The ELRA Newsletter

July - December 2005

Vol.10 n.3 & 4

Contents

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.org
Web sites:
http://www.elra.info or
http://www.elda.org

Dear Colleagues,

In this last double issue of 2005, we would like to focus on Evaluation and Machine Translation.

During the last quarter of 2005, a number of events related to Evaluation took place :

- HLT Evaluation workshop, held in Malta on December 1st and 2nd to celebrate the 10th anniversary of ELRA : a quick overview is being given here and follow-up actions through the organisation of a workshop at LREC 2006, for instance, will be taken in 2006.

- HLT Evaluation portal (http://www.hlt-evaluation.org): The portal has been opened and now contains a number of valuable information on our field.

Further to the MT Summit held in Thailand in September 2005, this newsletter contains several articles related to Machine Translation:

- An overview of the MT Summit: “MT Summit X and New Development of MT”

-The need of translation of huge volumes of documents has fostered the development of machine translation (MT) systems as described in “Speech in Machine Translation and Computer-Assisted Translation”

- “What is Statistical about Statistical Machine Translation?” In recent years, the use of statistics as an alternative approach for machine translation applications has been gaining more and more enthusiasts.

The organisation of our conference LREC 2006 in Genoa in May 2006 has continued and the last quarter of 2005 has been mainly dedicated to the scientific reviews and selections of the papers. The very high number of submissions (over 800) in all the fields covered by LREC (speech, written, multimodal and evaluation) shows the growing interest of the enlarged community in LREC.
All information can be found on our web site: http://www.lrec-conf.org/lrec2006/

New resources have been secured for distribution. These are announced in the last section of this newsletter and consist of :

- S0181 SALA II Spanish from Costa Rica database
- S0182 SALA II Spanish from Argentina database
- S0183 OrienTel Morocco MCA (Modern Colloquial Arabic) database
- S0184 OrienTel Morocco MSA (Modern Standard Arabic) database
- S0185 OrienTel French as spoken in Morocco database
- S0186 OrienTel Tunisia MCA (Modern Colloquial Arabic) database
- S0187 OrienTel Tunisia MSA (Modern Standard Arabic) database
- S0188 OrienTel French as spoken in Tunisia database
- S0189 OrienTel Hebrew database
- S0190 OrienTel Arabic as spoken in Israel database
- L0057 Euskararen Datu-Base Lexikala (EDBL) – Lexical Database for Basque

Once again, if you would like to join ELRA and benefit from its services (that are summarized at www.elra.info), please, do not hesitate to contact us.

Bente Maegaard, President

Khalid Choukri, CEO

MT Summit X and New Development of MT

Jun-ichi Tsujii

MT Summit X was held at Phuket, Thailand from September 13 to 15, 2005. It was the fourth Summit in Asia, following Hakone (1987), Kobe (1993) and Singapore (1999). It successfully attracted more than 250 participants from 30 countries.

While the USA has invested substantially in MT and multilingual NLP for a decade, Europe and Japan, who used to be major players in the field, have been rather inactive for quite a few years. However, the large number of participants of MT Summit X proves that there is renewed interest in the field in Asia. The fact that three invited speakers out of six were from China means that there is strong interest in MT treating Chinese language. The Digital Olympics project in China, which aims to provide multi-lingual real-time information services at the Beijing Olympics 2008 and in which Machine Translation will play a major role, motivated the participants. Many participants seem to be very enthusiastic about the special issue of AAMT (Asian Pacific Association of MT) journal on MT activities in Asia as well.

SMT (Statistical MT) has reached its optimum in research. One of the main speakers, Herman Ney, who gave an exciting invited talk on SMT, told me that, when he attended the MT Summit for the first time at Santiago de Compostela in Spain, 2001, he felt that SMT was on the verge of the MT community and that people looked curiously at their work as interesting research but did not really take it seriously. The situation has changed significantly since then. SMT has become the mainstream of MT research or at least one of the focus points. Other paradigms such as rule based MT systems have also adopted a more data-driven approach, such as the automatic acquisition of lexica or rules, while the example-based MT and the SMT are becoming more and more similar.

However, as Ed Hovy, former president of IAMT, eloquently said at one of the panels, MT has a long history in which "new" paradigms appeared to be promising, but before they achieved their much-expected breakthroughs they were replaced with another paradigm. While the SMT has contributed significantly to the field and shed new light, the limitations of the paradigm have become clear to many researchers, who have the feeling of déjà vu that, while one paradigm has reached its peak, we have yet to overcome major obstacles.

I was involved in rule-based and linguistics-based paradigms at MT during the 80's and early 90's. These paradigms failed to live up to their expectations, partly because we did not have the technology for it. Linguistics-based grammar was not robust enough for dealing with real-world text and the parsing technology based on such grammar was still in the early stages of development. For example, it used to take either a few minutes or a few hours to parse a sentence of normal length by unification-based grammar. However, the technology has made significant progress since then. Our group at the University of Tokyo recently succeeded in parsing the whole abstracts in Medline (1,418,949,650 words) by using HPSG-based grammar. The grammar is robust enough for dealing with actual sentences in abstracts in Bio-medical papers and the parser we have developed is efficient enough. A sentence whose average length is more than 20 words can be parsed in an average of less than one second.

