, development of language models for three Indian languages i.e. Indian English, Punjabi and Nepali. The main objective of this study is to find distinction among these languages and development of language models for their identification. Detailed statistical analysis have been done to compute the information about entropy, perplexity, vocabulary growth rate etc. Based on statistical features a comparative analysis has been done to find the similarities and differences among these languages. Subsequently an effort has been made to develop a trigram model of Indian English, Punjabi and Nepali. A corpus of 500000 words of each language has been collected and used to develop their models (unigram, bigram and trigram models). The models have been tried in two different databases- Parallel corpora of French and English and Non-parallel corpora of Indian English, Punjabi and Nepali. In the second case, the performance of the model is comparable. Usage of JAVA platform has provided a special effect for dealing with a very large database with high computational speed. Furthermore various enhancive concepts like Smoothing, Discounting, Back off, and Interpolation have been included for the designing of an effective model. The results obtained from this experiment have been described. The information can be useful for development of Automatic Speech Language Identification System.

Building a Dataset for Summarization and Keyword Extraction from Emails

Vanessa Loza, Shibamouli Lahiri, Rada Mihalcea and Po-Hsiang Lai

This paper introduces a new email dataset, consisting of both single and thread emails, manually annotated with summaries and keywords. A total of 349 emails and threads have been annotated. The dataset is our first step toward developing automatic methods for summarization and keyword extraction from emails. We describe the email corpus, along with the annotation interface, annotator guidelines, and agreement studies.

P35 - Grammar and Syntax

Thursday, May 29, 14:55

Chairperson: **Tamás Váradi** Poster Session

Language CoLLAGE: Grammatical Description with the LinGO Grammar Matrix

Emily M. Bender

Language CoLLAGE is a collection of grammatical descriptions developed in the context of a grammar engineering graduate course with the LinGO Grammar Matrix. These grammatical descriptions include testsuites in well-formed interlinear glossed text (IGT) format, high-level grammatical characterizations called 'choices files', HPSG grammar fragments (capable of parsing and generation), and documentation. As of this writing, Language CoLLAGE includes resources for 52 typologically and areally diverse languages and this number is expected to grow over time. The resources for each language cover a similar range of core grammatical phenomena and are implemented in a uniform framework, compatible with the DELPH-IN suite of processing tools.

To Pay or to Get Paid: Enriching a Valency Lexicon with Diatheses

Anna Vernerová, Václava Kettnerová and Marketa Lopatkova

Valency lexicons typically describe only unmarked usages of verbs (the active form); however, verbs prototypically enter different surface structures. In this paper, we focus on the so-called diatheses, i.e., the relations between different surface syntactic manifestations of verbs that are brought about by changes in the morphological category of voice, e.g., the passive diathesis. The change in voice of a verb is prototypically associated with shifts of some of its valency complementations in the surface structure. These shifts are implied by changes in morphemic forms of the involved valency complementations and are regular enough to be captured by syntactic rules. However, as diatheses are lexically conditioned, their applicability to an individual lexical unit of a verb is not predictable from its valency frame alone. In this work, we propose a representation of this linguistic phenomenon in a valency lexicon of Czech verbs, VALLEX, with the aim to enhance this lexicon with the information on individual types of Czech diatheses. In order to reduce the amount of necessary manual annotation, a semi-automatic method is developed. This method draws evidence from a large morphologically annotated corpus, relying on grammatical constraints on the applicability of individual types of diatheses.

The Ellogon Pattern Engine: Context-free Grammars over Annotations

Georgios Petasis

This paper presents the pattern engine that is offered by the Ellogon language engineering platform. This pattern engine allows the application of context-free grammars over annotations, which are metadata generated during the processing of documents by natural language tools. In addition, grammar development is aided by a graphical grammar editor, giving grammar authors the capability to test and debug grammars.

Extracting a bilingual semantic grammar from FrameNet-annotated corpora

Dana Dannells and Normunds Gruzitis

We present the creation of an English-Swedish FrameNet-based grammar in Grammatical Framework. The aim of this research is to make existing framenets computationally accessible for multilingual natural language applications via a common semantic grammar API, and to facilitate the porting of such grammar to other languages. In this paper, we describe the abstract syntax of the semantic grammar while focusing on its automatic extraction possibilities. We have extracted a shared abstract syntax from 58,500 annotated sentences in Berkeley FrameNet (BFN) and 3,500 annotated sentences in Swedish FrameNet (SweFN). The abstract syntax defines 769 frame-specific valence patterns that cover 77,8% examples in BFN and 74,9% in SweFN belonging to the shared set of 471 frames. As a side result, we provide a unified method for comparing semantic and syntactic valence patterns across framenets.

Relating Frames and Constructions in Japanese FrameNet

Kyoko Ohara

Relations between frames and constructions must be made explicit in FrameNet-style linguistic resources such as Berkeley FrameNet (Fillmore & Baker, 2010, Fillmore, Lee-Goldman & Rhomieux, 2012), Japanese FrameNet (Ohara, 2013), and Swedish Constructicon (Lyngfelt et al., 2013). On the basis of analyses of Japanese constructions for the purpose of building a constructicon in the Japanese FrameNet project, this paper argues that constructions can be classified based on whether they evoke frames or not. By recognizing such a distinction among constructions, it becomes possible for FrameNet-style linguistic resources to have a proper division of labor between frame annotations and construction annotations. In addition to the three kinds of "meaningless" constructions which have been proposed already, this paper suggests there may be yet another subtype

of constructions without meanings. Furthermore, the present paper adds support to the claim that there may be constructions without meanings (Fillmore, Lee-Goldman & Rhomieux, 2012) in a current debate concerning whether all constructions should be seen as meaning-bearing (Goldberg, 2006: 166-182).

MultiVal - Towards a Multilingual Valence Lexicon

Lars Hellan, Dorothee Beermann, Tore Bruland, Mary Esther Kropp Dakubu and Montserrat Marimón

MultiVal is a valence lexicon derived from lexicons of computational HPSG grammars for Norwegian, Spanish and Ga (ISO 639-3, gaa), with altogether about 22,000 verb entries and on average more than 200 valence types defined for each language. These lexical resources are mapped onto a common set of discriminants with a common array of values, and stored in a relational database linked to a web demo and a wiki presentation. Search discriminants are 'syntactic argument structure' (SAS), functional specification, situation type and aspect, for any subset of languages, as well as the verb type systems of the grammars. Search results are lexical entries satisfying the discriminants entered, exposing the specifications from the respective provenance grammars. The Ga grammar lexicon has in turn been converted from a Ga Toolbox lexicon. Aside from the creation of such a multilingual valence resource through converging or converting existing resources, the paper also addresses a tool for the creation of such a resource as part of corpus annotation for less resourced languages.

A Vector Space Model for Syntactic Distances Between Dialects

Emanuele di Buccio, Giorgio Maria di Nunzio and Gianmaria Silvello

Syntactic comparison across languages is essential in the research field of linguistics, e.g. when investigating the relationship among closely related languages. In IR and NLP, the syntactic information is used to understand the meaning of word occurrences according to the context in which their appear. In this paper, we discuss a mathematical framework to compute the distance between languages based on the data available in current state-of-the-art linguistic databases. This framework is inspired by approaches presented in IR and NLP.

Resources in Conflict: A Bilingual Valency Lexicon vs. a Bilingual Treebank vs. a Linguistic Theory

Jana Sindlerova, Zdenka Uresova and Eva Fucikova

In this paper, we would like to exemplify how a syntactically annotated bilingual treebank can help us in exploring and revising a developed linguistic theory. On the material of the Prague Czech-English Dependency Treebank we observe sentences in which an Addressee argument in one language is linked translationally to a Patient argument in the other one, and make generalizations about the theoretical grounds of the argument non-correspondences and its relations to the valency theory beyond the annotation practice. Exploring verbs of three semantic classes (Judgement verbs, Teaching verbs and Attempt Suasion verbs) we claim that the Functional Generative Description argument labelling is highly dependent on the morphosyntactic realization of the individual participants, which then results in valency frame differences. Nevertheless, most of the differences can be overcome without substantial changes to the linguistic theory itself.

P36 - Metaphors

Thursday, May 29, 14:55

Chairperson: **Walter Daelemans** Poster Session

A Multi-Cultural Repository of Automatically Discovered Linguistic and Conceptual Metaphors

Samira Shaikh, Tomek Strzalkowski, Ting Liu, George Aaron Broadwell, Boris Yamrom, Sarah Taylor, Laurie Feldman, Kit Cho, Umit Boz, Ignacio Cases, Yuliya Peshkova and Ching-Sheng Lin

In this article, we present details about our ongoing work towards building a repository of Linguistic and Conceptual Metaphors. This resource is being developed as part of our research effort into the large-scale detection of metaphors from unrestricted text. We have stored a large amount of automatically extracted metaphors in American English, Mexican Spanish, Russian and Iranian Farsi in a relational database, along with pertinent metadata associated with these metaphors. A substantial subset of the contents of our repository has been systematically validated via rigorous social science experiments. Using information stored in the repository, we are able to posit certain claims in a cross-cultural context about how peoples in these cultures (America, Mexico, Russia and Iran) view particular concepts related to Governance and Economic Inequality through the use of metaphor. Researchers in the field can use this resource as a reference of typical metaphors used across these cultures. In addition, it can be used to recognize metaphors of the same form or pattern, in other domains of research.

Two Approaches to Metaphor Detection

Brian MacWhinney and Davida Fromm

Methods for automatic detection and interpretation of metaphors have focused on analysis and utilization of the ways in

which metaphors violate selectional preferences (Martin, 2006). Detection and interpretation processes that rely on this method can achieve wide coverage and may be able to detect some novel metaphors. However, they are prone to high false alarm rates, often arising from imprecision in parsing and supporting ontological and lexical resources. An alternative approach to metaphor detection emphasizes the fact that many metaphors become conventionalized collocations, while still preserving their active metaphorical status. Given a large enough corpus for a given language, it is possible to use tools like SketchEngine (Kilgariff, Rychly, Smrz, & Tugwell, 2004) to locate these high frequency metaphors for a given target domain. In this paper, we examine the application of these two approaches and discuss their relative strengths and weaknesses for metaphors in the target domain of economic inequality in English, Spanish, Farsi, and Russian.

Mining Online Discussion Forums for Metaphors

Andrew Gargett and John Barnden

We present an approach to mining online forums for figurative language such as metaphor. We target in particular online discussions within the illness and the political conflict domains, with a view to constructing corpora of Metaphor in Illness Discussion, andMetaphor in Political Conflict Discussion. This paper reports on our ongoing efforts to combine manual and automatic detection strategies for labelling the corpora, and present some initial results from our work showing that metaphor use is not independent of illness domain.

P37 - Named Entity Recognition

Thursday, May 29, 14:55

Chairperson: **German Rigau** Poster Session

Simple Effective Microblog Named Entity Recognition: Arabic as an Example

Kareem Darwish and Wei Gao

Despite many recent papers on Arabic Named Entity Recognition (NER) in the news domain, little work has been done on microblog NER. NER on microblogs presents many complications such as informality of language, shortened named entities, brevity of expressions, and inconsistent capitalization (for cased languages). We introduce simple effective language-independent approaches for improving NER on microblogs, based on using large gazetteers, domain adaptation, and a two-pass semi-supervised method. We use Arabic as an example language to compare the relative effectiveness of the approaches and when best to use them. We also present a new dataset for the task. Results of combining the proposed approaches show an improvement of 35.3 F-measure points over a baseline system trained on news data and an improvement of 19.9 F-measure points over the same system but trained on microblog data.

Biomedical Entity Extraction using Machine Learning-based Approaches

Cyril Grouin

In this paper, we present the experiments we made to process entities from the biomedical domain. Depending on the task to process, we used two distinct supervised machine-learning techniques: Conditional Random Fields to perform both named entity identification and classification, and Maximum Entropy to classify given entities. Machine-learning approaches outperformed knowledge-based techniques on categories where sufficient annotated data was available. We showed that the use of external features (unsupervised clusters, information from ontology and taxonomy) improved the results significantly.

NoSta-D Named Entity Annotation for German: Guidelines and Dataset

Darina Benikova, Chris Biemann and Marc Reznicek

We describe the annotation of a new dataset for German Named Entity Recognition (NER). The need for this dataset is motivated by licensing issues and consistency issues of existing datasets. We describe our approach to creating annotation guidelines based on linguistic and semantic considerations, and how we iteratively refined and tested them in the early stages of annotation in order to arrive at the largest publicly available dataset for German NER, consisting of over 31,000 manually annotated sentences (over 591,000 tokens) from German Wikipedia and German online news. We provide a number of statistics on the dataset, which indicate its high quality, and discuss legal aspects of distributing the data as a compilation of citations. The data is released under the permissive CC-BY license, and will be fully available for download in September 2014 after it has been used for the GermEval 2014 shared task on NER. We further provide the full annotation guidelines and links to the annotation tool used for the creation of this resource.

Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese

Haibo Li, Masato Hagiwara, Qi Li and Heng Ji

Word Segmentation is usually considered an essential step for many Chinese and Japanese Natural Language Processing tasks, such as name tagging. This paper presents several new observations and analysis on the impact of word segmentation on name tagging; (1). Due to the limitation of current state-of-the-art

Chinese word segmentation performance, a character-based name tagger can outperform its word-based counterparts for Chinese but not for Japanese; (2). It is crucial to keep segmentation settings (e.g. definitions, specifications, methods) consistent between training and testing for name tagging; (3). As long as (2) is ensured, the performance of word segmentation does not have appreciable impact on Chinese and Japanese name tagging.

HFST-SweNER – A New NER Resource for Swedish

Dimitrios Kokkinakis, Jyrki Niemi, Sam Hardwick, Krister Lindén and Lars Borin

Named entity recognition (NER) is a knowledge-intensive information extraction task that is used for recognizing textual mentions of entities that belong to a predefined set of categories, such as locations, organizations and time expressions. NER is a challenging, difficult, yet essential preprocessing technology for many natural language processing applications, and particularly crucial for language understanding. NER has been actively explored in academia and in industry especially during the last years due to the advent of social media data. This paper describes the conversion, modeling and adaptation of a Swedish NER system from a hybrid environment, with integrated functionality from various processing components, to the Helsinki Finite-State Transducer Technology (HFST) platform. This new HFST-based NER (HFST-SweNER) is a full-fledged open source implementation that supports a variety of generic named entity types and consists of multiple, reusable resource layers, e.g., various n-gram-based named entity lists (gazetteers).

Crowdsourcing and Annotating NER for Twitter #drift

Hege Fromreide, Dirk Hovy and Anders Søgaard

We present two new NER datasets for Twitter; a manually annotated set of 1,467 tweets (kappa=0.942) and a set of 2,975 expert-corrected, crowdsourced NER annotated tweets from the dataset described in Finin et al. (2010). In our experiments with these datasets, we observe two important points: (a) language drift on Twitter is significant, and while off-the-shelf systems have been reported to perform well on in-sample data, they often perform poorly on new samples of tweets, (b) state-of-the-art performance across various datasets can be obtained from crowdsourced annotations, making it more feasible to "catch up" with language drift.

Clustering of Multi-Word Named Entity Variants: Multilingual Evaluation

Guillaume Jacquet, Maud Ehrmann and Ralf Steinberger

Multi-word entities, such as organisation names, are frequently written in many different ways. We have previously automatically identified over one million acronym pairs in 22 languages, consisting of their short form (e.g. EC) and their corresponding long forms (e.g. European Commission, European Union Commission). In order to automatically group such long form variants as belonging to the same entity, we cluster them, using bottom-up hierarchical clustering and pair-wise string similarity metrics. In this paper, we address the issue of how to evaluate the named entity variant clusters automatically, with minimal human annotation effort. We present experiments that make use of Wikipedia redirection tables and we show that this method produces good results.

Comparative Analysis of Portuguese Named Entities Recognition Tools

Daniela Amaral, Evandro Fonseca, Lucelene Lopes and Renata Vieira

This paper describes an experiment to compare four tools to recognize named entities in Portuguese texts. The experiment was made over the HAREM corpora, a golden standard for named entities recognition in Portuguese. The tools experimented are based on natural language processing techniques and also machine learning. Specifically, one of the tools is based on Conditional random fields, an unsupervised machine learning model that has being used to named entities recognition in several languages, while the other tools follow more traditional natural language approaches. The comparison results indicate advantages for different tools according to the different classes of named entities. Despite of such balance among tools, we conclude pointing out foreseeable advantages to the machine learning based tool.

Generating a Resource for Products and Brandnames Recognition. Application to the Cosmetic Domain.

Cédric Lopez, Frédérique Segond, Olivier Hondermarck, Paolo Curtoni and Luca Dini

Named Entity Recognition task needs high-quality and large-scale resources. In this paper, we present RENCO, a based-rules system focused on the recognition of entities in the Cosmetic domain (brandnames, product names, . . .). RENCO has two main objectives: 1) Generating resources for named entity recognition; 2) Mining new named entities relying on the previous generated resources. In order to build lexical resources for the cosmetic domain, we propose a system based on local lexico-syntactic rules complemented by a learning module. As the outcome of the system, we generate both a simple lexicon and a structured lexicon. Results of the evaluation show that even if RENCO

outperforms a classic Conditional Random Fields algorithm, both systems should combine their respective strengths.

Named Entity Corpus Construction using Wikipedia and DBpedia Ontology

Younggyun Hahm, Jungyeul Park, Kyungtae Lim, Youngsik Kim, Dosam Hwang and Key-Sun Choi

In this paper, we propose a novel method to automatically build a named entity corpus based on the DBpedia ontology. Since most of named entity recognition systems require time and effort consuming annotation tasks as training data. Work on NER has thus for been limited on certain languages like English that are resource-abundant in general. As an alternative, we suggest that the NE corpus generated by our proposed method, can be used as training data. Our approach introduces Wikipedia as a raw text and uses the DBpedia data set for named entity disambiguation. Our method is language-independent and easy to be applied to many different languages where Wikipedia and DBpedia are provided. Throughout the paper, we demonstrate that our NE corpus is of comparable quality even to the manually annotated NE corpus.

Exploring the Utility of Coreference Chains for Improved Identification of Personal Names

Andrea Glaser and Jonas Kuhn

Identifying the real world entity that a proper name refers to is an important task in many NLP applications. Context plays an important role in disambiguating entities with the same names. In this paper, we discuss a dataset and experimental set-up that allows us to systematically explore the effects of different sizes and types of context in this disambiguation task. We create context by first identifying coreferent expressions in the document and then combining sentences these expressions occur in to one informative context. We apply different filters to obtain different levels of coreference-based context. Since hand-labeling a dataset of a decent size is expensive, we investigate the usefulness of an automatically created pseudo-ambiguity dataset. The results on this pseudo-ambiguity dataset show that using coreference-based context performs better than using a fixed window of context around the entity. The insights taken from the pseudo data experiments can be used to predict how the method works with real data. In our experiments on real data we obtain comparable results.

Named Entity Tagging a Very Large Unbalanced Corpus: Training and Evaluating NE Classifiers

Joachim Bingel and Thomas Haider

We describe a systematic and application-oriented approach to training and evaluating named entity recognition and classification (NERC) systems, the purpose of which is to identify an optimal system and to train an optimal model for named entity tagging DeReKo, a very large general-purpose corpus of contemporary German (Kupietz et al., 2010). DeReKo 's strong dispersion wrt. genre, register and time forces us to base our decision for a specific NERC system on an evaluation performed on a representative sample of DeReKo instead of performance figures that have been reported for the individual NERC systems when evaluated on more uniform and less diverse data. We create and manually annotate such a representative sample as evaluation data for three different NERC systems, for each of which various models are learnt on multiple training data. The proposed sampling method can be viewed as a generally applicable method for sampling evaluation data from an unbalanced target corpus for any sort of natural language processing.

P38 - Question Answering

Thursday, May 29, 14:55

Chairperson: **António Branco** Poster Session

REFRACTIVE: An Open Source Tool to Extract Knowledge from Syntactic and Semantic Relations

Peter Exner and Pierre Nugues

The extraction of semantic propositions has proven instrumental in applications like IBM Watson and in Google's knowledge graph . One of the core components of IBM Watson is the PRISMATIC knowledge base consisting of one billion propositions extracted from the English version of Wikipedia and the New York Times. However, extracting the propositions from the English version of Wikipedia is a time-consuming process. In practice, this task requires multiple machines and a computation distribution involving a good deal of system technicalities. In this paper, we describe Refractive, an open-source tool to extract propositions from a parsed corpus based on the Hadoop variant of MapReduce. While the complete process consists of a parsing part and an extraction part, we focus here on the extraction from the parsed corpus and we hope this tool will help computational linguists speed up the development of applications.

Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving

Akira Fujita, Akihiro Kameda, Ai Kawazoe and Yusuke Miyao

We introduce the organization of the Todai Robot Project and discuss its achievements. The Todai Robot Project task focuses on benchmarking NLP systems for problem solving. This

task encourages NLP-based systems to solve real high-school examinations. We describe the details of the method to manage question resources and their correct answers, answering tools and participation by researchers in the task. We also analyse the answering accuracy of the developed systems by comparing the systems' answers with answers given by human test-takers.

Annotating Question Decomposition on Complex Medical Questions

Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu and Dina Demner-Fushman

This paper presents a method for annotating question decomposition on complex medical questions. The annotations cover multiple syntactic ways that questions can be decomposed, including separating independent clauses as well as recognizing coordinations and exemplifications. We annotate a corpus of 1,467 multi-sentence consumer health questions about genetic and rare diseases. Furthermore, we label two additional medical-specific annotations: (1) background sentences are annotated with a number of medical categories such as symptoms, treatments, and family history, and (2) the central focus of the complex question (a disease) is marked. We present simple baseline results for automatic classification of these annotations, demonstrating the challenging but important nature of this task.

JUST.ASK, a QA System that Learns to Answer New Questions from Previous Interactions

Sérgio Curto, Ana C. Mendes, Pedro Curto, Luísa Coheur and Angela Costa

We present JUST.ASK, a publicly available Question Answering system, which is freely available. Its architecture is composed of the usual Question Processing, Passage Retrieval and Answer Extraction components. Several details on the information generated and manipulated by each of these components are also provided to the user when interacting with the demonstration. Since JUST.ASK also learns to answer new questions based on users' feedback, (s)he is invited to identify the correct answers. These will then be used to retrieve answers to future questions.

Extraction of Daily Changing Words for Question Answering

Kugatsu Sadamitsu, Ryuichiro Higashinaka and Yoshihiro Matsuo

This paper proposes a method for extracting Daily Changing Words (DCWs), words that indicate which questions are real-time dependent. Our approach is based on two types of template matching using time and named entity slots from large size corpora and adding simple filtering methods from news corpora. Extracted DCWs are utilized for detecting and sorting real-time dependent questions. Experiments confirm that our DCW method achieves higher accuracy in detecting real-time dependent questions than existing word classes and a simple supervised machine learning approach.

LinkedHealthAnswers: Towards Linked Data-driven Question Answering for the Health Care Domain

Artem Ostankov, Florian Röhrbein and Ulli Waltinger

This paper presents Linked Health Answers, a natural language question answering systems that utilizes health data drawn from the Linked Data Cloud. The contributions of this paper are three-fold: Firstly, we review existing state-of-the-art NLP platforms and components, with a special focus on components that allow or support an automatic SPARQL construction. Secondly, we present the implemented architecture of the Linked Health Answers systems. Thirdly, we propose an statistical bootstrap approach for the identification and disambiguation of RDF-based predicates using a machine learning-based classifier. The evaluation focuses on predicate detection in sentence statements, as well as within the scenario of natural language questions.

A Tool Suite for Creating Question Answering Benchmarks

Axel-Cyrille Ngonga Ngomo, Norman Heino, René Speck and Prodromos Malakasiotis

We introduce the BIOASQ suite, a set of open-source Web tools for the creation, assessment and community-driven improvement of question answering benchmarks. The suite comprises three main tools: (1) the annotation tool supports the creation of benchmarks per se. In particular, this tool allows a team of experts to create questions and answers as well as to annotate the latter with documents, document snippets, RDF triples and ontology concepts. While the creation of questions is supported by different views and contextual information pertaining to the same question, the creation of answers is supported by the integration of several search engines and context information to facilitate the retrieval of the said answers as well as their annotation. (2) The assessment tool allows comparing several answers to the same question. Therewith, it can be used to assess the inter-annotator agreement as well as to manually evaluate automatically generated answers. (3) The third tool in the suite, the social network, aims to ensure the sustainability and iterative improvement of the benchmark by empowering communities of experts to provide insights on the questions in the benchmark. The BIOASQ suite has already been used successfully to create the 311 questions comprised in the BIOASQ question answering benchmark. It

has also been evaluated by the experts who used it to create the BIOASQ benchmark.

P39 - Speech Resources

Thursday, May 29, 14:55

Chairperson: **Henk van den Heuvel** Poster Session

The DIRHA simulated corpus

Luca Cristoforetti, Mirco Ravanelli, Maurizio Omologo, Alessandro Sosi, Alberto Abad, Martin Hagmueller and Petros Maragos

This paper describes a multi-microphone multi-language acoustic corpus being developed under the EC project Distant-speech Interaction for Robust Home Applications (DIRHA). The corpus is composed of several sequences obtained by convolution of dry acoustic events with more than 9000 impulse responses measured in a real apartment equipped with 40 microphones. The acoustic events include in-domain sentences of different typologies uttered by native speakers in four different languages and non-speech events representing typical domestic noises. To increase the realism of the resulting corpus, background noises were recorded in the real home environment and then added to the generated sequences. The purpose of this work is to describe the simulation procedure and the data sets that were created and used to derive the corpus. The corpus contains signals of different characteristics making it suitable for various multi-microphone signal processing and distant speech recognition tasks.

Euronews: a Multilingual Speech Corpus for ASR

Roberto Gretter

In this paper we present a multilingual speech corpus, designed for Automatic Speech Recognition (ASR) purposes. Data come from the portal Euronews and were acquired both from the Web and from TV. The corpus includes data in 10 languages (Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish) and was designed both to train AMs and to evaluate ASR performance. For each language, the corpus is composed of about 100 hours of speech for training (60 for Polish) and about 4 hours, manually transcribed, for testing. Training data include the audio, some reference text, the ASR output and their alignment. We plan to make public at least part of the benchmark in view of a multilingual ASR benchmark for IWSLT 2014.

Towards Multilingual Conversations in the Medical Domain: Development of Multilingual Medical Data and A Network-based ASR System

Sakriani Sakti, Keigo Kubo, Sho Matsumiya, Graham Neubig, Tomoki Toda, Satoshi Nakamura, Fumihiro Adachi and Ryosuke Isotani

This paper outlines the recent development on multilingual medical data and multilingual speech recognition system for network-based speech-to-speech translation in the medical domain. The overall speech-to-speech translation (S2ST) system was designed to translate spoken utterances from a given source language into a target language in order to facilitate multilingual conversations and reduce the problems caused by language barriers in medical situations. Our final system utilizes a weighted finite-state transducers with n-gram language models. Currently, the system successfully covers three languages: Japanese, English, and Chinese. The difficulties involved in connecting Japanese, English and Chinese speech recognition systems through Web servers will be discussed, and the experimental results in simulated medical conversation will also be presented.

The Slovene BNSI Broadcast News Database and Reference Speech Corpus GOS: Towards the Uniform Guidelines for Future Work

Andrej Zgank, Ana Zwitter Vitez and Darinka Verdonik

The aim of the paper is to search for common guidelines for the future development of speech databases for less resourced languages in order to make them the most useful for both main fields of their use, linguistic research and speech technologies. We compare two standards for creating speech databases, one followed when developing the Slovene speech database for automatic speech recognition – BNSI Broadcast News, the other followed when developing the Slovene reference speech corpus GOS, and outline possible common guidelines for future work. We also present an add-on for the GOS corpus, which enables its usage for automatic speech recognition.

Aix Map Task Corpus: the French Multimodal Corpus of Task-oriented Dialogue

Jan Gorisch, Corine Astésano, Ellen, Gurman Bard, Brigitte Bigi and Laurent Prévot

This paper introduces the Aix Map Task corpus, a corpus of audio and video recordings of task-oriented dialogues. It was modelled after the original HCRC Map Task corpus. Lexical material was designed for the analysis of speech and prosody, as described in Astésano et al. (2007). The design of the lexical material, the protocol and some basic quantitative features of the existing corpus are presented. The corpus was collected

under two communicative conditions, one audio-only condition and one face-to-face condition. The recordings took place in a studio and a sound attenuated booth respectively, with head-set microphones (and in the face-to-face condition with two video cameras). The recordings have been segmented into Inter-Pausal-Units and transcribed using transcription conventions containing actual productions and canonical forms of what was said. It is made publicly available online.

CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis

Carmen Garcia-Mateo, Antonio Cardenal, Xose Luis Regueira, Elisa Fernández Rei, Marta Martinez, Roberto Seara, Rocío Varela and Noemí Basanta

This paper describes the CORILGA ("Corpus Oral Informatizado da Lingua Galega"). CORILGA is a large high-quality corpus of spoken Galician from the 1960s up to present-day, including both formal and informal spoken language from both standard and non-standard varieties, and across different generations and social levels. The corpus will be available to the research community upon completion. Galician is one of the EU languages that needs further research before highly effective language technology solutions can be implemented. A software repository for speech resources in Galician is also described. The repository includes a structured database, a graphical interface and processing tools. The use of a database enables to perform search in a simple and fast way based in a number of different criteria. The web-based user interface facilitates users the access to the different materials. Last but not least a set of transcription-based modules for automatic speech recognition has been developed, thus facilitating the orthographic labelling of the recordings.

Basque Speecon-like and Basque SpeechDAT MDB-600: Speech Databases for the Development of ASR Technology for Basque

Igor Odriozola, Inma Hernaez, María Inés Torres, Luis Javier Rodriguez-Fuentes, Mikel Penagarikano and Eva Navas

This paper introduces two databases specifically designed for the development of ASR technology for the Basque language: the Basque Speecon-like database and the Basque SpeechDat MDB-600 database. The former was recorded in an office environment according to the Speecon specifications, whereas the later was recorded through mobile telephones according to the SpeechDat specifications. Both databases were created under an initiative that the Basque Government started in 2005, a program called ADITU, which aimed at developing speech technologies for Basque. The databases belong to the Basque Government. A comprehensive description of both databases is provided in this work, highlighting the differences with regard to their corresponding standard specifications. The paper also presents several initial experimental results for both databases with the purpose of validating their usefulness for the development of speech recognition technology. Several applications already developed with the Basque Speecon-like database are also described. Authors aim to make these databases widely known to the community as well, and foster their use by other groups.

New Bilingual Speech Databases for Audio Diarization

David Tavarez, Eva Navas, Daniel Erro, Ibon Saratxaga and Inma Hernaez

This paper describes the process of collecting and recording two new bilingual speech databases in Spanish and Basque. They are designed primarily for speaker diarization in two different application domains: broadcast news audio and recorded meetings. First, both databases have been manually segmented. Next, several diarization experiments have been carried out in order to evaluate them. Our baseline speaker diarization system has been applied to both databases with around 30% of DER for broadcast news audio and 40% of DER for recorded meetings. Also, the behavior of the system when different languages are used by the same speaker has been tested.

Erlangen-CLP: A Large Annotated Corpus of Speech from Children with Cleft Lip and Palate

Tobias Bocklet, Andreas Maier, Korbinian Riedhammer, Ulrich Eysholdt and Elmar Nöth

In this paper we describe Erlangen-CLP, a large speech database of children with Cleft Lip and Palate. More than 800 German children with CLP (most of them between 4 and 18 years old) and 380 age matched control speakers spoke the semi-standardized PLAKSS test that consists of words with all German phonemes in different positions. So far 250 CLP speakers were manually transcribed, 120 of these were analyzed by a speech therapist and 27 of them by four additional therapists. The tharapists marked 6 different processes/criteria like pharyngeal backing and hypernasality which typically occur in speech of people with CLP. We present detailed statistics about the the marked processes and the inter-rater agreement.

The Development of the Multilingual LUNA Corpus for Spoken Language System Porting

Evgeny Stepanov, Giuseppe Riccardi and Ali Orkan Bayer

The development of annotated corpora is a critical process in the development of speech applications for multiple target languages.

While the technology to develop a monolingual speech application has reached satisfactory results (in terms of performance and effort), porting an existing application from a source language to a target language is still a very expensive task. In this paper we address the problem of creating multilingual aligned corpora and its evaluation in the context of a spoken language understanding (SLU) porting task. We discuss the challenges of the manual creation of multilingual corpora, as well as present the algorithms for the creation of multilingual SLU via Statistical Machine Translation (SMT).

O29 - Sentiment Analysis (2)

Thursday, May 29, 16:55

Chairperson: **Frédérique Segond** Oral Session

CLIPS Stylometry Investigation (CSI) Corpus: a Dutch Corpus for the Detection of Age, Gender, Personality, Sentiment and Deception in Text

Ben Verhoeven and Walter Daelemans

We present the CLiPS Stylometry Investigation (CSI) corpus, a new Dutch corpus containing reviews and essays written by university students. It is designed to serve multiple purposes: detection of age, gender, authorship, personality, sentiment, deception, topic and genre. Another major advantage is its planned yearly expansion with each year's new students. The corpus currently contains about 305,000 tokens spread over 749 documents. The average review length is 128 tokens; the average essay length is 1126 tokens. The corpus will be made available on the CLiPS website (www.clips.uantwerpen.be/datasets) and can freely be used for academic research purposes. An initial deception detection experiment was performed on this data. Deception detection is the task of automatically classifying a text as being either truthful or deceptive, in our case by examining the writing style of the author. This task has never been investigated for Dutch before. We performed a supervised machine learning experiment using the SVM algorithm in a 10-fold cross-validation setup. The only features were the token unigrams present in the training data. Using this simple method, we reached a state-of-the-art F-score of 72.2%.

Building and Modelling Multilingual Subjective Corpora

Motaz Saad, David Langlois and Kamel Smaili

Building multilingual opinionated models requires multilingual corpora annotated with opinion labels. Unfortunately, such kind of corpora are rare. We consider opinions in this work as subjective or objective. In this paper, we introduce an annotation method that can be reliably transferred across topic domains and across languages. The method starts by building a classifier that annotates sentences into subjective/objective label using a training data from "movie reviews "domain which is in English language. The annotation can be transferred to another language by classifying English sentences in parallel corpora and transferring the same annotation to the same sentences of the other language. We also shed the light on the link between opinion mining and statistical language modelling, and how such corpora are useful for domain specific language modelling. We show the distinction between subjective and objective sentences which tends to be stable across domains and languages. Our experiments show that language models trained on objective (respectively subjective) corpus lead to better perplexities on objective (respectively subjective) test.

