Syntactic-aware language modeling for SMT

Varvara Logacheva, Tilde, Latvia

Outline

- The aim – more syntactically-motivated SMT output
- Ways:
  - Pre-processing
  - Post-processing
  - Translation model
  - Language model

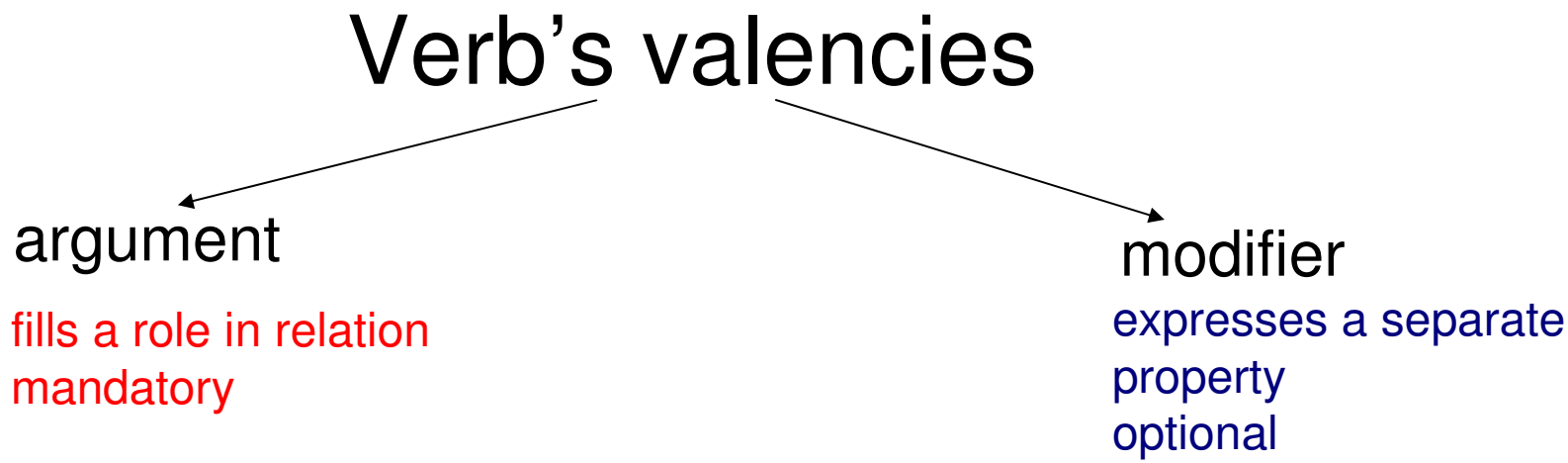
Previous works

- Syntax in **translation model**:
  - Tree structure isn't always preserved in parallel sentences
  - Syntactic variety within one language
- Parser as **language model**:
  - parsers are trained to work with consistent data, inconsistencies make the result unpredictable

Subcategorization frames (valencies)

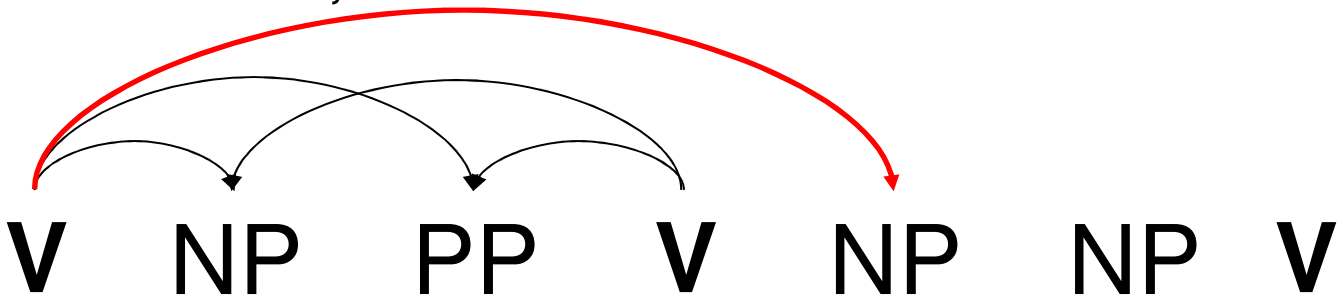
Ability of a lexical item to allow an argument

First approximation: consider only **verb** as lexical item, only **nouns** and **prepositional phrases** as arguments



Core concept

- Assumption: each noun or prepositional phrase can be governed by any verb in a sentence
- Extract information about all (presumed) subordinates, accumulate counts



- Arguments will occur more often, than errors and accidental matches

Results

- 1 000 000 sentences processed (0.1 of Russian part of UN corpus)
- All valencies filtered with tf-idf measure, threshold 0,03
- Subcategorization frames extracted for 2700 verbs (1-3 per verb)
- Quality (precision):
  - 55% arguments
  - 30% modifiers
  - 15% errors

Evaluation challenges

- Valencies ranking:
  - which measure to use (tf-idf, entropy, plain frequency)
  - more fine-grained counts
- Valencies lexicon evaluation:
  - **precision:**
    - distinguish between arguments and modifiers
    - compare with existing lexicons?
  - **recall:** gold standard?
  - switch to **automatic** evaluation
  - **overall**: what result is good?
- MT output evaluation

Drawbacks

- Unable to detect **subject** and **direct object** – too common, appear in all verbs' lists
- **Flawed measure**: valencies with rare prepositions get inadequately high rates

Further work

- Look for new measure
- Cluster verbs by subcategorization frames
- Apply extracted valencies lexicon to machine translation:
  - Language model
  - Translation model
- Distinguish automatically between arguments and modifiers
- Expand the method on other types of frames (verb + infinitive, noun + noun etc.)