Improving SMT by Using Parallel Data of a Closely Related Language

Petra Galuščáková and Ondřej Bojar
presented by Mark Fishel

Institute of Formal and Applied Linguistics
Charles University in Prague
{galusca,bojar}@ufal.mff.cuni.cz
Motivation

- The amount of training data in SMT critically affects translation quality.
- We demonstrate how to increase translation quality for one language pair by introducing parallel data from a closely related language.
- We improve English→Slovak translation using:
  - a large Czech–English parallel corpus and
  - a shallow MT system for Czech→Slovak translation.
Related Work – Pivoting

• Several concepts of using pivot or intermediate languages to improve MT quality:

  1) **combine the models** (phrase tables) of two translation systems (from the source to pivot language and from pivot to target language),

  2) “**triangulation**”, where the MT systems based on different pivot languages have to agree on the translation,

  3) **cascading** – combining the lists of the best translations – or creating artificial parallel data.

• Especially helpful if the pivot language is **closely related** to the source or target language and when only a small amount of parallel data is available for the source or target language [Babych et al., 2007].
MT Systems Used

- Česílko 1.0
  - Stand-alone MT system designed for closely related languages.
  - Supports only **Czech→Slovak** language pair.
    - Česílko 2.0 supports more pairs but performs worse for **cs→sk**.
  - Steps:
    1) Czech morphological analysis + statistical tagging,
    2) Simple dictionary for transfer,
    3) Slovak morphological generation.
  - Relies on the similarity of the languages in question,
    - e.g. it **does not change word order** during the translation.
  - Chosen because it performed well in a comparison of several **cs→sk** translation systems - fairly **robust** to various input text types.
MT Systems Used

- **Moses**
  - Open source phrase-based statistical machine translation system.
  - Used:
    - as the baseline direct \texttt{en$\rightarrow$sk} translation,
    - for the various configurations of pivoting.
Training Data

- **CzEng**
  - Freely available English–Czech parallel corpus.
  - Compiled from different type of sources.
  - We use version 0.9, but there is now a twice as big version 1.0.
  - We translated the Czech side into Slovak using Česílko 1.0.

- **English-Slovak Parallel Corpus**
  - http://hdl.handle.net/11858/00-097C-0000-0006-AADF-0
  - Compiled from freely available sources: Acquis, European Commission Website and parts of OPUS Corpus (EMEA, EUconst, KDE4 and PHP).
## Training Data Sizes

<table>
<thead>
<tr>
<th></th>
<th>CzEng</th>
<th>En-Sk Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Czech</td>
</tr>
<tr>
<td><strong>Sentences</strong></td>
<td>7.15 mil</td>
<td>7.15 mil</td>
</tr>
<tr>
<td><strong>Tokens</strong></td>
<td>85.09 mil</td>
<td>72.86 mil</td>
</tr>
</tbody>
</table>
Test Data (1/2)

- Derived from the **WMT 2011** shared task test data.
- Consists of **newspaper articles** covering a broad range of topics.
- Multi-parallel, available in Czech, English, German, Spanish and French.
- The **source** languages of the news articles **differ**
  - each article comes from one of the five languages and it is translated sentence by sentence to all the other languages.
Test Data (2/2)

- The extended the dataset to include Slovak version:
  - **Czech version was translated into Slovak,**
  - English version was provided to the translators only for reference in ambiguous or unclear cases.
- Many **discrepancies** between the English and Czech sentences in the original WMT data were found.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Czech</th>
<th>Slovak</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentences</strong></td>
<td>3 003</td>
<td>3 003</td>
<td>3 003</td>
</tr>
<tr>
<td><strong>Tokens</strong></td>
<td>77 086</td>
<td>68 108</td>
<td>63 730</td>
</tr>
</tbody>
</table>
Experiments
Setups Examined (1/2)

- **Direct Translation**
  - Statistical translation system Moses is trained and tuned on English–Slovak parallel data.
  - The resulting model is used for direct English→Slovak translation.

- **Moses+Česílko**
  - Simple MT system cascading with Czech used as the pivot language.
    - Moses is trained and tuned on the English–Czech corpus,
    - The resulting model is used for English→Czech translation, the output of which is further translated into Slovak by Česílko.
Setups Examined (2/2)

• **Česílko+Moses**
  - Synthetic parallel corpus:
    - The Czech part of the English–Czech corpus is automatically translated by Česílko into Slovak.
    - Moses is trained and tuned on this synthetic parallel corpus and the model is used for English→Slovak translation.