Author-Specific Sentiment Aggregation for Polarity Prediction of Reviews

Subhabrata Mukherjee and Sachindra Joshi

In this work, we propose an author-specific sentiment aggregation model for polarity prediction of reviews using an ontology. We propose an approach to construct a Phrase Annotated Author Specific Sentiment Ontology Tree (PASOT), where the facet nodes are annotated with opinion phrases of the author, used to describe the facets, as well as the author's preference for the facets. We show that an author-specific aggregation of sentiment over an ontology fares better than a flat classification model, which does not take the domain-specific facet importance or author-specific facet preference into account. We compare our approach to supervised classification using Support Vector Machines, as well as other baselines from previous works, where we achieve an accuracy improvement of 7.55% over the SVM baseline. Furthermore, we also show the effectiveness of our approach in capturing thwarting in reviews, achieving an accuracy improvement of 11.53% over the SVM baseline.

Meta-Classifiers Easily Improve Commercial Sentiment Detection Tools

Mark Cieliebak, Oliver Dürr and Fatih Uzdilli

In this paper, we analyze the quality of several commercial tools for sentiment detection. All tools are tested on nearly 30,000 short texts from various sources, such as tweets, news, reviews etc. The best commercial tools have average accuracy of 60%. We then apply machine learning techniques (Random Forests) to combine all tools, and show that this results in a meta-classifier that improves the overall performance significantly.

O30 - Multimodality

Thursday, May 29, 16:55

Chairperson: **Michael Kipp** Oral Session

AusTalk: an Audio-Visual Corpus of Australian English

Dominique Estival, Steve Cassidy, Felicity Cox and Denis Burnham

This paper describes the AusTalk corpus, which was designed and created through the Big ASC, a collaborative project with the two main goals of providing a standardised infrastructure for audio-visual recordings in Australia and of producing a large audio-visual corpus of Australian English, with 3 hours of AV recordings for 1000 speakers. We first present the overall project, then describe the corpus itself and its components, the strict data collection protocol with high levels of standardisation and automation, and the processes put in place for quality control. We also discuss the annotation phase of the project, along with its goals and challenges; a major contribution of the project has been to explore procedures for automating annotations and we present our solutions. We conclude with the current status of the corpus and with some examples of research already conducted with this new resource. AusTalk is one of the corpora included in the HCS vLab, which is briefly sketched in the conclusion.

From Synsets to Videos: Enriching ItalWordNet Multimodally

Roberto Bartolini, Valeria Quochi, Irene de Felice, Irene Russo and Monica Monachini

The paper describes the multimodal enrichment of ItalWordNet action verbs' entries by means of an automatic mapping with an ontology of action types instantiated by video scenes (ImagAct). The two resources present important differences as well as interesting complementary features, such that a mapping of these two resources can lead to a an enrichment of IWN, through the connection between synsets and videos apt to illustrate the meaning described by glosses. Here, we describe an approach inspired by ontology matching methods for the automatic mapping of ImagAct video scened onto ItalWordNet sense. The experiments described in the paper are conducted on Italian, but the same methodology can be extended to other languages for which WordNets have been created, since ImagAct is done also for English, Chinese and Spanish. This source of multimodal information can be exploited to design second language learning tools, as well as for language grounding in video action recognition and potentially for robotics.

A Multimodal Dataset for Deception Detection

Veronica Perez-Rosas, Rada Mihalcea, Alexis Narvaez and Mihai Burzo

This paper presents the construction of a multimodal dataset for deception detection, including physiological, thermal, and visual responses of human subjects under three deceptive scenarios. We present the experimental protocol, as well as the data acquisition process. To evaluate the usefulness of the dataset for the task of deception detection, we present a statistical analysis of the physiological and thermal modalities associated with the deceptive and truthful conditions. Initial results show that physiological and thermal responses can differentiate between deceptive and truthful states.

The Distress Analysis Interview Corpus of Human and Computer Interviews

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Albert "Skip" Rizzo and Louis-Philippe Morency

The Distress Analysis Interview Corpus (DAIC) contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post traumatic stress disorder. The interviews are conducted by humans, human controlled agents and autonomous agents, and the participants include both distressed and non-distressed individuals. Data collected include audio and video recordings and extensive questionnaire responses; parts of the corpus have been transcribed and annotated for a variety of verbal and non-verbal features. The corpus has been used to support the creation of an automated interviewer agent, and for research on the automatic identification of psychological distress.

O31 - Under-resourced Languages

Thursday, May 29, 16:55

Chairperson: **Andrejs Vasiljevs** Oral Session

Modern Chinese Helps Archaic Chinese Processing: Finding and Exploiting the Shared Properties

Yan Song and Fei Xia

Languages change over time and ancient languages have been studied in linguistics and other related fields. A main challenge in this research area is the lack of empirical data; for instance,

ancient spoken languages often leave little trace of their linguistic properties. From the perspective of natural language processing (NLP), while the NLP community has created dozens of annotated corpora, very few of them are on ancient languages. As an effort toward bridging the gap, we have created a word segmented and POS tagged corpus for Archaic Chinese using articles from Huainanzi, a book written during China's Western Han Dynasty (206 BC-9 AD). We then compare this corpus with the Chinese Penn Treebank (CTB), a well-known corpus for Modern Chinese, and report several interesting differences and similarities between the two corpora. Finally, we demonstrate that the CTB can be used to improve the performance of word segmenters and POS taggers for Archaic Chinese, but only through features that have similar behaviors in the two corpora.

Linguistic Landscaping of South Asia using Digital Language Resources: Genetic vs. Areal Linguistics

Lars Borin, Anju Saxena, Taraka Rama and Bernard Comrie

Like many other research fields, linguistics is entering the age of big data. We are now at a point where it is possible to see how new research questions can be formulated - and old research questions addressed from a new angle or established results verified - on the basis of exhaustive collections of data, rather than small, carefully selected samples. For example, South Asia is often mentioned in the literature as a classic example of a linguistic area, but there is no systematic, empirical study substantiating this claim. Examination of genealogical and areal relationships among South Asian languages requires a large-scale quantitative and qualitative comparative study, encompassing more than one language family. Further, such a study cannot be conducted manually, but needs to draw on extensive digitized language resources and state-of-the-art computational tools. We present some preliminary results of our large-scale investigation of the genealogical and areal relationships among the languages of this region, based on the linguistic descriptions available in the 19 tomes of Grierson's monumental "Linguistic Survey of India" (1903-1927), which is currently being digitized with the aim of turning the linguistic information in the LSI into a digital language resource suitable for a broad array of linguistic investigations.

PanLex: Building a Resource for Panlingual Lexical Translation

David Kamholz, Jonathan Pool and Susan Colowick

PanLex, a project of The Long Now Foundation, aims to enable the translation of lexemes among all human languages in the world. By focusing on lexemic translations, rather than grammatical or corpus data, it achieves broader lexical and language coverage than related projects. The PanLex database currently documents 20 million lexemes in about 9,000 language varieties, with 1.1 billion pairwise translations. The project primarily engages in content procurement, while encouraging outside use of its data for research and development. Its data acquisition strategy emphasizes broad, high-quality lexical and language coverage. The project plans to add data derived from 4,000 new sources to the database by the end of 2016. The dataset is publicly accessible via an HTTP API and monthly snapshots in CSV, JSON, and XML formats. Several online applications have been developed that query PanLex data. More broadly, the project aims to make a contribution to the preservation of global linguistic diversity.

Enriching ODIN

Fei Xia, William Lewis, Michael Wayne Goodman, Joshua Crowgey and Emily M. Bender

In this paper, we describe the expansion of the ODIN resource, a database containing many thousands of instances of Interlinear Glossed Text (IGT) for over a thousand languages harvested from scholarly linguistic papers posted to the Web. A database containing a large number of instances of IGT, which are effectively richly annotated and heuristically aligned bitexts, provides a unique resource for bootstrapping NLP tools for resource-poor languages. To make the data in ODIN more readily consumable by tool developers and NLP researchers, we propose a new XML format for IGT, called Xigt. We call the updated release ODIN-II.

O32 - Parallel Corpora

Thursday, May 29, 16:55

Chairperson: **Patrizia Paggio** Oral Session

Creating a Massively Parallel Bible Corpus

Thomas Mayer and Michael Cysouw

We present our ongoing effort to create a massively parallel Bible corpus. While an ever-increasing number of Bible translations is available in electronic form on the internet, there is no large-scale parallel Bible corpus that allows language researchers to easily get access to the texts and their parallel structure for a large variety of different languages. We report on the current status of the corpus, with over 900 translations in more than 830 language varieties. All translations are tokenized (e.g., separating punctuation marks) and Unicode normalized. Mainly due to copyright restrictions only portions of the texts are made publicly available. However, we provide co-occurrence information for each translation in a

(sparse) matrix format. All word forms in the translation are given together with their frequency and the verses in which they occur.

DCEP -Digital Corpus of the European Parliament

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger and Daniel Varga

We are presenting a new highly multilingual document-aligned parallel corpus called DCEP - Digital Corpus of the European Parliament. It consists of various document types covering a wide range of subject domains. With a total of 1.37 billion words in 23 languages (253 language pairs), gathered in the course of ten years, this is the largest single release of documents by a European Union institution. DCEP contains most of the content of the European Parliament's official Website. It includes different document types produced between 2001 and 2012, excluding only the documents already exist in the Europarl corpus to avoid overlapping. We are presenting the typical acquisition steps of the DCEP corpus: data access, document alignment, sentence splitting, normalisation and tokenisation, and sentence alignment efforts. The sentence-level alignment is still in progress but based on some first experiments; we showed that DCEP is very useful for NLP applications, in particular for Statistical Machine Translation.

Innovations in Parallel Corpus Search Tools

Martin Volk, Johannes Graën and Elena Callegaro

Recent years have seen an increased interest in and availability of parallel corpora. Large corpora from international organizations (e.g. European Union, United Nations, European Patent Office), or from multilingual Internet sites (e.g. OpenSubtitles) are now easily available and are used for statistical machine translation but also for online search by different user groups. This paper gives an overview of different usages and different types of search systems. In the past, parallel corpus search systems were based on sentence-aligned corpora. We argue that automatic word alignment allows for major innovations in searching parallel corpora. Some online query systems already employ word alignment for sorting translation variants, but none supports the full query functionality that has been developed for parallel treebanks. We propose to develop such a system for efficiently searching large parallel corpora with a powerful query language.

An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora

Valérie Hanoka and Benoît Sagot

This paper describes YaMTG (Yet another Multilingual Translation Graph), a new open-source heavily multilingual translation database (over 664 languages represented) built using several sources, namely various wiktionaries and the OPUS parallel corpora (Tiedemann, 2009). We detail the translation extraction process for 21 wiktionary language editions, and provide an evaluation of the translations contained in YaMTG.

P40 - Lexicons

Thursday, May 29, 16:55

Chairperson: **Yoshihiko Hayashi** Poster Session

Mapping the Lexique des Verbes du français (Lexicon of French Verbs) to a NLP Lexicon using Examples

Bruno Guillaume, Karën Fort, Guy Perrier and Paul Bédaride

This article presents experiments aiming at mapping the Lexique des Verbes du Français (Lexicon of French Verbs) to FRILEX, a Natural Language Processing (NLP) lexicon based on DICOVALENCE. The two resources (Lexicon of French Verbs and DICOVALENCE) were built by linguists, based on very different theories, which makes a direct mapping nearly impossible. We chose to use the examples provided in one of the resource to find implicit links between the two and make them explicit.

Text Readability and Word Distribution in Japanese

Satoshi Sato

This paper reports the relation between text readability and word distribution in the Japanese language. There was no similar study in the past due to three major obstacles: (1) unclear definition of Japanese "word", (2) no balanced corpus, and (3) no readability measure. Compilation of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and development of a readability predictor remove these three obstacles and enable this study. First, we have counted the frequency of each word in each text in the corpus. Then we have calculated the frequency rank of words both in the whole corpus and in each of three readability bands. Three major findings are: (1) the proportion of high-frequent words to tokens in Japanese is lower than that in English; (2) the type-coverage curve of words in the difficult-band draws an unexpected shape; (3) the size of the intersection between high-frequent words in the easy-band and these in the difficult-band is unexpectedly small.

High Quality Word Lists as a Resource for Multiple Purposes

Uwe Quasthoff, Dirk Goldhahn, Thomas Eckart, Erla Hallsteinsdóttir and Sabine Fiedler

Since 2011 the comprehensive, electronically available sources of the Leipzig Corpora Collection have been used consistently for the compilation of high quality word lists. The underlying corpora include newspaper texts, Wikipedia articles and other randomly collected Web texts. For many of the languages featured in this collection, it is the first comprehensive compilation to use a large-scale empirical base. The word lists have been used to compile dictionaries with comparable frequency data in the Frequency Dictionaries series. This includes frequency data of up to 1,000,000 word forms presented in alphabetical order. This article provides an introductory description of the data and the methodological approach used. In addition, language-specific statistical information is provided with regard to letters, word structure and structural changes. Such high quality word lists also provide the opportunity to explore comparative linguistic topics and such monolingual issues as studies of word formation and frequency-based examinations of lexical areas for use in dictionaries or language teaching. The results presented here can provide initial suggestions for subsequent work in several areas of research.

ISLEX – a Multilingual Web Dictionary

Þórdís Úlfarsdóttir

ISLEX is a multilingual Scandinavian dictionary, with Icelandic as a source language and Danish, Norwegian, Swedish, Faroese and Finnish as target languages. Within ISLEX are in fact contained several independent, bilingual dictionaries. While Faroese and Finnish are still under construction, the other languages were opened to the public on the web in November 2011. The use of the dictionary is free of charge and it has been extremely well received by its users. The result of the project is threefold. Firstly, some long awaited Icelandic-Scandinavian dictionaries have been published on the digital medium. Secondly, the project has been an important experience in Nordic language collaboration by jointly building such a work in six countries simultaneously, by academic institutions in Iceland, Denmark, Norway, Sweden, The Faroe Islands and Finland. Thirdly, the work has resulted in a compilation of structured linguistic data of the Nordic languages. This data is suitable for use in further lexicographic work and in various language technology projects.

Automatic Mapping Lexical Resources: A Lexical Unit as the Keystone

Eduard Bejček, Kettnerová Václava and Marketa Lopatkova

This paper presents the fully automatic linking of two valency lexicons of Czech verbs: VALLEX and PDT-VALLEX. Despite the same theoretical background adopted by these lexicons and the same linguistic phenomena they focus on, the fully automatic mapping of these resouces is not straightforward. We demonstrate that converting these lexicons into a common format represents a relatively easy part of the task whereas the automatic identification of pairs of corresponding valency frames (representing lexical units of verbs) poses difficulties. The overall achieved precision of 81% can be considered satisfactory. However, the higher number of lexical units a verb has, the lower the precision of their automatic mapping usually is. Moreover, we show that especially (i) supplementing further information on lexical units and (ii) revealing and reconciling regular discrepancies in their annotations can greatly assist in the automatic merging.

Towards Electronic SMS Dictionary Construction: An Alignment-based Approach

Cédric Lopez, Reda Bestandji, Mathieu Roche and Rachel Panckhurst

In this paper, we propose a method for aligning text messages (entitled AlignSMS) in order to automatically build an SMS dictionary. An extract of 100 text messages from the 88milSMS corpus (Panckhurst el al., 2013, 2014) was used as an initial test. More than 90,000 authentic text messages in French were collected from the general public by a group of academics in the south of France in the context of the sud4science project (http://www.sud4science.org). This project is itself part of a vast international SMS data collection project, entitled sms4science (http://www.sms4science.org, Fairon et al. 2006, Cougnon, 2014). After corpus collation, pre-processing and anonymisation (Accorsi et al., 2012, Patel et al., 2013), we discuss how "raw" anonymised text messages can be transcoded into normalised text messages, using a statistical alignment method. The future objective is to set up a hybrid (symbolic/statistic) approach based on both grammar rules and our statistical AlignSMS method.

Bilingual dictionaries for all EU languages

Ahmet Aker, Monica Paramita, Mārcis Pinnis and Robert Gaizauskas

Bilingual dictionaries can be automatically generated using the GIZA++ tool. However, these dictionaries contain a lot of noise, because of which the quality of outputs of tools relying on the dictionaries are negatively affected. In this work we present three different methods for cleaning noise from automatically generated bilingual dictionaries: LLR, pivot and translation based approach. We have applied these approaches on the GIZA++ dictionaries – dictionaries covering official EU languages – in order to remove noise. Our evaluation showed that all methods help to reduce noise. However, the best performance is achieved using the transliteration based approach. We provide all bilingual dictionaries (the original GIZA++ dictionaries and the cleaned ones) free for download. We also provide the cleaning tools and scripts for free download.

Automatic Acquisition of Urdu Nouns (along with Gender and Irregular Plurals))

Tafseer Ahmed Khan

The paper describes a set of methods to automatically acquire the Urdu nouns (and its gender) on the basis of inflectional and contextual clues. The algorithms used are a blend of computer's brute force on the corpus and careful design of distinguishing rules on the basis linguistic knowledge. As there are homograph inflections for Urdu nouns, adjectives and verbs, we compare potential inflectional forms with paradigms of inflections in strict order and gives best guess (of part of speech) for the word. We also worked on irregular plurals i.e. the plural forms that are borrowed from Arabic, Persian and English. Evaluation shows that not all the borrowed rules have same productivity in Urdu. The commonly used borrowed plural rules are shown in the result.

NomLex-PT: A Lexicon of Portuguese Nominalizations

Valeria de Paiva, Livy Real, Alexandre Rademaker and Gerard de Melo

This paper presents NomLex-PT, a lexical resource describing Portuguese nominalizations. NomLex-PT connects verbs to their nominalizations, thereby enabling NLP systems to observe the potential semantic relationships between the two words when analysing a text. NomLex-PT is freely available and encoded in RDF for easy integration with other resources. Most notably, we have integrated NomLex-PT with OpenWordNet-PT, an open Portuguese Wordnet.

P41 - Parsing

Thursday, May 29, 16:55

Chairperson: **Simonetta Montemagni** Poster Session

Sentence Rephrasing for Parsing Sentences with OOV Words

Hen-Hsen Huang, Huan-Yuan Chen, Chang-Sheng Yu, Hsin-Hsi Chen, Po-Ching Lee and Chun-Hsun Chen

This paper addresses the problems of out-of-vocabulary (OOV) words, named entities in particular, in dependency parsing. The OOV words, whose word forms are unknown to the learning-based parser, in a sentence may decrease the parsing performance. To deal with this problem, we propose a sentence rephrasing approach to replace each OOV word in a sentence with a popular word of the same named entity type in the training set, so that the knowledge of the word forms can be used for parsing. The highest-frequency-based rephrasing strategy and the information-retrieval-based rephrasing strategy are explored to select the word to replace, and the Chinese Treebank 6.0 (CTB6) corpus is adopted to evaluate the feasibility of the proposed sentence rephrasing strategies. Experimental results show that rephrasing some specific types of OOV words such as Corporation, Organization, and Competition increases the parsing performances. This methodology can be applied to domain adaptation to deal with OOV problems.

Pruning the Search Space of the Wolof LFG Grammar Using a Probabilistic and a Constraint Grammar Parser

Cheikh M. Bamba Dione

This paper presents a method for greatly reducing parse times in LFG by integrating a Constraint Grammar parser into a probabilistic context-free grammar. The CG parser is used in the pre-processing phase to reduce morphological and lexical ambiguity. Similarly, the c-structure pruning mechanism of XLE is used in the parsing phase to discard low-probability c-structures, before f-annotations are solved. The experiment results show a considerable increase in parsing efficiency and robustness in the annotation of Wolof running text. The Wolof CG parser indicated an f-score of 90% for morphological disambiguation and a speedup of ca. 40%, while the c-structure pruning method increased the speed of the Wolof grammar by over 36%. On a small amount of data, CG disambiguation and c-structure pruning

allowed for a speedup of 58%, however with a substantial drop in parse accuracy of 3.62.

How Could Veins Speed Up the Process of Discourse Parsing

Elena Mitocariu, Daniel Anechitei and Dan Cristea

In this paper we propose a method of reducing the search space of a discourse parsing process, while keeping unaffected its capacity to generate cohesive and coherent tree structures. The parsing method uses Veins Theory (VT), by developing incrementally a forest of parallel discourse trees, evaluating them on cohesion and coherence criteria and keeping only the most promising structures to go on with at each step. The incremental development is constrained by two general principles, well known in discourse parsing: sequentiality of the terminal nodes and attachment restricted to the right frontier. A set of formulas rooted on VT helps to guess the most promising nodes of the right frontier where an attachment can be made, thus avoiding an exhaustive generation of the whole search space and in the same time maximizing the coherence of the discourse structures. We report good results of applying this approach, representing a significant improvement in discourse parsing process.

Parsing Heterogeneous Corpora with a Rich Dependency Grammar

Achim Stein

Grammar models conceived for parsing purposes are often poorer than models that are motivated linguistically. We present a grammar model which is linguistically satisfactory and based on the principles of traditional dependency grammar. We show how a state-of-the-art dependency parser (mate tools) performs with this model, trained on the Syntactic Reference Corpus of Medieval French (SRCMF), a manually annotated corpus of medieval (Old French) texts. We focus on the problems caused by small and heterogeneous training sets typical for corpora of older periods. The result is the first publicly available dependency parser for Old French. On a 90/10 training/evaluation split of eleven OF texts (206000 words), we obtained an UAS of 89.68% and a LAS of 82.62%. Three experiments showed how heterogeneity, typical of medieval corpora, affects the parsing results: (a) a 'one-on-one' cross evaluation for individual texts, (b) a 'leave-one-out' cross evaluation, and (c) a prose/verse cross evaluation.

Treelet Probabilities for HPSG Parsing and Error Correction

Angelina Ivanova and Gertjan van Noord

Most state-of-the-art parsers take an approach to produce an analysis for any input despite errors. However, small grammatical mistakes in a sentence often cause parser to fail to build a correct syntactic tree. Applications that can identify and correct mistakes during parsing are particularly interesting for processing user-generated noisy content. Such systems potentially could take advantage of linguistic depth of broad-coverage precision grammars. In order to choose the best correction for an utterance, probabilities of parse trees of different sentences should be comparable which is not supported by discriminative methods underlying parsing software for processing deep grammars. In the present work we assess the treelet model for determining generative probabilities for HPSG parsing with error correction. In the first experiment the treelet model is applied to the parse selection task and shows superior exact match accuracy than the baseline and PCFG. In the second experiment it is tested for the ability to score the parse tree of the correct sentence higher than the constituency tree of the original version of the sentence containing grammatical error.

Self-training a Constituency Parser using n-gram Trees

Arda Celebi and Arzucan Özgür

In this study, we tackle the problem of self-training a feature-rich discriminative constituency parser. We approach the self-training problem with the assumption that while the full sentence parse tree produced by a parser may contain errors, some portions of it are more likely to be correct. We hypothesize that instead of feeding the parser the guessed full sentence parse trees of its own, we can break them down into smaller ones, namely n-gram trees, and perform self-training on them. We build an n-gram parser and transfer the distinct expertise of the n-gram parser to the full sentence parser by using the Hierarchical Joint Learning (HJL) approach. The resulting jointly self-trained parser obtains slight improvement over the baseline.

A Gold Standard Dependency Corpus for English

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer and Chris Manning

We present a gold standard annotation of syntactic dependencies in the English Web Treebank corpus using the Stanford Dependencies formalism. This resource addresses the lack of a gold standard dependency treebank for English, as well as the limited availability of gold standard syntactic annotations for English informal text genres. We also present experiments on the use of this resource, both for training dependency parsers and for evaluating the quality of different versions of the Stanford Parser, which includes a converter tool to produce dependency annotation from constituency trees. We show that training a dependency

parser on a mix of newswire and web data leads to better performance on that type of data without hurting performance on newswire text, and therefore gold standard annotations for non-canonical text can be a valuable resource for parsing. Furthermore, the systematic annotation effort has informed both the SD formalism and its implementation in the Stanford Parser's dependency converter. In response to the challenges encountered by annotators in the EWT corpus, the formalism has been revised and extended, and the converter has been improved.

Discosuite - A Parser Test Suite for German Discontinuous Structures

Wolfgang Maier, Miriam Kaeshammer, Peter Baumann and Sandra Kübler

Parser evaluation traditionally relies on evaluation metrics which deliver a single aggregate score over all sentences in the parser output, such as PARSEVAL. However, for the evaluation of parser performance concerning a particular phenomenon, a test suite of sentences is needed in which this phenomenon has been identified. In recent years, the parsing of discontinuous structures has received a rising interest. Therefore, in this paper, we present a test suite for testing the performance of dependency and constituency parsers on non-projective dependencies and discontinuous constituents for German. The test suite is based on the newly released TIGER treebank version 2.2. It provides a unique possibility of benchmarking parsers on non-local syntactic relationships in German, for constituents and dependencies. We include a linguistic analysis of the phenomena that cause discontinuity in the TIGER annotation, thereby closing gaps in previous literature. The linguistic phenomena we investigate include extraposition, a placeholder/repeated element construction, topicalization, scrambling, local movement, parentheticals, and fronting of pronouns.

P42 - Part-of-Speech Tagging

Thursday, May 29, 16:55

Chairperson: **Krister Linden** Poster Session

SWIFT Aligner, A Multifunctional Tool for Parallel Corpora: Visualization, Word Alignment, and (Morpho)-Syntactic Cross-Language Transfer

Timur Gilmanov, Olga Scrivner and Sandra Kübler

It is well known that word aligned parallel corpora are valuable linguistic resources. Since many factors affect automatic alignment quality, manual post-editing may be required in some applications. While there are several state-of-the-art word-aligners, such as GIZA++ and Berkeley, there is no simple visual tool that would enable correcting and editing aligned corpora of different formats. We have developed SWIFT Aligner, a free, portable software that allows for visual representation and editing of aligned corpora from several most commonly used formats: TALP, GIZA, and NAACL. In addition, our tool has incorporated part-of-speech and syntactic dependency transfer from an annotated source language into an unannotated target language, by means of word-alignment.

The CLE Urdu POS Tagset

Saba Urooj, Sarmad Hussain, Asad Mustafa, Rahila Parveen, Farah Adeeba, Tafseer Ahmed Khan, Miriam Butt and Annette Hautli

The paper presents a design schema and details of a new Urdu POS tagset. This tagset is designed due to challenges encountered in working with existing tagsets for Urdu. It uses tags that judiciously incorporate information about special morpho-syntactic categories found in Urdu. With respect to the overall naming schema and the basic divisions, the tagset draws on the Penn Treebank and a Common Tagset for Indian Languages. The resulting CLE Urdu POS Tagset consists of 12 major categories with subdivisions, resulting in 32 tags. The tagset has been used to tag 100k words of the CLE Urdu Digest Corpus, giving a tagging accuracy of 96.8%.

Using Stem-Templates to Improve Arabic POS and Gender/Number Tagging

Kareem Darwish, Ahmed Abdelali and Hamdy Mubarak

This paper presents an end-to-end automatic processing system for Arabic. The system performs: correction of common spelling errors pertaining to different forms of alef, ta marbouta and ha, and alef maqsoura and ya; context sensitive word segmentation into underlying clitics, POS tagging, and gender and number tagging of nouns and adjectives. We introduce the use of stem templates as a feature to improve POS tagging by 0.5% and to help ascertain the gender and number of nouns and adjectives. For gender and number tagging, we report accuracies that are significantly higher on previously unseen words compared to a state-of-the-art system.

The LIMA Multilingual Analyzer Made Free: FLOSS Resources Adaptation and Correction

Gaël de Chalendar

At CEA LIST, we have decided to release our multilingual analyzer LIMA as Free software. As we were not proprietary of all the language resources used we had to select and adapt free ones in order to attain results good enough and equivalent to those obtained with our previous ones. For English and French, we

found and adapted a full-form dictionary and an annotated corpus for learning part-of-speech tagging models.

A Tagged Corpus and a Tagger for Urdu

Bushra Jawaid, Amir Kamran and Ondrej Bojar

In this paper, we describe a release of a sizeable monolingual Urdu corpus automatically tagged with part-of-speech tags. We extend the work of Jawaid and Bojar (2012) who use three different taggers and then apply a voting scheme to disambiguate among the different choices suggested by each tagger. We run this complex ensemble on a large monolingual corpus and release the tagged corpus. Additionally, we use this data to train a single standalone tagger which will hopefully significantly simplify Urdu processing. The standalone tagger obtains the accuracy of 88.74% on test data.

Correcting Errors in a New Gold Standard for Tagging Icelandic Text

Sigrún Helgadóttir, Hrafn Loftsson and Eiríkur Rögnvaldsson

In this paper, we describe the correction of PoS tags in a new Icelandic corpus, MIM-GOLD, consisting of about 1 million tokens sampled from the Tagged Icelandic Corpus, MÍM, released in 2013. The goal is to use the corpus, among other things, as a new gold standard for training and testing PoS taggers. The construction of the corpus was first described in 2010 together with preliminary work on error detection and correction. In this paper, we describe further the correction of tags in the corpus. We describe manual correction and a method for semi-automatic error detection and correction. We show that, even after manual correction, the number of tagging errors in the corpus can be reduced significantly by applying our semi-automatic detection and correction method. After the semi-automatic error correction, preliminary evaluation of tagging accuracy shows very low error rates. We hope that the existence of the corpus will make it possible to improve PoS taggers for Icelandic text.

PoliTa: a Multitagger for Polish

Łukasz Kobyliński

Part-of-Speech (POS) tagging is a crucial task in Natural Language Processing (NLP). POS tags may be assigned to tokens in text manually, by trained linguists, or using algorithmic approaches. Particularly, in the case of annotated text corpora, the quantity of textual data makes it unfeasible to rely on manual tagging and automated methods are used extensively. The quality of such methods is of critical importance, as even 1% tagger error rate results in introducing millions of errors in a corpus consisting of a billion tokens. In case of Polish several POS taggers have been proposed to date, but even the best of the taggers achieves an accuracy of ca. 93%, as measured on the one million subcorpus of the National Corpus of Polish (NCP). As the task of tagging is an example of classification, in this article we introduce a new POS tagger for Polish, which is based on the idea of combining several classifiers to produce higher quality tagging results than using any of the taggers individually.

P43 - Semantics

Thursday, May 29, 16:55

Chairperson: **Marc Verhagen** Poster Session

Polysemy Index for Nouns: an Experiment on Italian using the PAROLE SIMPLE CLIPS Lexical Database

Francesca Frontini, Valeria Quochi, Sebastian Padó, Monica Monachini and Jason Utt

An experiment is presented to induce a set of polysemous basic type alternations (such as Animal-Food, or Building-Institution) by deriving them from the sense alternations found in an existing lexical resource. The paper builds on previous work and applies those results to the Italian lexicon PAROLE SIMPLE CLIPS. The new results show how the set of frequent type alternations that can be induced from the lexicon is partly different from the set of polysemy relations selected and explicitely applied by lexicographers when building it. The analysis of mismatches shows that frequent type alternations do not always correpond to prototypical polysemy relations, nevertheless the proposed methodology represents a useful tool offered to lexicographers to systematically check for possible gaps in their resource.

Comparing Similarity Measures for Distributional Thesauri

Muntsa Padró, Marco Idiart, Aline Villavicencio and Carlos Ramisch

Distributional thesauri have been applied for a variety of tasks involving semantic relatedness. In this paper, we investigate the impact of three parameters: similarity measures, frequency thresholds and association scores. We focus on the robustness and stability of the resulting thesauri, measuring inter-thesaurus agreement when testing different parameter values. The results obtained show that low-frequency thresholds affect thesaurus quality more than similarity measures, with more agreement found for increasing thresholds.These results indicate the sensitivity of distributional thesauri to frequency. Nonetheless, the observed differences do not transpose over extrinsic evaluation using

TOEFL-like questions. While this may be specific to the task, we argue that a careful examination of the stability of distributional resources prior to application is needed.

Reconstructing the Semantic Landscape of Natural Language Processing

Elisa Omodei, Jean-Philippe Cointet and Thierry Poibeau

This paper investigates the evolution of the computational linguistics domain through a quantitative analysis of the ACL Anthology (containing around 12,000 papers published between 1985 and 2008). Our approach combines complex system methods with natural language processing techniques. We reconstruct the socio-semantic landscape of the domain by inferring a co-authorship and a semantic network from the analysis of the corpus. First, keywords are extracted using a hybrid approach mixing linguistic patterns with statistical information. Then, the semantic network is built using a co-occurrence analysis of these keywords within the corpus. Combining temporal and network analysis techniques, we are able to examine the main evolutions of the field and the more active subfields over time. Lastly we propose a model to explore the mutual influence of the social and the semantic network over time, leading to a socio-semantic co-evolutionary system.

Compounds and Distributional Thesauri

Olivier Ferret

The building of distributional thesauri from corpora is a problem that was the focus of a significant number of articles, starting with (Grefenstette, 1994) and followed by (Lin, 1998), (Curran and Moens, 2002) or (Heylen and Peirsman, 2007). However, in all these cases, only single terms were considered. More recently, the topic of compositionality in the framework of distributional semantic representations has come to the surface and was investigated for building the semantic representation of phrases or even sentences from the representation of their words. However, this work was not done until now with the objective of building distributional thesauri. In this article, we investigate the impact of the introduction of compounds for achieving such building. More precisely, we consider compounds as undividable lexical units and evaluate their influence according to three different roles: as features in the distributional contexts of single terms, as possible neighbors of single term entries and finally, as entries of a thesaurus. This investigation was conducted through an intrinsic evaluation for a large set of nominal English single terms and compounds with various frequencies.

UnixMan Corpus: A Resource for Language Learning in the Unix Domain

Kyle Richardson and Jonas Kuhn

We present a new resource, the UnixMan Corpus, for studying language learning it the domain of Unix utility manuals. The corpus is built by mining Unix (and other Unix related) man pages for parallel example entries, consisting of English textual descriptions with corresponding command examples. The commands provide a grounded and ambiguous semantics for the textual descriptions, making the corpus of interest to work on Semantic Parsing and Grounded Language Learning. In contrast to standard resources for Semantic Parsing, which tend to be restricted to a small number of concepts and relations, the UnixMan Corpus spans a wide variety of utility genres and topics, and consists of hundreds of command and domain entity types. The semi-structured nature of the manuals also makes it easy to exploit other types of relevant information for Grounded Language Learning. We describe the details of the corpus and provide preliminary classification results.

Multilingual eXtended WordNet Knowledge Base: Semantic Parsing and Translation of Glosses

Tatiana Erekhinskaya, Meghana Satpute and Dan Moldovan

This paper presents a method to create WordNet-like lexical resources for different languages. Instead of directly translating glosses from one language to another, we perform first semantic parsing of WordNet glosses and then translate the resulting semantic representation. The proposed approach simplifies the machine translation of the glosses. The approach provides ready to use semantic representation of glosses in target languages instead of just plain text.

Relation Inference in Lexical Networks ... with Refinements

Manel Zarrouk and Mathieu Lafourcade

Improving lexical network's quality is an important issue in the creation process of these language resources. This can be done by automatically inferring new relations from already existing ones with the purpose of (1) densifying the relations to cover the eventual lack of information and (2) detecting errors. In this paper, we devise such an approach applied to the JeuxDeMots lexical network, which is a freely available lexical and semantic resource for French. We first present the principles behind the lexical network construction with crowdsourcing and games with a purpose and illustrated them with JeuxDeMots (JDM). Then, we present the outline of an elicitation engine based on an inference

engine using schemes like deduction, induction and abduction which will be referenced and briefly presented and we will especially highlight the new scheme (Relation Inference Scheme with Refinements) added to our system. An experiment showing the relevance of this scheme is then presented.

