This research was supported by the grants FP7-ICT-2011-7-288487 (MosesCore), FP7-ICT-2013-10-610516 (QTLeap), GAUK 2010013.

Google Translate

Eurotran

CU TectoMT

NNs
dep

voters

http://ufal.mff.cuni.cz/depfix

D trim	l

voters were

called

Verb Tense Translation
defining

Verb Tense Translation
defining

verb

are defining

brani

generadores are

VBP

Caught

CU Chimera English → Czech MT System

Input

The source is available to all components to minimize error propagation.

TectoMT (Popel)

- hybrid (rule-based/statistical) MT system
- transfer at tectogrammatical (deep syntactical) layer
- our combination: get an extra phrase table for Moses from TectoMT output

Moses (Bojar)

- phrase-based SMT
- large-scale data
- morphological tags as factors for a better grammatical coherence

Depfix (Rosa)

- automatic post-editing
- generates the final output

Evaluation of CU Chimera

CU Chimera is the state-of-the-art for English → Czech MT!

<table>
<thead>
<tr>
<th></th>
<th>WMT 2013</th>
<th>WMT 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU Chimera</td>
<td>19.2</td>
<td>20.0</td>
</tr>
<tr>
<td>UEDIN</td>
<td>18.6</td>
<td>19.1</td>
</tr>
<tr>
<td>JHU</td>
<td>18.9</td>
<td>20.1</td>
</tr>
<tr>
<td>Eurotran</td>
<td></td>
<td>19.5</td>
</tr>
<tr>
<td>Microsoft Bing</td>
<td></td>
<td>20.1</td>
</tr>
<tr>
<td>Google Translate</td>
<td>14.2</td>
<td>14.7</td>
</tr>
</tbody>
</table>

Get Depfix Today!

- implemented in Perl in Treex NLP framework
- needs Linux + 4GB RAM
- 3 seconds per sentence
- released under GNU/GPL
- source is commented

CU Chimera English → Czech Machine Translation

Moses/Google Translator

Source

Czech sentence from SMT

Treex Analysis

tokenizer (Treex)
lemmatizing tagger (MorphoDta)
word aligner (GIZA++)
dependency parser (MST)
dependency relations labeler (Treex)
named entity recognizer (Stanford)
conversion to tectogrammatical (Treex)

Czech sentence from SMT

Analytical (surface syntax) dependency trees
Named entities
Tectogrammatical (deep syntax) trees

Depfix Evaluation (Δ BLEU)

Depfix improves most MT systems!

<table>
<thead>
<tr>
<th></th>
<th>WMT 2011</th>
<th>WMT 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU Bojar</td>
<td>+0.47</td>
<td>+0.07</td>
</tr>
<tr>
<td>CU TectoMT</td>
<td>−0.10</td>
<td>−0.02</td>
</tr>
<tr>
<td>UEDIN</td>
<td>+0.64</td>
<td>+0.23</td>
</tr>
<tr>
<td>JHU</td>
<td>+0.42</td>
<td>+0.32</td>
</tr>
<tr>
<td>Eurotran</td>
<td>+0.21</td>
<td>+0.15</td>
</tr>
<tr>
<td>Microsoft Bing</td>
<td>?</td>
<td>+0.37</td>
</tr>
<tr>
<td>Google Translate</td>
<td>+0.23</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Rudolf Rosa
rosa@ufal.mff.cuni.cz
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics