CUNI in WMT14: Chimera Still Awaits Bellerophon

Ondřej Bojar, Martin Popel, Rudolf Rosa, Aleš Tamchyna  
{bojar,popel,rosa,tamchyna}@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague  
http://ufal.mff.cuni.cz

English → Czech

Final Results

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>TER</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>UEDIN-UNCONSTRAINED</td>
<td>21.1</td>
<td>0.670</td>
<td>0.375</td>
</tr>
<tr>
<td>GOOGLE TRANSLATE</td>
<td>20.2</td>
<td>0.687</td>
<td>0.168</td>
</tr>
<tr>
<td>CU-BOJAR</td>
<td>21.2</td>
<td>0.675</td>
<td>0.287</td>
</tr>
<tr>
<td>CU-TECTOMT</td>
<td>15.2</td>
<td>0.716</td>
<td>0.177</td>
</tr>
<tr>
<td>CU-DEPFIX</td>
<td>20.9</td>
<td>0.674</td>
<td>0.333</td>
</tr>
<tr>
<td>CU-FUNKY</td>
<td>20.7</td>
<td>0.674</td>
<td>0.333</td>
</tr>
</tbody>
</table>

Adapted Language Model

- WMT 12+13+14 source sides
- Baseline system
- Query corpus
- Large target-side monolingual corpus
- Lucene
- Domain-relevant LM
- Sents from doc. 1
- Sents from doc. 2
- Sents from doc. 3
- Sents from doc. N
- Lucene
- LM specific for doc. 1
- LM specific for doc. 2
- LM specific for doc. 3
- LM specific for doc. N

Document-specific Language Models

- WMT 13+14 source sides
- Baseline system
- Sents from doc. 1
- Lucene
- LM specific for doc. 1
- Sents from doc. 2
- Lucene
- LM specific for doc. 2
- Sents from doc. 3
- Lucene
- LM specific for doc. 3
- Sents from doc. N
- Lucene
- LM specific for doc. N

English → Hindi

Essentially baseline Moses.

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Hindi</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>HindEnCorp</td>
<td>276k</td>
<td>4.09M</td>
</tr>
<tr>
<td>NewsCrawl</td>
<td>1.27M</td>
<td>27.27M</td>
</tr>
<tr>
<td>HindMonoCorp</td>
<td>43.38M</td>
<td>945.43M</td>
</tr>
<tr>
<td>Total</td>
<td>44.93M</td>
<td>976.80M</td>
</tr>
</tbody>
</table>

Morphological Processing of Hindi

- Back-off % of vocab. size
  - stem4 30
  - lemma4 32
  - lemma 90
  - form 100

Impact of Word Alignment on BLEU

- English
  - stem4 22.96+1.17
  - lemma4 22.59+1.17
  - lemma 22.41+1.20

Other Failed Attempts

- Factored setup (form-form+POS) with 10-gram POS-LM
- Reverse Self-Training

Ondřej Bojar, Martin Popel, Rudolf Rosa, Aleš Tamchyna  
{bojar,popel,rosa,tamchyna}@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague  
http://ufal.mff.cuni.cz

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap), FP7-ICT-2011-7-288487 (MosesCore), SVV 260 104 and GAUK 1572314.

This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).