Extracting Semantic Relations from Portuguese Corpora using Lexical-Syntactic Patterns

Raquel Amaro

The growing investment on automatic extraction procedures, together with the need for extensive resources, makes semi-automatic construction a new viable and efficient strategy for developing of language resources, combining accuracy, size, coverage and applicability. These assumptions motivated the work depicted in this paper, aiming at the establishment and use of lexical-syntactic patterns for extracting semantic relations for Portuguese from corpora, part of a larger ongoing project for the semi-automatic extension of WordNet.PT. 26 lexical-syntactic patterns were established, covering hypernymy/hyponymy and holonymy/meronymy relations between nominal items, and over 34 000 contexts were manually analyzed to evaluate the productivity of each pattern. The set of patterns and respective examples are given, as well as data concerning the extraction of relations - right hits, wrong hits and related hits-, and the total of occurrences of each pattern in CPRC. Although language-dependent, and thus clearly of obvious interest for the development of lexical resources for Portuguese, the results depicted in this paper are also expected to be helpful as a basis for the establishment of patterns for related languages such as Spanish, Catalan, French or Italian.

An Analysis of Ambiguity in Word Sense Annotations

David Jurgens

Word sense annotation is a challenging task where annotators distinguish which meaning of a word is present in a given context. In some contexts, a word usage may elicit multiple interpretations, resulting either in annotators disagreeing or in allowing the usage to be annotated with multiple senses. While some works have allowed the latter, the extent to which multiple sense annotations are needed has not been assessed. The present work analyzes a dataset of instances annotated with multiple WordNet senses to assess the causes of the multiple interpretations and their relative frequencies, along with the effect of the multiple senses on the contextual interpretation. We show that contextual underspecification is the primary cause of multiple interpretations but that syllepsis still accounts for more than a third of the cases. In addition, we show that sense coarsening can only partially remove the need for labeling instances with multiple senses and we provide suggestions for how future sense annotation guidelines might be developed to account for this need.

PropBank: Semantics of New Predicate Types

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang and Martha Palmer

This research focuses on expanding PropBank, a corpus annotated with predicate argument structures, with new predicate types; namely, noun, adjective and complex predicates, such as Light Verb Constructions. This effort is in part inspired by a sister project to PropBank, the Abstract Meaning Representation project, which also attempts to capture "who is doing what to whom" in a sentence, but does so in a way that abstracts away from syntactic structures. For example, alternate realizations of a 'destroying' event in the form of either the verb 'destroy' or the noun 'destruction' would receive the same Abstract Meaning Representation. In order for PropBank to reach the same level of coverage and continue to serve as the bedrock for Abstract Meaning Representation, predicate types other than verbs, which have previously gone without annotation, must be annotated. This research describes the challenges therein, including the development of new annotation practices that walk the line between abstracting away from language-particular syntactic facts to explore deeper semantics, and maintaining the connection between semantics and syntactic structures that has proven to be very valuable for PropBank as a corpus of training data for Natural Language Processing applications.

Semi-Supervised Methods for Expanding Psycholinguistics Norms by Integrating Distributional Similarity with the Structure of WordNet

Michael Mohler, Marc Tomlinson, David Bracewell and Bryan Rink

In this work, we present two complementary methods for the expansion of psycholinguistics norms. The first method is a random-traversal spreading activation approach which transfers existing norms onto semantically related terms using notions of synonymy, hypernymy, and pertainymy to approach full coverage of the English language. The second method makes use of recent advances in distributional similarity representation to transfer existing norms to their closest neighbors in a high-dimensional vector space. These two methods (along with a naive hybrid approach combining the two) have been shown to significantly outperform a state-of-the-art resource expansion system at our pilot task of imageability expansion. We have evaluated these systems in a cross-validation experiment using 8,188 norms found

in existing pscholinguistics literature. We have also validated the quality of these combined norms by performing a small study using Amazon Mechanical Turk (AMT).

A Graph-based Approach for Computing Free Word Associations

Gemma Bel Enguix, Reinhard Rapp and Michael Zock

A graph-based algorithm is used to analyze the co-occurrences of words in the British National Corpus. It is shown that the statistical regularities detected can be exploited to predict human word associations. The corpus-derived associations are evaluated using a large test set comprising several thousand stimulus/response pairs as collected from humans. The finding is that there is a high agreement between the two types of data. The considerable size of the test set allows us to split the stimulus words into a number of classes relating to particular word properties. For example, we construct six saliency classes, and for the words in each of these classes we compare the simulation results with the human data. It turns out that for each class there is a close relationship between the performance of our system and human performance. This is also the case for classes based on two other properties of words, namely syntactic and semantic word ambiguity. We interpret these findings as evidence for the claim that human association acquisition must be based on the statistical analysis of perceived language and that when producing associations the detected statistical regularities are replicated.

A Hierarchical Taxonomy for Classifying Hardness of Inference Tasks

Martin Gleize and Brigitte Grau

Exhibiting inferential capabilities is one of the major goals of many modern Natural Language Processing systems. However, if attempts have been made to define what textual inferences are, few seek to classify inference phenomena by difficulty. In this paper we propose a hierarchical taxonomy for inferences, relatively to their hardness, and with corpus annotation and system design and evaluation in mind. Indeed, a fine-grained assessment of the difficulty of a task allows us to design more appropriate systems and to evaluate them only on what they are designed to handle. Each of seven classes is described and provided with examples from different tasks like question answering, textual entailment and coreference resolution. We then test the classes of our hierarchy on the specific task of question answering. Our annotation process of the testing data at the QA4MRE 2013 evaluation campaign reveals that it is possible to quantify the contrasts in types of difficulty on datasets of the same task.

P44 - Speech Recognition and Synthesis

Thursday, May 29, 16:55

Chairperson: **Denise DiPersio** Poster Session

Speech Recognition Web Services for Dutch

Joris Pelemans, Kris Demuynck, Hugo van Hamme and Patrick Wambacq

In this paper we present 3 applications in the domain of Automatic Speech Recognition for Dutch, all of which are developed using our in-house speech recognition toolkit SPRAAK. The speech-to-text transcriber is a large vocabulary continuous speech recognizer, optimized for Southern Dutch. It is capable to select components and adjust parameters on the fly, based on the observed conditions in the audio and was recently extended with the capability of adding new words to the lexicon. The grapheme-to-phoneme converter generates possible pronunciations for Dutch words, based on lexicon lookup and linguistic rules. The speech-text alignment system takes audio and text as input and constructs a time aligned output where every word receives exact begin and end times. All three of the applications (and others) are freely available, after registration, as a web application on http://www.spraak.org/webservice/ and in addition, can be accessed as a web service in automated tools.

Morpho-Syntactic Study of Errors from Speech Recognition System

Maria Goryainova, Cyril Grouin, Sophie Rosset and Ioana Vasilescu

The study provides an original standpoint of the speech transcription errors by focusing on the morpho-syntactic features of the erroneous chunks and of the surrounding left and right context. The typology concerns the forms, the lemmas and the POS involved in erroneous chunks, and in the surrounding contexts. Comparison with error free contexts are also provided. The study is conducted on French. Morpho-syntactic analysis underlines that three main classes are particularly represented in the erroneous chunks: (i) grammatical words (to, of, the), (ii) auxiliary verbs (has, is), and (iii) modal verbs (should, must). Such items are widely encountered in the ASR outputs as frequent candidates to transcription errors. The analysis of the context points out that some left 3-grams contexts (e.g., repetitions, that is disfluencies, bracketing formulas such as "c'est", etc.) may be better predictors than others. Finally, the surface analysis conducted through a Levensthein distance analysis, highlighted

that the most common distance is of 2 characters and mainly involves differences between inflected forms of a unique item.

Human Annotation of ASR Error Regions: is "gravity" a Sharable Concept for Human Annotators?

Daniel Luzzati, Cyril Grouin, Ioana Vasilescu, Martine Adda-Decker, Eric Bilinski, Nathalie Camelin, Juliette Kahn, Carole Lailler, Lori Lamel and Sophie Rosset

This paper is concerned with human assessments of the severity of errors in ASR outputs. We did not design any guidelines so that each annotator involved in the study could consider the "seriousness" of an ASR error using their own scientific background. Eight human annotators were involved in an annotation task on three distinct corpora, one of the corpora being annotated twice, hiding this annotation in duplicate to the annotators. None of the computed results (inter-annotator agreement, edit distance, majority annotation) allow any strong correlation between the considered criteria and the level of seriousness to be shown, which underlines the difficulty for a human to determine whether a ASR error is serious or not.

Development of a TV Broadcasts Speech Recognition System for Qatari Arabic

Mohamed Elmahdy, Mark Hasegawa-Johnson and Eiman Mustafawi

A major problem with dialectal Arabic speech recognition is due to the sparsity of speech resources. In this paper, a transfer learning framework is proposed to jointly use a large amount of Modern Standard Arabic (MSA) data and little amount of dialectal Arabic data to improve acoustic and language modeling. The Qatari Arabic (QA) dialect has been chosen as a typical example for an under-resourced Arabic dialect. A wide-band speech corpus has been collected and transcribed from several Qatari TV series and talk-show programs. A large vocabulary speech recognition baseline system was built using the QA corpus. The proposed MSA-based transfer learning technique was performed by applying orthographic normalization, phone mapping, data pooling, acoustic model adaptation, and system combination. The proposed approach can achieve more than 28% relative reduction in WER.

Automatic Long Audio Alignment and Confidence Scoring for Conversational Arabic Speech

Mohamed Elmahdy, Mark Hasegawa-Johnson and Eiman Mustafawi

In this paper, a framework for long audio alignment for conversational Arabic speech is proposed. Accurate alignments help in many speech processing tasks such as audio indexing, speech recognizer acoustic model (AM) training, audio summarizing and retrieving, etc. We have collected more than 1,400 hours of conversational Arabic besides the corresponding human generated non-aligned transcriptions. Automatic audio segmentation is performed using a split and merge approach. A biased language model (LM) is trained using the corresponding text after a pre-processing stage. Because of the dominance of non-standard Arabic in conversational speech, a graphemic pronunciation model (PM) is utilized. The proposed alignment approach is performed in two passes. Firstly, a generic standard Arabic AM is used along with the biased LM and the graphemic PM in a fast speech recognition pass. In a second pass, a more restricted LM is generated for each audio segment, and unsupervised acoustic model adaptation is applied. The recognizer output is aligned with the processed transcriptions using Levenshtein algorithm. The proposed approach resulted in an initial alignment accuracy of 97.8-99.0% depending on the amount of disfluencies. A confidence scoring metric is proposed to accept/reject aligner output. Using confidence scores, it was possible to reject the majority of mis-aligned segments resulting in alignment accuracy of 99.0-99.8% depending on the speech domain and the amount of disfluencies.

The WaveSurfer Automatic Speech Recognition Plugin

Giampiero Salvi and Niklas Vanhainen

This paper presents a plugin that adds automatic speech recognition (ASR) functionality to the WaveSurfer sound manipulation and visualisation program. The plugin allows the user to run continuous speech recognition on spoken utterances, or to align an already available orthographic transcription to the spoken material. The plugin is distributed as free software and is based on free resources, namely the Julius speech recognition engine and a number of freely available ASR resources for different languages. Among these are the acoustic and language models we have created for Swedish using the NST database.

A Toolkit for Efficient Learning of Lexical Units for Speech Recognition

Matti Varjokallio and mikko kurimo

String segmentation is an important and recurring problem in natural language processing and other domains. For morphologically rich languages, the amount of different word forms caused by morphological processes like agglutination, compounding and inflection, may be huge and causes problems for traditional word-based language modeling approach. Segmenting text into better modelable units is thus an important

part of the modeling task. This work presents methods and a toolkit for learning segmentation models from text. The methods may be applied to lexical unit selection for speech recognition and also other segmentation tasks.

Using Audio Books for Training a Text-to-Speech System

Aimilios Chalamandaris, Pirros Tsiakoulis, Sotiris Karabetsos and Spyros Raptis

Creating new voices for a TTS system often requires a costly procedure of designing and recording an audio corpus, a time consuming and effort intensive task. Using publicly available audiobooks as the raw material of a spoken corpus for such systems creates new perspectives regarding the possibility of creating new synthetic voices quickly and with limited effort. This paper addresses the issue of creating new synthetic voices based on audiobook data in an automated method. As an audiobook includes several types of speech, such as narration, character playing etc., special care is given in identifying the data subset that leads to a more neutral and general purpose synthetic voice. The main goal is to identify and address the effect the audiobook speech diversity on the resulting TTS system. Along with the methodology for coping with this diversity in the speech data, we also describe a set of experiments performed in order to investigate the efficiency of different approaches for automatic data pruning. Further plans for exploiting the diversity of the speech incorporated in an audiobook are also described in the final section and conclusions are drawn.

O33 - Linked Data and Semantic Web

Thursday, May 29, 18:20

Chairperson: **Guadalupe Aguado-de-Cea** Oral Session

TMO – The Federated Ontology of the TrendMiner Project

Hans-Ulrich Krieger and Thierry Declerck

This paper describes work carried out in the European project TrendMiner which partly deals with the extraction and representation of real time information from dynamic data streams. The focus of this paper lies on the construction of an integrated ontology, TMO, the TrendMiner Ontology, that has been assembled from several independent multilingual taxonomies and ontologies which are brought together by an interface specification, expressed in OWL. Within TrendMiner, TMO serves as a common language that helps to interlink data, delivered from both symbolic and statistical components of the TrendMiner system. Very often, the extracted data is supplied as quintuples, RDF triples that are extended by two further temporal arguments, expressing the temporal extent in which an atemporal statement is true. In this paper, we will also sneak a peek on the temporal entailment rules and queries that are built into the semantic repository hosting the data and which can be used to derive useful new information.

A Collection of Scholarly Book Reviews from the Platforms of Electronic Sources in Humanities and Social Sciences OpenEdition.org

Chahinez Benkoussas, Hussam Hamdan, Patrice Bellot, Frédéric Béchet and Elodie Faath

In this paper, we present our contribution for the automatic construction of the Scholarly Book Reviews corpora from two different sources, the OpenEdition platform which is dedicated to electronic resources in the humanities and social sciences, and the Web. The main target is the collect of reviews in order to provide automatic links between each review and its potential book in the future. For these purposes, we propose different document representations and we apply some supervised approaches for binary genre classification before evaluating their impact.

A Meta-data Driven Platform for Semi-automatic Configuration of Ontology Mediators

Manuel Fiorelli, Maria Teresa Pazienza and Armando Stellato

Ontology mediators often demand extensive configuration, or even the adaptation of the input ontologies for remedying unsupported modeling patterns. In this paper we propose MAPLE (MAPping Architecture based on Linguistic Evidences), an architecture and software platform that semi-automatically solves this configuration problem, by reasoning on metadata about the linguistic expressivity of the input ontologies, the available mediators and other components relevant to the mediation task. In our methodology mediators should access the input ontologies through uniform interfaces abstracting many low-level details, while depending on generic third-party linguistic resources providing external information. Given a pair of ontologies to reconcile, MAPLE ranks the available mediators according to their ability to exploit most of the input ontologies content, while coping with the exhibited degree of linguistic heterogeneity. MAPLE provides the chosen mediator with concrete linguistic resources and suitable implementations of the required interfaces. The resulting mediators are more robust, as they are isolated from many low-level issues, and their applicability and performance may increase over time as new and better resources and other

components are made available. To sustain this trend, we foresee the use of the Web as a large scale repository.

O34 - Dialogue (1)

Thursday, May 29, 18:20

Chairperson: **Francesco Cutugno** Oral Session

Twente Debate Corpus – A Multimodal Corpus for Head Movement Analysis

Bayu Rahayudi, Ronald Poppe and Dirk Heylen

This paper introduces a multimodal discussion corpus for the study into head movement and turn-taking patterns in debates. Given that participants either acted alone or in a pair, cooperation and competition and their nonverbal correlates can be analyzed. In addition to the video and audio of the recordings, the corpus contains automatically estimated head movements, and manual annotations of who is speaking and who is looking where. The corpus consists of over 2 hours of debates, in 6 groups with 18 participants in total. We describe the recording setup and present initial analyses of the recorded data. We found that the person who acted as single debater speaks more and also receives more attention compared to the other debaters, also when corrected for the time speaking. We also found that a single debater was more likely to speak after a team debater. Future work will be aimed at further analysis of the relation between speaking and looking patterns, the outcome of the debate and perceived dominance of the debaters.

A Multimodal Corpus of Rapid Dialogue Games

Maike Paetzel, David Nicolas Racca and David DeVault

This paper presents a multimodal corpus of spoken human-human dialogues collected as participants played a series of Rapid Dialogue Games (RDGs). The corpus consists of a collection of about 11 hours of spoken audio, video, and Microsoft Kinect data taken from 384 game interactions (dialogues). The games used for collecting the corpus required participants to give verbal descriptions of linguistic expressions or visual images and were specifically designed to engage players in a fast-paced conversation under time pressure. As a result, the corpus contains many examples of participants attempting to communicate quickly in specific game situations, and it also includes a variety of spontaneous conversational phenomena such as hesitations, filled pauses, overlapping speech, and low-latency responses. The corpus has been created to facilitate research in incremental speech processing for spoken dialogue systems. Potentially, the corpus could be used in several areas of speech and language research, including speech recognition, natural language understanding, natural language generation, and dialogue management.

The Tutorbot Corpus – A Corpus for Studying Tutoring Behaviour in Multiparty Face-to-Face Spoken Dialogue

Maria Koutsombogera, Samer Al Moubayed, Bajibabu Bollepalli, Ahmed Hussen Abdelaziz, Martin Johansson, José David Aguas Lopes, Jekaterina Novikova, Catharine Oertel, Kalin Stefanov and Gül Varol

This paper describes a novel experimental setup exploiting state-of-the-art capture equipment to collect a multimodally rich game-solving collaborative multiparty dialogue corpus. The corpus is targeted and designed towards the development of a dialogue system platform to explore verbal and nonverbal tutoring strategies in multiparty spoken interactions. The dialogue task is centered on two participants involved in a dialogue aiming to solve a card-ordering game. The participants were paired into teams based on their degree of extraversion as resulted from a personality test. With the participants sits a tutor that helps them perform the task, organizes and balances their interaction and whose behavior was assessed by the participants after each interaction. Different multimodal signals captured and auto-synchronized by different audio-visual capture technologies, together with manual annotations of the tutor's behavior constitute the Tutorbot corpus. This corpus is exploited to build a situated model of the interaction based on the participants' temporally-changing state of attention, their conversational engagement and verbal dominance, and their correlation with the verbal and visual feedback and conversation regulatory actions generated by the tutor.

O35 - Word Sense Annotation and Disambiguation

Thursday, May 29, 18:20

Chairperson: **Patrik Lambert** Oral Session

Exploiting Portuguese Lexical Knowledge Bases for Answering Open Domain Cloze Questions Automatically

Hugo Gonçalo Oliveira, Inês Coelho and Paulo Gomes

We present the task of answering cloze questions automatically and how it can be tackled by exploiting lexical knowledge bases (LKBs). This task was performed in what can be seen as an indirect evaluation of Portuguese LKB. We introduce the LKBs used and the algorithms applied, and then report on the obtained results and draw some conclusions: LKBs are definitely useful resources for this challenging task, and exploiting them, especially with PageRanking-based algorithms, clearly improves

the baselines. Moreover, larger LKB, created automatically and not sense-aware led to the best results, as opposed to handcrafted LKB structured on synsets.

Single Classifier Approach for Verb Sense Disambiguation based on Generalized Features

Daisuke Kawahara and Martha Palmer

We present a supervised method for verb sense disambiguation based on VerbNet. Most previous supervised approaches to verb sense disambiguation create a classifier for each verb that reaches a frequency threshold. These methods, however, have a significant practical problem that they cannot be applied to rare or unseen verbs. In order to overcome this problem, we create a single classifier to be applied to rare or unseen verbs in a new text. This single classifier also exploits generalized semantic features of a verb and its modifiers in order to better deal with rare or unseen verbs. Our experimental results show that the proposed method achieves equivalent performance to per-verb classifiers, which cannot be applied to unseen verbs. Our classifier could be utilized to improve the classifications in lexical resources of verbs, such as VerbNet, in a semi-automatic manner and to possibly extend the coverage of these resources to new verbs.

Annotating the MASC Corpus with BabelNet

Andrea Moro, Roberto Navigli, Francesco Maria Tucci and Rebecca J. Passonneau

In this paper we tackle the problem of automatically annotating, with both word senses and named entities, the MASC 3.0 corpus, a large English corpus covering a wide range of genres of written and spoken text. We use BabelNet 2.0, a multilingual semantic network which integrates both lexicographic and encyclopedic knowledge, as our sense/entity inventory together with its semantic structure, to perform the aforementioned annotation task. Word sense annotated corpora have been around for more than twenty years, helping the development of Word Sense Disambiguation algorithms by providing both training and testing grounds. More recently Entity Linking has followed the same path, with the creation of huge resources containing annotated named entities. However, to date, there has been no resource that contains both kinds of annotation. In this paper we present an automatic approach for performing this annotation, together with its output on the MASC corpus. We use this corpus because its goal of integrating different types of annotations goes exactly in our same direction. Our overall aim is to stimulate research on the joint exploitation and disambiguation of word senses and named entities. Finally, we estimate the quality of our annotations using both manually-tagged named entities and word senses, obtaining an accuracy of roughly 70% for both named entities and word sense annotations.

O36 - Legal and Ethical Issues

Thursday, May 29, 18:20

Chairperson: **Christopher Cieri** Oral Session

The Liability of Service Providers in e-Research Infrastructures: Killing the Messenger?

Pawel Kamocki

Hosting Providers play an essential role in the development of Internet services such as e-Research Infrastructures. In order to promote the development of such services, legislators on both sides of the Atlantic Ocean introduced "safe harbour" provisions to protect Service Providers (a category which includes Hosting Providers) from legal claims (e.g. of copyright infringement). Relevant provisions can be found in § 512 of the United States Copyright Act and in art. 14 of the Directive 2000/31/EC (and its national implementations). The cornerstone of this framework is the passive role of the Hosting Provider through which he has no knowledge of the content that he hosts. With the arrival of Web 2.0, however, the role of Hosting Providers on the Internet changed; this change has been reflected in court decisions that have reached varying conclusions in the last few years. The purpose of this article is to present the existing framework (including recent case law from the US, Germany and France).

Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter

Alain Couillault, Karën Fort, Gilles Adda and Hugues de Mazancourt

The authors have written the Ethic and Big Data Charter in collaboration with various agencies, private bodies and associations. This Charter aims at describing any large or complex resources, and in particular language resources, from a legal and ethical viewpoint and ensuring the transparency of the process of creating and distributing such resources. We propose in this article an analysis of the documentation coverage of the most frequently mentioned language resources with regards to the Charter, in order to show the benefit it offers.

Disclose Models, Hide the Data - How to Make Use of Confidential Corpora without Seeing Sensitive Raw Data

Erik Faessler, Johannes Hellrich and Udo Hahn

Confidential corpora from the medical, enterprise, security or intelligence domains often contain sensitive raw data which

lead to severe restrictions as far as the public accessibility and distribution of such language resources are concerned. The enforcement of strict mechanisms of data protection consitutes a serious barrier for progress in language technology (products) in such domains, since these data are extremely rare or even unavailable for scientists and developers not directly involved in the creation and maintenance of such resources. In order to bypass this problem, we here propose to distribute trained language models which were derived from such resources as a substitute for the original confidential raw data which remain hidden to the outside world. As an example, we exploit the access-protected German-language medical FRAMED corpus from which we generate and distribute models for sentence splitting, tokenization and POS tagging based on software taken from OPENNLP, NLTK and JCORE, our own UIMA-based text analytics pipeline.

P45 - Anaphora and Coreference

Thursday, May 29, 18:20

Chairperson: **Costanza Navarretta** Poster Session

Corpus for Coreference Resolution on Scientific Papers

Panot Chaimongkol, Akiko Aizawa and Yuka Tateisi

The ever-growing number of published scientific papers prompts the need for automatic knowledge extraction to help scientists keep up with the state-of-the-art in their respective fields. To construct a good knowledge extraction system, annotated corpora in the scientific domain are required to train machine learning models. As described in this paper, we have constructed an annotated corpus for coreference resolution in multiple scientific domains, based on an existing corpus. We have modified the annotation scheme from Message Understanding Conference to better suit scientific texts. Then we applied that to the corpus. The annotated corpus is then compared with corpora in general domains in terms of distribution of resolution classes and performance of the Stanford Dcoref coreference resolver. Through these comparisons, we have demonstrated quantitatively that our manually annotated corpus differs from a general-domain corpus, which suggests deep differences between general-domain texts and scientific texts and which shows that different approaches can be made to tackle coreference resolution for general texts and scientific texts.

ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann and Bonnie Webber

We present ParCor, a parallel corpus of texts in which pronoun coreference – reduced coreference in which pronouns are used as referring expressions – has been annotated. The corpus is intended to be used both as a resource from which to learn systematic differences in pronoun use between languages and ultimately for developing and testing informed Statistical Machine Translation systems aimed at addressing the problem of pronoun coreference in translation. At present, the corpus consists of a collection of parallel English-German documents from two different text genres: TED Talks (transcribed planned speech), and EU Bookshop publications (written text). All documents in the corpus have been manually annotated with respect to the type and location of each pronoun and, where relevant, its antecedent. We provide details of the texts that we selected, the guidelines and tools used to support annotation and some corpus statistics. The texts in the corpus have already been translated into many languages, and we plan to expand the corpus into these other languages, as well as other genres, in the future.

The DARE Corpus: A Resource for Anaphora Resolution in Dialogue-based Intelligent Tutoring Systems

Nobal Niraula, Vasile Rus, Rajendra Banjade, Dan Stefanescu, William Baggett and Brent Morgan

We describe the DARE corpus, an annotated data set focusing on pronoun resolution in tutorial dialogue. Although data sets for general purpose anaphora resolution exist, they are not suitable for dialogue based Intelligent Tutoring Systems. To the best of our knowledge, no data set is currently available for pronoun resolution in dialogue based intelligent tutoring systems. The described DARE corpus consists of 1,000 annotated pronoun instances collected from conversations between high-school students and the intelligent tutoring system DeepTutor. The data set is publicly available.

CROMER: a Tool for Cross-Document Event and Entity Coreference

Christian Girardi, Manuela Speranza, Rachele Sprugnoli and Sara Tonelli

In this paper we present CROMER (CROss-document Main Events and entities Recognition), a novel tool to manually annotate event and entity coreference across clusters of documents. The tool has been developed so as to handle large collections of documents, perform collaborative annotation (several annotators can work on the same clusters), and enable the linking of the annotated data to external knowledge sources. Given the availability of semantic information encoded in Semantic Web resources, this tool is designed to support annotators in linking entities and events to DBPedia and Wikipedia, so as to facilitate the automatic retrieval of additional

semantic information. In this way, event modelling and chaining is made easy, while guaranteeing the highest interconnection with external resources. For example, the tool can be easily linked to event models such as the Simple Event Model [van Hage et al , 2011] and the Grounded Annotation Framework [Fokkens et al. 2013].

Coreference Resolution for Latvian

Arturs Znotinš and Peteris Paikens

Coreference resolution (CR) is a current problem in natural language processing (NLP) research and it is a key task in applications such as question answering, text summarization and information extraction for which text understanding is of crucial importance. We describe an implementation of coreference resolution tools for Latvian language, developed as a part of a tool chain for newswire text analysis but usable also as a separate, publicly available module. LVCoref is a rule based CR system that uses entity centric model that encourages the sharing of information across all mentions that point to the same real-world entity. The system is developed to provide starting ground for further experiments and generate a reference baseline to be compared with more advanced rule-based and machine learning based future coreference resolvers. It now reaches 66.6 F-score using predicted mentions and 78.1% F-score using gold mentions. This paper describes current efforts to create a CR system and to improve NER performance for Latvian. Task also includes creation of the corpus of manually annotated coreference relations.

An Exercise in Reuse of Resources: Adapting General Discourse Coreference Resolution for Detecting Lexical Chains in Patent Documentation

Nadjet Bouayad-Agha, Alicia Burga, Gerard Casamayor, Joan Codina, Rogelio Nazar and Leo Wanner

The Stanford Coreference Resolution System (StCR) is a multi-pass, rule-based system that scored best in the CoNLL 2011 shared task on general discourse coreference resolution. We describe how the StCR has been adapted to the specific domain of patents and give some cues on how it can be adapted to other domains. We present a linguistic analysis of the patent domain and how we were able to adapt the rules to the domain and to expand coreferences with some lexical chains. A comparative evaluation shows an improvement of the coreference resolution system, denoting that (i) StCR is a valuable tool across different text genres; (ii) specialized discourse NLP may significantly benefit from general discourse NLP research.

The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Resolution

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler and Katrin Schweitzer

DIRNDL is a spoken and written corpus based on German radio news, which features coreference and information-status annotation (including bridging anaphora and their antecedents), as well as prosodic information. We have recently extended DIRNDL with a fine-grained two-dimensional information status labeling scheme. We have also applied a state-of-the-art part-of-speech and morphology tagger to the corpus, as well as highly accurate constituency and dependency parsers. In the light of this development we believe that DIRNDL is an interesting resource for NLP researchers working on automatic coreference and bridging resolution. In order to enable and promote usage of the data, we make it available for download in an accessible tabular format, compatible with the formats used in the CoNLL and SemEval shared tasks on automatic coreference resolution.

Multilingual Corpora with Coreferential Annotation of Person Entities

Marcos Garcia and Pablo Gamallo

This paper presents three corpora with coreferential annotation of person entities for Portuguese, Galician and Spanish. They contain coreference links between several types of pronouns (including elliptical, possessive, indefinite, demonstrative, relative and personal clitic and non-clitic pronouns) and nominal phrases (including proper nouns). Some statistics have been computed, showing distributional aspects of coreference both in journalistic and in encyclopedic texts. Furthermore, the paper shows the importance of coreference resolution for a task such as Information Extraction, by evaluating the output of an Open Information Extraction system on the annotated corpora. The corpora are freely distributed in two formats: (i) the SemEval-2010 and (ii) the brat rapid annotation tool, so they can be enlarged and improved collaboratively.

Polish Coreference Corpus in Numbers

Maciej Ogrodniczuk, Mateusz Kopeć and Agata Savary

This paper attempts a preliminary interpretation of the occurrence of different types of linguistic constructs in the manually-annotated Polish Coreference Corpus by providing analyses of various statistical properties related to mentions, clusters and near-identity links. Among others, frequency of mentions, zero subjects and singleton clusters is presented, as well as the average mention and cluster size. We also show that some coreference clustering constraints, such as gender or number agreement, are

frequently not valid in case of Polish. The need for lemmatization for automatic coreference resolution is supported by an empirical study. Correlation between cluster and mention count within a text is investigated, with short characteristics of outlier cases. We also examine this correlation in each of the 14 text domains present in the corpus and show that none of them has abnormal frequency of outlier texts regarding the cluster/mention ratio. Finally, we report on our negative experiences concerning the annotation of the near-identity relation. In the conclusion we put forward some guidelines for the future research in the area.

P46 - Information Extraction and Information Retrieval

Thursday, May 29, 18:20

Chairperson: **Dimitrios Kokkinakis** Poster Session

French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime

Véronique Moriceau and Xavier Tannier

In this paper, we describe the development of French resources for the extraction and normalization of temporal expressions with HeidelTime, a open-source multilingual, cross-domain temporal tagger. HeidelTime extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard. Several types of temporal expressions are extracted: dates, times, durations and temporal sets. French resources have been evaluated in two different ways: on the French TimeBank corpus, a corpus of newspaper articles in French annotated according to the ISO-TimeML standard, and on a user application for automatic building of event timelines. Results on the French TimeBank are quite satisfaying as they are comparable to those obtained by HeidelTime in English and Spanish on newswire articles. Concerning the user application, we used two temporal taggers for the preprocessing of the corpus in order to compare their performance and results show that the performances of our application on French documents are better with HeidelTime. The French resources and evaluation scripts are publicly available with HeidelTime.

Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain

Zdenka Uresova, Jan Hajic, Pavel Pecina and Ondrej Dusek

This paper presents development and test sets for machine translation of search queries in cross-lingual information retrieval in the medical domain. The data consists of the total of 1,508 real user queries in English translated to Czech, German, and French. We describe the translation and review process involving medical professionals and present a baseline experiment where our data sets are used for tuning and evaluation of a machine translation system.

Semantic Approaches to Software Component Retrieval with English Queries

Huijing Deng and Grzegorz Chrupała

Enabling code reuse is an important goal in software engineering, and it depends crucially on effective code search interfaces. We propose to ground word meanings in source code and use such language-code mappings in order to enable a search engine for programming library code where users can pose queries in English. We exploit the fact that there are large programming language libraries which are documented both via formally specified function or method signatures as well as descriptions written in natural language. Automatically learned associations between words in descriptions and items in signatures allows us to use queries formulated in English to retrieve methods which are not documented via natural language descriptions, only based on their signatures. We show that the rankings returned by our model substantially outperforms a strong term-matching baseline.

Annotating Relation Mentions in Tabloid Press

Hong Li, Sebastian Krause, Feiyu Xu, Hans Uszkoreit, Robert Hummel and Veselina Mironova

This paper presents a new resource for the training and evaluation needed by relation extraction experiments. The corpus consists of annotations of mentions for three semantic relations: marriage, parent–child, siblings, selected from the domain of biographic facts about persons and their social relationships. The corpus contains more than one hundred news articles from Tabloid Press. In the current corpus, we only consider the relation mentions occurring in the individual sentences. We provide multi-level annotations which specify the marked facts from relation, argument, entity, down to the token level, thus allowing for detailed analysis of linguistic phenomena and their interactions. A generic markup tool Recon developed at the DFKI LT lab has been utilised for the annotation task. The corpus has been annotated by two human experts, supported by additional conflict resolution conducted by a third expert. As shown in the evaluation, the annotation is of high quality as proved by the stated inter-annotator agreements both on sentence level and on relationmention level. The current corpus is already in active use in our research

for evaluation of the relation extraction performance of our automatically learned extraction patterns.

Construction of Diachronic Ontologies from People's Daily of Fifty Years

Shaoda He, Xiaojun Zou, Liumingjing Xiao and Junfeng Hu

This paper presents an Ontology Learning From Text (OLFT) method follows the well-known OLFT cake layer framework. Based on the distributional similarity, the proposed method generates multi-level ontologies from comparatively small corpora with the aid of HITS algorithm. Currently, this method covers terms extraction, synonyms recognition, concepts discovery and concepts hierarchical clustering. Among them, both concepts discovery and concepts hierarchical clustering are aided by the HITS authority, which is obtained from the HITS algorithm by an iteratively recommended way. With this method, a set of diachronic ontologies is constructed for each year based on People's Daily corpora of fifty years (i.e., from 1947 to 1996). Preliminary experiments show that our algorithm outperforms the Google's RNN and K-means based algorithm in both concepts discovery and concepts hierarchical clustering.

