Automatic Alignment of Czech and English Deep Syntactic Dependency Trees

David Mareček, Zdeněk Žabokrtský, Václav Novák

Institute of Formal and Applied Linguistics, ÚFAL MFF, Charles University in Prague

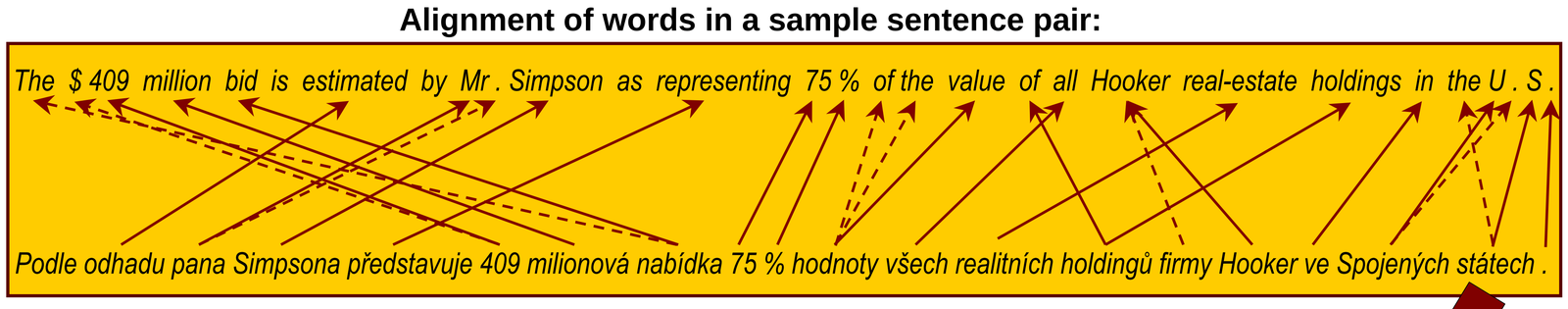
1. Introduction

- **Task:** to find correspondences between two tectogrammatical (deep-syntactic) trees that represent an English sentence and its Czech translation.
- **Motivation:** aligned tectogrammatical trees are needed for training tree-to-tree transfer models in our MT system.
- **Hypothesis:** tectogrammatical representations of Czech and English sentences are more similar compared to the similarity of the sentence surface shapes, thus higher agreement/precision in alignment should be achievable.

