Improving Word Alignment using Tectogrammatical Alignment

David Mareček
marecek@ufal.mff.cuni.cz

Text, Speech and Dialogue
Pilsen, September 14, 2009

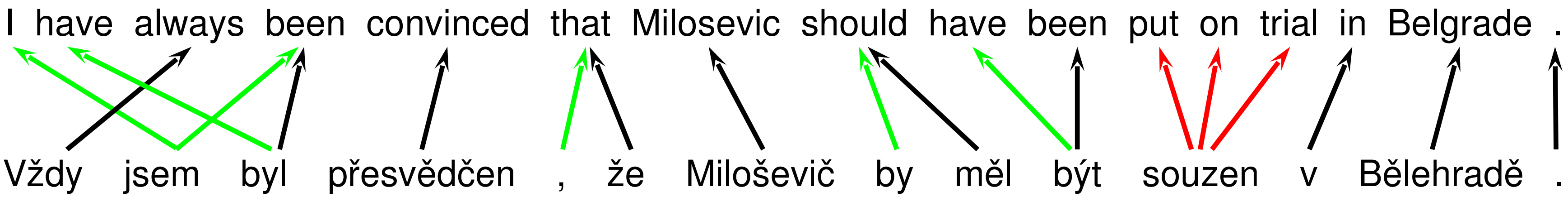
Outline

- Word alignment and its problems
- Tectogrammatical alignment
  - Advantages and disadvantages
- T-aligner
  - A tool for automatic alignment of tectogrammatical trees
- Combination of two aligners
  - GIZA++ and T-aligner
  - Application in SMT toolkit Moses
- Conclusions and references

Czech-English Word alignment

- Given an English sentence and its Czech translation:
- Word alignment is a set of connections between words of this two sentences that belongs together
  - One word is a translation of the other word
  - It is not a translation but somehow belongs to the other word

- An example of manually aligned sentence (*source: Project Syndicate corpus*)



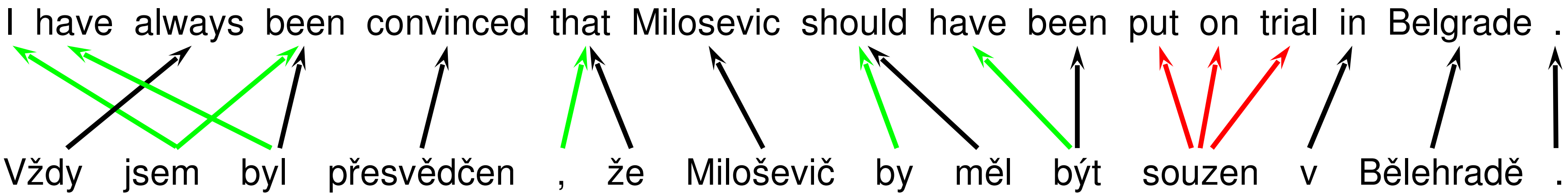
I have always been convinced that Milosevic should have been put on trial in Belgrade .

Vždy jsem byl přesvědčen , že Miloševič by měl být souzen v Bělehradě .

- Annotators used three types of connections:
  - sure: individual words match
  - possible: connect words that do not have a real equivalent in the other language but syntactically clearly belong to a word nearby
  - phrasal: whole phrases correspond but not literally

Czech-English Word alignment - problems

- Alignment of function words is sometimes problematic
  - It is not clear which word to choose as a counterpart



