Analysis and alignment of parallel data in TectoMT

David Mareček

marecek@ufal.mff.cuni.cz

MT Marathon 2009, January 26 - 30, Prague

Task and motivation

INPUT: set of English-Czech parallel sentences

OUTPUT: set of aligned tectogrammatical trees (+ lower layers)

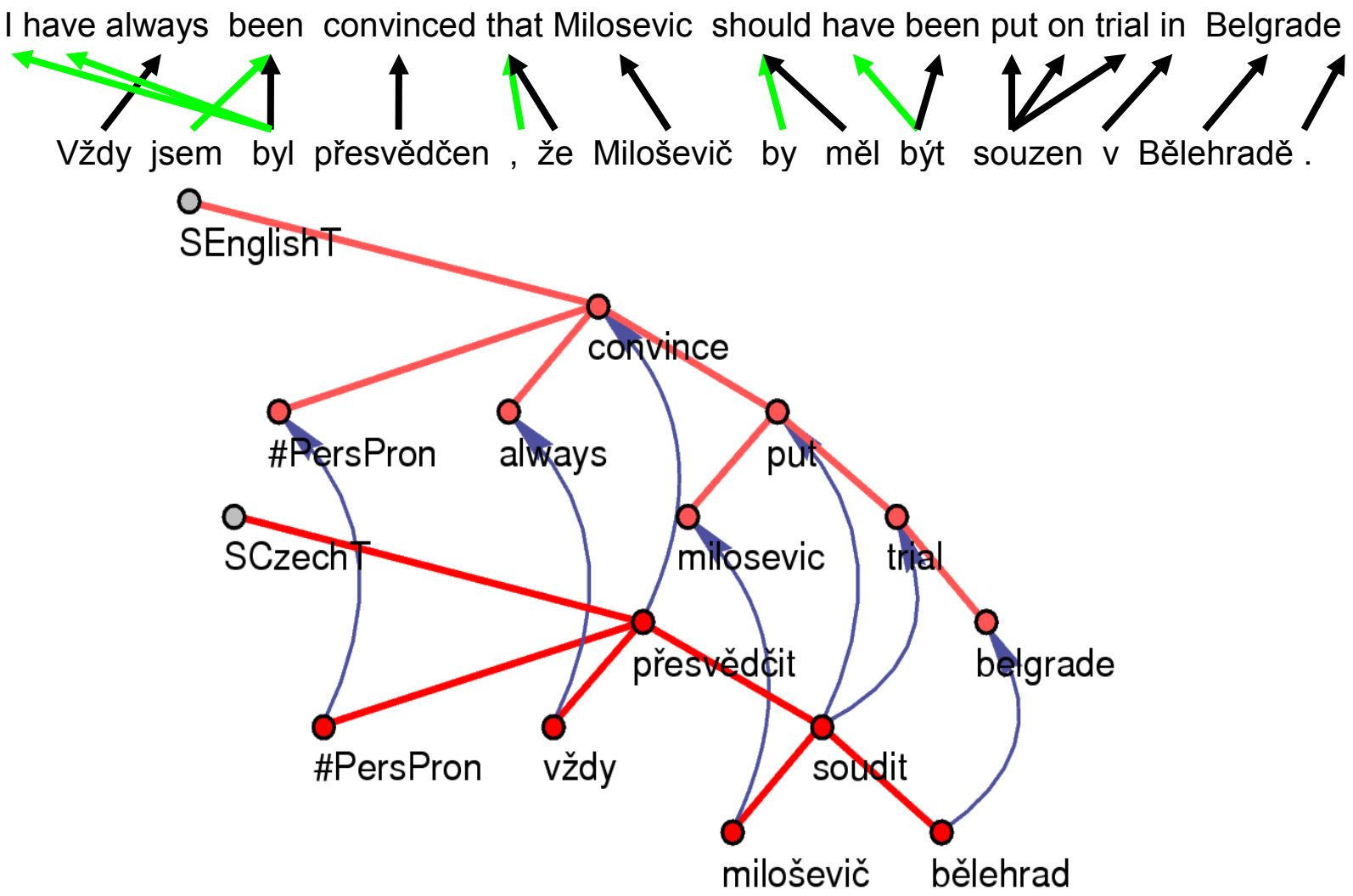
Advantage of tectogrammatical alignment over word alignment:

- Functional words (e.g. articles, prepositions, auxiliary verb ‘be’, 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 layer – we needn’t align them.
- The tree structure may help.