Even though we have yet to fully understand the integration of linguistics-based high-precision parsing technology with the SMT and other data-driven MT paradigms, I believe that the merging of several MT paradigms, including the SMT, Linguistics-based and Knowledge-based MT, will be possible sooner than expected. High-quality MT, which is more in demand in wider markets than fast but average-quality MT, cannot be attained only by the pure form of SMT.

The MT Summit X will be remembered as a milestone in the future development of MT.

Jun-ichi TSUJII
Professor of Natural Language Processing at the University of Tokyo
Director of National Centre for Text Mining, and Professor of Text Mining at the School of Informatics University of Manchester, UK
President of IAMT (International Association of MT, 2003-2005)

What is Statistical about Statistical Machine Translation?

Rafael E. Banchs

The problem of machine translation is one of the most complex human-task problems in the artificial intelligence's agenda. The generation of correct translations requires information and knowledge about morphology, syntax, grammar, semantics and also, in many cases, the cultural and social contexts.

In recent years, the use of statistics as an alternative approach for machine translation applications has been gaining more and more enthusiasts. It is not new for many people that statistics is one of the most controversial areas of mathematics, and consequently, as it has been the case in many other scientific applications, the use of statistics in machine translation has generated some interesting polemics among researches. For example, it is very outrageous for many people the idea that, in principle, a statistical machine translation system can be built without having any linguistic knowledge about the languages involved.

In this sense, researches in machine translation technologies are somehow divided into three great groups. The first group, led by linguists, is composed by those who think that only the appropriate use of specialized linguistic knowledge can provide translation systems capable of achieving human-like translation qualities. The second group, which is mainly integrated by mathematicians and related scientists, corresponds to those who think that, on the other hand, it is only a matter of time for being able to achieve human-like translations by using a pure statistical approach. The only requirements would be the availability of larger data sets and more computational capabilities.

Finally, the third group, which is mainly integrated by engineers, thinks that human-like translation quality will be only achieved by smartly combining the linguistic knowledge approach with the statistical approach. Although this third approach, at a first glance, seems to be the most reasonable one, the continuous failure of recent efforts for achieving better translation qualities by combining statistics and linguistic knowledge has strongly discouraged this third group.

From a theoretical perspective, both approaches, the one based on linguistic knowledge and the one based on statistics, should be able to provide human-like translations. The problem is that in practice, both approaches have serious limitations regarding the implementation of a translation system. In any of the two approaches, the problem is basically the same, it would be the construction of models or rules for decoding and recoding information from one language to another. In the case of the linguistic knowledge approach, the achievement of correct translations would then require the possibility of representing all rules and exceptions embedded in the interpretation and generation processes for both involved languages, source and target, as well as a very broad and complete representation of the real world. Otherwise, in the case of the statistical approach, the achievement of correct translations would require the possibility of estimating probabilities for each conceivable bilingual pair of sentences between two languages.

Therefore, the basic difference between these two approaches is the way in which the rules are constructed. In the first case, the linguistic rules, which are previously known, are "wired" directly into the machine translation system. In the second case, the information conveyed by the "rules of languages" is inferred from bilingual data and represented in the form of statistical models by using probabilities. In both cases, the exact solution of the machine translation problem is still unachievable for today's technologies and, in the practice, such a solution can be only approximated.

The actual reason of why statistics as an alternative approach for machine translation applications has been gaining more and more enthusiasts in recent years is simply a technical reason: the increasing availability of large amounts of bilingual data and a more powerful computational capacity. However, the fundamentals of statistical machine translation can be traced back to the message decoding problems studied during World War II.

The formulation of the translation problem from a statistical point of view is very simple and it is based on the concept of conditional probabilities. The conditional probability of an event A given an event B, which is denoted as P(A|B), is defined as the joint probability of both events divided by the probability of event B, i.e. P(A and B) / P(B). In other words, it corresponds to a normalization of the probability of event A with respect to the subspace defined by event B. For example, consider the experiment of throwing a dice. The probability of getting an output which is less than three -event A- given that the output is an even number -event B-, P(A|B), equals 1/3. This result is obtained by dividing the joint probability P(A and B), i.e. the probability of getting a two, which equals 1/6, by the probability of getting an even number P(B), which equals 3/6. So, P(A|B) = P(A and B) / P(B) = (1/6) / (3/6) = 1/3. Alternatively, in this example, P(A|B) corresponds to the probability of getting an output which is less than three computed with respect to the subspace defined by the dice's even outputs: 2, 4 and 6; which is actually one out of three possibilities. Then, P(A|B) = 1/3.

Taking into account the previous definition, the translation problem can be defined in terms of an optimization problem in the statistical framework as follows. Consider a source language sentence, S, which should be translated into a target language sentence T. The translation problem is then implemented as a search, over the space of all possible target language sentences Ts, for the target sentence which maximizes the conditional probability P(T|S). This problem, just as stated, is not solvable in the practice because of two principal difficulties. First, it implies the computation of conditional probabilities for each conceivable bilingual pair of sentences S and T, and second, the search must be performed

over a practically infinite search space.

In the practical implementation of a statistical machine translation system these two problems are handled as follows. In the case of the modeling problem, the conditional probability P(T|S) is decomposed into other probabilities which are more easily estimated from data. The most classical decomposition is the so called noisy channel approach, which is based on the Bayes theorem. According to this theorem P(T|S) equals P(S|T) P(T) / P(S), and when searching over the space of the target language sentences, the optimization problem is reduced to the search of the sentence T which maximizes the product P(S|T) P(T). These new probabilities are referred to as the translation model and the target language model, respectively. Similarly, these new model probabilities are also decomposed into probabilities which are easier to estimate, as for example conditional probabilities between bilingual pair of words (word-based SMT systems) or group of words (phrase-based SMT systems) in the case of the translation models; and conditional probabilities among sequences of words (n-gram-based models) in the case of the target language models. In a more recent decomposition approach, which is based on the maximum entropy framework, the conditional probability P(T|S) is decomposed into a weighted combination of many different models which are referred to as feature functions. In this new framework, the noisy channel approach constitutes a particular case for which only two feature functions are considered.

Regarding the search problem, which is also referred to as decoding, it actually constitutes an integer optimization problem for which the search must be performed over a practically infinite space. So, in the practice, the methods used for decoding restrict the search to "interesting" subregions of the search space. In this way, although they do not guarantee optimal solutions, they hugely reduce the computational costs of the actual search problem.

In summary, after this brief overview about the statistical approach to machine translation, the original question may be answered: the only thing that is actually statistical about statistical machine translation is the approach. And it seems to be the case that this approach will continue gaining enthusiasts over the years, unless the human-language translation task is finally undestood; because, as in any other area of science where a given phenomenum cannot be totally explained, as far as the phenomenum is observable it will be always possible to compute statistics. Meanwhile, many stubborn researchers in the machine translation community will continue trying to combine statistics with linguistic knowledge.

Rafael E. Banchs
rbanchs@gps.tsc.upc.edu
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla, Universitat Politècnica de Catalunya, Barcelona, Spain.

Speech in Machine Translation and Computer-Assisted Translation

Francisco Casacuberta, Enrique Vidal

Translation services are fundamental in the administrative organization of the European Union (EU).

Moreover, some EU members have more than one official language (e.g. Spain, Belgium, etc.). In these cases, the official writings, the Parliament speeches, etc. must be produced in all the official languages. On the other hand, the translation of many court documents becomes crucial to avoid a slow down of the proceedings. Finally, the translation to other non official EU languages is becoming increasingly important due to the existence of an important immigrant community (medical consultation, legal consulting, services, etc.)

The need of translation of huge volumes of documents has fostered the development of *machine translation* (MT) systems. Nowadays, there are many commercial MT systems for *text-to-text translation* (T2TT) available [8]. But T2TT systems are far to be perfect and high-quality translation is required in many cases. In practice, T2TT systems generally need human post-processing to correct the possible errors incurred by the system. However, this is a time consuming process due to the required human effort (generally, less than the effort needed to translate the document without any tool). An alternative to increase the productivity of the whole translation process (T2TT plus human work) is to incorporate the human correction activities within the translation process itself, in a *computer-assisted translation* (CAT) system [13]. The idea is to use a T2TT system in an iterative process where human translator activity is included in the loop [9, 13, 21, 11, 2].

In spite of the limitations of the present T2TT systems, they are widely accepted in many private and public organizations. Unfortunately, this is not the situation of real *speech-to-speech* (or *speech-to-text*) *translation* (S2ST), where the reliable systems are laboratory prototypes because the present technology only allows S2ST in very restricted domains [27, 1, 15, 5].

However, speech can play an important role in MT, since the human translator can use speech recognizers in addition to the keyboard and the mouse. He or she can dictate portions of target sentences that the system has produced with errors in order to correct them as well as to give commands to the CAT system [12, 24, 28].

Text translation

Corpus-based approaches have become very appealing for MT since they allow to exploit available bilingual data. In this framework, the *statistical* models can be estimated from a bilingual corpus using powerful training algorithms and the translation process is based on different families of efficient search algorithms [16]. Statistical models were initially based on an automatically derived word-to-word dictio-

nary (*statistical dictionaries*) and alignments between word positions (*statistical alignment models*) [3, 19]. In this case, the basic assumption is that each source word is generated by only one target word. This assumption often fails in natural language; in some cases, it is necessary to know the context of the word to be translated. One way to upgrade this simple assumption comes from the so called *alignment templates* (AT) [20] or the *phrase-based* (PB) models [25, 29]. In these approaches, an entire group of adjacent words in the source sentence may be aligned with an entire group of adjacent target words (*bilingual phrases* or *bilingual segments*). The main difference between AT and PB models is that in the later case, each bilingual segment is considered as a whole unit without an internal structure, while in an AT model, a bilingual segment is based on the words and the alignments that constitute it. In some tasks (and for some pairs of languages), the same order in the sequence of source segments and in the sequence of the corresponding target segments can be assumed (*monotone PB* models) [26].

Another model, closely related with monotone PB models is the *stochastic finite-state transducer* (SFST) [6], where a finite-state representation of the relationship between sequences of bilingual phrases is assumed [6].

Under the statistical approach, building MT systems require less human effort than under the more traditional *linguistic* approach [8] but, in this case, large bilingual corpora are needed to produce reliable MT systems.

A recent comparison of different MT systems has been published in [18], where a system based on the statistical approach reportedly achieved the best performance in several Arabic-to-English and Chinese-to-English tasks.

An important current research activity is how to incorporate linguistic information in the statistical MT systems in order to increase the performance of the system and to reduce the size of the required training corpus [7, 17].