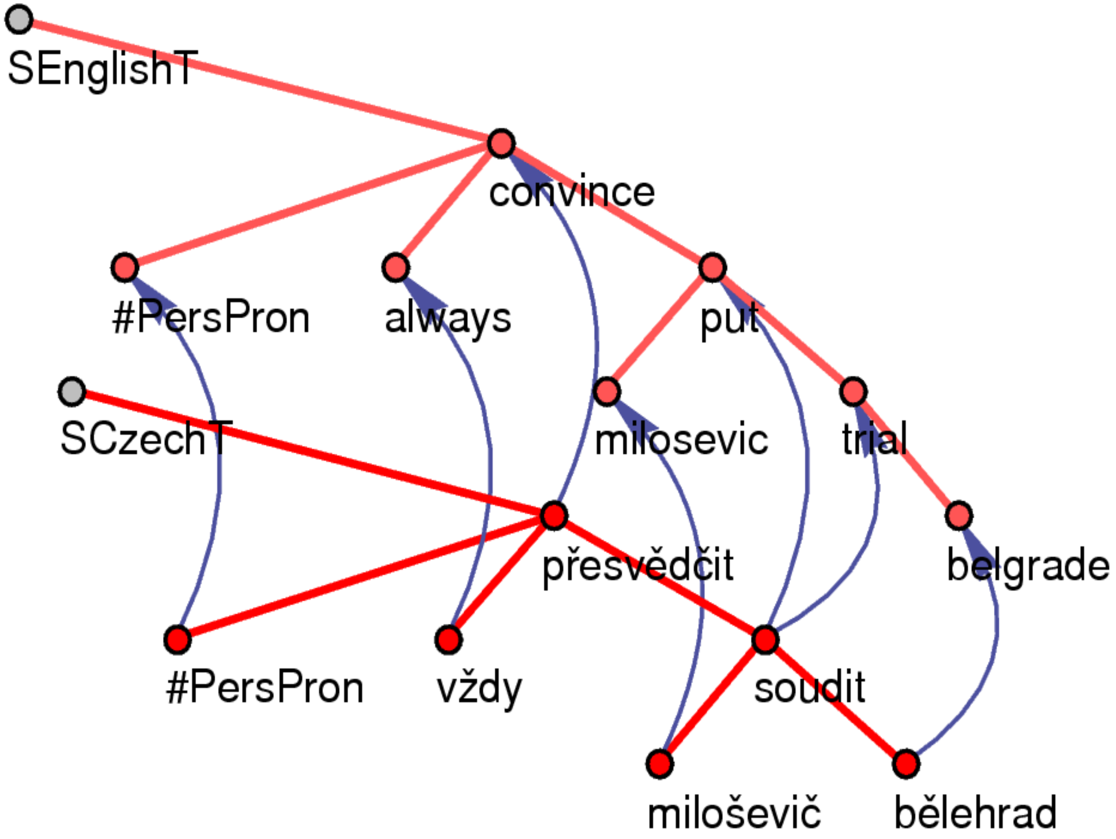
I have always been convinced that Milosevic should have been put on trial in Belgrade .

Vždy jsem byl přesvědčen , že Miloševič by měl být souzen v Bělehradě .

- The English word “I” does not have any equivalent in the Czech sentence (pro-drop, the first person is expressed by the verb)
  - Should it be aligned with “jsem” or “byl”?
- The comma
  - Should it be aligned with “that” or not?

Tectogrammatical alignment

- Alignment of tectogrammatical trees
  - Only content (autosemantic) words have their own nodes
  - Function words (articles, prepositions, auxiliary verbs, …) are hidden. They are attached to content words in the form of their attributes.
  - The words dropped in the surface shape of the sentence are added (#PersPron)



Tectogrammatical alignment (2)

- 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:
  - Function 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 nodes in the tectogrammatical trees – we needn't align them.
  - The tree structure may help

- Disadvantages:
  - We have to build the trees automatically. Errors in tagging and parsing often causes errors in the alignment.
  - Only content words are aligned.

Inter-annotator agreement

- Two annotators A and B aligned manually 2500 pairs of Czech-English sentences
  - They used three different types of connections: sure, possible and phrasal
  - Their agreement was computed using the following formula, where L_A and L_B are sets of connections made by annotator A and B

$$IAA(A,B) = \frac{2 \cdot |L_A \cap L_B|}{|L_A| + |L_B|}$$

IAA	all words	content words	function words
Types distinguished	83 %	90 %	76 %
Types not distinguished	89 %	94 %	84 %

- Therefore, the alignment of tectogrammatical trees is for annotators less problematic than the word alignment.

T-Aligner - algorithm

- Perceptron based algorithm for tectogrammatical alignment
  - A score is assigned to each possible connection between nodes of Czech and English tectogrammatical tree