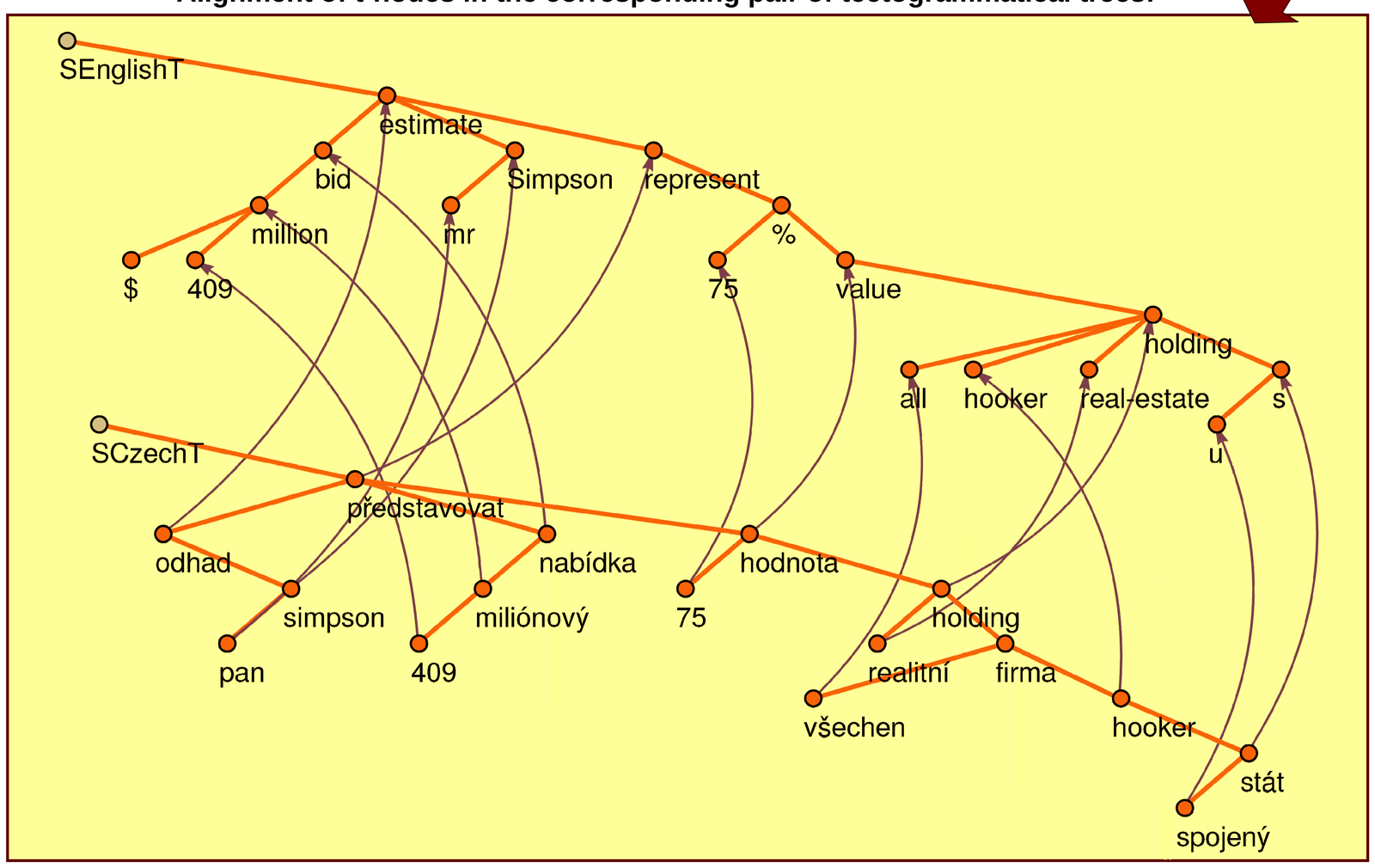
2. Manually aligned data

- 515 sentences (about 13,000 tokens) manually aligned on the word level, in parallel by two independent annotators.
- Three types of links distinguished: (a) sure links, (b) possible links, (c) phrasal links.
- The sentences were automatically parsed up to the tectogrammatical layer.
- Then the word alignment was transferred to the tectogrammatical trees in order to provide data for training and testing tectogrammatical aligners.

Alignment of words in a sample sentence pair:

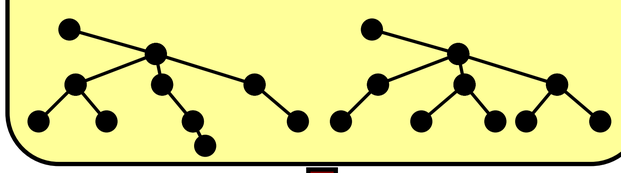


Alignment of t-nodes in the corresponding pair of tectogrammatical trees:



3. Alignment algorithm

INPUT: a pair of Czech and English tectogrammatical trees



Step 1: Greedy feature-based 1:1 alignment

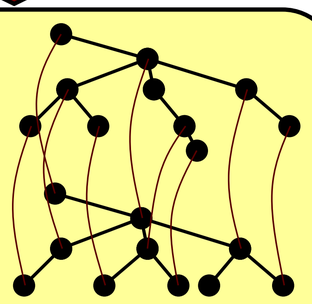
foreach (*cnode, enode*): *cnode* ∈ *CTree, enode* ∈ *ETree* **do**
  score(*cnode, enode*) = $\vec{w} \cdot \vec{f}$ (*cnode, enode*);
  Add *cnode* to *CNonUsed*;
  Add *enode* to *ENonUsed*;
while exist (*cnode, enode*): *cnode* ∈ *CNonUsed, enode* ∈ *ENonUsed* **do**
  Find (*cmax*, *emax*) with the highest *score*(*cmax*, *emax*);
  if *score*(*cmax*, *emax*) ≥ *threshold* **then**
    Align(*cmax*, *emax*);
    Delete *cnode* from *CNonUsed*;
    Delete *enode* from *ENonUsed*;
    foreach (*cnode, enode*): *cnode* ∈ *CNonUsed, enode* ∈ *ENonUsed* **do**
      if cnode = *parent*(*cmax*) **or** *cnode* ∈ *children*(*cmax*)
      or *enode* = *parent*(*emax*) **or** *enode* ∈ *children*(*emax*) **then**
        score(*cnode, enode*) = $\vec{w} \cdot \vec{f}$ (*cnode, enode*);
  else
    break;

Step 2: Completing 1:N relations

We align two t-nodes K, L if the following conditions are fulfilled:

- K is not yet aligned and its parent or child t-node is aligned to L
- The pair (K, L) was also aligned by GIZA++ (grow-final-diag sym.)
- The pair (K, L) occures in the probabilistic dictionary

OUTPUT: aligned trees



4. Function for scoring candidate node pairs

- Based on a set of manually designed features - feature vector **f**
- Vector of feature weights **w** found by perceptron using the annotated data
- Scalar product scoring function:

  score(cnode, enode) = **w** · **f** (cnode, enode)

Feature weights

feature name	*range*	*weight*
similarity in linear position	⟨0, 1⟩	2.81
aligned by GIZA++, intersection	0 or 1	2.78
the same digit prefix	0 or 1	2.63
the same 5-letter prefix	0 or 1	2.28
the same 4-letter prefix	0 or 1	1.81
translation probability from GIZA++	⟨0, 1⟩	1.49
identical t-lemmas	0 or 1	1.00
t-lemma pair in dictionary	0 or 1	0.95
aligned by GIZA++, grow-diag-final	0 or 1	0.64
both coord/apos. roots	0 or 1	0.51
the same 3-letter prefix	0 or 1	0.49
aligned parent	0 or 1	0.37
aligned child	0, 1, 2, ...	0.33
translation probability from dict.	⟨0, 1⟩	0.17
equal semantic POS	0 or 1	0.11

5. Evaluation

- **inter-annotator agreement** (f-measure)
  - on aligning words: **82.1 %**
  - on aligning t-nodes (i.e., after transferring the manual word alignment to t-trees): **94.7 %**
- performance of the **automatic t-trees aligners** (f-measure)
  - baseline: t-lemma sequences aligned by GIZA++: **82.6 %**
  - alignment of t-trees by our feature-based aligner: **90.4 %** (10-fold cross validation)

6. Conclusions

- Inter-annotator agreement on aligning t-nodes (≈ content words) is considerably higher than the agreement on aligning all words of the original sentences.
- Our feature-based tectogrammatical aligner outperforms the GIZA++ baseline.