The University of Illinois submission
to the WMT 2015 Shared Translation Task

Lane Schwartz, Bill Bryce, Chase Geigle, Sean Massung, Yisi Liu, Haoru Peng, Vignesh Raja, Subhro Roy, Shyam Upadhyay
University of Illinois at Urbana-Champaign

Abstract

In the 2015 WMT translation task, Finnish-English was introduced as a language pair of competition for the first time. We present experiments examining several variations on a morphologically-aware statistical phrase-based machine translation system for translating Finnish into English. Our system variations attempt to mitigate the issue of rich agglutinative morphology when translating from Finnish into English. Our WMT submission for Finnish-English preprocesses Finnish data with Omorfi (Pirinen, 2015), a Finnish morphological analyzer. We also present results for two other language pairs with morphologically interesting source languages, namely German-English and Czech-English.

1. Methodology

- Use current stable release (v3) of Moses, a state-of-the-art statistical phrase-based machine translation system.
- Train language models on the English Gigaword v5 corpus (Parker et al., 2011) using KenLM (Heafield et al., 2013).

2. Finnish-English

We tried various morphological tokenization schemes on the source language (Finnish) in order to mitigate its strong agglutination. The target language (English) was tokenized with the default Moses tokenizer script.

2.1 Finnish segmentation using Omorfi

- Adapt lattice technique of Dyer et al. (2009) to Finnish.
- Concatenate source side of training data with its one-best Morfessor (Creutz and Lagus, 2007) segmentation.
- Construct source lattice at test time using the top five Morfessor segmentations for each word.

2.2 Finnish segmentation using Omorfi

First word of Finnish Europarl, as processed by Omorfi:

- Isto:ntoukauden
- Isto:ntoukauden

We performed three experimental variations using Omorfi as the morphological segmenter:

1. Segment data using omorfi
2. Concatenate unsegmented and segmented data
3. Segment only out-of-vocabulary words

Table 1: Results for Finnish-English using Omorfi

<table>
<thead>
<tr>
<th>System</th>
<th>LM</th>
<th>TM</th>
<th>BLEU</th>
<th>BLEU-cased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>5</td>
<td>5</td>
<td>15.67</td>
<td>14.88</td>
</tr>
<tr>
<td>V1-omorfi</td>
<td>5</td>
<td>5</td>
<td>15.79</td>
<td>14.79</td>
</tr>
<tr>
<td>V2-omorfi</td>
<td>5</td>
<td>5</td>
<td>15.14</td>
<td>14.32</td>
</tr>
<tr>
<td>V3-omorfi</td>
<td>4</td>
<td>4</td>
<td>14.90</td>
<td>14.00</td>
</tr>
</tbody>
</table>

Table 2: Results for Finnish-English using Omorfi

For Czech-English, we performed experimental variations along two orthogonal dimensions:

- Morphologically segmentation of Czech data
  - no morphological segmentation
  - stem morphemes
  - morphological segmentation using Morfessor
- Part-of-speech (POS) intersection re-ranking feature
  - no POS intersection feature
  - use POS intersection feature to rerank

POS intersection was defined as follows. MorphoDiTa Straková et al. (2014) and the Stanford CoreNLP toolkit Manning et al. (2014) were used to POS tag the Czech and English sentences, respectively. A dictionary maps English and Czech POS tag values. The POS intersection score was defined as the number of identical POS tags between a Czech sentence and the hypothesized English translation.

Table 3: Results for Czech into English

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>BLEU-cased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses trained on Europarl</td>
<td>18.59</td>
<td>17.72</td>
</tr>
<tr>
<td>Moses trained on Europarl, Common Crawl and News Commentary</td>
<td>20.69</td>
<td>19.83</td>
</tr>
<tr>
<td>Stemming as pre-processing, Moses trained on Europarl</td>
<td>17.88</td>
<td>17.08</td>
</tr>
<tr>
<td>Morfessor trained on Europarl</td>
<td>16.48</td>
<td>15.74</td>
</tr>
<tr>
<td>POS intersection, Moses trained on Europarl</td>
<td>13.68</td>
<td>13.46</td>
</tr>
<tr>
<td>Morfessor trained on Europarl, POS intersection, Moses trained on Europarl</td>
<td>13.43</td>
<td>13.74</td>
</tr>
</tbody>
</table>

4. German-English

Following Holmqvist et al. (2011), we attempt to transform each German sentence de into a sentence de’ with a more English-like word order:

- Parse de using Stanford Parser (Manning et al., 2014)
- Restructure trees using Collins et al. (2005) rules 4 & 6
- Split German compounds using jWordSplitter

Table 4: Results for German and German’ into English

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>BLEU-cased</th>
</tr>
</thead>
<tbody>
<tr>
<td>de-en</td>
<td>24.9</td>
<td>23.8</td>
</tr>
<tr>
<td>de-en-’</td>
<td>24.9</td>
<td>23.8</td>
</tr>
</tbody>
</table>

References


Holmqvist, M., Stynne, S., and Ahrenberg, L. (2011). Experiments with word alignment, normalization and clause reordering for SMT between English and German. In WMT.