Speech translation

Speech translation is a great challenge in MT research. Most of the current S2ST systems are based on a two-step process: automatic speech recognition (ASR) and a T2TT (*decoupled* approach to S2ST). The main problem with such an approach is that the current ASR systems are not error free and most T2TT systems assume that the input does not contain errors. As an alternative, a full, tight *integration* of the speech recognition *and* translation processes has been pursued [27].

SFSTs are particularly appropriate for a complete integration of recognition-translation [5]. This integration can be carried out in a similar way as for speech recognition where the acoustic models (hidden Markov models) are embedded into the language model (*n*-grams). In this case, the acoustic models are embedded into the finite-state transducer. One of the appeals of these models is that search (the actual translation process) can be performed using the Viterbi algorithm. However, the models can be very large and, in practice, they seem to be adequate only for restricted tasks.

In the EuTrans project [27], results were obtained using both integrated and decoupled architectures to S2ST for several tasks in a tourist domain [5]. Integrated architecture seems more adequate in low perplexity tasks and provided that enough training data is available. Experimental translation word error rates ranged from 13%, in a quite realistic semi-artificial telephone-input Spanish-English task, to close to 38% in a more realistic Italian-English task. A subjective evaluation of the results allowed us to conclude that, in most cases, the translations preserved the original meaning of the sentences.

The approaches proposed in the EuTrans project were used to develop other S2ST systems for Portuguese-to-English [23] and Spanish-to-Basque [22] and for other tasks [14].

Computer-assisted translation

A typical solution to improve the quality of the translations supplied by a MT system is to perform a post-process by a human translator. In this case, however, the MT system does not take advantage of the human translator knowledge and the human translator does not take advantage of the potential adapting ability of a (statistical) MT system.

An alternative way to take advantage of the existing MT technologies is to allow the participation of the human translator during the process of translation in a CAT framework, rather than at the end of this process [13]. Historically, CAT and MT have been considered close but different technologies [8]. An innovative solution proposed recently in the *Transtype* [13] and *TransType2* [11] (TT2) projects is to embed a MT engine within an interactive translation environment. In this way, the system combines the best of two paradigms: the CAT paradigm, in which the human translator ensures high-quality output, and the MT paradigm, in which the machine ensures significant productivity gains. This approach is called *interactive* CAT.

In interactive CAT, a machine translation system produces portions of the target sentence that can be accepted or amended by a human translator and these correct portions are then used by the MT system as additional information to achieve further, hopefully improved suggestions. More specifically, in each iteration, a prefix of the target sentence has somehow been fixed by the human translator in the previous iteration and the MT system computes its best (or N-best) translation suffix hypothesis to complete this prefix.

Different techniques can be used for the MT engine: SFST, AT and PB models [2]. Existing search algorithms can be adapted in order to provide completions (rather than full translations) in a very efficient way [9, 21].

In the TT2 project [11, 2], two different tasks involving the translation of printer user manuals and the Bulletin of the European Union were considered to assess statistical approaches to interactive CAT. The pairs of languages were: English-Spanish, English-French and

English- German. Results showed that the use of such an approach can reduce the number of keystrokes needed to produce a translation more than 50% with respect to typing the whole translated document and about 15% with respect the keystrokes needed to correct the result of a MT system.

Speech recognition for computer-assisted translation

An important feature of the interactive CAT approach proposed in [2] is that the human translator can correct or accept the suggestions from the system using different interfaces. Typically, the keyboard and the mouse are the most widely used devices, but others can be used. Speech is one of the most natural ways of communication for human beings and it was explored in [28]. An idea, already proposed in previous works [10, 4], is that a human translator would dictate aloud the translation in the target language. Note the big difference with respect to S2ST. In speech translation, the system has to deal both with speech input *in the source language* and with the translation into the target language. Here, in contrast, the translation process would only be the responsibility of the human and the system should deal only with speech recognition *in the target language*. As the source text is known by the system, this knowledge can be used to reduce speech recognition errors. Unfortunately, however, current speech recognition technology does not allow for a sufficiently high accuracy and human post-processing would still be required.

An alternative to this idea is proposed in [28] within the interactive CAT framework. In this case, the human translator determines acceptable prefixes of the suggestions made by the system by reading (with possible modifications) parts of these suggestions. In this way, a much lower degree of freedom is possible and the correspondingly lower perplexity allows for sufficiently high recognition accuracy. Moreover, as this is fully integrated within the CAT paradigm, the user can make use of the conventional means (keyboard and/or mouse) to guarantee that the produced text achieves the required level of quality. Preliminary empirical results, presented in [28], support the potential usefulness of using speech within the CAT paradigm.

Conclusions

The present state-of-the-art of (statistical) MT allows us to obtain fully-automated T2TT with an acceptable quality in some cases and high quality translations using interactive CAT. However, this is not true for S2ST where acceptable results can only be obtained for very restricted tasks. Nevertheless, speech will play other important roles in MT, since it can be used by the human translators to dictate portions of the target sentences in interactive CAT systems. In the future, we speculate that the ideas behind the use of speech in interactive CAT can be exported to interpretation, i.e. a CAT system for S2ST with a small time delay.