• **Česílko+Moses+Direct**
  - A combination of the direct and synthetic corpus approaches.
  - The training data are acquired as the concatenation of the manual English–Slovak corpus (as used in Direct Translation) and the synthetic English–Slovak corpus from Česílko+Moses.
  - This combined corpus is used for training of Moses and the model is used for English→Slovak translation.
Stemming for Word Alignment

- To overcome data sparseness.
- Only the first 4 letters of each word in both source and target languages were used for word alignment in all experiments.

<table>
<thead>
<tr>
<th>Preprocessing for word alignment</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Form</td>
<td>11.65 [11.04, 12.27]</td>
<td>71.43 [70.52, 71.43]</td>
</tr>
<tr>
<td>First 4 Characters</td>
<td>12.11 [11.51, 12.75]</td>
<td>70.71 [69.81, 71.60]</td>
</tr>
</tbody>
</table>
Tuning Data (1/2)

- Should we tune Moses on Slovak sentences translated:
  - from English manually, or
  - from Czech automatically using Česílko?
- A preparatory experiment using **WMT 2011** test set:
  - The first half serves for **tuning**, either in its **manual** Slovak version, or an **automatic** version obtained by Česílko.
  - The second half (always manual translation) used for **evaluation**.

<table>
<thead>
<tr>
<th>Reference of the tuning set</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Automatic</strong></td>
<td>12.73 [12.15, 13.32]</td>
<td>68.80 [67.94, 69.66]</td>
</tr>
<tr>
<td><strong>Manual</strong></td>
<td>12.61 [12.01, 13.19]</td>
<td>68.88 [68.03, 69.77]</td>
</tr>
</tbody>
</table>
Scores achieved using the automatically translated tuning data were slightly better than the results of the experiment which used manually translated data.

May be caused by the properties of Česílko and BLEU:

- Česílko translates word for word and does not change the word order → could lead to the higher scores when calculated by BLEU.

We opted for the automatic translation because it allows us to use larger tuning and test sets for the main experiments:

- For tuning we use WMT 2010 test set automatically translated from Czech into Slovak using Česílko.
- For testing we use the whole WMT 2011 test set (with manual Slovak).
## Pivoting Experiments

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Translation</td>
<td>10.83 [10.39, 11.25]</td>
<td>72.48 [71.89, 73.14]</td>
</tr>
<tr>
<td>Moses+Česílko</td>
<td>11.31 [10.89, 11.71]</td>
<td>71.11 [70.49, 71.71]</td>
</tr>
<tr>
<td>Česílko+Moses</td>
<td>11.89 [11.43, 12.30]</td>
<td>70.49 [69.86, 71.13]</td>
</tr>
<tr>
<td>Česílko+Moses+Direct</td>
<td>12.61 [12.13, 13.05]</td>
<td>69.14 [68.51, 69.79]</td>
</tr>
</tbody>
</table>
Pivoting Experiments

• **Direct Translation** is **significantly worse** than the results of all the other translation schemes.

• The result of Česílko+Moses, in which the English–Czech corpus is translated into Slovak and then used for training, performs significantly better than the converse Moses+Česílko when Moses operates on English→Czech and the resulting Czech is then translated into Slovak by Česílko.

• The **best result** was achieved when **both corpora**, the smaller manual English–Slovak and the larger English–Czech automatically translated to Slovak, were used.
Detailed BLEU

- We examined the n-gram components of BLEU scores.
- The tendency is the same for all en→sk translations:
  - the n-gram precision decreases exponentially with n.
- Česílko cs→sk translation:
  - = the 2nd step in simple cascading if the 1st step were ideal,
  - reaches BLEU of 42.45,
  - n-gram precision drop flatter.
- In line with Babych et al.:
  - a linear decrease of the n-gram precision for closely related languages, and
  - an exponential decrease for distant languages.
Conclusion

- We examined techniques for improving **English→Slovak** MT.
  - employing language resources of a closely related language, **Czech**.
- Pivoting via a closely related language performs well.
- Creating a **synthetic parallel corpus** by translating the Czech side of an English–Czech parallel corpus gave **results superior** to a simple cascading of the **en→cs** and **cs→sk** translation systems.
- The **best result** was obtained using **all available data**:
  - the parallel corpus for the direct **en→sk** translation, and
  - the synthetic **en-sk** constructed using shallow **cs→sk** MT.
Remark on Czech → Slovak

- **BLEU** score for the Česílko cs→sk translation is **42.45**, with the confidence interval [41.67,43.18].
  - (Measured on the very same WMT 2011 Slovak reference translations as our main en→sk experiments.)

- This high score may reflect:
  - **text source and translation direction**:
    - The Slovak version was created by translating from Czech.
    - The English version comes from various source languages.
  - **properties of Česílko, manual translation and BLEU**:
    - Česílko preserves the word order,
    - The translators may have pursued the same approach because they were also translating from Czech,
    - BLEU may thus give a high credit to matching n-grams