Using Tectogrammatical Alignment in Phrase-Based Machine Translation

David Mareček
marecek@ufal.mff.cuni.cz

Week of Doctoral Students,
Prague, June 2, 2009

Motivation

- Phrase-based machine translation is state-of-the-art in the field of statistical machine translation.
  - OpenSource SMT toolkit MOSES
- Phrases are learned from parallel corpora, which has to be first aligned on the word level.
  - Alignment = Connections of corresponding words between the two languages in the parallel corpus.
  - GIZA++ - standard tool for word alignment.
- Tectogrammatical alignment works on content (autosemantic) words only, nevertheless, it outperforms GIZA++ on them.
- Our goal is to use tectogrammatical alignment to improve GIZA++ word alignment and explore whether it can improve also the phrase-based machine translation.

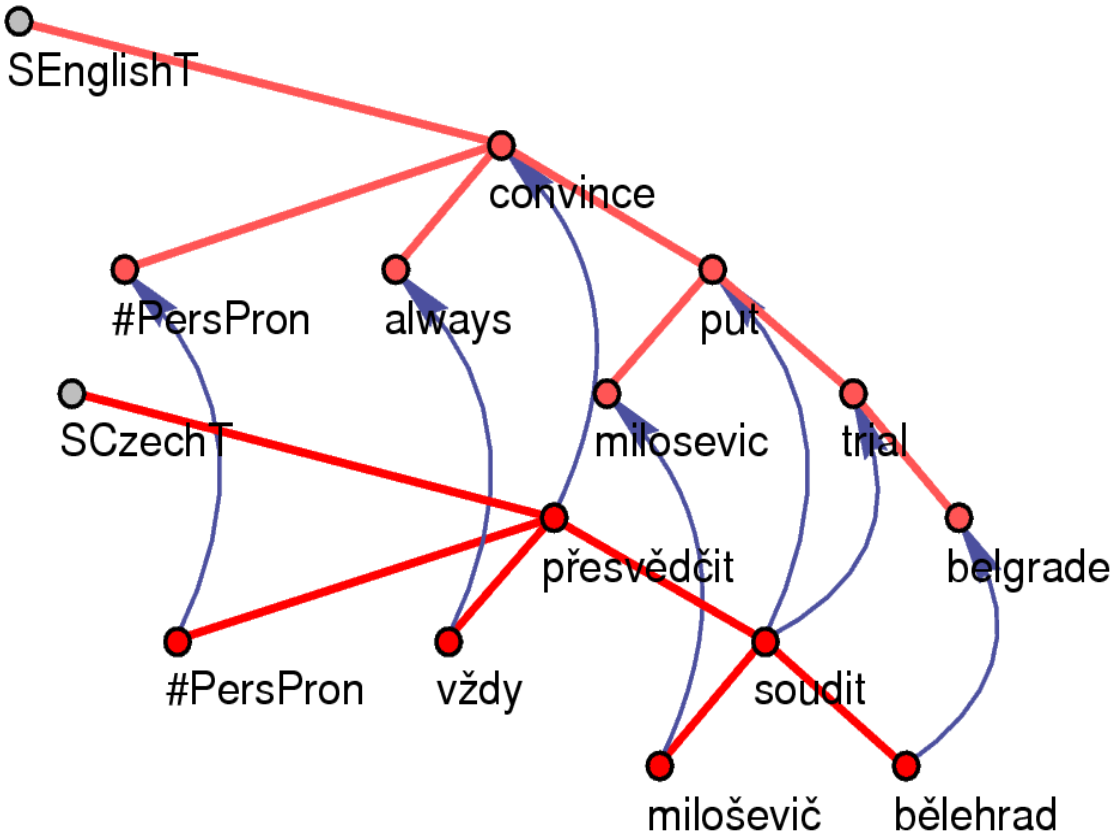
Outline

- Tectogrammatical alignment vs. word alignment
  - Advantages and disadvantages
  - Alignment error rate (AER)
- Word alignment
  - GIZA++
  - Symmetrization methods
  - T-aligner on words
- Combined alignment
  - Combination of previous two alignments
- Applying combined alignment in phrase-based machine translation
  - Comparing BLEU scores
  - Comparing SemPOS scores

Tectogrammatical alignment

- Tectogrammatical tree:
  - Deep syntactic dependency tree, where only content (autosemantic) words have their own nodes.
  - Functional words are hidden.
- Tectogrammatical alignment:
  - Given a sentence and its translation to another language and tectogrammatical representations of this two sentences:
  - Tectogrammatical alignment is a set of links between the two trees that connect the corresponding nodes.
- Advantages over word alignment:
  - Functional words (e.g. articles, prepositions, auxiliary verbs, modal verbs …), that are often problematic to align (they can have different functions in different languages), don't have their own node in the tectogrammatical trees – we needn't align them.
  - The tree structure may help
- Disadvantages:
  - Errors in tagging and parsing often causes errors in the alignment.
  - Only content words are aligned.

Tectogrammatical alignment vs. word alignment



Alignment error rate

- For evaluation purposes, 2500 sentences were manually aligned
  - Texts from newspapers, commentaries, E-books, EU-law
  - Each sentence was aligned independently by two annotators
- Alignment error rate (AER) (Och and Ney, 2003)
  - A metric for measuring alignment quality comparing to the people annotations
  - Lower AER → better alignment

alignment tool	AER	
	all words	content words only
GIZA++	13.2	10.6
T-aligner	–	7.3
GIZA++ with alignment correction of content words using T-aligner	10.7	–

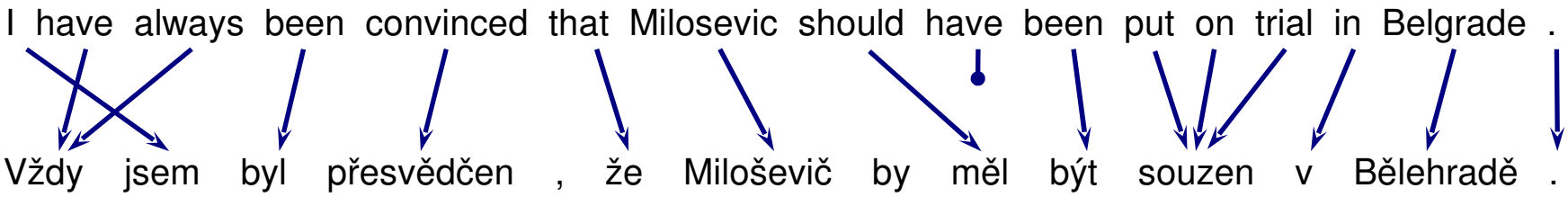
Hypothesis

- We know, how to produce a better word alignment, then GIZA++ does.

- Will be the machine translation better if we use this “better” alignment?
  - In several works (e.g. Fraser and Marcu, 2006) was shown that lower AER doesn’t imply better translations.
  - In addition, it seems that word-alignment made by people is not exactly the alignment that phrase-based translation needs.
  - Howewer, we can somehow improve the word alignment using an other knowledge (tectogrammatical structure), so we should test it.

GIZA++

- Statistical word-alignment toolkit (Och and Ney, 2003)
  - Based on IBM Models and HMM
  - Unsupervised (no manually aligned data needed)
- For the English-Czech pair, GIZA++ has much better results on lemmatized sentences.
  - Since Czech is morphologically very rich language, the word forms are too sparse, compared to the English.
- For each word in the source sentence at most one counterpart in the target sentence is found.
  - Asymetric output



GIZA++ Symmetrization Methods

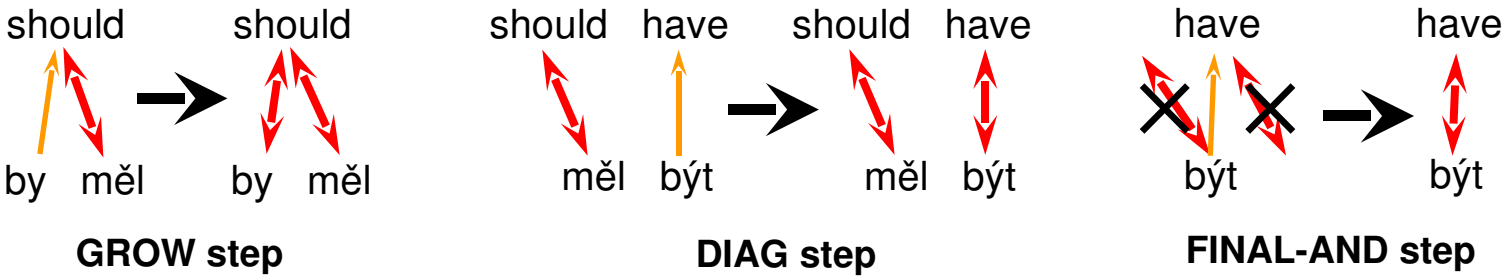
- To symmetrize the output, we run GIZA++ in both directions.
  - English → Czech

  - Czech → English

  - Now we can make intersection or union of the previous two alignments.

Other GIZA++ Symmetrizations

- Other symmetrizations (*grow*, *grow-diag*, *grow-diag-final*, and *grow-diag-final-and*) are somewhere between *intersection* and *union.*
  - All of them include all links from *intersection* symmetrization and add some links from the *union.*
  - *intersection* ⊂ *grow* ⊂ *grow-diag* ⊂ *grow-diag-final-and* ⊂ *grow-diag-final* ⊂ *union*
  - In ***grow*** and ***diag*** steps we add links from *union* that neighbour with any of already added links.
  - In ***final*** step we add links from *union* connecting words that have been not aligned yet.



T-aligner

- A tool for alignment of tectogrammatical trees (Mareček, 2008)
  - Classifier based on averaged perceptron.
  - It uses translation dictionary of t-lemmas, similarities of t-lemmas, positions of nodes in the trees, their semantic part-of-speech, …
  - T-aligner can be run only on languages where tectogrammatics is developed (Czech, English)
- For each tectogrammatical node the most probable counterpart is found.
  - The output is thus asymetric and similarly as for GIZA++ it has to be run twice in both directions and then it can be symmetrized.

Combined word alignment

- The T-aligner outperforms GIZA++ on content words, but it can not align functional words.
- GIZA++ aligns all words.
- Therefore, the combined word-alignment is made as follows:
  - Content words are aligned by T-aligner.
  - Other (functional) words are aligned by GIZA++.
- The combined alignment is made in both the directions (English → Czech, Czech → English) and then it is symmetrized by one of the presented methods.
  - intersection
  - grow
  - grow-diag
  - grow-diag-final-and
  - grow-diag-final
  - union

Applying combined alignment in MOSES

- Direction of translation:
  - English → Czech
- Training data:
  - WMT08 (about 80,000 parallel sentences from Project Syndicate corpus)
- Tuning and evaluation data
  - WMT08 (about 1,000 tuning and 2,000 evaluation parallel sentences)
- Tuning
  - Minimum error-rate training (MERT) for tuning the parameters

Results (BLEU)

- We measure the quality of translations using BLEU score.
  - Based on count of matching n-grams against the reference translations
  - The higher BLEU → the better translation

symmetrization method	BLEU	
	GIZA++ alignment	Combined alignment
intersection	12.37	12.46
grow	12.53	12.60
grow-diag	12.80	12.82
grow-diag-final-and	12.93	**13.00**
grow-diag-final	12.91	12.64
union	12.96	12.64

Results (SemPOS)

- SemPos (Kos and Bojar, 2008) is MT metric, which has better correlation with human judgements than BLEU, especially for English-Czech language pair
  - Computes overlapping t-lemmas with respect to their semantic part-of-speech

symmetrization method	SemPOS	
	GIZA++ alignment	Combined alignment
intersection	44.34	44.86
grow-diag-final-and	45.52	**46.20**
union	45.99	45.40

Conclusions

- Using tectogrammatical alignment slightly improved the machine translation.
- Although the word-alignment error rate was decreased from 13.2 to 10.7, the differences in MT scores are very small.
  - Only 0.07 BLEU points.
- Training on larger corpus (CzEng) has not been tested yet, but it is very probable that the differences will be far smaller.
- Tectogrammatical alignment (in the presented way) is therefore not much usable for phrase-based MT, because of its high computational cost and very low improvement in MT quality.

Thank you for your attention