References

[1] J. C. Amengual, J. C., J. M. Benedí, A. Castaño, A. Castellanos, V. M. Jiménez, D. Llorens, A. Marzal, M. Pastor, F. Prat, E. Vidal, and J. M. Vilar, "The EuTrans-I speech translation system", in *Machine Translation*, 15:75–103. 2000.

[2] S. Barrachina, O. Bender, J. Civera, E. Cubel. S. Khadivi, A. Lagarda, H. Ney, J. Tomás, E. Vidal and J.M. Vilar "Statistical and finite-state approaches to computer-assisted translation", to be published. 2006.

[3] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. "The mathematics of statistical machine translation: Parameter estimation", in *Computational Linguistics*, 19(2):263–310. 1993.

[4] P. Brown, S. Chen, S. D. Pietra, V. D. Pietra, S. Kehler, and R. Mercer, "Automatic speech recognition in machine aided translation", in *Computer Speech and Language*, 8:177–187, 1994.

[5] F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, and C. Tillmann. "Some approaches to statistical and finite-state speech-to-speech translation", in *Computer Speech and Language*, 18:25–47. 2004.

[6] F. Casacuberta, F., E. Vidal. "Machine Translation with Inferred Stochastic Finite-State Transducers", in *Computational Linguistics*, 30(2):205–225, 2004.

[7] E. Charniak, K. Knight and K. Yamada, "Syntax-based language models for statistical machine translation", in *Proceedings of MT Summit IX* (New Orleans, USA), September 23–27, 2003.

[8] O. Craciunescu, C. Gerding-Salas and S. Stringer-O'Keeffe: "Machine Translation and Computer- Assisted Translation: a New Way of Translating?", in *Translation Journal*, 8(3), 2004.

[9] E. Cubel, E., J. González, A. Lagarda, F. Casacuberta, A. Juan and E. Vidal. "Adapting finitestate translation to the TransType2 project," in *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop* (Dublin, Ireland), May 15–17, 2003.

[10] M. Dymetman, J. Brousseau, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon, "Towards an automatic dictation system for translators: the TransTalk project," in *Proceedings of International Conference on Spoken Language Processing* (ICSLP94) (Yokohama, Japan), September 18–22, 1994.

[11] J. Esteban, J. Lorenzo, A. S. Valderrábanos and Guy Lapalme. "TransType2 - An Innovative Computer-Assisted Translation System", in *Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics* (Barcelona, Spain), July 21–26, 2004.

[12] S. Khadivi, A. Zolnay, and H. Ney, "Automatic Text Dictation in Computer-Assisted Translation", in *Proceedings of the European Conference on Speech Communication and Technology, Interspeech* (Lisboa, Portugal), September 4–8, 2005.

[13] P. Langlais, G. Foster and G. Lapalme. "TransType: a computer–aided translation typing system," in *Proceedings of the Workshop on Embedded Machine Translation Systems (NAACL/ANLP2000)* (Seattle, Washington), May 4, 2000.

[14] E. Matusov, S. Kanthak, and H. Ney: "On the Integration of Speech Recognition and Statistical Machine Translation," in *Proceedings of the European Conference on Speech Communication and Technology, Interspeech* (Lisboa, Portugal), September 4-8, 2005.

[15] H. Ney, S. Nießen, F. Och, H. Sawaf, C. Tillmann, and S. Vogel. "Algorithms for statistical translation of spoken language," in *IEEE Transactions on Speech and Audio Processing*, 8(1):24–36, 2000.

[16] H. Ney. "One Decade of Statistical Machine Translation: 1996-2005," in *Proceedings of the MT Summit X*, Phuket Thailand, September 12–16, 2005.

[17] S. Nießen and H. Ney. "Statistical Machine Translation with Scarce Resources Using Morpho– syntactic Information," in

Computational Linguistics, 30(2):181–204, 2004.

[18] NIST: "NIST 2005 Machine translation Evaluation Official Results," in *http://www.nist.gov/speech/tests/mt/mt05eval official results release 20050801 v3.html*, August 1, 2005.

[19] F. J. Och and H. Ney. "A Systematic Comparison of Various Statistical Alignment Models", In *Computational Linguistics*, 29(1), pp. 19-51, 2003.

[20] F. J. Och and H. Ney. "The Alignment Template Approach to Statistical Machine Translation". In *Computational Linguistics*, 30(4):417–449, 2004.

[21] F.J. Och, R. Zens and H. Ney. "Efficient search for interactive statistical machine translation," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (Budapest, Hungary), April 12-17, 2003.

[22] A. Pérez, F. Casacuberta, M.I. Torres and V. Guijarrubia, "Finite–state transducers based on ktss grammars for speech translation", in *Proceedings of Finite-State Methods and Natural Language Processing (FSMNLP 2005)* (Helsinki, Finland), September 1–2, 2005.

[23] D. Picó, J. González, F. Casacuberta, D. Caseiro and I. Trancoso. "Finite-state transducer inference for a speech-input Portuguese-to-English machine translation system", in *Proceedings of Interspeech'05* (Lisboa, Portugal), September 4–8, 2005.

[24] L. Rodríguez, J. Civera, E. Vidal, F. Casacuberta and C. Martínez. "On the use of speech recognition in computer assisted translation", in *Proceedings of Interspeech'05* (Lisboa, Portugal), September 4–8, 2005.

