Formemes in English–Czech Deep Syntactic MT

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majlíš, Michal Novák, David Mareček – {odusek,zabokrtsky,popel,majlis,mnovak,marecek}@ufal.mff.cuni.cz
Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

The TectoMT System

Analysis
She was taken to hospital to treat a laceration.

Transfer
• On the deep (tectogrammatical) layer
• Separated into three virtually independent subtasks

Grammatemes
• Rule-based modules
decact 4
[noun lemma] n.k+3 = sg

Lemma
• Maximum-Entropy model: P(cs-lemma/en-context-features)
• Simple model: P(cs-lemma/en-lemma)
• Derivation back-off models

Target-Language Tree Model
• Viterbi algorithm adapted for trees
• Selects best lemma+formeme

Target Language

Synthesis
• Rule-based addition of function words from formemes

What is a formeme
v:to+inf

Syntactic part-of-speech (verb, noun, adjective, adverb)
Preposition or subordinate conjunction

Morphosyntactic form

Examples
So far, everything was just right.
The tram was forced to break abruptly on Herzbegalee when a vehicle stopped suddenly in front of it.

Visualizations

Formeme Improvements
• Reducing redundancy
• Reducing data sparsity
• Increasing inter-language consistency
• Maintaining linguistic adequacy

He is one of the best at school.

Many of them were late.

Evaluation

Czech-to-Czech round trip
TectoMT translation of the WMT’12 test set (trained on 1/2 of CzEng 1.0)

TectoMT Performance

WMT’12 Human Evaluation

WMT’12 BLEU Evaluation

(x = Overall winner, ⋆ = Constrained winner)