Use of Unsupervised Word Classes for Entity Recognition: Application to the Detection of Disorders in Clinical Reports

Maria Evangelia Chatzimina, Cyril Grouin and Pierre Zweigenbaum

Unsupervised word classes induced from unannotated text corpora are increasingly used to help tasks addressed by supervised classification, such as standard named entity detection. This paper studies the contribution of unsupervised word classes to a medical entity detection task with two specific objectives: How do unsupervised word classes compare to available knowledge-based semantic classes? Does syntactic information help produce unsupervised word classes with better properties? We design and test two syntax-based methods to produce word classes: one applies the Brown clustering algorithm to syntactic dependencies, the other collects latent categories created by a PCFG-LA parser. When added to non-semantic features, knowledge-based semantic classes gain 7.28 points of F-measure. In the same context, basic unsupervised word classes gain 4.16pt, reaching 60% of the contribution of knowledge-based semantic classes and outperforming Wikipedia, and adding PCFG-LA unsupervised word classes gain one more point at 5.11pt, reaching 70%. Unsupervised word classes could therefore provide a useful semantic back-off in domains where no knowledge-based semantic classes are available. The combination of both knowledge-based and basic unsupervised classes gains 8.33pt. Therefore, unsupervised classes are still useful even when rich knowledge-based classes exist.

The Weltmodell: A Data-Driven Commonsense Knowledge Base

Alan Akbik and Thilo Michael

We present the Weltmodell, a commonsense knowledge base that was automatically generated from aggregated dependency parse fragments gathered from over 3.5 million English language books. We leverage the magnitude and diversity of this dataset to arrive at close to ten million distinct N-ary commonsense facts using techniques from open-domain Information Extraction (IE). Furthermore, we compute a range of measures of association and distributional similarity on this data. We present the results of our efforts using a browsable web demonstrator and publicly release all generated data for use and discussion by the research community. In this paper, we give an overview of our knowledge acquisition method and representation model, and present our web demonstrator.

Discovering and Visualising Stories in News

Marieke van Erp, Gleb Satyukov, Piek Vossen and Marit Nijsen

Daily news streams often revolve around topics that span over a longer period of time such as the global financial crisis or the healthcare debate in the US. The length and depth of these stories can be such that they become difficult to track for information specialists who need to reconstruct exactly what happened for policy makers and companies. We present a framework to model stories from news: we describe the characteristics that make up interesting stories, how these translate to filters on our data and we present a first use case in which we detail the steps to visualising story lines extracted from news articles about the global automotive industry.

A Large Scale Database of Strongly-related Events in Japanese

Tomohide Shibata, Shotaro Kohama and Sadao Kurohashi

The knowledge about the relation between events is quite useful for coreference resolution, anaphora resolution, and several NLP applications such as dialogue system. This paper presents a large scale database of strongly-related events in Japanese, which has been acquired with our proposed method (Shibata and Kurohashi, 2011). In languages, where omitted arguments or zero anaphora are often utilized, such as Japanese, the coreference-based event extraction methods are hard to be applied, and so our method extracts strongly-related events in

a two-phrase construct. This method first calculates the co-occurrence measure between predicate-arguments (events), and regards an event pair, whose mutual information is high, as strongly-related events. To calculate the co-occurrence measure efficiently, we adopt an association rule mining method. Then, we identify the remaining arguments by using case frames. The database contains approximately 100,000 unique events, with approximately 340,000 strongly-related event pairs, which is much larger than an existing automatically-constructed event database. We evaluated randomly-chosen 100 event pairs, and the accuracy was approximately 68%.

ClearTK 2.0: Design Patterns for Machine Learning in UIMA

Steven Bethard, Philip Ogren and Lee Becker

ClearTK adds machine learning functionality to the UIMA framework, providing wrappers to popular machine learning libraries, a rich feature extraction library that works across different classifiers, and utilities for applying and evaluating machine learning models. Since its inception in 2008, ClearTK has evolved in response to feedback from developers and the community. This evolution has followed a number of important design principles including: conceptually simple annotator interfaces, readable pipeline descriptions, minimal collection readers, type system agnostic code, modules organized for ease of import, and assisting user comprehension of the complex UIMA framework.

P47 - Language Identification

Thursday, May 29, 18:20

Chairperson: **Michael Rosner** Poster Session

Vocabulary-based Language Similarity using Web Corpora

Dirk Goldhahn and Uwe Quasthoff

This paper will focus on the evaluation of automatic methods for quantifying language similarity. This is achieved by ascribing language similarity to the similarity of text corpora. This corpus similarity will first be determined by the resemblance of the vocabulary of languages. Thereto words or parts of them such as letter n-grams are examined. Extensions like transliteration of the text data will ensure the independence of the methods from text characteristics such as the writing system used. Further analyzes will show to what extent knowledge about the distribution of words in parallel text can be used in the context of language similarity.

Automatic Language Identity Tagging on Word and Sentence-Level in Multilingual Text Sources: a Case-Study on Luxembourgish

Thomas Lavergne, Gilles Adda, Martine Adda-Decker and Lori Lamel

Luxembourgish, embedded in a multilingual context on the divide between Romance and Germanic cultures, remains one of Europe's under-described languages. This is due to the fact that the written production remains relatively low, and linguistic knowledge and resources, such as lexica and pronunciation dictionaries, are sparse. The speakers or writers will frequently switch between Luxembourgish, German, and French, on a per-sentence basis, as well as on a sub-sentence level. In order to build resources like lexicons, and especially pronunciation lexicons, or language models needed for natural language processing tasks such as automatic speech recognition, language used in text corpora should be identified. In this paper, we present the design of a manually annotated corpus of mixed language sentences as well as the tools used to select these sentences. This corpus of difficult sentences was used to test a word-based language identification system. This language identification system was used to select textual data extracted from the web, in order to build a lexicon and language models. This lexicon and language model were used in an Automatic Speech Recognition system for the Luxembourgish language which obtain a 25% WER on the Quaero development data.

VarClass: An Open-source Language Identification Tool for Language Varieties

Marcos Zampieri and Binyam Gebre

This paper presents VarClass, an open-source tool for language identification available both to be downloaded as well as through a graphical user-friendly interface. The main difference of VarClass in comparison to other state-of-the-art language identification tools is its focus on language varieties. General purpose language identification tools do not take language varieties into account and our work aims to fill this gap. VarClass currently contains language models for over 27 languages in which 10 of them are language varieties. We report an average performance of over 90.5% accuracy in a challenging dataset. More language models will be included in the upcoming months.

Native Language Identification Using Large, Longitudinal Data

Xiao Jiang, Yufan Guo, Jeroen Geertzen, Dora Alexopoulou, Lin Sun and Anna Korhonen

Native Language Identification (NLI) is a task aimed at determining the native language (L1) of learners of second

language (L2) on the basis of their written texts. To date, research on NLI has focused on relatively small corpora. We apply NLI to the recently released EFCamDat corpus which is not only multiple times larger than previous L2 corpora but also provides longitudinal data at several proficiency levels. Our investigation using accurate machine learning with a wide range of linguistic features reveals interesting patterns in the longitudinal data which are useful for both further development of NLI and its application to research on L2 acquisition.

On the Romance Languages Mutual Intelligibility

Liviu Dinu and Alina Maria Ciobanu

We propose a method for computing the similarity of natural languages and for clustering them based on their lexical similarity. Our study provides evidence to be used in the investigation of the written intelligibility, i.e., the ability of people writing in different languages to understand one another without prior knowledge of foreign languages. We account for etymons and cognates, we quantify lexical similarity and we extend our analysis from words to languages. Based on the introduced methodology, we compute a matrix of Romance languages intelligibility.

P48 - Morphology

Thursday, May 29, 18:20

Chairperson: **Karel Pala** Poster Session

Heuristic Hyper-minimization of Finite State Lexicons

Senka Drobac, Krister Lindén, Tommi Pirinen and Miikka Silfverberg

Flag diacritics, which are special multi-character symbols executed at runtime, enable optimising finite-state networks by combining identical sub-graphs of its transition graph. Traditionally, the feature has required linguists to devise the optimisations to the graph by hand alongside the morphological description. In this paper, we present a novel method for discovering flag positions in morphological lexicons automatically, based on the morpheme structure implicit in the language description. With this approach, we have gained significant decrease in the size of finite-state networks while maintaining reasonable application speed. The algorithm can be applied to any language description, where the biggest achievements are expected in large and complex morphologies. The most noticeable reduction in size we got with a morphological transducer for Greenlandic, whose original size is on average about 15 times larger than other morphologies. With the presented hyper-minimization method, the transducer is reduced to 10,1% of the original size, with lookup speed decreased only by 9,5%.

Crowd-sourcing Evaluation of Automatically Acquired, Morphologically Related Word Groupings

Claudia Borg and Albert Gatt

The automatic discovery and clustering of morphologically related words is an important problem with several practical applications. This paper describes the evaluation of word clusters carried out through crowd-sourcing techniques for the Maltese language. The hybrid (Semitic-Romance) nature of Maltese morphology, together with the fact that no large-scale lexical resources are available for Maltese, make this an interesting and challenging problem.

Morphological Parsing of Swahili using Crowdsourced Lexical Resources

Patrick Littell, Kaitlyn Price and Lori Levin

We describe a morphological analyzer for the Swahili language, written in an extension of XFST/LEXC intended for the easy declaration of morphophonological patterns and importation of lexical resources. Our analyzer was supplemented extensively with data from the Kamusi Project (kamusi.org), a user-contributed multilingual dictionary. Making use of this resource allowed us to achieve wide lexical coverage quickly, but the heterogeneous nature of user-contributed content also poses some challenges when adapting it for use in an expert system.

Chasing the Perfect Splitter: A Comparison of Different Compound Splitting Tools

Carla Parra Escartín

This paper reports on the evaluation of two compound splitters for German. Compounding is a very frequent phenomenon in German and thus efficient ways of detecting and correctly splitting compound words are needed for natural language processing applications. This paper presents different strategies for compound splitting, focusing on German. Four compound splitters for German are presented. Two of them were used in Statistical Machine Translation (SMT) experiments, obtaining very similar qualitative scores in terms of BLEU and TER and therefore a thorough evaluation of both has been carried out.

Generating and using Probabilistic Morphological Resources for the Biomedical Domain

Vincent Claveau and Ewa Kijak

In most Indo-European languages, many biomedical terms are rich morphological structures composed of several constituents

mainly originating from Greek or Latin. The interpretation of these compounds are keystones to access information. In this paper, we present morphological resources aiming at coping with these biomedical morphological compounds. Following previous work (Claveau et al. 2011,Claveau et al. 12), these resources are automatically built using Japanese terms in Kanjis as a pivot language and alignment techniques. We show how these alignment information can be used for segmenting compounds, attaching semantic interpretation to each part, proposing definitions (gloses) of the compounds... When possible, these tasks are compared with state-of-the-art tools, and the results show the interest of our automatically built probabilistic resources.

Using Resource-Rich Languages to Improve Morphological Analysis of Under-Resourced Languages

Peter Baumann and Janet Pierrehumbert

The world-wide proliferation of digital communications has created the need for language and speech processing systems for under-resourced languages. Developing such systems is challenging if only small data sets are available, and the problem is exacerbated for languages with highly productive morphology. However, many under-resourced languages are spoken in multi-lingual environments together with at least one resource-rich language and thus have numerous borrowings from resource-rich languages. Based on this insight, we argue that readily available resources from resource-rich languages can be used to bootstrap the morphological analyses of under-resourced languages with complex and productive morphological systems. In a case study of two such languages, Tagalog and Zulu, we show that an easily obtainable English wordlist can be deployed to seed a morphological analysis algorithm from a small training set of conversational transcripts. Our method achieves a precision of 100% and identifies 28 and 66 of the most productive affixes in Tagalog and Zulu, respectively.

Turkish Treebank as a Gold Standard for Morphological Disambiguation and Its Influence on Parsing

Ozlem Cetinoglu

So far predicted scenarios for Turkish dependency parsing have used a morphological disambiguator that is trained on the data distributed with the tool(Sak et al., 2008). Although models trained on this data have high accuracy scores on the test and development data of the same set, the accuracy drastically drops when the model is used in the preprocessing of Turkish Treebank parsing experiments. We propose to use the Turkish Treebank(Oflazer et al., 2003) as a morphological resource to overcome this problem and convert the treebank to the morphological disambiguator's format. The experimental results show that we achieve improvements in disambiguating the Turkish Treebank and the results also carry over to parsing. With the help of better morphological analysis, we present the best labelled dependency parsing scores to date on Turkish.

CroDeriV: a New Resource for Processing Croatian Morphology

Krešimir Šojat, Matea Srebačić, Marko Tadić and Tin Pavelić

The paper deals with the processing of Croatian morphology and presents CroDeriV – a newly developed language resource that contains data about morphological structure and derivational relatedness of verbs in Croatian. In its present shape, CroDeriV contains 14 192 Croatian verbs. Verbs in CroDeriV are analyzed for morphemes and segmented into lexical, derivational and inflectional morphemes. The structure of CroDeriV enables the detection of verbal derivational families in Croatian as well as the distribution and frequency of particular affixes and lexical morphemes. Derivational families consist of a verbal base form and all prefixed or suffixed derivatives detected in available machine readable Croatian dictionaries and corpora. Language data structured in this way was further used for the expansion of other language resources for Croatian, such as Croatian WordNet and the Croatian Morphological Lexicon. Matching the data from CroDeriV on one side and Croatian WordNet and the Croatian Morphological Lexicon on the other resulted in significant enrichment of Croatian WordNet and enlargement of the Croatian Morphological Lexicon.

DerivBase.hr: A High-Coverage Derivational Morphology Resource for Croatian

Jan Šnajder

Knowledge about derivational morphology has been proven useful for a number of natural language processing (NLP) tasks. We describe the construction and evaluation of DerivBase.hr, a large-coverage morphological resource for Croatian. DerivBase.hr groups 100k lemmas from web corpus hrWaC into 56k clusters of derivationally related lemmas, so-called derivational families. We focus on suffixal derivation between and within nouns, verbs, and adjectives. We propose two approaches: an unsupervised approach and a knowledge-based approach based on a hand-crafted morphology model but without using any additional lexico-semantic resources The resource acquisition procedure consists of three steps: corpus preprocessing, acquisition of an inflectional lexicon, and the induction of derivational families.

We describe an evaluation methodology based on manually constructed derivational families from which we sample and annotate pairs of lemmas. We evaluate DerivBase.hr on the so-obtained sample, and show that the knowledge-based version attains good clustering quality of 81.2% precision, 76.5% recall, and 78.8% F1 -score. As with similar resources for other languages, we expect DerivBase.hr to be useful for a number of NLP tasks.

Finite-State Morphological Transducers for Three Kypchak Languages

Jonathan Washington, Ilnar Salimzyanov and Francis Tyers

This paper describes the development of free/open-source finite-state morphological transducers for three Turkic languages–Kazakh, Tatar, and Kumyk–representing one language from each of the three sub-branches of the Kypchak branch of Turkic. The finite-state toolkit used for the work is the Helsinki Finite-State Toolkit (HFST). This paper describes how the development of a transducer for each subsequent closely-related language took less development time. An evaluation is presented which shows that the transducers all have a reasonable coverage–around 90%–on freely available corpora of the languages, and high precision over a manually verified test set.

P49 - Multimodality

Thursday, May 29, 18:20

Chairperson: **Volker Steinbiss** Poster Session

Representing Multimodal Linguistic Annotated Data

Brigitte Bigi, Tatsuya Watanabe and Laurent Prévot

The question of interoperability for linguistic annotated resources covers different aspects. First, it requires a representation framework making it possible to compare, and eventually merge, different annotation schema. In this paper, a general description level representing the multimodal linguistic annotations is proposed. It focuses on time representation and on the data content representation: This paper reconsiders and enhances the current and generalized representation of annotations. An XML schema of such annotations is proposed. A Python API is also proposed. This framework is implemented in a multi-platform software and distributed under the terms of the GNU Public License.

Single-Person and Multi-Party 3D Visualizations for Nonverbal Communication Analysis

Michael Kipp, Levin Freiherr von Hollen, Michael Christopher Hrstka and Franziska Zamponi

The qualitative analysis of nonverbal communication is more and more relying on 3D recording technology. However, the human analysis of 3D data on a regular 2D screen can be challenging as 3D scenes are difficult to visually parse. To optimally exploit the full depth of the 3D data, we propose to enhance the 3D view with a number of visualizations that clarify spatial and conceptual relationships and add derived data like speed and angles. In this paper, we present visualizations for directional body motion, hand movement direction, gesture space location, and proxemic dimensions like interpersonal distance, movement and orientation. The proposed visualizations are available in the open source tool JMocap and are planned to be fully integrated into the ANVIL video annotation tool. The described techniques are intended to make annotation more efficient and reliable and may allow the discovery of entirely new phenomena.

The AV-LASYN Database: a Synchronous Corpus of Audio and 3D Facial Marker Data for Audio-Visual Laughter Synthesis

Huseyin Cakmak, Jerome Urbain, Thierry Dutoit and Joelle Tilmanne

A synchronous database of acoustic and 3D facial marker data was built for audio-visual laughter synthesis. Since the aim is to use this database for HMM-based modeling and synthesis, the amount of collected data from one given subject had to be maximized. The corpus contains 251 utterances of laughter from one male participant. Laughter was elicited with the help of humorous videos. The resulting database is synchronous between modalities (audio and 3D facial motion capture data). Visual 3D data is available in common formats such as BVH and C3D with head motion and facial deformation independently available. Data is segmented and audio has been annotated. Phonetic transcriptions are available in the HTK-compatible format. Principal component analysis has been conducted on visual data and has shown that a dimensionality reduction might be relevant. The corpus may be obtained under a research license upon request to authors.

Linking Pictographs to Synsets: Sclera2Cornetto

Vincent Vandeghinste and Ineke Schuurman

Social inclusion of people with Intellectual and Developmental Disabilities can be promoted by offering them ways to independently use the internet. People with reading or writing disabilities can use pictographs instead of text. We present a

resource in which we have linked a set of 5710 pictographs to lexical-semantic concepts in Cornetto, a Wordnet-like database for Dutch. We show that, by using this resource in a text-to-pictograph translation system, we can greatly improve the coverage comparing with a baseline where words are converted into pictographs only if the word equals the filename.

The MMASCS Multi-Modal Annotated Synchronous Corpus of Audio, Video, Facial Motion and Tongue Motion Data of Normal, Fast and Slow Speech

Dietmar Schabus, Michael Pucher and Phil Hoole

In this paper, we describe and analyze a corpus of speech data that we have recorded in multiple modalities simultaneously: facial motion via optical motion capturing, tongue motion via electro-magnetic articulography, as well as conventional video and high-quality audio. The corpus consists of 320 phonetically diverse sentences uttered by a male Austrian German speaker at normal, fast and slow speaking rate. We analyze the influence of speaking rate on phone durations and on tongue motion. Furthermore, we investigate the correlation between tongue and facial motion. The data corpus is available free of charge for research use, including phonetic annotations and a playback software which visualizes the 3D data, from the website http://cordelia.ftw.at/mmascs

Mining a Multimodal Corpus for Non-Verbal Behavior Sequences Conveying Attitudes

Mathieu Chollet, Magalie Ochs and Catherine Pelachaud

Interpersonal attitudes are expressed by non-verbal behaviors on a variety of different modalities. The perception of these behaviors is influenced by how they are sequenced with other behaviors from the same person and behaviors from other interactants. In this paper, we present a method for extracting and generating sequences of non-verbal signals expressing interpersonal attitudes. These sequences are used as part of a framework for non-verbal expression with Embodied Conversational Agents that considers different features of non-verbal behavior: global behavior tendencies, interpersonal reactions, sequencing of non-verbal signals, and communicative intentions. Our method uses a sequence mining technique on an annotated multimodal corpus to extract sequences characteristic of different attitudes. New sequences of non-verbal signals are generated using a probabilistic model, and evaluated using the previously mined sequences.

The IMAGACT Visual Ontology. an Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action

Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini and Alessandro Panunzi

Action verbs have many meanings, covering actions in different ontological types. Moreover, each language categorizes action in its own way. One verb can refer to many different actions and one action can be identified by more than one verb. The range of variations within and across languages is largely unknown, causing trouble for natural language processing tasks. IMAGACT is a corpus-based ontology of action concepts, derived from English and Italian spontaneous speech corpora, which makes use of the universal language of images to identify the different action types extended by verbs referring to action in English, Italian, Chinese and Spanish. This paper presents the infrastructure and the various linguistic information the user can derive from it. IMAGACT makes explicit the variation of meaning of action verbs within one language and allows comparisons of verb variations within and across languages. Because the action concepts are represented with videos, extension into new languages beyond those presently implemented in IMAGACT is done using competence-based judgments by mother-tongue informants without intense lexicographic work involving underdetermined semantic description

Multimodal Dialogue Segmentation with Gesture Post-Processing

Kodai Takahashi and Masashi Inoue

We investigate an automatic dialogue segmentation method using both verbal and non-verbal modalities. Dialogue contents are used for the initial segmentation of dialogue; then, gesture occurrences are used to remove the incorrect segment boundaries. A unique characteristic of our method is to use verbal and non-verbal information separately. We use a three-party dialogue that is rich in gesture as data. The transcription of the dialogue is segmented into topics without prior training by using the TextTiling and U00 algorithm. Some candidates for segment boundaries - where the topic continues - are irrelevant. Those boundaries can be found and removed by locating gestures that stretch over the boundary candidates. This ltering improves the segmentation accuracy of text-only segmentation.

The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech

Shannon Hennig, Ryad Chellali and Nick Campbell

Biosignals, such as electrodermal activity (EDA) and heart rate, are increasingly being considered as potential data sources to

provide information about the temporal fluctuations in affective experience during human interaction. This paper describes an English-speaking, multiple session corpus of small groups of people engaged in informal, unscripted conversation while wearing wireless, wrist-based EDA sensors. Additionally, one participant per recording session wore a heart rate monitor. This corpus was collected in order to observe potential interactions between various social and communicative phenomena and the temporal dynamics of the recorded biosignals. Here we describe the communicative context, technical set-up, synchronization process, and challenges in collecting and utilizing such data. We describe the segmentation and annotations to date, including laughter annotations, and how the research community can access and collaborate on this corpus now and in the future. We believe this corpus is particularly relevant to researchers interested in unscripted social conversation as well as to researchers with a specific interest in observing the dynamics of biosignals during informal social conversation rich with examples of laughter, conversational turn-taking, and non-task-based interaction.

Keynote Speech 2

Friday, May 30, 9:00

Chairperson: **Jan Odijk**

When Will Robots Speak like You and Me?

luc Steels

The incredible growth in language resources has lead to unprecedented opportunities for language research and a lot can still be done by exploiting existing corpora and statistical language processing techniques. Nevertheless we should also remain ambitious. We should try to keep forging ahead with fundamental research, trying to tackle new application areas and improving existing applications by more sophisticated linguistic theories and language processing systems. This talk reports on work in our group on grounded language interaction between humans and robots. This problem is extraordinarily difficult because we need to figure out how to achieve true language understanding, i.e. deep language parsing coupled to a semantics grounded in the sensori-motor embodiment of robots, and true language production, i.e. planning what to say, conceptualising the world for language and translation into utterances. We also need to figure out how artificial agents can cope with highly ungrammatical and fragmentary input by full exploitation of the context. On top of that, we can no longer view language as a static system of conventions but as a living system that is always changing and evolving, with new or shifting word senses and new or shifting usage of grammatical constructions. This implies that artificial speakers and listeners need to constantly learn, expand their language when needed, align themselves to the language use of others, and act as tutors to help others understand and acquire language. I will present some of the key ideas that we are currently exploring to tackle these enormously challenging issues. They include a novel computational formalism called Fluid Construction Grammar, which is an attempt to operationalise key insights from construction grammar, cognitive linguistics and embodied semantics. Flexible language processing and learning is implemented using a meta-level in which diagnostics detect anomalies or gaps and repair strategies try to cope with them by ignoring ungrammaticalities or expanding the language system. We have also developed techniques for studying language as a complex adaptive system and done several experiments how vocabularies and grammars can emerge in situated embodied interactions between robotic agents. The talk is illustrated with live demos and videoclips of robots playing language games.

O37 - Sentiment Analysis and Social Media (2)

Friday, May 30, 9:45

Chairperson: **Piek Vossen** Oral Session

Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis

Diana Maynard and Mark Greenwood

Sarcasm is a common phenomenon in social media, and is inherently difficult to analyse, not just automatically but often for humans too. It has an important effect on sentiment, but is usually ignored in social media analysis, because it is considered too tricky to handle. While there exist a few systems which can detect sarcasm, almost no work has been carried out on studying the effect that sarcasm has on sentiment in tweets, and on incorporating this into automatic tools for sentiment analysis. We perform an analysis of the effect of sarcasm scope on the polarity of tweets, and have compiled a number of rules which enable us to improve the accuracy of sentiment analysis when sarcasm is known to be present. We consider in particular the effect of sentiment and sarcasm contained in hashtags, and have developed a hashtag tokeniser for GATE, so that sentiment and sarcasm found within hashtags can be detected more easily. According to our experiments, the hashtag tokenisation achieves 98% Precision, while the sarcasm detection achieved 91% Precision and polarity detection 80%.

SenTube: A Corpus for Sentiment Analysis on YouTube Social Media

Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi and Alessandro Moschitti

In this paper we present SenTube – a dataset of user-generated comments on YouTube videos annotated for information content and sentiment polarity. It contains annotations that allow to develop classifiers for several important NLP tasks: (i) sentiment analysis, (ii) text categorization (relatedness of a comment to video and/or product), (iii) spam detection, and (iv) prediction of comment informativeness. The SenTube corpus favors the development of research on indexing and searching YouTube videos exploiting information derived from comments. The corpus will cover several languages: at the moment, we focus on English and Italian, with Spanish and Dutch parts scheduled for the later stages of the project. For all the languages, we collect videos for the same set of products, thus offering possibilities for multi- and cross-lingual experiments. The paper provides annotation guidelines, corpus statistics and annotator agreement details.

Getting Reliable Annotations for Sarcasm in Online Dialogues

Reid Swanson, Stephanie Lukin, Luke Eisenberg, Thomas Corcoran and Marilyn Walker

The language used in online forums differs in many ways from that of traditional language resources such as news. One difference is the use and frequency of nonliteral, subjective dialogue acts such as sarcasm. Whether the aim is to develop a theory of sarcasm in dialogue, or engineer automatic methods for reliably detecting sarcasm, a major challenge is simply the difficulty of getting enough reliably labelled examples. In this paper we describe our work on methods for achieving highly reliable sarcasm annotations from untrained annotators on Mechanical Turk. We explore the use of a number of common statistical reliability measures, such as Kappa, Karger's, Majority Class, and EM. We show that more sophisticated measures do not appear to yield better results for our data than simple measures such as assuming that the correct label is the one that a majority of Turkers apply.

Modelling Irony in Twitter: Feature Analysis and Evaluation

Francesco Barbieri and Horacio Saggion

Irony, a creative use of language, has received scarce attention from the computational linguistics research point of view. We propose an automatic system capable of detecting irony with good accuracy in the social network Twitter. Twitter allows users to post short messages (140 characters) which usually do not follow the expected rules of the grammar, users tend to truncate words and use particular punctuation. For these reason automatic detection of Irony in Twitter is not trivial and requires specific linguistic tools. We propose in this paper a new set of experiments to assess the relevance of the features included in our model. Our model does not include words or sequences of words as features, aiming to detect inner characteristic of Irony.

Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts

Alexandra Balahur, Marco Turchi, Ralf Steinberger, Jose Manuel Perea-Ortega, Guillaume Jacquet, Dilek Kucuk, Vanni Zavarella and Adil El Ghali

This paper presents an evaluation of the use of machine translation to obtain and employ data for training multilingual sentiment classifiers. We show that the use of machine translated data obtained similar results as the use of native-speaker translations of the same data. Additionally, our evaluations pinpoint to the fact that the use of multilingual data, including that obtained through machine translation, leads to improved results in sentiment classification. Finally, we show that the performance of the sentiment classifiers built on machine translated data can be improved using original data from the target language and that even a small amount of such texts can lead to significant growth in the classification performance.

O38 - Paraphases

Friday, May 30, 9:45

Chairperson: **Bernardo Magnini** Oral Session

Semantic Clustering of Pivot Paraphrases

Marianna Apidianaki, Emilia Verzeni and Diana McCarthy

Paraphrases extracted from parallel corpora by the pivot method (Bannard and Callison-Burch, 2005) constitute a valuable resource for multilingual NLP applications. In this study, we analyse the semantics of unigram pivot paraphrases and use a graph-based sense induction approach to unveil hidden sense distinctions in the paraphrase sets. The comparison of the acquired senses to gold data from the Lexical Substitution shared task (McCarthy and Navigli, 2007) demonstrates that sense distinctions exist in the paraphrase sets and highlights the need for a disambiguation step in applications using this resource.

The Multilingual Paraphrase Database

Juri Ganitkevitch and Chris Callison-Burch

We release a massive expansion of the paraphrase database (PPDB) that now includes a collection of paraphrases in 23 different languages. The resource is derived from large volumes of bilingual parallel data. Our collection is extracted and ranked using state of the art methods. The multilingual PPDB has over a billion paraphrase pairs in total, covering the following languages: Arabic, Bulgarian, Chinese, Czech, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portugese, Romanian, Russian, Slovak, Slovenian, and Swedish.

Multiple Choice Question Corpus Analysis for Distractor Characterization

Van-Minh Pho, Thibault André, Anne-Laure Ligozat, Brigitte Grau, Gabriel Illouz and Thomas Francois

In this paper, we present a study of MCQ aiming to define criteria in order to automatically select distractors. We are aiming to show that distractor editing follows rules like syntactic and semantic homogeneity according to associated answer, and the possibility to automatically identify this homogeneity. Manual analysis shows that homogeneity rule is respected to edit distractors and automatic analysis shows the possibility to reproduce these criteria. These ones can be used in future works to automatically select distractors, with the combination of other criteria.

Creating and Using Large Monolingual Parallel Corpora for Sentential Paraphrase Generation

Sander Wubben, Antal van den Bosch and Emiel Krahmer

In this paper we investigate the automatic generation of paraphrases by using machine translation techniques. Three contributions we make are the construction of a large paraphrase corpus for English and Dutch, a re-ranking heuristic to use machine translation for paraphrase generation and a proper evaluation methodology. A large parallel corpus is constructed by aligning clustered headlines that are scraped from a news aggregator site. To generate sentential paraphrases we use a standard phrase-based machine translation (PBMT) framework modified with a re-ranking component (henceforth PBMT-R). We demonstrate this approach for Dutch and English and evaluate by using human judgements collected from 76 participants. The judgments are compared to two automatic machine translation evaluation metrics. We observe that as the paraphrases deviate more from the source sentence, the performance of the PBMT-R system degrades less than that of the word substitution baseline system.

Aligning Predicate-Argument Structures for Paraphrase Fragment Extraction

Michaela Regneri, Rui Wang and Manfred Pinkal

Paraphrases and paraphrasing algorithms have been found of great importance in various natural language processing tasks. While most paraphrase extraction approaches extract equivalent sentences, sentences are an inconvenient unit for further processing, because they are too specific, and often not exact paraphrases. Paraphrase fragment extraction is a technique that post-processes sentential paraphrases and prunes them to more convenient phrase-level units. We present a new approach that uses semantic roles to extract paraphrase fragments from sentence pairs that share semantic content to varying degrees, including full paraphrases. In contrast to previous systems, the use of semantic parses allows for extracting paraphrases with high wording variance and different syntactic categories. The approach is tested on four different input corpora and compared to two previous systems for extracting paraphrase fragments. Our system finds three times as many good paraphrase fragments per sentence pair as the baselines, and at the same time outputs 30% fewer unrelated fragment pairs.

O39 - Information Extraction (2)

Friday, May 30, 9:45

Chairperson: **Eduard Hovy** Oral Session

Combining Dependency Information and Generalization in a Pattern-based Approach to the Classification of Lexical-Semantic Relation Instances

Silvia Necsulescu, Sara Mendes and Núria Bel

This work addresses the classification of word pairs as instances of lexical-semantic relations. The classification is approached by leveraging patterns of co-occurrence contexts from corpus data. The significance of using dependency information, of augmenting the set of dependency paths provided to the system, and of generalizing patterns using part-of-speech information for the classification of lexical-semantic relation instances is analyzed. Results show that dependency information is decisive to achieve better results both in precision and recall, while generalizing features based on dependency information by replacing lexical forms with their part-of-speech increases the coverage of classification systems. Our experiments also make apparent that approaches based on the context where word pairs co-occur are upper-bound-limited by the times these appear in the same sentence. Therefore strategies to use information across sentence boundaries are necessary.

Corpus and Method for Identifying Citations in Non-Academic Text

Yifan He and Adam Meyers

We attempt to identify citations in non-academic text such as patents. Unlike academic articles which often provide bibliographies and follow consistent citation styles, non-academic text cites scientific research in a more ad-hoc manner. We manually annotate citations in 50 patents, train a CRF classifier to find new citations, and apply a reranker to incorporate non-local information. Our best system achieves 0.83 F-score on 5-fold cross validation.

Language Resources and Annotation Tools for Cross-Sentence Relation Extraction

Sebastian Krause, Hong Li, Feiyu Xu, Hans Uszkoreit, Robert Hummel and Luise Spielhagen

In this paper, we present a novel combination of two types of language resources dedicated to the detection of relevant relations (RE) such as events or facts across sentence boundaries. One of the two resources is the sar-graph, which aggregates for each target relation ten thousands of linguistic patterns of semantically associated relations that signal instances of the target relation (Uszkoreit and Xu, 2013). These have been learned from the Web by intra-sentence pattern extraction (Krause et al., 2012) and after semantic filtering and enriching have been automatically combined into a single graph. The other resource is cockrACE, a specially annotated corpus for the training and evaluation of cross-sentence RE. By employing our powerful annotation tool Recon, annotators mark selected entities and relations (including events), coreference relations among these entities and events, and also terms that are semantically related to the relevant relations and events. This paper describes how the two resources are created and how they complement each other.

Creative Language Explorations through a high-Expressivity N-grams Query Language

Carlo Strapparava, Lorenzo Gatti, Marco Guerini and Oliviero Stock

In computation linguistics a combination of syntagmatic and paradigmatic features is often exploited. While the first aspects are typically managed by information present in large n-gram databases, domain and ontological aspects are more properly modeled by lexical ontologies such as WordNet and semantic similarity spaces. This interconnection is even stricter when we are dealing with creative language phenomena, such as metaphors, prototypical properties, puns generation, hyperbolae and other rhetorical phenomena. This paper describes a way to focus on and accomplish some of these tasks by exploiting NgramQuery, a generalized query language on Google N-gram database. The expressiveness of this query language is boosted by plugging semantic similarity acquired both from corpora (e.g. LSA) and from WordNet, also integrating operators for phonetics and sentiment analysis. The paper reports a number of examples of usage in some creative language tasks.