[25] J. Tomás, J. and F. Casacuberta. "Monotone statistical translation using word groups", in *Proceedings of the Machine Translation Summit VIII* (Santiago de Compostela, Spain), September 18–22, 2001.

[26] J. Tomás and F. Casacuberta. "Monotone and non-monotone phrase-based statistical machine translation", to be published, 2006.

[27] E. Vidal "The EuTrans Speech-to-Speech Translation", in *ELRA Newsletter, 5(4)*, October–December 2000.

[28] E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera and C. Martínez. "Computer-assisted translation using speech recognition", in *IEEE Transaction on Speech and Audio Processing*, in press, 2006.

[29] R. Zens, and H. Ney. "Improvements in phrase-based statistical machine translation", in *Proceedings of the Human Language Technology Conference (HLT-NAACL)* (Boston, USA).

Francisco Casacuberta
Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática,
Universidad Politécnica de Valencia,
Spain
fcn@dsic.upv.es

Enrique Vidal
Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática,
Universidad Politécnica de Valencia,
Spain

HLT-Evaluation.org: a Portal for Human Language Technology Evaluation

Yun-Chuang Chiao, Khalid Choukri, ELRA

The general mission of HLT evaluation is to assist in improving the quality of language engineering products. It is essential for validating research hypotheses, for assessing progress and for choosing between research alternatives. The language technology and neighbouring technology communities range from academic to industrial partners who share an interest in the evaluation.

Building a portal for these communities serving as a platform for communication between partners, and a permanent infrastructure for the development of evaluation activities in Europe are the main goals of the HLT Evaluation Portal.

In addition to providing all kinds of information related to the evaluation for the language technology community and also for the general public and helping users who need to have quick and easily understandable information on evaluation protocols including evaluation methodologies, metrics, evaluation tasks, resources and the worldwide evaluation activities such as research projects, campaigns, permanent and generic protocols and packages for the major language technologies such as machine translation, information extraction and retrieval, speech processing including speech recognition, speech synthesis and speech translation will be provided.

The first version of the online HLT evaluation web portal available at http://www.hlt-evaluation.org is now structured into three main sections:

- **Overview: Information** on HLT evaluation (projects, activities,...)

- **Evaluation services**: Evaluation services offered by ELRA/ELDA.and their partners.

- **Evaluation resources**: List of evaluation data and evaluation packages (e.g. methodologies, scoring software, test data).

HLT evaluation information

In the first section, the portal provides a free information service to the R&D community, potential users of language technologies and other interested parties.

Future development of the information section will consist of updating and acquiring new information for each subsection, increasing the number of studied technologies.

HLT evaluation services

In 2006, the main activity behind the HLT evaluation portal will be focused on the development of evaluation services. An initiative has been carried out for the definition of evaluation services offered by language technology. For a wide rage of language technologies, the portal would be available to offer a WWW online automatic evaluation and a customized evaluation. A specifications document of detailed service descriptions is being prepared for the each of the following technology:

HLT evaluation resources

The Resources section offers pointers to available commercial and research toolkits which allow one to perform comparative evaluation. An evaluation resource catalogue including evaluation packages for a wide range of language technologies is under development and is expected to be available in the 3rd trimester of 2006.

HLT Evaluation Workshop in Malta

Victoria Arranz, ELRA

To celebrate ELRA's 10th anniversary, a 2-day workshop dedicated to Human Language Technology (HLT) Evaluation was held in Malta on December 1st and 2nd, 2005. About 40 experts attended the workshop. Among the participants we had both industrial and academic players, together with representatives from the European Commission, several national agencies and other research, evaluation and LR distribution related institutions. This resulted in very fruitful discussions on evaluation methodologies and approaches, principles, purposes, initiatives and future actions to be followed.

The workshop started with an introduction to the event by Bente Maegaard (ELRA President) and Khalid Choukri (ELRA CEO), emphasizing the growing importance of evaluation. This received the support from the European Commission, through its representative, Mats Ljungqvist. He agreed on the great importance of joint evaluation and of shared data, tools and annotation schemes, and detailed the European Commission's increasing efforts towards these all throughout FP4, FP5, FP6 and the coming FP7. He also encouraged the community to help the EC to define what needs to be done as well as to work towards the setting of its own challenges.

Along the 2 days of workshop, discussions ranged from general principles and objectives to be targeted, to specific approaches already implemented in some initiatives, or with respect to some technologies/components or to some currently running projects.

Furthermore, representatives of CELCT, CST, DFKI, ELRA, French Ministry of Research, ISSCO and LDC described their already existing initiatives in terms of evaluation.

Most of the participants agreed that another "recommendation report" to enhance and boost the activities on HLT evaluation should be drafted after the workshop with the contribution and support from all the attendees. This action would be coordinated by ELRA. A half-day workshop is planned to take place at the LREC-2006 Conference (http://www.lrec-conf.org/lrec2006/).

The program of the workshop is provided below.

