The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2008

Coşkun Mermer, Hamza Kaya, Ömer Farukhan Gunes, Mehmet Uğur Doğan
National Research Institute of Electronics and Cryptology (UEKAE), The Scientific and Technological Research Council of Turkey (TÜBİTAK), Gebze, Kocaeli 41470, Türkiye
(coskun,hamzaky,farukhan,mugur)@uekae.tubitak.gov.tr

1. Abstract
In this study, the TÜBİTAK-UEKAE statistical machine translation system based on the open-source phrase-based statistical machine translation software, Moses, is presented. Additionally, phrase-table augmentation is applied to maximize source language coverage; lexical approximation is applied to replace out-of-vocabulary words with known words prior to decoding; and automatic punctuation insertion is improved. We describe the preprocessing and postprocessing steps and our training and decoding procedures.

2. Introduction
Among the six translation tasks in IWSLT 2008, we participated in the following:
- Arabic-to-English (BTEC Task)
- Chinese-to-English (BTEC Task)
- Chinese-to-Spanish (BTEC Task)
- Chinese-to-English-to-Spanish (Pivot Task)

Used Resources:
- Supplied training data
- Buckwalter Arabic Morphological Analyzer (for BTEC AR-EN Task)

Lexical Approximation:
- To handle previously unseen words during decoding, the run-time lexical approximation method is used.
- An out-of-vocabulary word is replaced with the closest known word having the same feature.

This system obtained the best translation results in last year’s evaluation campaign (both AR-EN and JP-EN).

3. Training

Inclusion of Development Sets in Training
devsets1-3 were included in training (with references) in order to:
- Obtain better phrase alignments
- Increase the systems target phrase coverage

Sentence Splitting
Before translation model training, multi-sentence segments are split so as to prevent erroneous word alignments across sentence boundaries.

Orthographical Normalisation
One of our goals from last year was to investigate the striking discrepancy between the performance of our system in correct recognition result (CRR) and ASR output conditions in the Arabic-to-English task.

Orthographical Normalization

<table>
<thead>
<tr>
<th>Task</th>
<th>BTEC</th>
<th>DS1</th>
<th>DS2</th>
<th>DS3</th>
<th>DS4</th>
<th>DS5</th>
<th>DS6</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>36.79</td>
<td>29.46</td>
<td>27.06</td>
<td>28.75</td>
<td>27.80</td>
<td>27.17</td>
<td>26.41</td>
</tr>
</tbody>
</table>

Number of segments in the training corpora before and after automatic splitting

Orthographic Normalization

Effect of orthographical normalization on ASR output and CRR translation BLEU scores in the BTEC task

We also tried this normalization for CRR translation. The table above shows the results for both ASR and CRR conditions. BLEU scores were improved in all development sets.

Phrase Table Augmentation
There may be some source-language words in the training corpus without a one-word entry in the phrase table. These words are treated as out-of-vocabulary in previously unseen contexts.

Add such words as new phrase-pairs to the list of extracted phrases.

The target phrases in these phrase-pairs are selected from GIZA++ word alignments.

Word pairs with lexical translation probabilities above a relative threshold are selected.

Figure 1: IWSLT 2008 - TÜBİTAK UEKAE System.

Figure 2: Example Translation Output of the TÜBİTAK-UEKAE System.

4. Decoding

For decoding, Moses is used, which is a phrase-based beam-search decoder that uses a log-linear model with default scoring functions.

Run-time Lexical Approximation
The basic premise of lexical approximation is to replace a previously unseen word with a known word that has the same feature.

(LA-1) The feature function returns the morphological root(s) of the word.

(LA-2) Still-remaining unknown words go through a second step in which the feature function returns an orthographical normalization of the word obtained by removing all the vowels and diacritics.

Case Restoration
After decoding, target language case information is automatically restored using the Moses recasing tool. A lowercase-to-truecase translation model is trained and applied on the translation outputs, together with a few simple rules.

5. Results and Discussion
It is surprising to note that the Chinese-to-Spanish translation with English as the intermediate language (pivot translation) achieves better BLEU scores than the direct translation. We suspect this is due to the similarity of the 2008 test set to the pivot training corpora.

6. Conclusion
We have presented our Arabic-to-English, Chinese-to-English, Chinese-to-Spanish, and Chinese-to-English-to-Spanish statistical machine translation systems based on publicly-available software. We described our modifications to translation model generation, automatic punctuation insertion, and treatment of OOV words and presented our training and decoding procedures. Official evaluation results with correct recognition result and ASR output conditions were reported and discussed.