Semantic Technologies for Querying Linguistic Annotations: An Experiment Focusing on Graph-Structured Data

Milen Kouylekov and Stephan Oepen

With growing interest in the creation and search of linguistic annotations that form general graphs (in contrast to formally simpler, rooted trees), there also is an increased need for infrastructures that support the exploration of such representations, for example logical-form meaning representations or semantic dependency graphs. In this work, we heavily lean on semantic technologies and in particular the data model of the Resource Description Framework (RDF) to represent, store, and efficiently query very large collections of text annotated with graph-structured representations of sentence meaning.

O40 - Lexicons and Ontologies

Friday, May 30, 9:45

Chairperson: **Gudrun Magnusdottir** Oral Session

From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers

Ingrid Falk, Delphine Bernhard and Christophe Gérard

In this paper we present a statistical machine learning approach to formal neologism detection going some way beyond the use of exclusion lists. We explore the impact of three groups of features: form related, morpho-lexical and thematic features. The latter type of features has not yet been used in this kind of application and represents a way to access the semantic context of new words. The results suggest that form related features are helpful at the overall classification task, while morpho-lexical and thematic features better single out true neologisms.

How to Construct a Multi-Lingual Domain Ontology

Nitsan Chrizman and Alon Itai

The research focuses on automatic construction of multi-lingual domain-ontologies, i.e., creating a DAG (directed acyclic graph) consisting of concepts relating to a specific domain and the

relations between them. The domain example on which the research performed is "Organized Crime". The contribution of the work is the investigation of and comparison between several data sources and methods to create multi-lingual ontologies. The first subtask was to extract the domain's concepts. The best source turned out to be Wikepedia's articles that are under the catgegory. The second task was to create an English ontology, i.e., the relationships between the concepts. Again the relationships between concepts and the hierarchy were derived from Wikipedia. The final task was to create an ontology for a language with far fewer resources (Hebrew). The task was accomplished by deriving the concepts from the Hebrew Wikepedia and assessing their relevance and the relationships between them from the English ontology.

Ruled-based, Interlingual Motivated Mapping of plWordNet onto SUMO Ontology

Paweł Kędzia and Maciej Piasecki

In this paper we study a rule-based approach to mapping plWordNet onto SUMO Upper Ontology on the basis of the already existing mappings: plWordNet – the Princeton WordNet – SUMO. Information acquired from the inter-lingual relations between plWordNet and Princeton WordNet and the relations between Princeton WordNet and SUMO ontology are used in the proposed rules. Several mapping rules together with the matching examples are presented. The automated mapping results were evaluated in two steps, (i) we automatically checked formal correctness of the mappings for the pairs of plWordNet synset and SUMO concept, (ii) a subset of 160 mapping examples was manually checked by two+one linguists. We analyzed types of the mapping errors and their causes. The proposed rules expressed very high precision, especially when the errors in the resources are taken into account. Because both wordnets were constructed independently and as a result the obtained rules are not trivial and they reveal the differences between both wordnets and both languages.

Augmenting English Adjective Senses with Supersenses

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui and Chris Dyer

We develop a supersense taxonomy for adjectives, based on that of GermaNet, and apply it to English adjectives in WordNet using human annotation and supervised classification. Results show that accuracy for automatic adjective type classification is high, but synsets are considerably more difficult to classify, even for trained human annotators. We release the manually annotated data, the classifier, and the induced supersense labeling of 12,304 WordNet adjective synsets.

Choosing which to Use? A Study of Distributional Models for Nominal Lexical Semantic Classification

Lauren Romeo, Gianluca Lebani, Núria Bel and Alessandro Lenci

This paper empirically evaluates the performances of different state-of-the-art distributional models in a nominal lexical semantic classification task. We consider models that exploit various types of distributional features, which thereby provide different representations of nominal behavior in context. The experiments presented in this work demonstrate the advantages and disadvantages of each model considered. This analysis also considers a combined strategy that we found to be capable of leveraging the bottlenecks of each model, especially when large robust data is not available.

P50 - Crowdsourcing

Friday, May 30, 9:45

Chairperson: **Cristina Vertan** Poster Session

A Crowdsourcing Smartphone Application for Swiss German: Putting Language Documentation in the Hands of the Users

Jean-Philippe Goldman, Adrian Leeman, Marie-José Kolly, Ingrid Hove, Ibrahim Almajai, Volker Dellwo and Steven Moran

This contribution describes an on-going projects a smartphone application called Voice Äpp, which is a follow-up of a previous application called Dialäkt Äpp. The main purpose of both apps is to identify the user's Swiss German dialect on the basis of the dialectal variations of 15 words. The result is returned as one or more geographical points on a map. In Dialäkt Äpp, launched in 2013, the user provides his or her own pronunciation through buttons, while the Voice Äpp, currently in development, asks users to pronounce the word and uses speech recognition techniques to identify the variants and localize the user. This second app is more challenging from a technical point of view but nevertheless recovers the nature of dialect variation of spoken language. Besides, the Voice Äpp takes its users on a journey in which they explore the individuality of their own voices, answering questions such as: How high is my voice? How fast do I speak? Do I speak faster than users in the neighbouring city?

TagNText: a Parallel Corpus for the Induction of Resource-specific non-Taxonomical Relations from Tagged Images

Theodosia Togia and Ann Copestake

When producing textual descriptions, humans express propositions regarding an object; but what do they express when

annotating a document with simple tags? To answer this question, we have studied what users of tagging systems would have said if they were to describe a resource with fully fledged text. In particular, our work attempts to answer the following questions: if users were to use full descriptions, would their current tags be words present in these hypothetical sentences? If yes, what kind of language would connect these words? Such questions, although central to the problem of extracting binary relations between tags, have been sidestepped in the existing literature, which has focused on a small subset of possible inter-tag relations, namely hierarchical ones (e.g. "car" –is-a– "vehicle"), as opposed to non-taxonomical relations (e.g. "woman" –wears– "hat"). TagNText is the first attempt to construct a parallel corpus of tags and textual descriptions with respect to particular resources. The corpus provides enough data for the researcher to gain an insight into the nature of underlying relations, as well as the tools and methodology for constructing larger-scale parallel corpora that can aid non-taxonomical relation extraction.

Crowdsourcing for Evaluating Machine Translation Quality

Shinsuke Goto, Donghui Lin and Toru Ishida

The recent popularity of machine translation has increased the demand for the evaluation of translations. However, the traditional evaluation approach, manual checking by a bilingual professional, is too expensive and too slow. In this study, we confirm the feasibility of crowdsourcing by analyzing the accuracy of crowdsourcing translation evaluations. We compare crowdsourcing scores to professional scores with regard to three metrics: translation-score, sentence-score, and system-score. A Chinese to English translation evaluation task was designed using around the NTCIR-9 PATENT parallel corpus with the goal being 5-range evaluations of adequacy and fluency. The experiment shows that the average score of crowdsource workers well matches professional evaluation results. The system-score comparison strongly indicates that crowdsourcing can be used to find the best translation system given the input of 10 source sentence.

NOMAD: Linguistic Resources and Tools Aimed at Policy Formulation and Validation

George Kiomourtzis, George Giannakopoulos, Georgios Petasis, Pythagoras Karampiperis and Vangelis Karkaletsis

The NOMAD project (Policy Formulation and Validation through non Moderated Crowd-sourcing) is a project that supports policy making, by providing rich, actionable information related to how citizens perceive different policies. NOMAD automatically analyzes citizen contributions to the informal web (e.g. forums, social networks, blogs, newsgroups and wikis) using a variety of tools. These tools comprise text retrieval, topic classification, argument detection and sentiment analysis, as well as argument summarization. NOMAD provides decision-makers with a full arsenal of solutions starting from describing a domain and a policy to applying content search and acquisition, categorization and visualization. These solutions work in a collaborative manner in the policy-making arena. NOMAD, thus, embeds editing, analysis and visualization technologies into a concrete framework, applicable in a variety of policy-making and decision support settings In this paper we provide an overview of the linguistic tools and resources of NOMAD.

sloWCrowd: a Crowdsourcing Tool for Lexicographic Tasks

Darja Fišer, Aleš Tavčar and Tomaž Erjavec

The paper presents sloWCrowd, a simple tool developed to facilitate crowdsourcing lexicographic tasks, such as error correction in automatically generated wordnets and semantic annotation of corpora. The tool is open-source, language-independent and can be adapted to a broad range of crowdsourcing tasks. Since volunteers who participate in our crowdsourcing tasks are not trained lexicographers, the tool has been designed to obtain multiple answers to the same question and compute the majority vote, making sure individual unreliable answers are discarded. We also make sure unreliable volunteers, who systematically provide unreliable answers, are not taken into account. This is achieved by measuring their accuracy against a gold standard, the questions from which are posed to the annotators on a regular basis in between the real question. We tested the tool in an extensive crowdsourcing task, i.e. error correction of the Slovene wordnet, the results of which are encouraging, motivating us to use the tool in other annotation tasks in the future as well.

P51 - Emotion Recognition and Generation

Friday, May 30, 9:45

Chairperson: **Patrick Paroubek** Poster Session

Comparison of Gender- and Speaker-adaptive Emotion Recognition

Maxim Sidorov, Stefan Ultes and Alexander Schmitt

Deriving the emotion of a human speaker is a hard task, especially if only the audio stream is taken into account. While state-of-the-art approaches already provide good results, adaptive methods have been proposed in order to further improve the recognition accuracy. A recent approach is to add characteristics of the speaker, e.g., the gender of the speaker. In this contribution,

we argue that adding information unique for each speaker, i.e., by using speaker identification techniques, improves emotion recognition simply by adding this additional information to the feature vector of the statistical classification algorithm. Moreover, we compare this approach to emotion recognition adding only the speaker gender being a non-unique speaker attribute. We justify this by performing adaptive emotion recognition using both gender and speaker information on four different corpora of different languages containing acted and non-acted speech. The final results show that adding speaker information significantly outperforms both adding gender information and solely using a generic speaker-independent approach.

Speech-based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm

Maxim Sidorov, Christina Brester, Wolfgang Minker and Eugene Semenkin

Automated emotion recognition has a number of applications in Interactive Voice Response systems, call centers, etc. While employing existing feature sets and methods for automated emotion recognition has already achieved reasonable results, there is still a lot to do for improvement. Meanwhile, an optimal feature set, which should be used to represent speech signals for performing speech-based emotion recognition techniques, is still an open question. In our research, we tried to figure out the most essential features with self-adaptive multi-objective genetic algorithm as a feature selection technique and a probabilistic neural network as a classifier. The proposed approach was evaluated using a number of multi-languages databases (English, German), which were represented by 37- and 384-dimensional feature sets. According to the obtained results, the developed technique allows to increase the emotion recognition performance by up to 26.08 relative improvement in accuracy. Moreover, emotion recognition performance scores for all applied databases are improved.

Emilya: Emotional Body Expression in Daily Actions Database

Nesrine Fourati and Catherine Pelachaud

The studies of bodily expression of emotion have been so far mostly focused on body movement patterns associated with emotional expression. Recently, there is an increasing interest on the expression of emotion in daily actions, called also non-emblematic movements (such as walking or knocking at the door). Previous studies were based on database limited to a small range of movement tasks or emotional states. In this paper, we describe our new database of emotional body expression in daily actions, where 11 actors express 8 emotions in 7 actions. We use motion capture technology to record body movements, but we recorded as well synchronized audio-visual data to enlarge the use of the database for different research purposes. We investigate also the matching between the expressed emotions and the perceived ones through a perceptive study. The first results of this study are discussed in this paper.

TexAFon 2.0: a Text Processing Tool for the Generation of Expressive Speech in TTS Applications

Juan-María Garrido, Yesika Laplaza, Benjamin Kolz and Miquel Cornudella

This paper presents TexAfon 2.0, an improved version of the text processing tool TexAFon, specially oriented to the generation of synthetic speech with expressive content. TexAFon is a text processing module in Catalan and Spanish for TTS systems, which performs all the typical tasks needed for the generation of synthetic speech from text: sentence detection, pre-processing, phonetic transcription, syllabication, prosodic segmentation and stress prediction. These improvements include a new normalisation module for the standardisation on chat text in Spanish, a module for the detection of the expressed emotions in the input text, and a module for the automatic detection of the intended speech acts, which are briefly described in the paper. The results of the evaluations carried out for each module are also presented.

EMOVO Corpus: an Italian Emotional Speech Database

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni and Massimiliano Todisco

This article describes the first emotional corpus, named EMOVO, applicable to Italian language,. It is a database built from the voices of up to 6 actors who played 14 sentences simulating 6 emotional states (disgust, fear, anger, joy, surprise, sadness) plus the neutral state. These emotions are the well-known Big Six found in most of the literature related to emotional speech. The recordings were made with professional equipment in the Fondazione Ugo Bordoni laboratories. The paper also describes a subjective validation test of the corpus, based on emotion-discrimination of two sentences carried out by two different groups of 24 listeners. The test was successful because it yielded an overall recognition accuracy of 80%. It is observed that emotions less easy to recognize are joy and disgust, whereas the most easy to detect are anger, sadness and the neutral state.

A Database of Full Body Virtual Interactions Annotated with Expressivity Scores

Demulier Virginie, Elisabetta Bevacqua, Florian Focone, Tom Giraud, Pamela Carreno, Brice Isableu, Sylvie Gibet, Pierre de Loor and Jean-Claude Martin

Recent technologies enable the exploitation of full body expressions in applications such as interactive arts but are still limited in terms of dyadic subtle interaction patterns. Our project aims at full body expressive interactions between a user and an autonomous virtual agent. The currently available databases do not contain full body expressivity and interaction patterns via avatars. In this paper, we describe a protocol defined to collect a database to study expressive full-body dyadic interactions. We detail the coding scheme for manually annotating the collected videos. Reliability measures for global annotations of expressivity and interaction are also provided.

Annotating Events in an Emotion Corpus

Sophia Lee, Shoushan Li and Chu-Ren Huang

This paper presents the development of a Chinese event-based emotion corpus. It specifically describes the corpus design, collection and annotation. The proposed annotation scheme provides a consistent way of identifying some emotion-associated events (namely pre-events and post-events). Corpus data show that there are significant interactions between emotions and pre-events as well as that of between emotion and post-events. We believe that emotion as a pivot event underlies an innovative approach towards a linguistic model of emotion as well as automatic emotion detection and classification.

P52 - Linked Data

Friday, May 30, 9:45

Chairperson: **John Philip McCrae** Poster Session

Towards Linked Hypernyms Dataset 2.0: Complementing DBpedia with Hypernym Discovery

Tomáš Kliegr and Ondřej Zamazal

This paper presents a statistical type inference algorithm for ontology alignment, which assigns DBpedia entities with a new type (class). To infer types for a specific entity, the algorithm first identifies types that co-occur with the type the entity already has, and subsequently prunes the set of candidates for the most confident one. The algorithm has one parameter for balancing specificity/reliability of the resulting type selection. The proposed algorithm is used to complement the types in the LHD dataset, which is RDF knowledge base populated by identifying hypernyms from the free text of Wikipedia articles. The majority of types assigned to entities in LHD 1.0 are DBpedia resources. Through the statistical type inference, the number of entities with a type from DBpedia Ontology is increased significantly: by 750 thousand entities for the English dataset, 200.000 for Dutch and 440.000 for German. The accuracy of the inferred types is at 0.65 for English (as compared to 0.86 for LHD 1.0 types). A byproduct of the mapping process is a set of 11.000 mappings from DBpedia resources to DBpedia Ontology classes with associated confidence values. The number of the resulting mappings is an order of magnitude larger than what can be achieved with standard ontology alignment algorithms (Falcon, LogMapLt and YAM++), which do not utilize the type co-occurrence information. The presented algorithm is not restricted to the LHD dataset, it can be used to address generic type inference problems in presence of class membership information for a large number of instances.

NIF4OGGD - NLP Interchange Format for Open German Governmental Data

Mohamed Sherif, Sandro Coelho, Ricardo Usbeck, Sebastian Hellmann, Jens Lehmann, Martin Brümmer and Andreas Both

In the last couple of years the amount of structured open government data has increased significantly. Already now, citizens are able to leverage the advantages of open data through increased transparency and better opportunities to take part in governmental decision making processes. Our approach increases the interoperability of existing but distributed open governmental datasets by converting them to the RDF-based NLP Interchange Format (NIF). Furthermore, we integrate the converted data into a geodata store and present a user interface for querying this data via a keyword-based search. The language resource generated in this project is publicly available for download and also via a dedicated SPARQL endpoint.

N^3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format

Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber and Andreas Both

Extracting Linked Data following the Semantic Web principle from unstructured sources has become a key challenge for scientific research. Named Entity Recognition and Disambiguation are two basic operations in this extraction process. One step towards the realization of the Semantic Web vision and the development of highly accurate tools is the availability of data for validating the quality of processes for Named Entity Recognition and Disambiguation as well as for

algorithm tuning. This article presents three novel, manually curated and annotated corpora (N3). All of them are based on a free license and stored in the NLP Interchange Format to leverage the Linked Data character of our datasets.

The LRE Map disclosed

Riccardo del Gratta, Gabriella Pardelli and Sara Goggi

This paper describes a serialization of the LRE Map database according to the RDF model. Due to the peculiar nature of the LRE Map, many ontologies are necessary to model the map in RDF, including newly created and reused ontologies. The importance of having the LRE Map in RDF and its connections to other open resources is also addressed.

Accommodations in Tuscany as Linked Data

Clara Bacciu, Angelica Lo Duca, Andrea Marchetti and Maurizio Tesconi

The OpeNER Linked Dataset (OLD) contains 19.140 entries about accommodations in Tuscany (Italy). For each accommodation, it describes the type, e.g. hotel, bed and breakfast, hostel, camping etc., and other useful information, such as a short description, the Web address, its location and the features it provides. OLD is the linked data version of the open dataset provided by Fondazione Sistema Toscana, the representative system for tourism in Tuscany. In addition, to the original dataset, OLD provides also the link of each accommodation to the most common social media (Facebook, Foursquare, Google Places and Booking). OLD exploits three common ontologies of the accommodation domain: Acco, Hontology and GoodRelations. The idea is to provide a flexible dataset, which speaks more than one ontology. OLD is available as a SPARQL node and is released under the Creative Commons release. Finally, OLD is developed within the OpeNER European project, which aims at building a set of ready to use tools to recognize and disambiguate entity mentions and perform sentiment analysis and opinion detection on texts. Within the project, OLD provides a named entity repository for entity disambiguation.

Global Intelligent Content: Active Curation of Language Resources using Linked Data

David Lewis, Rob Brennan, Leroy Finn, Dominic Jones, Alan Meehan, Declan O'sullivan, Sebastian Hellmann and Felix Sasaki

As language resources start to become available in linked data formats, it becomes relevant to consider how linked data interoperability can play a role in active language processing workflows as well as for more static language resource publishing. This paper proposes that linked data may have a valuable role to play in tracking the use and generation of language resources in such workflows in order to assess and improve the performance of the language technologies that use the resources, based on feedback from the human involvement typically required within such processes. We refer to this as Active Curation of the language resources, since it is performed systematically over language processing workflows to continuously improve the quality of the resource in specific applications, rather than via dedicated curation steps. We use modern localisation workflows, i.e. assisted by machine translation and text analytics services, to explain how linked data can support such active curation. By referencing how a suitable linked data vocabulary can be assembled by combining existing linked data vocabularies and meta-data from other multilingual content processing annotations and tool exchange standards we aim to demonstrate the relative ease with which active curation can be deployed more broadly.

P53 - Machine Translation

Friday, May 30, 9:45

Chairperson: **Mikel Forcada** Poster Session

HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation

Ondrej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Stranak, Vit Suchomel, Aleš Tamchyna and Daniel Zeman

We present HindEnCorp, a parallel corpus of Hindi and English, and HindMonoCorp, a monolingual corpus of Hindi in their release version 0.5. Both corpora were collected from web sources and preprocessed primarily for the training of statistical machine translation systems. HindEnCorp consists of 274k parallel sentences (3.9 million Hindi and 3.8 million English tokens). HindMonoCorp amounts to 787 million tokens in 44 million sentences. Both the corpora are freely available for non-commercial research and their preliminary release has been used by numerous participants of the WMT 2014 shared translation task.

Online Optimisation of Log-linear Weights in Interactive Machine Translation

Mara Chinea-Rios, Germán Sanchis Trilles, Daniel Daniel Ortiz-Martínez and Francisco Casacuberta

Whenever the quality provided by a machine translation system is not enough, a human expert is required to correct the sentences provided by the machine translation system. In such a setup, it is crucial that the system is able to learn from the errors

that have already been corrected. In this paper, we analyse the applicability of discriminative ridge regression for learning the log-linear weights of a state-of-the-art machine translation system underlying an interactive machine translation framework, with encouraging results.

An Efficient and User-friendly Tool for Machine Translation Quality Estimation

Kashif Shah, Marco Turchi and Lucia Specia

We present a new version of QUEST – an open source framework for machine translation quality estimation – which brings a number of improvements: (i) it provides a Web interface and functionalities such that non-expert users, e.g. translators or lay-users of machine translations, can get quality predictions (or internal features of the framework) for translations without having to install the toolkit, obtain resources or build prediction models; (ii) it significantly improves over the previous runtime performance by keeping resources (such as language models) in memory; (iii) it provides an option for users to submit the source text only and automatically obtain translations from Bing Translator; (iv) it provides a ranking of multiple translations submitted by users for each source text according to their estimated quality. We exemplify the use of this new version through some experiments with the framework.

Word Alignment-based Reordering of Source Chunks in PB-SMT

Santanu Pal, Sudip Kumar Naskar and Sivaji Bandyopadhyay

Reordering poses a big challenge in statistical machine translation between distant language pairs. The paper presents how reordering between distant language pairs can be handled efficiently in phrase-based statistical machine translation. The problem of reordering between distant languages has been approached with prior reordering of the source text at chunk level to simulate the target language ordering. Prior reordering of the source chunks is performed in the present work by following the target word order suggested by word alignment. The testset is reordered using monolingual MT trained on source and reordered source. This approach of prior reordering of the source chunks was compared with pre-ordering of source words based on word alignments and the traditional approach of prior source reordering based on language-pair specific reordering rules. The effects of these reordering approaches were studied on an English–Bengali translation task, a language pair with different word order. From the experimental results it was found that word alignment based reordering of the source chunks is more effective than the other reordering approaches, and it produces statistically significant improvements over the baseline system on BLEU. On manual inspection we found significant improvements in terms of word alignments.

Comparing the Quality of Focused Crawlers and of the Translation Resources Obtained from them

Bruno Laranjeira, Viviane Moreira, Aline Villavicencio, Carlos Ramisch and Maria José Finatto

Comparable corpora have been used as an alternative for parallel corpora as resources for computational tasks that involve domain-specific natural language processing. One way to gather documents related to a specific topic of interest is to traverse a portion of the web graph in a targeted way, using focused crawling algorithms. In this paper, we compare several focused crawling algorithms using them to collect comparable corpora on a specific domain. Then, we compare the evaluation of the focused crawling algorithms to the performance of linguistic processes executed after training with the corresponding generated corpora. Also, we propose a novel approach for focused crawling, exploiting the expressive power of multiword expressions.

N-gram Counts and Language Models from the Common Crawl

Christian Buck, Kenneth Heafield and Bas van Ooyen

We contribute 5-gram counts and language models trained on the Common Crawl corpus, a collection over 9 billion web pages. This release improves upon the Google n-gram counts in two key ways: the inclusion of low-count entries and deduplication to reduce boilerplate. By preserving singletons, we were able to use Kneser-Ney smoothing to build large language models. This paper describes how the corpus was processed with emphasis on the problems that arise in working with data at this scale. Our unpruned Kneser-Ney English 5-gram language model, built on 975 billion deduplicated tokens, contains over 500 billion unique n-grams. We show gains of 0.5-1.4 BLEU by using large language models to translate into various languages.

A Corpus of Machine Translation Errors Extracted from Translation Students Exercises

Guillaume Wisniewski, Natalie Kübler and François Yvon

In this paper, we present a freely available corpus of automatic translations accompanied with post-edited versions, annotated with labels identifying the different kinds of errors made by the MT system. These data have been extracted from translation students exercises that have been corrected by a senior professor.

This corpus can be useful for training quality estimation tools and for analyzing the types of errors made MT system.

Pre-ordering of Phrase-based Machine Translation Input in Translation Workflow

Alexandru Ceausu and Sabine Hunsicker

Word reordering is a difficult task for decoders when the languages involved have a significant difference in syntax. Phrase-based statistical machine translation (PBSMT), preferred in commercial settings due to its maturity, is particularly prone to errors in long range reordering. Source sentence pre-ordering, as a pre-processing step before PBSMT, proved to be an efficient solution that can be achieved using limited resources. We propose a dependency-based pre-ordering model with parameters optimized using a reordering score to pre-order the source sentence. The source sentence is then translated using an existing phrase-based system. The proposed solution is very simple to implement. It uses a hierarchical phrase-based statistical machine translation system (HPBSMT) for pre-ordering, combined with a PBSMT system for the actual translation. We show that the system can provide alternate translations of less post-editing effort in a translation workflow with German as the source language.

A Wikipedia-based Corpus for Contextualized Machine Translation

Jennifer Drexler, Pushpendre Rastogi, Jacqueline Aguilar, Benjamin van Durme and Matt Post

We describe a corpus for target-contextualized machine translation (MT), where the task is to improve the translation of source documents using language models built over presumably related documents in the target language. The idea presumes a situation where most of the information about a topic is in a foreign language, yet some related target-language information is known to exist. Our corpus comprises a set of curated English Wikipedia articles describing news events, along with (i) their Spanish counterparts and (ii) some of the Spanish source articles cited within them. In experiments, we translated these Spanish documents, treating the English articles as target-side context, and evaluate the effect on translation quality when including target-side language models built over this English context and interpolated with other, separately-derived language model data. We find that even under this simplistic baseline approach, we achieve significant improvements as measured by BLEU score.

P54 - Multimodality

Friday, May 30, 9:45

Chairperson: **Kristiina Jokinen** Poster Session

Transfer Learning of Feedback Head Expressions in Danish and Polish Comparable Multimodal Corpora

Costanza Navarretta and Magdalena Lis

The paper is an investigation of the reusability of the annotations of head movements in a corpus in a language to predict the feedback functions of head movements in a comparable corpus in another language. The two corpora consist of naturally occurring triadic conversations in Danish and Polish, which were annotated according to the same scheme. The intersection of common annotation features was used in the experiments. A Naïve Bayes classifier was trained on the annotations of a corpus and tested on the annotations of the other corpus. Training and test datasets were then reversed and the experiments repeated. The results show that the classifier identifies more feedback behaviours than the majority baseline in both cases and the improvements are significant. The performance of the classifier decreases significantly compared with the results obtained when training and test data belong to the same corpus. Annotating multimodal data is resource consuming, thus the results are promising. However, they also confirm preceding studies that have identified both similarities and differences in the use of feedback head movements in different languages. Since our datasets are small and only regard a communicative behaviour in two languages, the experiments should be tested on more data types.

Improving the Exploitation of Linguistic Annotations in ELAN

Onno Crasborn and Han Sloetjes

This paper discusses some improvements in recent and planned versions of the multimodal annotation tool ELAN, which are targeted at improving the usability of annotated files. Increased support for multilingual documents is provided, by allowing for multilingual vocabularies and by specifying a language per document, annotation layer (tier) or annotation. In addition, improvements in the search possibilities and the display of the results have been implemented, which are especially relevant in the interpretation of the results of complex multi-tier searches.

Web-imageability of the Behavioral Features of Basic-level Concepts

Yoshihiko Hayashi

The recent research direction toward multimodal semantic representation would be further advanced, if we could have a

machinery to collect adequate images from the Web, given a target concept. With this motivation, this paper particularly investigates into the Web imageabilities of the behavioral features (e.g. "beaver builds dams") of a basic-level concept (beaver). The term Web-imageability denotes how adequately the images acquired from the Web deliver the intended meaning of a complex concept. The primary contributions made in this paper are twofold: (1) "beaver building dams"-type queries can better yield relevant Web images, suggesting that the present participle form ("-ing" form) of a verb ("building"), as a query component, is more effective than the base form; (2) the behaviors taken by animate beings are likely to be more depicted on the Web, particularly if the behaviors are, in a sense, inherent to animate beings (e.g.,motion, consumption), while the creation-type behaviors of inanimate beings are not. The paper further analyzes linguistic annotations that were independently given to some of the images, and discusses an aspect of the semantic gap between image and language.

A Model to Generate Adaptive Multimodal Job Interviews with a Virtual Recruiter

Zoraida Callejas, Brian Ravenet, Magalie Ochs and Catherine Pelachaud

This paper presents an adaptive model of multimodal social behavior for embodied conversational agents. The context of this research is the training of youngsters for job interviews in a serious game where the agent plays the role of a virtual recruiter. With the proposed model the agent is able to adapt its social behavior according to the anxiety level of the trainee and a predefined difficulty level of the game. This information is used to select the objective of the system (to challenge or comfort the user), which is achieved by selecting the complexity of the next question posed and the agent's verbal and non-verbal behavior. We have carried out a perceptive study that shows that the multimodal behavior of an agent implementing our model successfully conveys the expected social attitudes.

A Multimodal Interpreter for 3D Visualization and Animation of Verbal Concepts

Coline Claude-Lachenaud, Eric Charton, Benoit Ozell and Michel Gagnon

We present an algorithm intended to visually represent the sense of verb related to an object described in a text sequence, as a movement in 3D space. We describe a specific semantic analyzer, based on a standard verbal ontology, dedicated to the interpretation of action verbs as spatial actions. Using this analyzer, our system build a generic 3D graphical path for verbal concepts allowing space representation, listed as SelfMotion concepts in the FrameNet ontology project. The object movement is build by first extracting the words and enriching them with the semantic analyzer. Then, weight tables, necessary to obtain characteristics values (orientation, shape, trajectory...) for the verb are used in order to get a 3D path, as realist as possible. The weight tables were created to make parallel between features defined for SelfMotion verbal concept (some provided by FrameNet, other determined during the project) and values used in the final algorithm used to create 3D moving representations from input text. We evaluate our analyzer on a corpus of short sentences and presents our results.

New Functions for a Multipurpose Multimodal Tool for Phonetic and Linguistic Analysis of Very Large Speech Corpora

Philippe MARTIN

The increased interest for linguistic analysis of spontaneous (i.e. non-prepared) speech from various points of view (semantic, syntactic, morphologic, phonologic and intonative) lead to the development of ever more sophisticated dedicated tools. Although the software Praat emerged as the de facto standard for the analysis of spoken data, its use for intonation studies is often felt as not optimal, notably for its limited capabilities in fundamental frequency tracking. This paper presents some of the recently implemented features of the software WinPitch, developed with the analysis of spontaneous speech in mind (and notably for the C-ORAL-ROM project 10 years ago). Among many features, WinPitch includes a set of multiple pitch tracking algorithms aimed to obtain reliable pitch curves in adverse recording conditions (echo, filtering, poor signal to noise ratio, etc.). Others functions of WinPitch incorporate an integrated concordancer, an on the fly text-sound aligner, and routines for EEG analysis.

Smile and Laughter in Human-Machine Interaction: a Study of Engagement

Mariette Soury and Laurence Devillers

This article presents a corpus featuring adults playing games in interaction with machine trying to induce laugh. This corpus was collected during Interspeech 2013 in Lyon to study behavioral differences correlated to different personalities and cultures. We first present the collection protocol, then the corpus obtained and finally different quantitative and qualitative measures. Smiles and laughs are types of affect bursts which are defined as short emotional "non-speech" expressions. Here we correlate smile and laugh with personality traits and cultural background. Our final objective is to propose a measure of engagement deduced from those affect bursts.

ALICO: a Multimodal Corpus for the Study of Active Listening

Hendrik Buschmeier, Zofia Malisz, Joanna Skubisz, Marcin Wlodarczak, Ipke Wachsmuth, Stefan Kopp and Petra Wagner

The Active Listening Corpus (ALICO) is a multimodal database of spontaneous dyadic conversations with diverse speech and gestural annotations of both dialogue partners. The annotations consist of short feedback expression transcription with corresponding communicative function interpretation as well as segmentation of interpausal units, words, rhythmic prominence intervals and vowel-to-vowel intervals. Additionally, ALICO contains head gesture annotation of both interlocutors. The corpus contributes to research on spontaneous human–human interaction, on functional relations between modalities, and timing variability in dialogue. It also provides data that differentiates between distracted and attentive listeners. We describe the main characteristics of the corpus and present the most important results obtained from analyses in recent years.

The DWAN Framework: Application of a Web Annotation Framework for the General Humanities to the Domain of Language Resources

Przemyslaw Lenkiewicz, Olha Shkaravska, Twan Goosen, Daan Broeder, Menzo Windhouwer, Stephanie Roth and Olof Olsson

Researchers share large amounts of digital resources, which offer new chances for cooperation. Collaborative annotation systems are meant to support this. Often these systems are targeted at a specific task or domain, e.g., annotation of a corpus. The DWAN framework for web annotation is generic and can support a wide range of tasks and domains. A key feature of the framework is its support for caching representations of the annotated resource. This allows showing the context of the annotation even if the resource has changed or has been removed. The paper describes the design and implementation of the framework. Use cases provided by researchers are well in line with the key characteristics of the DWAN annotation framework.

Co-Training for Classification of Live or Studio Music Recordings

Nicolas Auguin and Pascale Fung

The fast-spreading development of online streaming services has enabled people from all over the world to listen to music. However, it is not always straightforward for a given user to find the "right" song version he or she is looking for. As streaming services may be affected by the potential dissatisfaction among their customers, the quality of songs and the presence of tags (or labels) associated with songs returned to the users are very important. Thus, the need for precise and reliable metadata becomes paramount. In this work, we are particularly interested in distinguishing between live and studio versions of songs. Specifically, we tackle the problem in the case where very little-annotated training data are available, and demonstrate how an original co-training algorithm in a semi-supervised setting can alleviate the problem of data scarcity to successfully discriminate between live and studio music recordings.

P55 - Ontologies

Friday, May 30, 9:45

Chairperson: **Monica Monachini** Poster Session

Efficient Reuse of Structured and Unstructured Resources for Ontology Population

Chetana Gavankar, Ashish Kulkarni and Ganesh Ramakrishnan

We study the problem of ontology population for a domain ontology and present solutions based on semi-automatic techniques. A domain ontology for an organization, often consists of classes whose instances are either specific to, or independent of the organization. E.g. in an academic domain ontology, classes like Professor, Department could be organization (university) specific, while Conference, Programming languages are organization independent. This distinction allows us to leverage data sources both—within the organization and those in the Internet — to extract entities and populate an ontology. We propose techniques that build on those for open domain IE. Together with user input, we show through comprehensive evaluation, how these semi-automatic techniques achieve high precision. We experimented with the academic domain and built an ontology comprising of over 220 classes. Intranet documents from five universities formed our organization specific corpora and we used open domain knowledge bases like Wikipedia, Linked Open Data, and web pages from the Internet as the organization independent data sources. The populated ontology that we built for one of the universities comprised of over 75,000 instances. We adhere to the semantic web standards and tools and make the resources available in the OWL format. These could be useful for applications such as information extraction, text annotation, and information retrieval.