1st December 2005

Introduction
Bente Maegaard, Khalid Choukri, ELRA President & CEO

Invited Speaker
Mats Ljungqvist, European Commission - Luxembourg

Towards HLT Evaluation at University of Malta
Mike Rosner, University of Malta

General Presentation on Evaluation
Margaret King, ISSCO Multilingual Information Processing Unit (TIM), School of Translation and Interpretation (ETI), University of Geneva - Switzerland

Evaluation Principles and Objectives (ISLE, EAGLES)
Keith Miller, The MITRE Corporation - USA

Information Retrieval, Cross Lingual, Question Answering
Carol Peters, Coordinator of the CLEF initiative, Istituto di Elaborazione della Informazione (IEI-CNR) - PISA - Italy
Further discussion by Christian Fluhr, CEA - France

Text Analysers (Morphology, syntax, ...)
Patrick Paroubek, LIMSI-CNRS - France

NLG Evaluation
Anja Belz, Brighton University, United Kingdom
Ehud Reiter, Aberdeen University, United Kingdom

Machine Translation
Tony Hartley, Centre for Translation Studies, University of Leeds - United Kingdom
Further discussion by Andrei Popescu-Belis, ISSCO - Switzerland

Speech Synthesis
Nick Campbell, Cognitive Media Informatics, Media Information Science Laboratories, ATR - Japan
Further discussion by Harald Höge, Siemens AG - Germany

Speech Recognition
Edouard Geoffrois, DGA/CEP - France

2nd December 2005

Speech-to-Speech translation
Gianni Lazzari, Interactive Sensory Systems Division ITC-Irst Centro per la Ricerca Scientifica e Tecnologica, Trento - Italy

Multimodal Interfaces
Jean Carletta, Human Communication Research Centre & Institute for Communicating and Collaborative Systems, the School of Informatics, The University of Edinburgh, United Kingdom

Evaluation Purposes - Technological Components
Gregor Thurmair, Linguatec GmbH, Language technologies, München - Germany

Evaluation initiatives
Representatives of DFKI, CELCT, ISSCO, ELRA, French Ministry of Research, LDC, CST

Corpora/Data for Evaluation Campaigns
Role of ELRA as the HLT Evaluation Institution
Khalid Choukri
ELRA CEO , Paris - France
LDC contribution to Corpora/data for evaluation campaigns
Christopher Cieri, LDC - USA

Wrap-up and Future Actions
Panel

For further information on the workshop (access to speakers' presentations, etc.) please visit our workshop site on the Events page at:

www.elra.info

New Resources

ELRA-S0181 SALA II Spanish from Costa Rica database

The SALA II Spanish from Costa Rica database collected in Costa Rica was recorded within the scope of the SALA II project.The database has been collected jointly by the Universidad de Costa Rica (UCR) and Applied Technologies on Language and Speech, S.L. (ATLAS) from Spain. The owner of the database is Telisma from France. The SALA II Spanish from Costa Rica database contains the recordings of 1,165 Costa Rican speakers (574 males and 591 females) recorded over the Costa Rican mobile telephone network.

This database is distributed as 1 DVD-ROM The speech files are stored as sequences of 8-bit, 8kHz a-law speech files and are not compressed, according to the specifications of SALA II. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SALA II format and content specifications.

	ELRA members	Non-members
For research use	15,000 Euro	18,000 Euro
For commercial use	18,000 Euro	22,500 Euro

ELRA-S0182: SALA II Spanish from Argentina database

The SALA II Spanish from Argentina database collected in Argentina was recorded within the scope of the SALA II project. The database has been collected jointly by the Instituto de Investigaciones Lingüísticas (LIS) of Universidad de Buenos Aires (UBA) and Applied Technologies on Language and Speech, S.L. (ATLAS) from Spain. The owner of the database is Siemens Aktiengesellschaft, Berlin und München (Siemens). The SALA II Spanish from Argentina database contains the recordings of 1,076 Argentinian speakers (534 males and 542 females) recorded over the Argentinian mobile telephone network.

This database is distributed as 1 DVD-ROM. The speech files are stored as sequences of 8-bit, 8kHz a-law speech files and are not compressed, according to the specifications of SALA II. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SALA II format and content specifications.

	ELRA members	Non-members
For research use	34,000 Euro	40,000 Euro
For commercial use	45,000 Euro	51,000 Euro

ELRA-S0183: OrienTel Morocco MCA (Modern Colloquial Arabic) database

The OrienTel Morocco MCA (Modern Colloquial Arabic) database comprises 772 Moroccan speakers (383 males, 389 females) recorded over the Moroccan fixed and mobile telephone network. Corpus was jointly designed by ELDA (France) and Universitat Politècnica de Catalunya (UPC, Barcelona, Spain). Recordings and recruitment was performed by ELDA. Transcription and formatting was done at UPC. The owner of the database is UPC. This database is partitioned into 1 CD and 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	18,000 Euro	22,500 Euro
For commercial use	24,000 Euro	30,000 Euro
Special price for combined purchase of S0183, S0184 and S0185 (see prices page 12)		

ELRA- S0184: OrienTel Morocco MSA (Modern Standard Arabic) database

The OrienTel Morocco MSA (Modern Standard Arabic) database comprises 530 Moroccan speakers (264 males, 266 females) recorded over the Moroccan fixed and mobile telephone network. Corpus was jointly designed by ELDA (France) and Universitat Politècnica de Catalunya (UPC, Barcelona, Spain). Recordings and recruitment was performed by ELDA. Transcription and formatting was done at UPC. The owner of the database is UPC. This database is partitioned into 1 CD and 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	12,000 Euro	15,000 Euro
For commercial use	16,000 Euro	20,000 Euro
Special price for combined purchase of S0183, S0184 and S0185 (see prices page 12)		