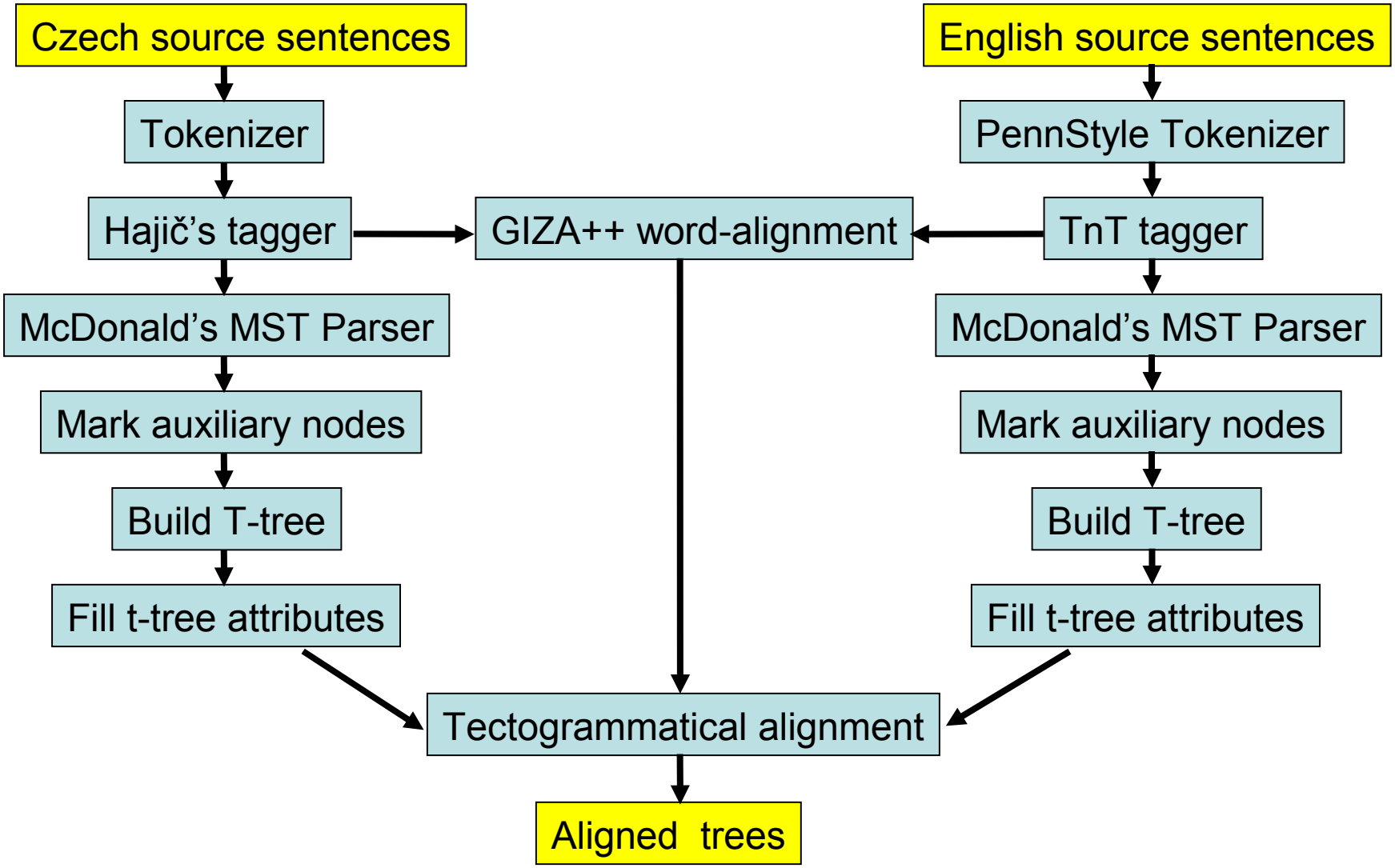
Usage:

- Extracting probabilistic translation dictionary from tectogramatically aligned parallel corpora.

Tecto-alignment x word-alignment



Schema


Czech source sentences
Tokenizer
Hajič's tagger
McDonald's MST Parser
Mark auxiliary nodes
Build T-tree
Fill t-tree attributes
GIZA++ word-alignment
English source sentences
PennStyle Tokenizer
TnT tagger
McDonald's MST Parser
Mark auxiliary nodes
Build T-tree
Fill t-tree attributes
Tectogrammatical alignment
Aligned trees

T-Aligner

- Greedy algorithm based on features

- A score is assigned to each possible connection (pair of Czech and English node)

$$score(en, cs) = \sum w_i \cdot f_i(en, cs)$$

- The weights ***w*** of the features ***f*** were obtained by perceptron learning
- Examples of features:
  - translation probability between tectogrammatical lemmas
  - similar position of nodes in the tree
  - similarities in other attributes
  - child/parent nodes similarities

- In each step, the algorithm finds the pair with the highest score.
- If both the nodes are free and the score is higher than a threshold, we connect them. (only on-to-one connections are allowed)

Alignment evaluation

- 2500 parallel sentences (E-Books, newspaper articles, EU-laws) were manually aligned on the word level, each by two annotators.
- The acquired word-alignment was then transferred to the tectogrammatical layer through the **lex.rf** references
  - **lex.rf** – attribute of a tectogrammatical node, refers to the analytical node from which it acquired its lexical meaning.

Aligner	F-measure
Our T-aligner	88.5 %
GIZA++ word-alignment transferred to t-trees	85.7 %
Our T-aligner using also GIZA++ word-alignment	91.0 %
(Inter-annotator agreement)	*94.8 %*

Tectogrammatical alignment results

References

David Mareček, Zdeněk Žabokrtský, Václav Novák: *Automatic Alignment of Czech and English Deep Syntactic Dependency Trees.* In Proceedings of EAMT08, Hamburg, Germany, 2008

Franz Josef Och, Hermann Ney: *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics 29(1), p.19-51, 2003