From Natural Language to Ontology Population in the Cultural Heritage Domain. A Computational Linguistics-based approach.

Maria Pia di Buono and Mario Monteleone

This paper presents an on-going Natural Language Processing (NLP) research based on Lexicon-Grammar (LG) and aimed at improving knowledge management of Cultural Heritage (CH) domain. We intend to demonstrate how our language formalization technique can be applied for both processing and populating a domain ontology. We also use NLP techniques for text extraction and mining to fill information gaps and improve access to cultural resources. The Linguistic Resources (LRs, i.e. electronic dictionaries) we built can be used in the structuring of effective Knowledge Management Systems (KMSs). In order to apply to Parts of Speech (POS) the classes and properties defined by the Conseil Interational des Musees (CIDOC) Conceptual Reference Model (CRM), we use Finite State Transducers/Automata (FSTs/FSA) and their variables built in the form of graphs. FSTs/FSA are also used for analysing corpora in order to retrieve recursive sentence structures, in which combinatorial and semantic constraints identify properties and denote relationship. Besides, FSTs/FSA are also used to match our electronic dictionary entries (ALUs, or Atomic Linguistic Units) to RDF subject, object and predicate (SKOS Core Vocabulary). This matching of linguistic data to RDF and their translation into SPARQL/SERQL path expressions allows the use ALUs to process natural-language queries.

A Gold Standard for CLIR evaluation in the Organic Agriculture Domain

Alessio Bosca, Matteo Casu, Matteo Dragoni and Nikolaos Marianos

We present a gold standard for the evaluation of Cross Language Information Retrieval systems in the domain of Organic Agriculture and AgroEcology. The presented resource is free to use for research purposes and it includes a collection of multilingual documents annotated with respect to a domain ontology, the ontology used for annotating the resources, a set of 48 queries in 12 languages and a gold standard with the correct resources for the proposed queries. The goal of this work consists in contributing to the research community with a resource for evaluating multilingual retrieval algorithms, with particular focus on domain adaptation strategies for "general purpose" multilingual information retrieval systems and on the effective exploitation of semantic annotations. Domain adaptation is in fact an important activity for tuning the retrieval system, reducing the ambiguities and improving the precision of information retrieval. Domain ontologies constitute a diffuse practice for defining the conceptual space of a corpus and mapping resources to specific topics and in our lab we propose as well to investigate and evaluate the impact of this information in enhancing the retrieval of contents. An initial experiment is described, giving a baseline for further research with the proposed gold standard.

VOAR: A Visual and Integrated Ontology Alignment Environment

Bernardo Severo, Cassia Trojahn and Renata Vieira

Ontology alignment is a key process for enabling interoperability between ontology-based systems in the Linked Open Data age. From two input ontologies, this process generates an alignment (set of correspondences) between them. In this paper we present VOAR, a new web-based environment for ontology alignment visualization and manipulation. Within this graphical environment, users can manually create/edit correspondences and apply a set of operations on alignments (filtering, merge, difference, etc.). VOAR allows invoking external ontology matching systems that implement a specific alignment interface, so that the generated alignments can be manipulated within the environment. Evaluating multiple alignments together against a reference one can also be carried out, using classical evaluation metrics (precision, recall and f-measure). The status of each correspondence with respect to its presence or absence in reference alignment is visually represented. Overall, the main new aspect of VOAR is the visualization and manipulation of alignments at schema level, in an integrated, visual and web-based environment.

O41 - Machine Translation

Friday, May 30, 11:45

Chairperson: **Alan Melby** Oral Session

On the Annotation of TMX Translation Memories for Advanced Leveraging in Computer-aided Translation

Mikel Forcada

The term advanced leveraging refers to extensions beyond the current usage of translation memory (TM) in computer-aided translation (CAT). One of these extensions is the ability to identify and use matches on the sub-segment level — for instance, using sub-sentential elements when segments are sentences— to help the translator when a reasonable fuzzy-matched proposal is not available; some such functionalities have started to become available in commercial CAT tools. Resources such as statistical word aligners, external machine translation systems, glossaries and term bases could be used to identify and annotate segment-level translation units at the sub-segment level, but there is

currently no single, agreed standard supporting the interchange of sub-segmental annotation of translation memories to create a richer translation resource. This paper discusses the capabilities and limitations of some current standards, envisages possible alternatives, and ends with a tentative proposal which slightly abuses (repurposes) the usage of existing elements in the TMX standard.

Manual Analysis of Structurally Informed Reordering in German-English Machine Translation

Teresa Herrmann, Jan Niehues and Alex Waibel

Word reordering is a difficult task for translation. Common automatic metrics such as BLEU have problems reflecting improvements in target language word order. However, it is a crucial aspect for humans when deciding on translation quality. This paper presents a detailed analysis of a structure-aware reordering approach applied in a German-to-English phrase-based machine translation system. We compare the translation outputs of two translation systems applying reordering rules based on parts-of-speech and syntax trees on a sentence-by-sentence basis. For each sentence-pair we examine the global translation performance and classify local changes in the translated sentences. This analysis is applied to three data sets representing different genres. While the improvement in BLEU differed substantially between the data sets, the manual evaluation showed that both global translation performance as well as individual types of improvements and degradations exhibit a similar behavior throughout the three data sets. We have observed that for 55-64% of the sentences with different translations, the translation produced using the tree-based reordering was considered to be the better translation. As intended by the investigated reordering model, most improvements are achieved by improving the position of the verb or being able to translate a verb that could not be translated before.

Conceptual Transfer: Using Local Classifiers for Transfer Selection

Gregor Thurmair

A key challenge for Machine Translation is transfer selection, i.e. to find the right translation for a given word from a set of alternatives (1:n). This problem becomes the more important the larger the dictionary is, as the number of alternatives increases. The contribution presents a novel approach for transfer selection, called conceptual transfer, where selection is done using classifiers based on the conceptual context of a translation candidate on the source language side. Such classifiers are built automatically by parallel corpus analysis: Creating subcorpora for each translation of a 1:n package, and identifying correlating concepts in these subcorpora as features of the classifier. The resulting resource can easily be linked to transfer components of MT systems as it does not depend on internal analysis structures. Tests show that conceptual transfer outperforms the selection techniques currently used in operational MT systems.

Sharing Resources Between Free/Open-Source Rule-based Machine Translation Systems: Grammatical Framework and Apertium

Grégoire Détrez, Víctor M. Sánchez-Cartagena and Aarne Ranta

In this paper, we describe two methods developed for sharing linguistic data between two free and open source rule based machine translation systems: Apertium, a shallow-transfer system; and Grammatical Framework (GF), which performs a deeper syntactic transfer. In the first method, we describe the conversion of lexical data from Apertium to GF, while in the second one we automatically extract Apertium shallow-transfer rules from a GF bilingual grammar. We evaluated the resulting systems in a English-Spanish translation context, and results showed the usefulness of the resource sharing and confirmed the a-priori strong and weak points of the systems involved.

Missed Opportunities in Translation Memory Matching

Friedel Wolff, Laurette Pretorius and Paul Buitelaar

A translation memory system stores a data set of source-target pairs of translations. It attempts to respond to a query in the source language with a useful target text from the data set to assist a human translator. Such systems estimate the usefulness of a target text suggestion according to the similarity of its associated source text to the source text query. This study analyses two data sets in two language pairs each to find highly similar target texts, which would be useful mutual suggestions. We further investigate which of these useful suggestions can not be selected through source text similarity, and we do a thorough analysis of these cases to categorise and quantify them. This analysis provides insight into areas where the recall of translation memory systems can be improved. Specifically, source texts with an omission, and semantically very similar source texts are some of the more frequent cases with useful target text suggestions that are not selected with the baseline approach of simple edit distance between the source texts.

O42 - Dialogue (2)

Friday, May 30, 11:45

Chairperson: **Shyam Agrawal** Oral Session

Interoperability of Dialogue Corpora through ISO 24617-2-based Querying

Volha Petukhova, Andrei Malchanau and Harry Bunt

This paper explores a way of achieving interoperability: developing a query format for accessing existing annotated corpora whose expressions make use of the annotation language defined by the standard. The interpretation of expressions in the query implements a mapping from ISO 24617-2 concepts to those of the annotation scheme used in the corpus. We discuss two possible ways to query existing annotated corpora using DiAML. One way is to transform corpora into DiAML compliant format, and subsequently query these data using XQuery or XPath. The second approach is to define a DiAML query that can be directly used to retrieve requested information from the annotated data. Both approaches are valid. The first one presents a standard way of querying XML data. The second approach is a DiAML-oriented querying of dialogue act annotated data, for which we designed an interface. The proposed approach is tested on two important types of existing dialogue corpora: spoken two-person dialogue corpora collected and annotated within the HCRC Map Task paradigm, and multiparty face-to-face dialogues of the AMI corpus. We present the results and evaluate them with respect to accuracy and completeness through statistical comparisons between retrieved and manually constructed reference annotations.

Comparative Analysis of Verbal Alignment in Human-Human and Human-Agent Interactions

Sabrina Campano, Jessica Durand and Chloé Clavel

Engagement is an important feature in human-human and human-agent interaction. In this paper, we investigate lexical alignment as a cue of engagement, relying on two different corpora : CID and SEMAINE. Our final goal is to build a virtual conversational character that could use alignment strategies to maintain user's engagement. To do so, we investigate two alignment processes : shared vocabulary and other-repetitions. A quantitative and qualitative approach is proposed to characterize these aspects in human-human (CID) and human-operator (SEMAINE) interactions. Our results show that these processes are observable in both corpora, indicating a stable pattern that can be further modelled in conversational agents.

Free English and Czech Telephone Speech Corpus Shared Under the CC-BY-SA 3.0 License

Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka and Filip Jurčíček

We present a dataset of telephone conversations in English and Czech, developed for training acoustic models for automatic speech recognition (ASR) in spoken dialogue systems (SDSs). The data comprise 45 hours of speech in English and over 18 hours in Czech. Large part of the data, both audio and transcriptions, was collected using crowdsourcing, the rest are transcriptions by hired transcribers. We release the data together with scripts for data pre-processing and building acoustic models using the HTK and Kaldi ASR toolkits. We publish also the trained models described in this paper. The data are released under the CC-BY-SA 3.0 license, the scripts are licensed under Apache 2.0. In the paper, we report on the methodology of collecting the data, on the size and properties of the data, and on the scripts and their use. We verify the usability of the datasets by training and evaluating acoustic models using the presented data and scripts.

Japanese Conversation Corpus for Training and Evaluation of Backchannel Prediction Model

Hiroaki Noguchi, Yasuhiro Katagiri and Yasuharu Den

In this paper, we propose an experimental method for building a specialized corpus for training and evaluating backchannel prediction models of spoken dialogue. To develop a backchannel prediction model using a machine learning technique, it is necessary to discriminate between the timings of the interlocutor 's speech when more listeners commonly respond with backchannels and the timings when fewer listeners do so. The proposed corpus indicates the normative timings for backchannels in each speech with millisecond accuracy. In the proposed method, we first extracted each speech comprising a single turn from recorded conversation. Second, we presented these speeches as stimuli to 89 participants and asked them to respond by key hitting whenever they thought it appropriate to respond with a backchannel. In this way, we collected 28983 responses. Third, we applied the Gaussian mixture model to the temporal distribution of the responses and estimated the center of Gaussian distribution, that is, the backchannel relevance place (BRP), in each case. Finally, we synthesized 10 pairs of stereo speech stimuli and asked 19 participants to rate each on a 7-point scale of naturalness. The results show that backchannels inserted at BRPs were significantly higher than those in the original condition.

DiVE-Arabic: Gulf Arabic Dialogue in a Virtual Environment

Andrew Gargett, Sam Hellmuth and Ghazi AlGethami

Documentation of communicative behaviour across languages seems at a crossroads. While methods for collecting data on spoken or written communication, backed up by computational techniques, are evolving, the actual data being collected remain largely the same. Inspired by the efforts of some innovative researchers who are directly tackling the various obstacles to investigating language in the field (e.g. see various papers collected in Enfield & Stivers 2007), we report here about ongoing work to solve the general problem of collecting in situ data for situated linguistic interaction. The initial stages of this project have involved employing a portable format designed to increase range and flexibility of doing such collections in the field. Our motivation is to combine this with a parallel data set for a typologically distinct language, in order to contribute a parallel corpus of situated language use.

O43 - Semantics (2)

Friday, May 30, 11:45

Chairperson: **James Pustejovsky** Oral Session

Evaluation of Simple Distributional Compositional Operations on Longer Texts

Tamara Polajnar, Laura Rimell and Stephen Clark

Distributional semantic models have been effective at representing linguistic semantics at the word level, and more recently research has moved on to the construction of distributional representations for larger segments of text. However, it is not well understood how the composition operators that work well on short phrase-based models scale up to full-length sentences. In this paper we test several simple compositional methods on a sentence-length similarity task and discover that their performance peaks at fewer than ten operations. We also introduce a novel sentence segmentation method that reduces the number of compositional operations.

A Character-based Approach to Distributional Semantic Models: Exploiting Kanji Characters for Constructing JapaneseWord Vectors

Akira Utsumi

Many Japanese words are made of kanji characters, which themselves represent meanings. However traditional word-based distributional semantic models (DSMs) do not benefit from the useful semantic information of kanji characters. In this paper, we propose a method for exploiting the semantic information of kanji characters for constructing Japanese word vectors in DSMs. In the proposed method, the semantic representations of kanji characters (i.e, kanji vectors) are constructed first using the techniques of DSMs, and then word vectors are computed by combining the vectors of constituent kanji characters using vector composition methods. The evaluation experiment using a synonym identification task demonstrates that the kanji-based DSM achieves the best performance when a kanji-kanji matrix is weighted by positive pointwise mutual information and word vectors are composed by weighted multiplication. Comparison between kanji-based DSMs and word-based DSMs reveals that our kanji-based DSMs generally outperform latent semantic analysis, and also surpasses the best score word-based DSM for infrequent words comprising only frequent kanji characters. These findings clearly indicate that kanji-based DSMs are beneficial in improvement of quality of Japanese word vectors.

A Cascade Approach for Complex-type Classification

Lauren Romeo, Sara Mendes and Núria Bel

The work detailed in this paper describes a 2-step cascade approach for the classification of complex-type nominals. We describe an experiment that demonstrates how a cascade approach performs when the task consists in distinguishing nominals from a given complex-type from any other noun in the language. Overall, our classifier successfully identifies very specific and not highly frequent lexical items such as complex-types with high accuracy, and distinguishes them from those instances that are not complex types by using lexico-syntactic patterns indicative of the semantic classes corresponding to each of the individual sense components of the complex type. Although there is still room for improvement with regard to the coverage of the classifiers developed, the cascade approach increases the precision of classification of the complex-type nouns that are covered in the experiment presented.

A Rank-based Distance Measure to Detect Polysemy and to Determine Salient Vector-Space Features for German Prepositions

Maximilian Köper and Sabine Schulte im Walde

This paper addresses vector space models of prepositions, a notoriously ambiguous word class. We propose a rank-based distance measure to explore the vector-spatial properties of the ambiguous objects, focusing on two research tasks: (i) to distinguish polysemous from monosemous prepositions in vector space; and (ii) to determine salient vector-space features for a classification of preposition senses. The rank-based measure predicts the polysemy vs. monosemy of prepositions with a

precision of up to 88%, and suggests preposition-subcategorised nouns as more salient preposition features than preposition-subcategorising verbs.

Focusing Annotation for Semantic Role Labeling

Daniel Peterson, Martha Palmer and Shumin Wu

Annotation of data is a time-consuming process, but necessary for many state-of-the-art solutions to NLP tasks, including semantic role labeling (SRL). In this paper, we show that language models may be used to select sentences that are more useful to annotate. We simulate a situation where only a portion of the available data can be annotated, and compare language model based selection against a more typical baseline of randomly selected data. The data is ordered using an off-the-shelf language modeling toolkit. We show that the least probable sentences provide dramatic improved system performance over the baseline, especially when only a small portion of the data is annotated. In fact, the lion's share of the performance can be attained by annotating only 10-20% of the data. This result holds for training a model based on new annotation, as well as when adding domain-specific annotation to a general corpus for domain adaptation.

O44 - Grammar and Parsing (2)

Friday, May 30, 11:45

Chairperson: **Sadao Kurohashi** Oral Session

When POS Data Sets Don't Add Up: Combatting Sample Bias

Dirk Hovy, Barbara Plank and Anders Søgaard

Several works in Natural Language Processing have recently looked into part-of-speech annotation of Twitter data and typically used their own data sets. Since conventions on Twitter change rapidly, models often show sample bias. Training on a combination of the existing data sets should help overcome this bias and produce more robust models than any trained on the individual corpora. Unfortunately, combining the existing corpora proves difficult: many of the corpora use proprietary tag sets that have little or no overlap. Even when mapped to a common tag set, the different corpora systematically differ in their treatment of various tags and tokens. This includes both pre-processing decisions, as well as default labels for frequent tokens, thus exhibiting data bias and label bias, respectively. Only if we address these biases can we combine the existing data sets to also overcome sample bias. We present a systematic study of several Twitter POS data sets, the problems of label and data bias, discuss their effects on model performance, and show how to overcome them to learn models that perform well on various test sets, achieving relative error reduction of up to 21%.

Using C5.0 and Exhaustive Search for Boosting Frame-Semantic Parsing Accuracy

Guntis Barzdins, Didzis Gosko, Laura Rituma and Peteris Paikens

Frame-semantic parsing is a kind of automatic semantic role labeling performed according to the FrameNet paradigm. The paper reports a novel approach for boosting frame-semantic parsing accuracy through the use of the C5.0 decision tree classifier, a commercial version of the popular C4.5 decision tree classifier, and manual rule enhancement. Additionally, the possibility to replace C5.0 by an exhaustive search based algorithm (nicknamed C6.0) is described, leading to even higher frame-semantic parsing accuracy at the expense of slightly increased training time. The described approach is particularly efficient for languages with small FrameNet annotated corpora as it is for Latvian, which is used for illustration. Frame-semantic parsing accuracy achieved for Latvian through the C6.0 algorithm is on par with the state-of-the-art English frame-semantic parsers. The paper includes also a frame-semantic parsing use-case for extracting structured information from unstructured newswire texts, sometimes referred to as bridging of the semantic gap.

ML-Optimization of Ported Constraint Grammars

Eckhard Bick

In this paper, we describe how a Constraint Grammar with linguist-written rules can be optimized and ported to another language using a Machine Learning technique. The effects of rule movements, sorting, grammar-sectioning and systematic rule modifications are discussed and quantitatively evaluated. Statistical information is used to provide a baseline and to enhance the core of manual rules. The best-performing parameter combinations achieved part-of-speech F-scores of over 92 for a grammar ported from English to Danish, a considerable advance over both the statistical baseline (85.7), and the raw ported grammar (86.1). When the same technique was applied to an existing native Danish CG, error reduction was 10% (F=96.94).

A Deep Context Grammatical Model For Authorship Attribution

Simon Fuller, Phil Maguire and Philippe Moser

We define a variable-order Markov model, representing a Probabilistic Context Free Grammar, built from the sentence-level, de-lexicalized parse of source texts generated by a standard lexicalized parser, which we apply to the authorship attribution task. First, we motivate this model in the context of previous research on syntactic features in the area, outlining some of the

general strengths and limitations of the overall approach. Next we describe the procedure for building syntactic models for each author based on training cases. We then outline the attribution process - assigning authorship to the model which yields the highest probability for the given test case. We demonstrate the efficacy for authorship attribution over different Markov orders and compare it against syntactic features trained by a linear kernel SVM. We find that the model performs somewhat less successfully than the SVM over similar features. In the conclusion, we outline how we plan to employ the model for syntactic evaluation of literary texts.

Mapping Between English Strings and Reentrant Semantic Graphs

Fabienne Braune, Daniel Bauer and Kevin Knight

We investigate formalisms for capturing the relation between semantic graphs and English strings. Semantic graph corpora have spurred recent interest in graph transduction formalisms, but it is not yet clear whether such formalisms are a good fit for natural language data–in particular, for describing how semantic reentrancies correspond to English pronouns, zero pronouns, reflexives, passives, nominalizations, etc. We introduce a data set that focuses on these problems, we build grammars to capture the graph/string relation in this data, and we evaluate those grammars for conciseness and accuracy.

P56 - Corpora and Annotation

Friday, May 30, 11:45

Chairperson: **Tomaž Erjavec** Poster Session

HiEve: A Corpus for Extracting Event Hierarchies from News Stories

Goran Glavaš, Jan Šnajder, Marie-Francine Moens and Parisa Kordjamshidi

In news stories, event mentions denote real-world events of different spatial and temporal granularity. Narratives in news stories typically describe some real-world event of coarse spatial and temporal granularity along with its subevents. In this work, we present HiEve, a corpus for recognizing relations of spatiotemporal containment between events. In HiEve, the narratives are represented as hierarchies of events based on relations of spatiotemporal containment (i.e., superevent–subevent relations). We describe the process of manual annotation of HiEve. Furthermore, we build a supervised classifier for recognizing spatiotemporal containment between events to serve as a baseline for future research. Preliminary experimental results are encouraging, with classifier performance reaching 58% F1-score, only 11% less than the inter annotator agreement.

Building a Database of Japanese Adjective Examples from Special Purpose Web Corpora

Masaya Yamaguchi

It is often difficult to collect many examples for low-frequency words from a single general purpose corpus. In this paper, I present a method of building a database of Japanese adjective examples from special purpose Web corpora (SPW corpora) and investigates the characteristics of examples in the database by comparison with examples that are collected from a general purpose Web corpus (GPW corpus). My proposed method construct a SPW corpus for each adjective considering to collect examples that have the following features: (i) non-bias, (ii) the distribution of examples extracted from every SPW corpus bears much similarity to that of examples extracted from a GPW corpus. The results of experiments shows the following: (i) my proposed method can collect many examples rapidly. The number of examples extracted from SPW corpora is more than 8.0 times (median value) greater than that from the GPW corpus. (ii) the distributions of co-occurrence words for adjectives in the database are similar to those taken from the GPW corpus.

TLAXCALA: a Multilingual Corpus of Independent News

Antonio Toral

We acquire corpora from the domain of independent news from the Tlaxcala website. We build monolingual corpora for 15 languages and parallel corpora for all the combinations of those 15 languages. These corpora include languages for which only very limited such resources exist (e.g. Tamazight). We present the acquisition process in detail and we also present detailed statistics of the produced corpora, concerning mainly quantitative dimensions such as the size of the corpora per language (for the monolingual corpora) and per language pair (for the parallel corpora). To the best of our knowledge, these are the first publicly available parallel and monolingual corpora for the domain of independent news. We also create models for unsupervised sentence splitting for all the languages of the study.

Votter Corpus: A Corpus of Social Polling Language

Nathan Green and Septina Dian Larasati

The Votter Corpus is a new annotated corpus of social polling questions and answers. The Votter Corpus is novel in its use of the mobile application format and novel in its coverage of specific demographics. With over 26,000 polls and close to

1 millions votes, the Votter Corpus covers everyday question and answer language, primarily for users who are female and between the ages of 13-24. The corpus is annotated by topic and by popularity of particular answers. The corpus contains many unique characteristics such as emoticons, common mobile misspellings, and images associated with many of the questions. The corpus is a collection of questions and answers from The Votter App on the Android operating system. Data is created solely on this mobile platform which differs from most social media corpora. The Votter Corpus is being made available online in XML format for research and non-commercial use. The Votter android app can be downloaded for free in most android app stores.

Developing Text Resources for Ten South African Languages

Roald Eiselen and Martin Puttkammer

The development of linguistic resources for use in natural language processing is of utmost importance for the continued growth of research and development in the field, especially for resource-scarce languages. In this paper we describe the process and challenges of simultaneously developing multiple linguistic resources for ten of the official languages of South Africa. The project focussed on establishing a set of foundational resources that can foster further development of both resources and technologies for the NLP industry in South Africa. The development efforts during the project included creating monolingual unannotated corpora, of which a subset of the corpora for each language was annotated on token, orthographic, morphological and morphosyntactic layers. The annotated subsets includes both development and test sets and were used in the creation of five core-technologies, viz. a tokeniser, sentenciser, lemmatiser, part of speech tagger and morphological decomposer for each language. We report on the quality of these tools for each language and discuss the importance of the resources within the South African context.

Momresp: A Bayesian Model for Multi-Annotator Document Labeling

Paul Felt, Robbie Haertel, Eric Ringger and Kevin Seppi

Data annotation in modern practice often involves multiple, imperfect human annotators. Multiple annotations can be used to infer estimates of the ground-truth labels and to estimate individual annotator error characteristics (or reliability). We introduce MomResp, a model that incorporates information from both natural data clusters as well as annotations from multiple annotators to infer ground-truth labels and annotator reliability for the document classification task. We implement this model and show dramatic improvements over majority vote in situations where both annotations are scarce and annotation quality is low as well as in situations where annotators disagree consistently. Because MomResp predictions are subject to label switching, we introduce a solution that finds nearly optimal predicted class reassignments in a variety of settings using only information available to the model at inference time. Although MomResp does not perform well in annotation-rich situations, we show evidence suggesting how this shortcoming may be overcome in future work.

The Polish Summaries Corpus

Maciej Ogrodniczuk and Mateusz Kopeć

This article presents the Polish Summaries Corpus, a new resource created to support the development and evaluation of the tools for automated single-document summarization of Polish. The Corpus contains a large number of manual summaries of news articles, with many independently created summaries for a single text. Such approach is supposed to overcome the annotator bias, which is often described as a problem during the evaluation of the summarization algorithms against a single gold standard. There are several summarizers developed specifically for Polish language, but their in-depth evaluation and comparison was impossible without a large, manually created corpus. We present in detail the process of text selection, annotation process and the contents of the corpus, which includes both abstract free-word summaries, as well as extraction-based summaries created by selecting text spans from the original document. Finally, we describe how that resource could be used not only for the evaluation of the existing summarization tools, but also for studies on the human summarization process in Polish language.

P57 - Information Extraction and Information Retrieval

Friday, May 30, 11:45

Chairperson: **Feiyu Xu** Poster Session

Evaluating Web-as-corpus Topical Document Retrieval with an Index of the OpenDirectory

Clément de Groc and Xavier Tannier

This article introduces a novel protocol and resource to evaluate Web-as-corpus topical document retrieval. To the contrary of previous work, our goal is to provide an automatic, reproducible and robust evaluation for this task. We rely on the OpenDirectory (DMOZ) as a source of topically annotated webpages and index

them in a search engine. With this OpenDirectory search engine, we can then easily evaluate the impact of various parameters such as the number of seed terms, queries or documents, or the usefulness of various term selection algorithms. A first fully automatic evaluation is described and provides baseline performances for this task. The article concludes with practical information regarding the availability of the index and resource files.

Improving Open Relation Extraction via Sentence Re-Structuring

Jordan Schmidek and Denilson Barbosa

Information Extraction is an important task in Natural Language Processing, consisting of finding a structured representation for the information expressed in natural language text. Two key steps in information extraction are identifying the entities mentioned in the text, and the relations among those entities. In the context of Information Extraction for the World Wide Web, unsupervised relation extraction methods, also called Open Relation Extraction (ORE) systems, have become prevalent, due to their effectiveness without domain-specific training data. In general, these systems exploit part-of-speech tags or semantic information from the sentences to determine whether or not a relation exists, and if so, its predicate. This paper discusses some of the issues that arise when even moderately complex sentences are fed into ORE systems. A process for re-structuring such sentences is discussed and evaluated. The proposed approach replaces complex sentences by several others that, together, convey the same meaning and are more amenable to extraction by current ORE systems. The results of an experimental evaluation show that this approach succeeds in reducing the processing time and increasing the accuracy of the state-of-the-art ORE systems.

Semantic Search in Documents Enriched by LOD-based Annotations

Pavel Smrz and Jan Kouril

This paper deals with information retrieval on semantically enriched web-scale document collections. It particularly focuses on web-crawled content in which mentions of entities appearing in Freebase, DBpedia and other Linked Open Data resources have been identified. A special attention is paid to indexing structures and advanced query mechanisms that have been employed into a new semantic retrieval system. Scalability features are discussed together with performance statistics and results of experimental evaluation of presented approaches. Examples given to demonstrate key features of the developed solution correspond to the cultural heritage domain in which the results of our work have been primarily applied.

BiographyNet: Methodological Issues when NLP Supports Historical Research

Antske Fokkens, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne and Guus Schreiber

When NLP is used to support research in the humanities, new methodological issues come into play. NLP methods may introduce a bias in their analysis that can influence the results of the hypothesis a humanities scholar is testing. This paper addresses this issue in the context of BiographyNet a multi-disciplinary project involving NLP, Linked Data and history. We introduce the project to the NLP community. We argue that it is essential for historians to get insight into the provenance of information, including how information was extracted from text by NLP tools.

Using Large Biomedical Databases as Gold Annotations for Automatic Relation Extraction

Tilia Ellendorff, Fabio Rinaldi and Simon Clematide

We show how to use large biomedical databases in order to obtain a gold standard for training a machine learning system over a corpus of biomedical text. As an example we use the Comparative Toxicogenomics Database (CTD) and describe by means of a short case study how the obtained data can be applied. We explain how we exploit the structure of the database for compiling training material and a testset. Using a Naive Bayes document classification approach based on words, stem bigrams and MeSH descriptors we achieve a macro-average F-score of 61% on a subset of 8 action terms. This outperforms a baseline system based on a lookup of stemmed keywords by more than 20%. Furthermore, we present directions of future work, taking the described system as a vantage point. Future work will be aiming towards a weakly supervised system capable of discovering complete biomedical interactions and events.

A Method for Building Burst-Annotated Co-Occurrence Networks for Analysing Trends in Textual Data

Yutaka Mitsuishi, Vit Novacek and Pierre-Yves Vandenbussche

This paper presents a method for constructing a specific type of language resources that are conveniently applicable to analysis of trending topics in time-annotated textual data. More specifically, the method consists of building a co-occurrence network from the on-line content (such as New York Times articles) that conform to key words selected by users (e.g., 'Arab Spring'). Within

the network, burstiness of the particular nodes (key words) and edges (co-occurrence relations) is computed. A service deployed on the network then facilitates exploration of the underlying text in order to identify trending topics. Using the graph structure of the network, one can assess also a broader context of the trending events. To limit the information overload of users, we filter the edges and nodes displayed by their burstiness scores to show only the presumably more important ones. The paper gives details on the proposed method, including a step-by-step walk through with plenty of real data examples. We report on a specific application of our method to the topic of 'Arab Spring' and make the language resource applied therein publicly available for experimentation. Last but not least, we outline a methodology of an ongoing evaluation of our method.

P58 - Lexicons

Friday, May 30, 11:45

Chairperson: **Kiril Simov** Poster Session

Definition Patterns for Predicative Terms in Specialized Lexical Resources

Antonio San Martín and Marie-Claude L' Homme

The research presented in this paper is part of a larger project on the semi-automatic generation of definitions of semantically-related terms in specialized resources. The work reported here involves the formulation of instructions to generate the definitions of sets of morphologically-related predicative terms, based on the definition of one of the members of the set. In many cases, it is assumed that the definition of a predicative term can be inferred by combining the definition of a related lexical unit with the information provided by the semantic relation (i.e. lexical function) that links them. In other words, terminographers only need to know the definition of "pollute" and the semantic relation that links it to other morphologically-related terms ("polluter", "polluting", "pollutant", etc.) in order to create the definitions of the set. The results show that rules can be used to generate a preliminary set of definitions (based on specific lexical functions). They also show that more complex rules would need to be devised for other morphological pairs.

ColLex.en: Automatically Generating and Evaluating a Full-form Lexicon for English

Tim vor der Brück, Alexander Mehler and Zahurul Islam

The paper describes a procedure for the automatic generation of a large full-form lexicon of English. We put emphasis on two statistical methods to lexicon extension and adjustment: in terms of a letter-based HMM and in terms of a detector of spelling variants and misspellings. The resulting resource, ColLex.en, is evaluated with respect to two tasks: text categorization and lexical coverage by example of the SUSANNE corpus and the Open ANC.

Enrichment of Bilingual Dictionary through News Stream Data

Ajay Dubey, Parth Gupta, Vasudeva Varma and Paolo Rosso

Bilingual dictionaries are the key component of the cross-lingual similarity estimation methods. Usually such dictionary generation is accomplished by manual or automatic means. Automatic generation approaches include to exploit parallel or comparable data to derive dictionary entries. Such approaches require large amount of bilingual data in order to produce good quality dictionary. Many time the language pair does not have large bilingual comparable corpora and in such cases the best automatic dictionary is upper bounded by the quality and coverage of such corpora. In this work we propose a method which exploits continuous quasi-comparable corpora to derive term level associations for enrichment of such limited dictionary. Though we propose our experiments for English and Hindi, our approach can be easily extendable to other languages. We evaluated dictionary by manually computing the precision. In experiments we show our approach is able to derive interesting term level associations across languages.

FLELex: a graded Lexical Resource for French Foreign Learners

Thomas Francois, Nùria Gala, Patrick Watrin and Cédrick Fairon

In this paper we present FLELex, the first graded lexicon for French as a foreign language (FFL) that reports word frequencies by difficulty level (according to the CEFR scale). It has been obtained from a tagged corpus of 777,000 words from available textbooks and simplified readers intended for FFL learners. Our goal is to freely provide this resource to the community to be used for a variety of purposes going from the assessment of the lexical difficulty of a text, to the selection of simpler words within text simplification systems, and also as a dictionary in assistive tools for writing.

OpenLogos Semantico-Syntactic Knowledge-Rich Bilingual Dictionaries

Anabela Barreiro, Fernando Batista, Ricardo Ribeiro, Helena Moniz and Isabel Trancoso

This paper presents 3 sets of OpenLogos resources, namely the English-German, the English-French, and the English-Italian

bilingual dictionaries. In addition to the usual information on part-of-speech, gender, and number for nouns, offered by most dictionaries currently available, OpenLogos bilingual dictionaries have some distinctive features that make them unique: they contain cross-language morphological information (inflectional and derivational), semantico-syntactic knowledge, indication of the head word in multiword units, information about whether a source word corresponds to an homograph, information about verb auxiliaries, alternate words (i.e., predicate or process nouns), causatives, reflexivity, verb aspect, among others. The focal point of the paper will be the semantico-syntactic knowledge that is important for disambiguation and translation precision. The resources are publicly available at the METANET platform for free use by the research community.