S0185: OrienTel French as spoken in Morocco database

The OrienTel French as spoken in Morocco database comprises 530 Moroccan speakers of French (264 males, 266 females) recorded over the Moroccan fixed and mobile telephone network. Corpus was jointly designed by ELDA (France) and Universitat Politècnica de Catalunya (UPC, Barcelona, Spain). Recordings and recruitment was performed by ELDA. Transcription and formatting was done at UPC. The owner of the database is UPC. This database is partitioned into 1 CD and 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	9,600 Euro	12,000 Euro
For commercial use	12,800 Euro	16,000 Euro

Special price for combined purchase of S0183, S0184 and S0185
(see prices page 12)

ELRA-S0186: OrienTel Tunisia MCA (Modern Colloquial Arabic) database

The OrienTel Tunisia MCA (Modern Colloquial Arabic) database comprises 792 Tunisian speakers (426 males, 366 females) recorded over the Tunisian fixed and mobile telephone network. Corpus was jointly designed by ELDA (France) and Universitat Politècnica de Catalunya (UPC, Barcelona, Spain). Recordings and recruitment was performed by ELDA. Transcription and formatting was done at UPC. The owner of the database is ELDA. This database is partitioned into 1 CD and 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	18,000 Euro	22,500 Euro
For commercial use	24,000 Euro	30,000 Euro

Special price for combined purchase of S0186, S0187 and S0188
(see prices page 12)

ELRA-S0187: OrienTel Tunisia MSA (Modern Standard Arabic) database

The OrienTel Tunisia MSA (Modern Standard Arabic) database comprises 598 Tunisian speakers (359 males, 239 females) recorded over the Tunisian fixed and mobile telephone network. Corpus was jointly designed by ELDA (France) and Universitat Politècnica de Catalunya (UPC, Barcelona, Spain). Recordings and recruitment was performed by ELDA. Transcription and formatting was done at UPC. The owner of the database is ELDA. This database is partitioned into 1 CD and 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	12,000 Euro	15,000 Euro
For commercial use	16,000 Euro	20,000 Euro

Special price for combined purchase of S0186, S0187 and S0188
(see prices page 12)

ELRA-S0188: OrienTel French as spoken in Tunisia database

The OrienTel French as spoken in Tunisia database comprises 576 Tunisian speakers of French (290 males, 286 females) recorded over the Tunisian fixed and mobile telephone network. Corpus was jointly designed by ELDA (France) and Universitat Politècnica de Catalunya (UPC, Barcelona, Spain). Recordings and recruitment was performed by ELDA. Transcription and formatting was done at UPC. The owner of the database is ELDA. This database is partitioned into 1 CD and 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	9,600 Euro	12,000 Euro
For commercial use	12,800 Euro	16,000 Euro

Special price for combined purchase of S0186, S0187 and S0188
(see prices page 12)

Special price for combined purchase of S0183, S0184 and S0185		
	ELRA members	Non-members
For research use	33,750 Euro	42,187 Euro
For commercial use	45,000 Euro	56,250 Euro
Special price for combined purchase of S0186, S0187 and S0188		
	ELRA members	Non-members
For research use	33,750 Euro	42,187 Euro
For commercial use	45,000 Euro	56,250 Euro

ELRA-S0189: OrienTel Hebrew database

The OrienTel Hebrew database comprises 1000 Hebrew speakers (500 males, 500 females) recorded over the Israeli fixed and mobile telephone network. The database has been collected and is owned by NSC Natural Speech Communication Ltd., Israel. This database is partitioned into 2 DVDs. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	43,125 Euro	46,405 Euro
For commercial use	47,500 Euro	51,875 Euro

ELRA-S0190: OrienTel Arabic as spoken in Israel database

The OrienTel Arabic as spoken in Israel database comprises 750 Arabic speakers (375 males, 375 females) recorded over the Israeli fixed and mobile telephone network. The database has been collected and is owned by NSC Natural Speech Communication Ltd., Israel. This database is partitioned into 2 DVDs. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

	ELRA members	Non-members
For research use	37,875 Euro	39,843 Euro
For commercial use	40,000 Euro	43,125 Euro

ELRA-L0057: Euskararen Datu-Base Lexikala (EDBL) - Lexical Database for Basque

EDBL (Lexical database for Basque) is the lexical basis needed for the automatic treatment of Basque. It was first developed as a lexical support for the spelling checker and corrector XUXEN, but in the course of the time it has proved to be a multipurpose tool. Nowadays, it is not only the lexical support of the speller but also of the morphological analyser MORFEUS and the lemmatiser EUSLEM. In the future, it will be also used for syntactic and semantic analysis.

Being neutral in relation to linguistic formalisms, flexible, open and easy to use, EDBL is, along with corpora, an essential tool for the Natural Language Processing. It is made up of about 75,000 entries divided into dictionary entries (the same you can find in a conventional dictionary), verb forms and dependent morphemes, all of them with their respective morphological information.

Currently, it is organized in a hierarchical structure, according to a category-system adapted to Basque. It aims to reflect the general lexicon of standard Basque (Euskara Batua) and it is the essential lexical information-store for Basque NLP.

	ELRA members	Non-members
For research use by academic organisations	1,500 Euro	3,000 Euro
For research use by commercial organisations	3,000 Euro	6,000 Euro
For commercial use	7,500 Euro	15,000 Euro