Merged bilingual trees based on Universal Dependencies in Machine Translation

David Mareček
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague

1. Universal Dependencies (UD)

Collection of treebanks with cross-linguistically consistent annotation (common part-of-speech tagset, common dependency relation set, common annotation guidelines)

There are 54 treebanks and 40 languages in UD version 1.3

http://universaldependencies.org

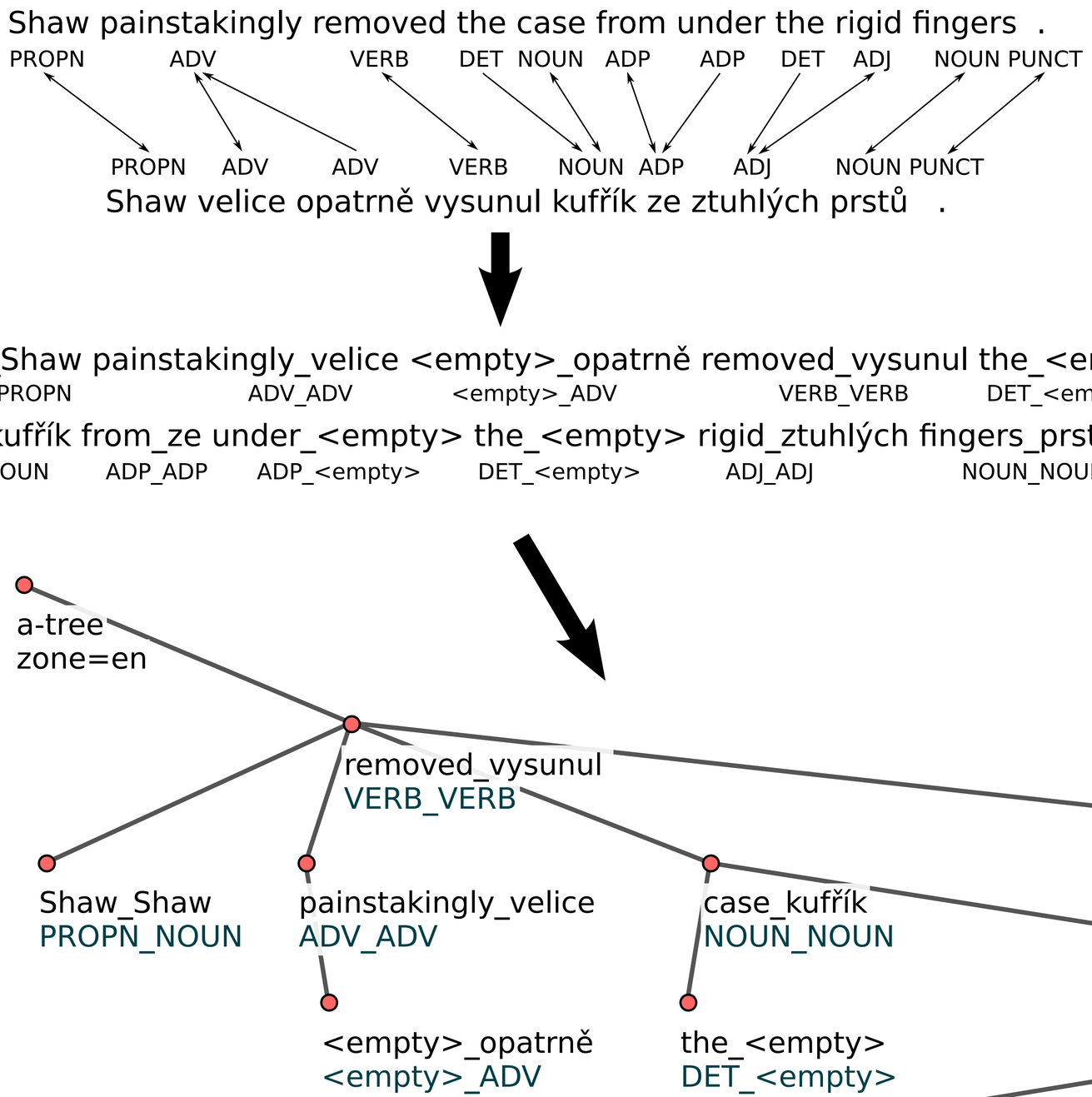
Function words are represented by leaf nodes and therefore the grammatical differences between two languages does not much affect the common dependency structure.

2. Merged trees

Parallel sentences from two languages are represented by a single dependency tree.

Each node of the tree consists of two word-forms and two POS tags.

Words that do not have their counterparts in the other sentence (1-0 or 0-1 alignment) are also represented by nodes and the missing counterpart is marked by label <empty>. All such nodes are leaves.



3. Merging parallel sentences

We use GIZA++ 1:n and n:1 alignments

The algorithm traverses through the source sentence and for each word, it collects all its target counterparts.

Where the alignments intersect the source word is merged with the target one.

The other target words stay alone and are completed with the <empty> label.

If there is no intersection counterpart for the English word, it is also completed with the <empty> label.

3. Minimally supervised parallel parsing

Based on Unsupervised Dependency Parser (http://ufal.mff.cuni.cz/udp)

Dependency Model with Valence + external prior probabilities to define grammatical rules for POS tags based on UD annotation style.

4. Machine Translation Experiments

TRAINING:

- CzEng parallel corpus v1.0

- merging and parsing algorithm (steps 1 and 2) applied to the whole corpus

RESULTS:

language pair	BLEU	BLEU cased
English to Czech	9.5	8.3
Czech to English	15.6	13.2

This work was supported by the grant 14-06548P of the Czech Science Foundation and the 7th Framework Programme of the EU grant QTLeap (No. 610516). This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project No. LM2015071 of the Ministry of Education, Youth and Sports of the Czech Republic.