Tharwa: A Large Scale Dialectal Arabic - Standard Arabic - English Lexicon

Mona Diab, Mohamed AlBadrashiny, Maryam Aminian, Mohammed Attia, Heba Elfardy, Nizar Habash, Abdelati Hawwari, Wael Salloum, Pradeep Dasigi and Ramy Eskander

We introduce an electronic three-way lexicon, Tharwa, comprising Dialectal Arabic, Modern Standard Arabic and English correspondents. The paper focuses on Egyptian Arabic as the first pilot dialect for the resource, with plans to expand to other dialects of Arabic in later phases of the project. We describe Tharwa's creation process and report on its current status. The lexical entries are augmented with various elements of linguistic information such as POS, gender, rationality, number, and root and pattern information. The lexicon is based on a compilation of information from both monolingual and bilingual existing resources such as paper dictionaries and electronic, corpus-based dictionaries. Multiple levels of quality checks are performed on the output of each step in the creation process. The importance of this lexicon lies in the fact that it is the first resource of its kind bridging multiple variants of Arabic with English. Furthermore, it is a wide coverage lexical resource containing over 73,000 Egyptian entries. Tharwa is publicly available. We believe it will have a significant impact on both Theoretical Linguistics as well as Computational Linguistics research.

Automatic Methods for the Extension of a Bilingual Dictionary using Comparable Corpora

Michael Rosner and Kurt Sultana

Bilingual dictionaries define word equivalents from one language to another, thus acting as an important bridge between languages. No bilingual dictionary is complete since languages are in a constant state of change. Additionally, dictionaries are unlikely to achieve complete coverage of all language terms. This paper investigates methods for extending dictionaries using non-aligned corpora, by finding translations through context similarity. Most methods compute word contexts from general corpora. This can lead to errors due to data sparsity. We investigate the hypothesis that this problem can be addressed by carefully choosing smaller corpora in which domain-specific terms are more predominant. We also introduce the notion of efficiency which we consider as the effort required to obtain a set of dictionary entries from a given corpus

Evaluating Lemmatization Models for Machine-Assisted Corpus-Dictionary Linkage

Kevin Black, Eric Ringger, Paul Felt, Kevin Seppi, Kristian Heal and Deryle Lonsdale

The task of corpus-dictionary linkage (CDL) is to annotate each word in a corpus with a link to an appropriate dictionary entry that documents the sense and usage of the word. Corpus-dictionary linked resources include concordances, dictionaries with word usage examples, and corpora annotated with lemmas or word-senses. Such CDL resources are essential in learning a language and in linguistic research, translation, and philology. Lemmatization is a common approximation to automating corpus-dictionary linkage, where lemmas are treated as dictionary entry headwords. We intend to use data-driven lemmatization models to provide machine assistance to human annotators in the form of pre-annotations, and thereby reduce the costs of CDL annotation. In this work we adapt the discriminative string transducer DirecTL+ to perform lemmatization for classical Syriac, a low-resource language. We compare the accuracy of DirecTL+ with the Morfette discriminative lemmatizer. DirecTL+ achieves 96.92% overall accuracy but only by a margin of 0.86% over Morfette at the cost of a longer time to train the model. Error analysis on the models provides guidance on how to apply these models in a machine assistance setting for corpus-dictionary linkage.

P59 - Language Resource Infrastructures

Friday, May 30, 11:45

Chairperson: **Martin Wynne** Poster Session

Linguistic Resources and Cats: How to Use ISOcat, RELcat and SCHEMAcat

Menzo Windhouwer and Ineke Schuurman

Within the European CLARIN infrastructure ISOcat is used to enable both humans and computer programs to find specific resources even when they use different terminology or data

structures. In order to do so, it should be clear which concepts are used in these resources, both at the level of metadata for the resource as well as its content, and what is meant by them. The concepts can be specified in ISOcat. SCHEMAcat enables us to relate the concepts used by a resource, while RELcat enables to type these relationships and add relationships beyond resource boundaries. This way these three registries together allow us (and the programs) to find what we are looking for.

Language Processing Infrastructure in the XLike Project

Lluís Padró, Zeljko Agic, Xavier Carreras, Blaz Fortuna, Esteban García-Cuesta, Zhixing Li, Tadej Stajner and Marko Tadić

This paper presents the linguistic analysis tools and its infrastructure developed within the XLike project. The main goal of the implemented tools is to provide a set of functionalities for supporting some of the main objectives of XLike, such as enabling cross-lingual services for publishers, media monitoring or developing new business intelligence applications. The services cover seven major and minor languages: English, German, Spanish, Chinese, Catalan, Slovenian, and Croatian. These analyzers are provided as web services following a lightweight SOA architecture approach, and they are publically callable and are catalogued in META-SHARE.

Access Control by Query Rewriting: the Case of KorAP

Piotr Banski, Nils Diewald, Michael Hanl, Marc Kupietz and Andreas Witt

We present an approach to an aspect of managing complex access scenarios to large and heterogeneous corpora that involves handling user queries that, intentionally or due to the complexity of the queried resource, target texts or annotations outside of the given user's permissions. We first outline the overall architecture of the corpus analysis platform KorAP, devoting some attention to the way in which it handles multiple query languages, by implementing ISO CQLF (Corpus Query Lingua Franca), which in turn constitutes a component crucial for the functionality discussed here. Next, we look at query rewriting as it is used by KorAP and zoom in on one kind of this procedure, namely the rewriting of queries that is forced by data access restrictions.

IXA pipeline: Efficient and Ready to Use Multilingual NLP tools

Rodrigo Agerri, Josu Bermudez and German Rigau

IXA pipeline is a modular set of Natural Language Processing tools (or pipes) which provide easy access to NLP technology. It offers robust and efficient linguistic annotation to both researchers and non-NLP experts with the aim of lowering the barriers of using NLP technology either for research purposes or for small industrial developers and SMEs. IXA pipeline can be used "as is" or exploit its modularity to pick and change different components. Given its open-source nature, it can also be modified and extended for it to work with other languages. This paper describes the general data-centric architecture of IXA pipeline and presents competitive results in several NLP annotations for English and Spanish.

Integration of Workflow and Pipeline for Language Service Composition

Trang Mai Xuan, Yohei Murakami, Donghui Lin and Toru Ishida

Integrating language resources and language services is a critical part of building natural language processing applications. Service workflow and processing pipeline are two approaches for sharing and combining language resources. Workflow languages focus on expressive power of the languages to describe variety of workflow patterns to meet users' needs. Users can combine those language services in service workflows to meet their requirements. The workflows can be accessible in distributed manner and can be invoked independently of the platforms. However, workflow languages lack of pipelined execution support to improve performance of workflows. Whereas, the processing pipeline provides a straightforward way to create a sequence of linguistic processing to analyze large amounts of text data. It focuses on using pipelined execution and parallel execution to improve throughput of pipelines. However, the resulting pipelines are standalone applications, i.e., software tools that are accessible only via local machine and that can only be run with the processing pipeline platforms. In this paper we propose an integration framework of the two approaches so that each offests the disadvantages of the other. We then present a case study wherein two representative frameworks, the Language Grid and UIMA, are integrated.

Interoperability and Customisation of Annotation Schemata in Argo

Rafal Rak, Jacob Carter, Andrew Rowley, Riza Theresa Batista-Navarro and Sophia Ananiadou

The process of annotating text corpora involves establishing annotation schemata which define the scope and depth of an annotation task at hand. We demonstrate this activity in Argo, a Web-based workbench for the analysis of textual resources, which facilitates both automatic and manual annotation. Annotation tasks in the workbench are defined by building workflows

consisting of a selection of available elementary analytics developed in compliance with the Unstructured Information Management Architecture specification. The architecture accommodates complex annotation types that may define primitive as well as referential attributes. Argo aids the development of custom annotation schemata and supports their interoperability by featuring a schema editor and specialised analytics for schemata alignment. The schema editor is a self-contained graphical user interface for defining annotation types. Multiple heterogeneous schemata can be aligned by including one of two type mapping analytics currently offered in Argo. One is based on a simple mapping syntax and, although limited in functionality, covers most common use cases. The other utilises a well established graph query language, SPARQL, and is superior to other state-of-the-art solutions in terms of expressiveness. We argue that the customisation of annotation schemata does not need to compromise their interoperability.

P60 - Metadata

Friday, May 30, 11:45

Chairperson: **Gil Francopoulo** Poster Session

Developing a Framework for Describing Relations among Language Resources

Penny Labropoulou, Christopher Cieri and Maria Gavrilidou

In this paper, we study relations holding between language resources as implemented in activities concerned with their documentation. We envision the term "language resources" with an inclusive definition covering datasets (corpora, lexica, ontologies, grammars, etc.), tools (including web services, workflows, platforms etc.), related publications and documentation, specifications and guidelines. However, the scope of the paper is limited to relations holding for datasets and tools. The study fosuses on the META-SHARE infrastructure and the Linguistic Data Consortium and takes into account the ISOcat DCR relations. Based on this study, we propose a taxonomy of relations, discuss their semantics and provide specifications for their use in order to cater for semantic interoperability. Issues of granularity, redundancy in codification, naming conventions and semantics of the relations are presented.

Towards Automatic Quality Assessment of Component Metadata

Thorsten Trippel, Daan Broeder, Matej Durco and Oddrun Ohren

Measuring the quality of metadata is only possible by assessing the quality of the underlying schema and the metadata instance. We propose some factors that are measurable automatically for metadata according to the CMD framework, taking into account the variability of schemas that can be defined in this framework. The factors include among others the number of elements, the (re-)use of reusable components, the number of filled in elements. The resulting score can serve as an indicator of the overall quality of the CMD instance, used for feedback to metadata providers or to provide an overview of the overall quality of metadata within a reposi-tory. The score is independent of specific schemas and generalizable. An overall assessment of harvested metadata is provided in form of statistical summaries and the distribution, based on a corpus of harvested metadata. The score is implemented in XQuery and can be used in tools, editors and repositories.

P61 - Opinion Mining and Sentiment Analysis

Friday, May 30, 11:45

Chairperson: **Gerard de Melo** Poster Session

Hope and Fear: How Opinions Influence Factuality

Chantal van Son, Marieke van Erp, Antske Fokkens and Piek Vossen

Both sentiment and event factuality are fundamental information levels for our understanding of events mentioned in news texts. Most research so far has focused on either modeling opinions or factuality. In this paper, we propose a model that combines the two for the extraction and interpretation of perspectives on events. By doing so, we can explain the way people perceive changes in (their belief of) the world as a function of their fears of changes to the bad or their hopes of changes to the good. This study seeks to examine the effectiveness of this approach by applying factuality annotations, based on FactBank, on top of the MPQA Corpus, a corpus containing news texts annotated for sentiments and other private states. Our findings suggest that this approach can be valuable for the understanding of perspectives, but that there is still some work to do on the refinement of the integration.

A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words

Nathan Hartmann, Lucas Avanço, Pedro Balage, Magali Duran, Maria das Graças Volpe Nunes, Thiago Pardo and Sandra Aluísio

Web 2.0 has allowed a never imagined communication boom. With the widespread use of computational and mobile devices, anyone, in practically any language, may post comments in

the web. As such, formal language is not necessarily used. In fact, in these communicative situations, language is marked by the absence of more complex syntactic structures and the presence of internet slang, with missing diacritics, repetitions of vowels, and the use of chat-speak style abbreviations, emoticons and colloquial expressions. Such language use poses severe new challenges for Natural Language Processing (NLP) tools and applications, which, so far, have focused on well-written texts. In this work, we report the construction of a large web corpus of product reviews in Brazilian Portuguese and the analysis of its lexical phenomena, which support the development of a lexical normalization tool for, in future work, subsidizing the use of standard NLP products for web opinion mining and summarization purposes.

Harmonization of German Lexical Resources for Opinion Mining

Thierry Declerck and Hans-Ulrich Krieger

We present on-going work on the harmonization of existing German lexical resources in the field of opinion and sentiment mining. The input of our harmonization effort consisted in four distinct lexicons of German word forms, encoded either as lemmas or as full forms, marked up with polarity features, at distinct granularity levels. We describe how the lexical resources have been mapped onto each other, generating a unique list of entries, with unified Part-of-Speech information and basic polarity features. Future work will be dedicated to the comparison of the harmonized lexicon with German corpora annotated with polarity information. We are further aiming at both linking the harmonized German lexical resources with similar resources in other languages and publishing the resulting set of lexical data in the context of the Linguistic Linked Open Data cloud.

Evaluation of Different Strategies for Domain Adaptation in Opinion Mining

Anne Garcia-Fernandez, Olivier Ferret and Marco Dinarelli

The work presented in this article takes place in the field of opinion mining and aims more particularly at finding the polarity of a text by relying on machine learning methods. In this context, it focuses on studying various strategies for adapting a statistical classifier to a new domain when training data only exist for one or several other domains. This study shows more precisely that a self-training procedure consisting in enlarging the initial training corpus with texts from the target domain that were reliably classified by the classifier is the most successful and stable strategy for the tested domains. Moreover, this strategy gets better results in most cases than (Blitzer et al., 2007)'s method on the same evaluation corpus while it is more simple.

Toward a Unifying Model for Opinion, Sentiment and Emotion Information Extraction

Amel Fraisse and Patrick Paroubek

This paper presents a logical formalization of a set 20 semantic categories related to opinion, emotion and sentiment. Our formalization is based on the BDI model (Belief, Desire and Intetion) and constitues a first step toward a unifying model for subjective information extraction. The separability of the subjective classes that we propose was assessed both formally and on two subjective reference corpora.

P62 - Speech Resources

Friday, May 30, 11:45

Chairperson: **Christoph Draxler** Poster Session

Exploiting the Large-Scale German Broadcast Corpus to Boost the Fraunhofer IAIS Speech Recognition System

Michael Stadtschnitzer, Jochen Schwenninger, Daniel Stein and Joachim Koehler

In this paper we describe the large-scale German broadcast corpus (GER-TV1000h) containing more than 1,000 hours of transcribed speech data. This corpus is unique in the German language corpora domain and enables significant progress in tuning the acoustic modelling of German large vocabulary continuous speech recognition (LVCSR) systems. The exploitation of this huge broadcast corpus is demonstrated by optimizing and improving the Fraunhofer IAIS speech recognition system. Due to the availability of huge amount of acoustic training data new training strategies are investigated. The performance of the automatic speech recognition (ASR) system is evaluated on several datasets and compared to previously published results. It can be shown that the word error rate (WER) using a larger corpus can be reduced by up to 9.1 % relative. By using both larger corpus and recent training paradigms the WER was reduced by up to 35.8 % relative and below 40 % absolute even for spontaneous dialectal speech in noisy conditions, making the ASR output a useful resource for subsequent tasks like named entity recognition also in difficult acoustic situations.

Macrosyntactic Segmenters of a French Spoken Corpus

Ilaine Wang, Sylvain Kahane and Isabelle Tellier

The aim of this paper is to describe an automated process to segment spoken French transcribed data into macrosyntactic units. While sentences are delimited by punctuation marks for written data, there is no obvious hint nor limit to major units for speech. As a reference, we used the manual annotation of macrosyntactic units based on illocutionary as well as syntactic criteria and developed for the Rhapsodie corpus, a 33.000 words prosodic and syntactic treebank. Our segmenters were built using machine learning methods as supervised classifiers : segmentation is about identifying the boundaries of units, which amounts to classifying each interword space. We trained six different models on Rhapsodie using different sets of features, including prosodic and morphosyntactic cues, on the assumption that their combination would be relevant for the task. Both types of cues could be resulting either from manual annotation/correction or from fully automated processes, which comparison might help determine the cost of manual effort, especially for the 3M words of spoken French of the Orfeo project those experiments are contributing to.

VOLIP: a Corpus of Spoken Italian and a Virtuous Example of Reuse of Linguistic Resources

Iolanda Alfano, Francesco Cutugno, Aurelio de Rosa, Claudio Iacobini, Renata Savy and Miriam Voghera

The corpus VoLIP (The Voice of LIP) is an Italian speech resource which associates the audio signals to the orthographic transcriptions of the LIP Corpus. The LIP Corpus was designed to represent diaphasic, diatopic and diamesic variation. The Corpus was collected in the early '90s to compile a frequency lexicon of spoken Italian and its size was tailored to produce a reliable frequency lexicon for the first 3,000 lemmas. Therefore, it consists of about 500,000 word tokens for 60 hours of recording. The speech materials belong to five different text registers and they were collected in four different cities. Thanks to a modern technological approach VoLIP web service allows users to search the LIP corpus using IMDI metadata, lexical or morpho-syntactic entry keys, receiving as result the audio portions aligned to the corresponding required entry. The VoLIP corpus is freely available at the URL http://www.parlaritaliano.it.

DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech

George Christodoulides, Mathieu Avanzi and Jean-Philippe Goldman

We present DisMo, a multi-level annotator for spoken language corpora that integrates part-of-speech tagging with basic disfluency detection and annotation, and multi-word unit recognition. DisMo is a hybrid system that uses a combination of lexical resources, rules, and statistical models based on Conditional Random Fields (CRF). In this paper, we present the first public version of DisMo for French. The system is trained and its performance evaluated on a 57k-token corpus, including different varieties of French spoken in three countries (Belgium, France and Switzerland). DisMo supports a multi-level annotation scheme, in which the tokenisation to minimal word units is complemented with multi-word unit groupings (each having associated POS tags), as well as separate levels for annotating disfluencies and discourse phenomena. We present the system's architecture, linguistic resources and its hierarchical tag-set. Results show that DisMo achieves a precision of 95% (finest tag-set) to 96.8% (coarse tag-set) in POS-tagging non-punctuated, sound-aligned transcriptions of spoken French, while also offering substantial possibilities for automated multi-level annotation.

Revising the Annotation of a Broadcast News Corpus: a Linguistic Approach

Vera Cabarrão, Helena Moniz, Fernando Batista, Ricardo Ribeiro, Nuno Mamede, Hugo Meinedo, Isabel Trancoso, Ana Isabel Mata and David Martins de Matos

This paper presents a linguistic revision process of a speech corpus of Portuguese broadcast news focusing on metadata annotation for rich transcription, and reports on the impact of the new data on the performance for several modules. The main focus of the revision process consisted on annotating and revising structural metadata events, such as disfluencies and punctuation marks. The resultant revised data is now being extensively used, and was of extreme importance for improving the performance of several modules, especially the punctuation and capitalization modules, but also the speech recognition system, and all the subsequent modules. The resultant data has also been recently used in disfluency studies across domains.

Teenage and Adult Speech in School Context: Building and Processing a Corpus of European Portuguese

Ana Isabel Mata, Helena Moniz, Fernando Batista and Julia Hirschberg

We present a corpus of European Portuguese spoken by teenagers and adults in school context, CPE-FACES, with an overview of the differential characteristics of high school oral presentations and the challenges this data poses to automatic speech processing. The CPE-FACES corpus has been created with two main goals: to provide a resource for the study of prosodic patterns in both spontaneous and prepared unscripted speech, and to capture inter-speaker and speaking style variations common at school, for research on oral presentations. Research on speaking styles is still largely based on adult speech. References to teenagers are sparse and cross-analyses of speech types comparing teenagers and adults are rare. We expect CPE-FACES, currently a unique resource in this domain, will contribute to filling this gap in European Portuguese. Focusing on disfluencies and phrase-final phonetic-phonological processes we show the impact of teenage speech on the automatic segmentation of oral presentations. Analyzing fluent final intonation contours in declarative utterances, we also show that communicative situation specificities, speaker status and cross-gender differences are key factors in speaking style variation at school.

Croatian Memories

Arjan van Hessen, Franciska de Jong, Stef Scagliola and Tanja Petrovic

In this contribution we describe a collection of approximately 400 video interviews recorded in the context of the project Croatian Memories (CroMe) with the objective of documenting personal war-related experiences. The value of this type of sources is threefold: they contain information that is missing in written sources, they can contribute to the process of reconciliation, and they provide a basis for reuse of data in disciplines with an interest in narrative data. The CroMe collection is not primarily designed as a linguistic corpus, but is the result of an archival effort to collect so-called oral history data. For researchers in the fields of natural language processing and speech analysis this type of life-stories may function as an object trouvé containing real-life language data that can prove to be useful for the purpose of modelling specific aspects of human expression and communication.

The KiezDeutsch Korpus (KiDKo) Release 1.0

Ines Rehbein, Sören Schalowski and Heike Wiese

This paper presents the first release of the KiezDeutsch Korpus (KiDKo), a new language resource with multiparty spoken dialogues of Kiezdeutsch, a newly emerging language variety spoken by adolescents from multiethnic urban areas in Germany. The first release of the corpus includes the transcriptions of the data as well as a normalisation layer and part-of-speech annotations. In the paper, we describe the main features of the new resource and then focus on automatic POS tagging of informal spoken language. Our tagger achieves an accuracy of nearly 97% on KiDKo. While we did not succeed in further improving the tagger using ensemble tagging, we present our approach to using the tagger ensembles for identifying error patterns in the automatically tagged data.

Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks

Anthony Rousseau, Paul Deléglise and Yannick Estève

In this paper, we present improvements made to the TED-LIUM corpus we released in 2012. These enhancements fall into two categories. First, we describe how we filtered publicly available monolingual data and used it to estimate well-suited language models (LMs), using open-source tools. Then, we describe the process of selection we applied to new acoustic data from TED talks, providing additions to our previously released corpus. Finally, we report some experiments we made around these improvements.

Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS

Jan Strunk, Florian Schiel and Frank Seifart

Language documentation projects supported by recent funding intiatives have created a large number of multimedia corpora of typologically diverse languages. Most of these corpora provide a manual alignment of transcription and audio data at the level of larger units, such as sentences or intonation units. Their usefulness both for corpus-linguistic and psycholinguistic research and for the development of tools and teaching materials could, however, be increased by achieving a more fine-grained alignment of transcription and audio at the word or even phoneme level. Since most language documentation corpora contain data on small languages, there usually do not exist any speech recognizers or acoustic models specifically trained on these languages. We therefore investigate the feasibility of untrained forced alignment for such corpora. We report on an evaluation of the tool (Web)MAUS (Kisler, 2012) on several language documentation corpora and discuss practical issues in the application of forced alignment. Our evaluation shows that (Web)MAUS with its existing acoustic models combined with simple grapheme-to-phoneme conversion can be successfully used for word-level

forced alignment of a diverse set of languages without additional training, especially if a manual prealignment of larger annotation units is already avaible.

O45 - Environment and Machine Interactions - Special Session

Friday, May 30, 14:55

Chairperson: **Laurence Devillers** Oral Session

The Sweet-Home Speech and Multimodal Corpus for Home Automation Interaction

Michel Vacher, Benjamin Lecouteux, Pedro Chahuara, François Portet, Brigitte Meillon and Nicolas Bonnefond

Ambient Assisted Living aims at enhancing the quality of life of older and disabled people at home thanks to Smart Homes and Home Automation. However, many studies do not include tests in real settings, because data collection in this domain is very expensive and challenging and because of the few available data sets. The S WEET-H OME multimodal corpus is a dataset recorded in realistic conditions in D OMUS, a fully equipped Smart Home with microphones and home automation sensors, in which participants performed Activities of Daily living (ADL). This corpus is made of a multimodal subset, a French home automation speech subset recorded in Distant Speech conditions, and two interaction subsets, the first one being recorded by 16 persons without disabilities and the second one by 6 seniors and 5 visually impaired people. This corpus was used in studies related to ADL recognition, context aware interaction and distant speech recognition applied to home automation controled through voice.

Multimodal Corpora for Silent Speech Interaction

João Freitas, António Teixeira and Miguel Dias

A Silent Speech Interface (SSI) allows for speech communication to take place in the absence of an acoustic signal. This type of interface is an alternative to conventional Automatic Speech Recognition which is not adequate for users with some speech impairments or in the presence of environmental noise. The work presented here produces the conditions to explore and analyze complex combinations of input modalities applicable in SSI research. By exploring non-invasive and promising modalities, we have selected the following sensing technologies used in human-computer interaction: Video and Depth input, Ultrasonic Doppler sensing and Surface Electromyography. This paper describes a novel data collection methodology where these independent streams of information are synchronously acquired with the aim of supporting research and development of a multimodal SSI. The reported recordings were divided into two rounds: a first one where the acquired data was silently uttered and a second round where speakers pronounced the scripted prompts in an audible and normal tone. In the first round of recordings, a total of 53.94 minutes were captured where 30.25% was estimated to be silent speech. In the second round of recordings, a total of 30.45 minutes were obtained and 30.05% of the recordings were audible speech.

3D Face Tracking and Multi-Scale, Spatio-temporal Analysis of Linguistically Significant Facial Expressions and Head Positions in ASL

Bo Liu, Jingjing Liu, Xiang Yu, Dimitris Metaxas and Carol Neidle

Essential grammatical information is conveyed in signed languages by clusters of events involving facial expressions and movements of the head and upper body. This poses a significant challenge for computer-based sign language recognition. Here, we present new methods for the recognition of nonmanual grammatical markers in American Sign Language (ASL) based on: (1) new 3D tracking methods for the estimation of 3D head pose and facial expressions to determine the relevant low-level features; (2) methods for higher-level analysis of component events (raised/lowered eyebrows, periodic head nods and head shakes) used in grammatical markings—with differentiation of temporal phases (onset, core, offset, where appropriate), analysis of their characteristic properties, and extraction of corresponding features; (3) a 2-level learning framework to combine low- and high-level features of differing spatio-temporal scales. This new approach achieves significantly better tracking and recognition results than our previous methods.

HuRIC: a Human Robot Interaction Corpus

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Luca Iocchi, Roberto Basili and Daniele Nardi

Recent years show the development of large scale resources (e.g. FrameNet for the Frame Semantics) that supported the definition of several state-of-the-art approaches in Natural Language Processing. However, the reuse of existing resources in heterogeneous domains such as Human Robot Interaction is not straightforward. The generalization offered by many data driven methods is strongly biased by the employed data, whose performance in out-of-domain conditions exhibit large drops. In this paper, we present the Human Robot Interaction Corpus (HuRIC). It is made of audio files paired with their transcriptions referring to commands for a robot, e.g. in a home environment. The recorded sentences are annotated with different kinds of linguistic information, ranging from morphological and syntactic

information to rich semantic information, according to the Frame Semantics, to characterize robot actions, and Spatial Semantics, to capture the robot environment. All texts are represented through the Abstract Meaning Representation, to adopt a simple but expressive representation of commands, that can be easily translated into the internal representation of the robot.

O46 - Event Extraction and Event Coreference

Friday, May 30, 14:55

Chairperson: **Martha Palmer** Oral Session

Event Extraction Using Distant Supervision

Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher D. Manning and Daniel Jurafsky

Distant supervision is a successful paradigm that gathers training data for information extraction systems by automatically aligning vast databases of facts with text. Previous work has demonstrated its usefulness for the extraction of binary relations such as a person's employer or a film's director. Here, we extend the distant supervision approach to template-based event extraction, focusing on the extraction of passenger counts, aircraft types, and other facts concerning airplane crash events. We present a new publicly available dataset and event extraction task in the plane crash domain based on Wikipedia infoboxes and newswire text. Using this dataset, we conduct a preliminary evaluation of four distantly supervised extraction models which assign named entity mentions in text to entries in the event template. Our results indicate that joint inference over sequences of candidate entity mentions is beneficial. Furthermore, we demonstrate that the Searn algorithm outperforms a linear-chain CRF and strong baselines with local inference.

SinoCoreferencer: An End-to-End Chinese Event Coreference Resolver

Chen Chen and Vincent Ng

Compared to entity coreference resolution, there is a relatively small amount of work on event coreference resolution. Much work on event coreference was done for English. In fact, to our knowledge, there are no publicly available results on Chinese event coreference resolution. This paper describes the design, implementation, and evaluation of SinoCoreferencer, an end-to-end state-of-the-art ACE-style Chinese event coreference system. We have made SinoCoreferencer publicly available, in hope to facilitate the development of high-level Chinese natural language applications that can potentially benefit from event coreference information.

Supervised Within-Document Event Coreference using Information Propagation

Zhengzhong Liu, Jun Araki, Eduard Hovy and Teruko Mitamura

Event coreference is an important task for full text analysis. However, previous work uses a variety of approaches, sources and evaluation, making the literature confusing and the results incommensurate. We provide a description of the differences to facilitate future research. Second, we present a supervised method for event coreference resolution that uses a rich feature set and propagates information alternatively between events and their arguments, adapting appropriately for each type of argument.

Using a Sledgehammer to Crack a Nut? Lexical Diversity and Event Coreference Resolution

Agata Cybulska and Piek Vossen

In this paper we examine the representativeness of the EventCorefBank (ECB, Bejan and Harabagiu, 2010) with regards to the language population of large-volume streams of news. The ECB corpus is one of the data sets used for evaluation of the task of event coreference resolution. Our analysis shows that the ECB in most cases covers one seminal event per domain, what considerably simplifies event and so language diversity that one comes across in the news. We augmented the corpus with a new corpus component, consisting of 502 texts, describing different instances of event types that were already captured by the 43 topics of the ECB, making it more representative of news articles on the web. The new "ECB+" corpus is available for further research.

Detecting Subevent Structure for Event Coreference Resolution

Jun Araki, Zhengzhong Liu, Eduard Hovy and Teruko Mitamura

In the task of event coreference resolution, recent work has shown the need to perform not only full coreference but also partial coreference of events. We show that subevents can form a particular hierarchical event structure. This paper examines a novel two-stage approach to finding and improving subevent structures. First, we introduce a multiclass logistic regression model that can detect subevent relations in addition to full coreference. Second, we propose a method to improve subevent structure based on subevent clusters detected by the model. Using a corpus in the Intelligence Community domain, we show that the

method achieves over 3.2 BLANC F1 gain in detecting subevent relations against the logistic regression model.

O47 - Standards and Interoperability

Friday, May 30, 14:55

Chairperson: **Key-Sun Choi** Oral Session

Three Dimensions of the so-called "Interoperability" of Annotation Schemes

Eva Hajičová

Interoperability" of annotation schemes is one of the key words in the discussions about annotation of corpora. In the present contribution, we propose to look at the so-called interoperability from (at least) three angles, namely (i) as a relation (and possible interaction or cooperation) of different annotation schemes for different layers or phenomena of a single language, (ii) the possibility to annotate different languages by a single (modified or not) annotation scheme, and (iii) the relation between different annotation schemes for a single language, or for a single phenomenon or layer of the same language. The pros and cons of each of these aspects are discussed as well as their contribution to linguistic studies and natural language processing. It is stressed that a communication and collaboration between different annotation schemes requires an explicit specification and consistency of each of the schemes.

Experiences with the ISOcat Data Category Registry

Daan Broeder, Ineke Schuurman and Menzo Windhouwer

The ISOcat Data Category Registry has been a joint project of both ISO TC 37 and the European CLARIN infrastructure. In this paper the experiences of using ISOcat in CLARIN are described and evaluated. This evaluation clarifies the requirements of CLARIN with regard to a semantic registry to support its semantic interoperability needs. A simpler model based on concepts instead of data cate-gories and a simpler workflow based on community recommendations will address these needs better and offer the required flexibility.

Towards Interoperable Discourse Annotation. Discourse Features in the Ontologies of Linguistic Annotation

Christian Chiarcos

This paper describes the extension of the Ontologies of Linguistic Annotation (OLiA) with respect to discourse features. The OLiA ontologies provide a a terminology repository that can be employed to facilitate the conceptual (semantic) interoperability of annotations of discourse phenomena as found in the most important corpora available to the community, including OntoNotes, the RST Discourse Treebank and the Penn Discourse Treebank. Along with selected schemes for information structure and coreference, discourse relations are discussed with special emphasis on the Penn Discourse Treebank and the RST Discourse Treebank. For an example contained in the intersection of both corpora, I show how ontologies can be employed to generalize over divergent annotation schemes.

Off-Road LAF: Encoding and Processing Annotations in NLP Workflows

Emanuele Lapponi, Erik Velldal, Stephan Oepen and Rune Lain Knudsen

The Linguistic Annotation Framework (LAF) provides an abstract data model for specifying interchange representations to ensure interoperability among different annotation formats. This paper describes an ongoing effort to adapt the LAF data model as the interchange representation in complex workflows as used in the Language Analysis Portal (LAP), an on-line and large-scale processing service that is developed as part of the Norwegian branch of the Common Language Resources and Technology Infrastructure (CLARIN) initiative. Unlike several related on-line processing environments, which predominantly instantiate a distributed architecture of web services, LAP achives scalability to potentially very large data volumes through integration with the Norwegian national e-Infrastructure, and in particular job sumission to a capacity compute cluster. This setup leads to tighter integration requirements and also calls for efficient, low-overhead communication of (intermediate) processing results with workflows. We meet these demands by coupling the LAF data model with a lean, non-redundant JSON-based interchange format and integration of an agile and performant NoSQL database, allowing parallel access from cluster nodes, as the central repository of linguistic annotation.

Universal Stanford Dependencies: a Cross-Linguistic Typology

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre and Christopher D. Manning

Revisiting the now de facto standard Stanford dependency representation, we propose an improved taxonomy to capture grammatical relations across languages, including morphologically rich ones. We suggest a two-layered taxonomy: a set of broadly attested universal grammatical relations, to which language-specific relations can be added. We emphasize the lexicalist stance of the Stanford Dependencies, which leads to a

particular, partially new treatment of compounding, prepositions, and morphology. We show how existing dependency schemes for several languages map onto the universal taxonomy proposed here and close with consideration of practical implications of dependency representation choices for NLP applications, in particular parsing.

O48 - Information Extraction and Text Structure

Friday, May 30, 14:55

Chairperson: **Mark Liberman** Oral Session

Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web

Giuseppe Rizzo, Marieke van Erp and Raphaël Troncy

Named entity recognition and disambiguation are of primary importance for extracting information and for populating knowledge bases. Detecting and classifying named entities has traditionally been taken on by the natural language processing community, whilst linking of entities to external resources, such as those in DBpedia, has been tackled by the Semantic Web community. As these tasks are treated in different communities, there is as yet no oversight on the performance of these tasks combined. We present an approach that combines the state-of-the art from named entity recognition in the natural language processing domain and named entity linking from the semantic web community. We report on experiments and results to gain more insights into the strengths and limitations of current approaches on these tasks. Our approach relies on the numerous web extractors supported by the NERD framework, which we combine with a machine learning algorithm to optimize recognition and linking of named entities. We test our approach on four standard data sets that are composed of two diverse text types, namely newswire and microposts.

Annotating Relations in Scientific Articles

Adam Meyers, Giancarlo Lee, Angus Grieve-Smith, Yifan He and Harriet Taber

Relations (ABBREVIATE, EXEMPLIFY, ORIGINATE, REL_ WORK, OPINION) between entities (citations, jargon, people, organizations) are annotated for PubMed scientific articles. We discuss our specifications, pre-processing and evaluation

Improving Entity Linking using Surface Form Refinement

Eric Charton, Marie-Jean Meurs, Ludovic Jean-Louis and Michel Gagnon

In this paper, we present an algorithm for improving named entity resolution and entity linking by using surface form generation and rewriting. Surface forms consist of a word or a group of words that matches lexical units like Paris or New York City. Used as matching sequences to select candidate entries in a knowledge base, they contribute to the disambiguation of those candidates through similarity measures. In this context, misspelled textual sequences (entities) can be impossible to identify due to the lack of available matching surface forms. To address this problem, we propose an algorithm for surface form refinement based on Wikipedia resources. The approach extends the surface form coverage of our entity linking system, and rewrites or reformulates misspelled mentions (entities) prior to starting the annotation process. The algorithm is evaluated on the corpus associated with the monolingual English entity linking task of NIST KBP 2013. We show that the algorithm improves the entity linking system performance.