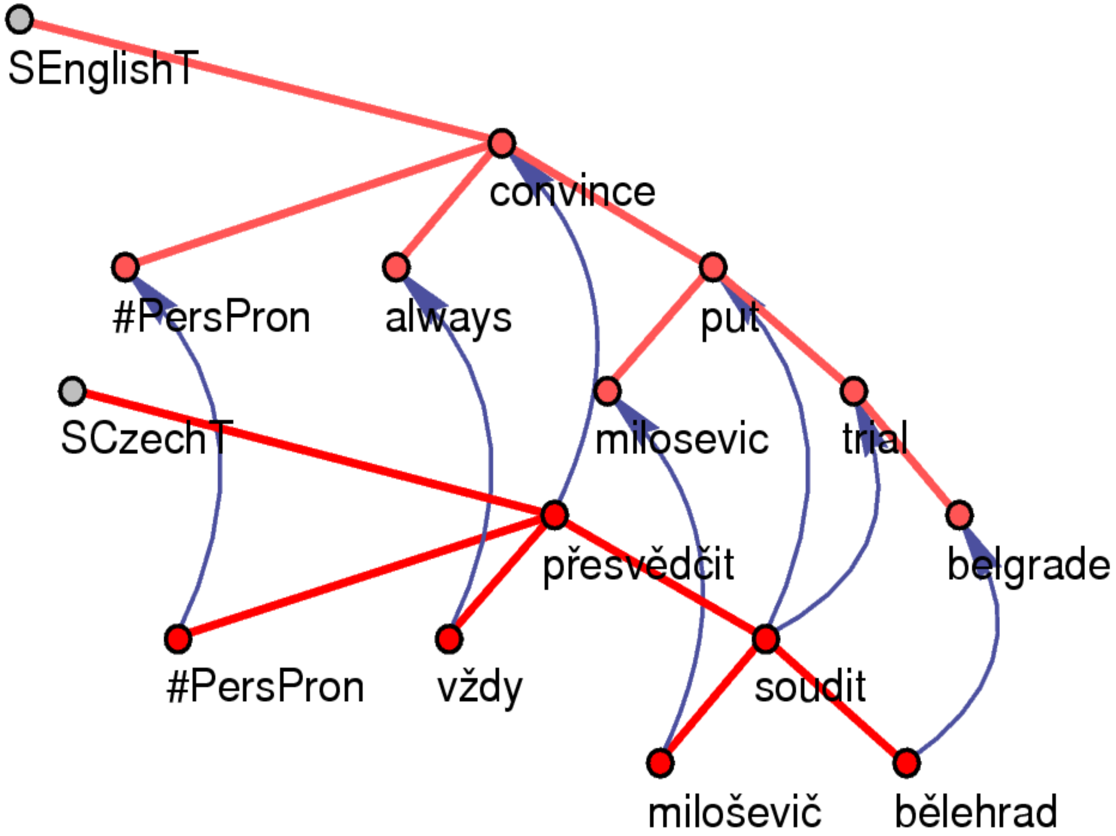
$$score(ennode, csnode) = \qquad w_i \cdot f_i(ennode, csnode)$$

  - *en, cs* ..… English and Czech tectogrammatical node
  - f_i(*en, cs*) … value of i-th feature of the connectrion (en, cs)
  - w_i .………. weight of i-th feature obtained by training the perceptron. This weights are learned using training part of the manually aligned data set.

- EN→CZ alignment: For each English node the Czech counterpart with the highest score is found.
- CZ→EN alignment: For each Czech node the English counterpart with the highest score is found.
- Output alignment: Intersection of previous two alignments

T-Aligner - Features

- translation probability between tectogrammatical lemmas
  - Probabilistic dictionary compiled from various sources (electronic dictionaries, parallel corpora)
- similar linear position of nodes in the tree
- similarities in other attributes
- child/parent similarities
- equal t-lemma prefix
- …
- (total of 15 features)



T-Aligner - evaluation

- We used the testing part of the 2500 manually aligned pairs of sentence from different sources
  - EU law (Acquis communauatire parallel corpus)
  - Commentaries (Project Syndicate parallel corpus)
  - Newspapers (Wall Street Journal and its Czech translation)
  - Short stories (Reader's Digest)
- Precision, recall and alignment error-rate was computed
- Alignment-error rate (AER)
  - Och and Ney, 2003
  - The lower AER, the better alignment

	precision	recall	AER
T-aligner	96,0 %	89,7 %	7,3 %

Combined alignment

- Word alignment made by GIZA++ (Och, Ney, 2003) may be improved using tectogrammatical alignment
  - T-aligner has better results for content words
  - GIZA++ aligns all words
  - → content words are aligned by T-aligner, other words by GIZA++

alignment tool	alignment error rate	
	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.
  - However, we can somehow improve the word alignment using an other knowledge (tectogrammatical structure), so we should test it.

Applying combined alignment in MOSES

- SMT toolkit Moses (Koehn et al., 2007)
  - Phrase based machine translation system
- Direction of translation:
  - English → Czech
- Training data:
  - WMT08 (about 80,000 parallel sentences from Project Syndicate)
- 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

MOSES 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

Conclusions

- Tectogrammatical alignment
  - It is less problematic for people (inter-annotator agreement on content words is 5% higher than for word-alignment)
  - Alignment tools (GIZA++, T-aligner) have better results on content words (2,6% improvement for GIZA++).

- If we combine alignment outputs from GIZA++ and T-aligner, AER of the resulting word alignment decrease from 13.2 to 10.7 %
  - However, the improvement in phrase-based MT (Moses) trained on this two different alignments is very small (only 0.07 BLEU points).

- Tectogrammatical alignment is used for training the transfer step in TectoMT (machine translation with transfer on tectogrammatical layer.

References

- Och, F.J., Ney, H.: *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics29(1) (2003) 19–51.
- Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, Zdeněk Žabokrtský: *English-Czech MT in 2008*. In Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece. Association for Computational Linguistics, 2009.
- Bojar, O., Prokopová, M.: *Czech-English Word Alignment*. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, ELRA (May 2006) 1236–1239.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. *Moses: Open source toolkit for statistical machine translation*. InACL,demonstration session, Prague, Czech Republic.
- Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M.: *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia (2006).

Thank you for your attention