Evaluating Improvised Hip Hop Lyrics - Challenges and Observations

Karteek Addanki and Dekai Wu

We investigate novel challenges involved in comparing model performance on the task of improvising responses to hip hop lyrics and discuss observations regarding inter-evaluator agreement on judging improvisation quality. We believe the analysis serves as a first step toward designing robust evaluation strategies for improvisation tasks, a relatively neglected area to date. Unlike most natural language processing tasks, improvisation tasks suffer from a high degree of subjectivity, making it difficult to design discriminative evaluation strategies to drive model development. We propose a simple strategy with fluency and rhyming as the criteria for evaluating the quality of generated responses, which we apply to both our inversion transduction grammar based FREESTYLE hip hop challenge-response improvisation system, as well as various contrastive systems. We report inter-evaluator agreement for both English and French hip hop lyrics, and analyze correlation with challenge length. We also compare the extent of agreement in evaluating fluency with that of rhyming, and quantify the difference in agreement with and without precise definitions of evaluation criteria.

Towards Automatic Detection of Narrative Structure

Jessica Ouyang and Kathy McKeown

We present novel computational experiments using William Labov's theory of narrative analysis. We describe his six elements of narrative structure and construct a new corpus based on his most recent work on narrative. Using this corpus, we explore the correspondence between Labov's elements of narrative structure and the implicit discourse relations of the Penn Discourse

Treebank, and we construct a mapping between the elements of narrative structure and the discourse relation classes of the PDTB. We present first experiments on detecting Complicating Actions, the most common of the elements of narrative structure, achieving an f-score of 71.55. We compare the contributions of features derived from narrative analysis, such as the length of clauses and the tenses of main verbs, with those of features drawn from work on detecting implicit discourse relations. Finally, we suggest directions for future research on narrative structure, such as applications in assessing text quality and in narrative generation.

P63 - Computer-Assisted Language Learning (CALL)

Friday, May 30, 14:55

Chairperson: **Keith Miller** Poster Session

Phoneme Set Design Using English Speech Database by Japanese for Dialogue-based English CALL Systems

Xiaoyun Wang, Jinsong Zhang, Masafumi Nishida and Seiichi Yamamoto

This paper describes a method of generating a reduced phoneme set for dialogue-based computer assisted language learning (CALL)systems. We designed a reduced phoneme set consisting of classified phonemes more aligned with the learners' speech characteristics than the canonical set of a target language. This reduced phoneme set provides an inherently more appropriate model for dealing with mispronunciation by second language speakers. In this study, we used a phonetic decision tree (PDT)-based top-down sequential splitting method to generate the reduced phoneme set and then applied this method to a translation-game type English CALL system for Japanese to determine its effectiveness. Experimental results showed that the proposed method improves the performance of recognizing non-native speech.

Generating a Lexicon of Errors in Portuguese to Support an Error Identification System for Spanish Native Learners

Lianet Sepúlveda Torres, Magali Sanches Duran and Sandra Aluísio

Portuguese is a less resourced language in what concerns foreign language learning. Aiming to inform a module of a system designed to support scientific written production of Spanish native speakers learning Portuguese, we developed an approach to automatically generate a lexicon of wrong words, reproducing language transfer errors made by such foreign learners. Each item of the artificially generated lexicon contains, besides the wrong word, the respective Spanish and Portuguese correct words. The wrong word is used to identify the interlanguage error and the correct Spanish and Portuguese forms are used to generate the suggestions. Keeping control of the correct word forms, we can provide correction or, at least, useful suggestions for the learners. We propose to combine two automatic procedures to obtain the error correction: i) a similarity measure and ii) a translation algorithm based on aligned parallel corpus. The similarity-based method achieved a precision of 52%, whereas the alignment-based method achieved a precision of 90%. In this paper we focus only on interlanguage errors involving suffixes that have different forms in both languages. The approach, however, is very promising to tackle other types of errors, such as gender errors.

Automatic Error Detection Concerning the Definite and Indefinite Conjugation in the HunLearner Corpus

Veronika Vincze, János Zsibrita, Péter Durst and Martina Katalin Szabó

In this paper we present the results of automatic error detection, concerning the definite and indefinite conjugation in the extended version of the HunLearner corpus, the learners' corpus of the Hungarian language. We present the most typical structures that trigger definite or indefinite conjugation in Hungarian and we also discuss the most frequent types of errors made by language learners in the corpus texts. We also illustrate the error types with sentences taken from the corpus. Our results highlight grammatical structures that might pose problems for learners of Hungarian, which can be fruitfully applied in the teaching and practicing of such constructions from the language teacher's or learners' point of view. On the other hand, these results may be exploited in extending the functionalities of a grammar checker, concerning the definiteness of the verb. Our automatic system was able to achieve perfect recall, i.e. it could find all the mismatches between the type of the object and the conjugation of the verb, which is promising for future studies in this area.

Presenting a System of Human-Machine Interaction for Performing Map Tasks.

Gabriele Pallotti, Francesca Frontini, Fabio Affè, Monica Monachini and Stefania Ferrari

A system for human machine interaction is presented, that offers second language learners of Italian the possibility of assessing their competence by performing a map task, namely by guiding the a virtual follower through a map with written instructions in natural language. The underlying natural language processing

algorithm is described, and the map authoring infrastructure is presented.

Assessment of Non-native Prosody for Spanish as L2 using Quantitative Scores and Perceptual Evaluation

Valentín Cardeñoso-Payo, César González-Ferreras and David Escudero

In this work we present SAMPLE, a new pronunciation database of Spanish as L2, and first results on the automatic assessment of Non-native prosody. Listen and repeat and read tasks are carried out by native and foreign speakers of Spanish. The corpus has been designed to support comparative studies and evaluation of automatic pronunciation error assessment both at phonetic and prosodic level. Four expert evaluators have annotated utterances with perceptual scores related to prosodic aspects of speech, intelligibility, phonetic quality and global proficiency level in Spanish. From each utterance, we computed several prosodic features and ASR scores. A correlation study over subjective and quantitative measures is carried out. An estimation of the prediction of perceptual scores from speech features is shown.

A Flexible Language Learning Platform based on Language Resources and Web Services

Elena Volodina, Ildikó Pilán, Lars Borin and Therese Lindström Tiedemann

We present Lärka, the language learning platform of Språkbanken (the Swedish Language Bank). It consists of an exercise generator which reuses resources available through Språkbanken: mainly Korp, the corpus infrastructure, and Karp, the lexical infrastructure. Through Lärka we reach new user groups – students and teachers of Linguistics as well as second language learners and their teachers – and this way bring Språkbanken's resources in a relevant format to them. Lärka can therefore be viewed as an case of real-life language resource evaluation with end users. In this article we describe Lärka's architecture, its user interface, and the five exercise types that have been released for users so far. The first user evaluation following in-class usage with students of linguistics, speech therapy and teacher candidates are presented. The outline of future work concludes the paper.

MAT: a Tool for L2 Pronunciation Errors Annotation

Renlong Ai and Marcela Charfuelan

In the area of Computer Assisted Language Learning(CALL), second language (L2) learners' spoken data is an important resource for analysing and annotating typical L2 pronunciation errors. The annotation of L2 pronunciation errors in spoken data is not an easy task though, normally it requires manual annotation from trained linguists or phoneticians. In order to facilitate this task, in this paper, we present the MAT tool, a web-based tool intended to facilitate the annotation of L2 learners' pronunciation errors at various levels. The tool has been designed taking into account recent studies on error detection in pronunciation training. It also aims at providing an easy and fast annotation process via a comprehensive and friendly user interface. The tool is based on the MARY TTS open source platform, from which it uses the components: text analyser (tokeniser, syllabifier, phonemiser), phonetic aligner and speech signal processor. Annotation results at sentence, word, syllable and phoneme levels are stored in XML format. The tool is currently under evaluation with a L2 learners' spoken corpus recorded in the SPRINTER (Language Technology for Interactive, Multi-Media Online Language Learning) project.

Modeling Language Proficiency Using Implicit Feedback

Chris Hokamp, Rada Mihalcea and Peter Schuelke

We describe the results of several experiments with interactive interfaces for native and L2 English students, designed to collect implicit feedback from students as they complete a reading activity. In this study, implicit means that all data is obtained without asking the user for feedback. To test the value of implicit feedback for assessing student proficiency, we collect features of user behavior and interaction, which are then used to train classification models. Based upon the feedback collected during these experiments, a student's performance on a quiz and proficiency relative to other students can be accurately predicted, which is a step on the path to our goal of providing automatic feedback and unintrusive evaluation in interactive learning environments.

P64 - Evaluation Methodologies

Friday, May 30, 14:55

Chairperson: **Kevin Bretonnel Cohen** Poster Session

ETER: a New Metric for the Evaluation of Hierarchical Named Entity Recognition

Mohamed Ben Jannet, Martine Adda-Decker, Olivier Galibert, Juliette Kahn and Sophie Rosset

This paper addresses the question of hierarchical named entity evaluation. In particular, we focus on metrics to deal with complex named entity structures as those introduced within the QUAERO project. The intended goal is to propose a smart way of evaluating partially correctly detected complex entities, beyond the scope

of traditional metrics. None of the existing metrics are fully adequate to evaluate the proposed QUAERO task involving entity detection, classification and decomposition. We are discussing the strong and weak points of the existing metrics. We then introduce a new metric, the Entity Tree Error Rate (ETER), to evaluate hierarchical and structured named entity detection, classification and decomposition. The ETER metric builds upon the commonly accepted SER metric, but it takes the complex entity structure into account by measuring errors not only at the slot (or complex entity) level but also at a basic (atomic) entity level. We are comparing our new metric to the standard one using first some examples and then a set of real data selected from the ETAPE evaluation results.

The ETAPE Speech Processing Evaluation

Olivier Galibert, Jeremy Leixa, Gilles Adda, Khalid Choukri and Guillaume Gravier

The ETAPE evaluation is the third evaluation in automatic speech recognition and associated technologies in a series which started with ESTER. This evaluation proposed some new challenges, by proposing TV and radio shows with prepared and spontaneous speech, annotation and evaluation of overlapping speech, a cross-show condition in speaker diarization, and new, complex but very informative named entities in the information extraction task. This paper presents the whole campaign, including the data annotated, the metrics used and the anonymized system results. All the data created in the evaluation, hopefully including system outputs, will be distributed through the ELRA catalogue in the future.

RECSA: Resource for Evaluating Cross-lingual Semantic Annotation

Achim Rettinger, Lei Zhang, Daša Berović, Danijela Merkler, Matea Srebačić and Marko Tadić

In recent years large repositories of structured knowledge (DBpedia, Freebase, YAGO) have become a valuable resource for language technologies, especially for the automatic aggregation of knowledge from textual data. One essential component of language technologies, which leverage such knowledge bases, is the linking of words or phrases in specific text documents with elements from the knowledge base (KB). We call this semantic annotation. In the same time, initiatives like Wikidata try to make those knowledge bases less language dependent in order to allow cross-lingual or language independent knowledge access. This poses a new challenge to semantic annotation tools which typically are language dependent and link documents in one language to a structured knowledge base grounded in the same language. Ultimately, the goal is to construct cross-lingual semantic annotation tools that can link words or phrases in one language to a structured knowledge database in any other language or to a language independent representation. To support this line of research we developed what we believe could serve as a gold standard Resource for Evaluating Cross-lingual Semantic Annotation (RECSA). We compiled a hand-annotated parallel corpus of 300 news articles in three languages with cross-lingual semantic groundings to the English Wikipedia and DBPedia. We hope that this new language resource, which is freely available, will help to establish a standard test set and methodology to comparatively evaluate cross-lingual semantic annotation technologies.

A Comparative Evaluation Methodology for NLG in Interactive Systems

Helen Hastie and Anja Belz

Interactive systems have become an increasingly important type of application for deployment of NLG technology over recent years. At present, we do not yet have commonly agreed terminology or methodology for evaluating NLG within interactive systems. In this paper, we take steps towards addressing this gap by presenting a set of principles for designing new evaluations in our comparative evaluation methodology. We start with presenting a categorisation framework, giving an overview of different categories of evaluation measures, in order to provide standard terminology for categorising existing and new evaluation techniques. Background on existing evaluation methodologies for NLG and interactive systems is presented. The comparative evaluation methodology is presented. Finally, a methodology for comparative evaluation of NLG components embedded within interactive systems is presented in terms of the comparative evaluation methodology, using a specific task for illustrative purposes.

Terminology Localization Guidelines for the National Scenario

Juris Borzovs, Ilze Ilzina, Iveta Keiša, Mārcis Pinnis and Andrejs Vasiljevs

This paper presents a set of principles and practical guidelines for terminology work in the national scenario to ensure a harmonized approach in term localization. These linguistic principles and guidelines are elaborated by the Terminology Commission in Latvia in the domain of Information and Communication Technology (ICT). We also present a novel approach in a corpus-based selection and an evaluation of the most frequently used terms. Analysis of the terms proves that, in general, in the normative terminology work in Latvia localized terms are coined according to these guidelines. We further evaluate how terms included in the database of official terminology are adopted in the

general use such as newspaper articles, blogs, forums, websites etc. Our evaluation shows that in a non-normative context the official terminology faces a strong competition from other variations of localized terms. Conclusions and recommendations from lexical analysis of localized terms are provided. We hope that presented guidelines and approach in evaluation will be useful to terminology institutions, regulative authorities and researchers in different countries that are involved in the national terminology work.

P65 - MultiWord Expressions and Terms

Friday, May 30, 14:55

Chairperson: **Valia Kordoni** Poster Session

TermWise: A CAT-tool with Context-Sensitive Terminological Support.

Kris Heylen, Stephen Bond, Dirk de Hertog, Ivan Vulić and Hendrik Kockaert

Increasingly, large bilingual document collections are being made available online, especially in the legal domain. This type of Big Data is a valuable resource that specialized translators exploit to search for informative examples of how domain-specific expressions should be translated. However, general purpose search engines are not optimized to retrieve previous translations that are maximally relevant to a translator. In this paper, we report on the TermWise project, a cooperation of terminologists, corpus linguists and computer scientists, that aims to leverage big online translation data for terminological support to legal translators at the Belgian Federal Ministry of Justice. The project developed dedicated knowledge extraction algorithms and a server-based tool to provide translators with the most relevant previous translations of domain-specific expressions relative to the current translation assignment. The functionality is implemented an extra database, a Term&Phrase Memory, that is meant to be integrated with existing Computer Assisted Translation tools. In the paper, we give an overview of the system, give a demo of the user interface, we present a user-based evaluation by translators and discuss how the tool is part of the general evolution towards exploiting Big Data in translation.

Extending the Coverage of a MWE Database for Persian CPs Exploiting Valency Alternations

Pollet Samvelian, Pegah Faghiri and Sarra El Ayari

PersPred is a manually elaborated multilingual syntactic and semantic Lexicon for Persian Complex Predicates (CPs), referred to also as "Light Verb Constructions" (LVCs) or "Compound Verbs". CPs constitutes the regular and the most common way of expressing verbal concepts in Persian, which has only around 200 simplex verbs. CPs can be defined as multi-word sequences formed by a verb and a non-verbal element and functioning in many respects as a simplex verb. Bonami & Samvelain (2010) and Samvelian & Faghiri (to appear) extendedly argue that Persian CPs are MWEs and consequently must be listed. The first delivery of PersPred, contains more than 600 combinations of the verb zadan 'hit' with a noun, presented in a spreadsheet. In this paper we present a semi-automatic method used to extend the coverage of PersPred 1.0, which relies on the syntactic information on valency alternations already encoded in the database. Given the importance of CPs in the verbal lexicon of Persian and the fact that lexical resources cruelly lack for Persian, this method can be further used to achieve our goal of making PersPred an appropriate resource for NLP applications.

Evaluation of Technology Term Recognition with Random Indexing

Behrang Zadeh and Siegfried Handschuh

In this paper, we propose a method that combines the principles of automatic term recognition and the distributional hypothesis to identify technology terms from a corpus of scientific publications. We employ the random indexing technique to model terms' surrounding words, which we call the context window, in a vector space at reduced dimension. The constructed vector space and a set of reference vectors, which represents manually annotated technology terms, in a k-nearest-neighbour voting classification scheme are used for term classification. In this paper, we examine a number of parameters that influence the obtained results. First, we inspect several context configurations, i.e. the effect of the context window size, the direction in which co-occurrence counts are collected, and information about the order of words within the context windows. Second, in the k-nearest-neighbour voting scheme, we study the role that neighbourhood size selection plays, i.e. the value of k. The obtained results are similar to word space models. The performed experiments suggest the best performing context are small (i.e. not wider than 3 words), are extended in both directions and encode the word order information. Moreover, the accomplished experiments suggest that the obtained results, to a great extent, are independent of the value of k.

Collaboratively Annotating Multilingual Parallel Corpora in the Biomedical Domain—some MANTRAs

Johannes Hellrich, Simon Clematide, Udo Hahn and Dietrich Rebholz-Schuhmann

The coverage of multilingual biomedical resources is high for the English language, yet sparse for non-English languages—an

observation which holds for seemingly well-resourced, yet still dramatically low-resourced ones such as Spanish, French or German but even more so for really under-resourced ones such as Dutch. We here present experimental results for automatically annotating parallel corpora and simultaneously acquiring new biomedical terminology for these under-resourced non-English languages on the basis of two types of language resources, namely parallel corpora (i.e. full translation equivalents at the document unit level) and (admittedly deficient) multilingual biomedical terminologies, with English as their anchor language. We automatically annotate these parallel corpora with biomedical named entities by an ensemble of named entity taggers and harmonize non-identical annotations the outcome of which is a so-called silver standard corpus. We conclude with an empirical assessment of this approach to automatically identify both known and new terms in multilingual corpora.

Aggregation Methods for Efficient Collocation Detection

Anca Dinu, Liviu Dinu and Ionut Sorodoc

In this article we propose a rank aggregation method for the task of collocations detection. It consists of applying some well-known methods (e.g. Dice method, chi-square test, z-test and likelihood ratio) and then aggregating the resulting collocations rankings by rank distance and Borda score. These two aggregation methods are especially well suited for the task, since the results of each individual method naturally forms a ranking of collocations. Combination methods are known to usually improve the results, and indeed, the proposed aggregation method performs better then each individual method taken in isolation.

An Evaluation of the Role of Statistical Measures and Frequency for MWE Identification

Sandra Antunes and Amália Mendes

We report on an experiment to evaluate the role of statistical association measures and frequency for the identification of MWE. We base our evaluation on a lexicon of 14.000 MWE comprising different types of word combinations: collocations, nominal compounds, light verbs + predicate, idioms, etc. These MWE were manually validated from a list of n-grams extracted from a 50 million word corpus of Portuguese (a subcorpus of the Reference Corpus of Contemporary Portuguese), using several criteria: syntactic fixedness, idiomaticity, frequency and Mutual Information measure, although no threshold was established, either in terms of group frequency or MI. We report on MWE that were selected on the basis of their syntactic and semantics properties while the MI or both the MI and the frequency show low values, which would constitute difficult cases to establish a cutting point. We analyze the MI values of the MWE selected in our gold dataset and, for some specific cases, compare these values with two other statistical measures.

P66 - Parsing

Friday, May 30, 14:55

Chairperson: **Giuseppe Attardi** Poster Session

The CMU METAL Farsi NLP Approach

Weston Feely, Mehdi Manshadi, Robert Frederking and Lori Levin

While many high-quality tools are available for analyzing major languages such as English, equivalent freely-available tools for important but lower-resourced languages such as Farsi are more difficult to acquire and integrate into a useful NLP front end. We report here on an accurate and efficient Farsi analysis front end that we have assembled, which may be useful to others who wish to work with written Farsi. The pre-existing components and resources that we incorporated include the Carnegie Mellon TurboParser and TurboTagger (Martins et al., 2010) trained on the Dadegan Treebank (Rasooli et al., 2013), the Uppsala Farsi text normalizer PrePer (Seraji, 2013), the Uppsala Farsi tokenizer (Seraji et al., 2012a), and Jon Dehdari's PerStem (Jadidinejad et al., 2010). This set of tools (combined with additional normalization and tokenization modules that we have developed and made available) achieves a dependency parsing labeled attachment score of 89.49%, unlabeled attachment score of 92.19%, and label accuracy score of 91.38% on a held-out parsing test data set. All of the components and resources used are freely available. In addition to describing the components and resources, we also explain the rationale for our choices.

Constituency Parsing of Bulgarian: Word- vs Class-based Parsing

Masood Ghayoomi, Kiril Simov and Petya Osenova

In this paper, we report the obtained results of two constituency parsers trained with BulTreeBank, an HPSG-based treebank for Bulgarian. To reduce the data sparsity problem, we propose using the Brown word clustering to do an off-line clustering and map the words in the treebank to create a class-based treebank. The observations show that when the classes outnumber the POS tags, the results are better. Since this approach adds on another dimension of abstraction (in comparison to the lemma), its coarse-grained representation can be used further for training statistical parsers.

A System for Experiments with Dependency Parsers

Kiril Simov, Iliana Simova, Ginka Ivanova, Maria Mateva and Petya Osenova

In this paper we present a system for experimenting with combinations of dependency parsers. The system supports initial training of different parsing models, creation of parsebank(s) with these models, and different strategies for the construction of ensemble models aimed at improving the output of the individual models by voting. The system employs two algorithms for construction of dependency trees from several parses of the same sentence and several ways for ranking of the arcs in the resulting trees. We have performed experiments with state-of-the-art dependency parsers including MaltParser, MSTParser, TurboParser, and MATEParser, on the data from the Bulgarian treebank – BulTreeBank. Our best result from these experiments is slightly better then the best result reported in the literature for this language.

An Out-of-Domain Test Suite for Dependency Parsing of German

Wolfgang Seeker and Jonas Kuhn

We present a dependency conversion of five German test sets from five different genres. The dependency representation is made as similar as possible to the dependency representation of TiGer, one of the two big syntactic treebanks of German. The purpose of these test sets is to enable researchers to test dependency parsing models on several different data sets from different text genres. We discuss some easy to compute statistics to demonstrate the variation and differences in the test sets and provide some baseline experiments where we test the effect of additional lexical knowledge on the out-of-domain performance of two state-of-the-art dependency parsers. Finally, we demonstrate with three small experiments that text normalization may be an important step in the standard processing pipeline when applied in an out-of-domain setting.

Dependency Parsing Representation Effects on the Accuracy of Semantic Applications - an Example of an Inflective Language

Lauma Pretkalnina, Arturs Znotinš, Laura Rituma and Didzis Goško

In this paper we investigate how different dependency representations of a treebank influence the accuracy of the dependency parser trained on this treebank and the impact on several parser applications: named entity recognition, coreference resolution and limited semantic role labeling. For these experiments we use Latvian Treebank, whose native annotation format is dependency based hybrid augmented with phrase-like elements. We explore different representations of coordinations, complex predicates and punctuation mark attachment. Our experiments shows that parsers trained on the variously transformed treebanks vary significantly in their accuracy, but the best-performing parser as measured by attachment score not always leads to best accuracy for an end application.

Validation Issues induced by an Automatic Pre-Annotation Mechanism in the Building of Non-projective Dependency Treebanks

Ophélie Lacroix and Denis Béchet

In order to build large dependency treebanks using the CDG Lab, a grammar-based dependency treebank development tool, an annotator usually has to fill a selection form before parsing. This step is usually necessary because, otherwise, the search space is too big for long sentences and the parser fails to produce at least one solution. With the information given by the annotator on the selection form the parser can produce one or several dependency structures and the annotator can proceed by adding positive or negative annotations on dependencies and launching iteratively the parser until the right dependency structure has been found. However, the selection form is sometimes difficult and long to fill because the annotator must have an idea of the result before parsing. The CDG Lab proposes to replace this form by an automatic pre-annotation mechanism. However, this model introduces some issues during the annotation phase that do not exist when the annotator uses a selection form. The article presents those issues and proposes some modifications of the CDG Lab in order to use effectively the automatic pre-annotation mechanism.

Automatic Refinement of Syntactic Categories in Chinese Word Structures

Jianqiang Ma

Annotated word structures are useful for various Chinese NLP tasks, such as word segmentation, POS tagging and syntactic parsing. Chinese word structures are often represented by binary trees, the nodes of which are labeled with syntactic categories, due to the syntactic nature of Chinese word formation. It is desirable to refine the annotation by labeling nodes of word structure trees with more proper syntactic categories so that the combinatorial properties in the word formation process are better captured. This can lead to improved performances on the tasks that exploit word structure annotations. We propose syntactically inspired algorithms to automatically induce syntactic categories of word structure trees using POS tagged corpus and branching in existing Chinese word structure trees. We evaluate the quality of our

annotation by comparing the performances of models based on our annotation and another publicly available annotation, respectively. The results on two variations of Chinese word segmentation task show that using our annotation can lead to significant performance improvements.

P67 - Part-of-Speech Tagging

Friday, May 30, 14:55

Chairperson: **Daniel Flickinger** Poster Session

Experiences with Parallelisation of an Existing NLP Pipeline: Tagging Hansard

Stephen Wattam, Paul Rayson, Marc Alexander and Jean Anderson

This poster describes experiences processing the two-billion-word Hansard corpus using a fairly standard NLP pipeline on a high performance cluster. Herein we report how we were able to parallelise and apply a traditional single-threaded batch-oriented application to a platform that differs greatly from that for which it was originally designed. We start by discussing the tagging toolchain, its specific requirements and properties, and its performance characteristics. This is contrasted with a description of the cluster on which it was to run, and specific limitations are discussed such as the overhead of using SAN-based storage. We then go on to discuss the nature of the Hansard corpus, and describe which properties of this corpus in particular prove challenging for use on the system architecture used. The solution for tagging the corpus is then described, along with performance comparisons against a naive run on commodity hardware. We discuss the gains and benefits of using high-performance machinery rather than relatively cheap commodity hardware. Our poster provides a valuable scenario for large scale NLP pipelines and lessons learnt from the experience.

Adapting a Part-of-Speech Tagset to Non-Standard Text: the Case of STTS

Heike Zinsmeister, Ulrich Heid and Kathrin Beck

The Stuttgart-Tübingen TagSet (STTS) is a de-facto standard for the part-of-speech tagging of German texts. Since its first publication in 1995, STTS has been used in a variety of annotation projects, some of which have adapted the tagset slightly for their specific needs. Recently, the focus of many projects has shifted from the analysis of newspaper text to that of non-standard varieties such as user-generated content, historical texts, and learner language. These text types contain linguistic phenomena that are missing from or are only suboptimally covered by STTS; in a community effort, German NLP researchers have therefore proposed additions to and modifications of the tagset that will handle these phenomena more appropriately. In addition, they have discussed alternative ways of tag assignment in terms of bipartite tags (stem, token) for historical texts and tripartite tags (lexicon, morphology, distribution) for learner texts. In this article, we report on this ongoing activity, addressing methodological issues and discussing selected phenomena and their treatment in the tagset adaptation process.

TALC-Sef a Manually-revised POS-Tagged Literary Corpus in Serbian, English and French

Antonio Balvet, Dejan Stosic and Aleksandra Miletic

In this paper, we present a parallel literary corpus for Serbian, English and French, the TALC-sef corpus. The corpus includes a manually-revised pos-tagged reference Serbian corpus of over 150,000 words. The initial objective was to devise a reference parallel corpus in the three languages, both for literary and linguistic studies. The French and English sub-corpora had been pos-tagged from the onset, using TreeTagger (Schmid, 1994), but the corpus lacked, until now, a tagged version of the Serbian sub-corpus. Here, we present the original parallel literary corpus, then we address issues related to pos-tagging a large collection of Serbian text: from the conception of an appropriate tagset for Serbian, to the choice of an automatic pos-tagger adapted to the task, and then to some quantitative and qualitative results. We then move on to a discussion of perspectives in the near future for further annotations of the whole parallel corpus.

An Open Source Part-of-Speech Tagger for Norwegian: Building on Existing Language Resources

Cristina Sánchez Marco

This paper presents an open source part-of-speech tagger for the Norwegian language. It describes how an existing language processing library (FreeLing) was used to build a new part-of-speech tagger for this language. This part-of-speech tagger has been built on already available resources, in particular a Norwegian dictionary and gold standard corpus, which were partly customized for the purposes of this paper. The results of a careful evaluation show that this tagger yields an accuracy close to state-of-the-art taggers for other languages.

Standardisation and Interoperation of Morphosyntactic and Syntactic Annotation Tools for Spanish and their Annotations

Antonio Pareja-Lora, Guillermo Cárcamo-Escorza and Alicia Ballesteros-Calvo

Linguistic annotation tools and linguistic annotations are scarcely syntactically and/or semantically interoperable. Their low

interoperability usually results from the number of factors taken into account in their development and design. These include (i) the type of phenomena annotated (either morphosyntactic, syntactic, semantic, etc.); (ii) how these phenomena are annotated (e.g., the particular guidelines and/or schema used to encode the annotations); and (iii) the languages (Java, C++, etc.) and technologies (as standalone programs, as APIs, as web services, etc.) used to develop them. This low level of interoperability makes it difficult to reuse both the linguistic annotation tools and their annotations in new scenarios, e.g., in natural language processing (NLP) pipelines. In spite of this, developing new linguistic tools from scratch is quite a high time-consuming task that also entails a very high cost. Therefore, cost-effective ways to systematically reuse linguistic tools and annotations must be found urgently. A traditional way to overcome reuse and/or interoperability problems is standardisation. In this paper, we present a web service version of FreeLing that provides standard-compliant morpho-syntactic and syntactic annotations for Spanish, according to several ISO linguistic annotation standards and standard drafts.

P68 - Tools, Systems, Applications

Friday, May 30, 14:55

Chairperson: **Lluís Padró** Poster Session

Exploring and Visualizing Variation in Language Resources

Peter Fankhauser, Jörg Knappen and Elke Teich

Language resources are often compiled for the purpose of variational analysis, such as studying differences between genres, registers, and disciplines, regional and diachronic variation, influence of gender, cultural context, etc. Often the sheer number of potentially interesting contrastive pairs can get overwhelming due to the combinatorial explosion of possible combinations. In this paper, we present an approach that combines well understood techniques for visualization heatmaps and word clouds with intuitive paradigms for exploration drill down and side by side comparison to facilitate the analysis of language variation in such highly combinatorial situations. Heatmaps assist in analyzing the overall pattern of variation in a corpus, and word clouds allow for inspecting variation at the level of words.

Introducing a Web Application for Labeling, Visualizing Speech and Correcting Derived Speech Signals

Raphael Winkelmann and Georg Raess

The advent of HTML5 has sparked a great increase in interest in the web as a development platform for a variety of different research applications. Due to its ability to easily deploy software to remote clients and the recent development of standardized browser APIs, we argue that the browser has become a good platform to develop a speech labeling tool for. This paper introduces a preliminary version of an open-source client-side web application for labeling speech data, visualizing speech and segmentation information and manually correcting derived speech signals such as formant trajectories. The user interface has been designed to be as user-friendly as possible in order to make the sometimes tedious task of transcribing as easy and efficient as possible. The future integration into the next iteration of the EMU speech database management system and its general architecture will also be outlined, as the work presented here is only one of several components contributing to the future system.

AraNLP: a Java-based Library for the Processing of Arabic Text

Maha Althobaiti, Udo Kruschwitz and Massimo Poesio

We present a free, Java-based library named "AraNLP" that covers various Arabic text preprocessing tools. Although a good number of tools for processing Arabic text already exist, integration and compatibility problems continually occur. AraNLP is an attempt to gather most of the vital Arabic text preprocessing tools into one library that can be accessed easily by integrating or accurately adapting existing tools and by developing new ones when required. The library includes a sentence detector, tokenizer, light stemmer, root stemmer, part-of speech tagger (POS-tagger), word segmenter, normalizer, and a punctuation and diacritic remover.

Applying Accessibility-Oriented Controlled Language (CL) Rules to Improve Appropriateness of Text Alternatives for Images: an Exploratory Study

Silvia Rodríguez Vázquez, Pierrette Bouillon and Anton Bolfing

At present, inappropriate text alternatives for images in the Web continue to pose web accessibility barriers for people with special needs. Although research efforts have been devoted to define how to write text equivalents for visual content in websites, existing guidelines often lack direct linguistic-oriented recommendations. Similarly, most web accessibility evaluation tools just provide users with an automated functionality to check the presence of text alternatives within the <img> element, rather than a platform to verify their content. This paper presents an overview of the findings from an exploratory study carried out to investigate if the appropriateness level of text alternatives for images in French can be improved when applying controlled language (CL)

rules. Results gathered suggest that using accessibility-oriented alt style rules can have a significant impact on text alternatives' appropriateness. Although more data would be needed to draw further conclusions about our proposal, this preliminary study already offers an interest insight into the potential use of CL checkers such as Acrolinx for language-based web accessibility evaluation.

GraPAT: a Tool for Graph Annotations

Jonathan Sonntag and Manfred Stede

We introduce GraPAT, a web-based annotation tool for building graph structures over text. Graphs have been demonstrated to be relevant in a variety of quite diverse annotation efforts and in different NLP applications, and they serve to model annotators' intuitions quite closely. In particular, in this paper we discuss the implementation of graph annotations for sentiment analysis, argumentation structure, and rhetorical text structures. All of these scenarios can create certain problems for existing annotation tools, and we show how GraPAT can help to overcome such difficulties.

Discovering the Italian Literature: Interactive Access to Audio-indexed Text Resources

Vincenzo Galatà, Alberto Benin, Piero Cosi, Giuseppe Riccardo Leone, Giulio Paci, Giacomo Sommavilla and Fabio Tesser

In this paper we present a web interface to study Italian through the access to read Italian literature. The system allows to browse the content, search for specific words and listen to the correct pronunciation produced by native speakers in a given context. This work aims at providing people who are interested in learning Italian with a new way of exploring the Italian culture and literature through a web interface with a search module. By submitting a query, users may browse and listen to the results through several modalities including: a) the voice of a native speaker: if an indexed audio track is available, the user can listen either to the query terms or to the whole context in which they appear (sentence, paragraph, verse); b) a synthetic voice: the user can listen to the results read by a text-to-speech system; c) an avatar: the user can listen to and look at a talking head reading the paragraph and visually reproducing real speech articulatory movements. In its up to date version, different speech technologies currently being developed at ISTC-CNR are implemented into a single framework. The system will be described in detail and hints for future work are discussed.

Creating Summarization Systems with SUMMA

Horacio Saggion

Automatic text summarization, the reduction of a text to its essential content is fundamental for an on-line information society. Although many summarization algorithms exist, there are few tools or infrastructures providing capabilities for developing summarization applications. This paper presents a new version of SUMMA, a text summarization toolkit for the development of adaptive summarization applications. SUMMA includes algorithms for computation of various sentence relevance features and functionality for single and multidocument summarization in various languages. It also offers methods for content-based evaluation of summaries.

Authors Index