ABSTRACT

This paper describes the Barcelona Media SMT system in the IWSLT 2009 evaluation campaign. The Barcelona Media system is an statistical phrase-based system enriched with source context information. Adding source context in an SMT system is interesting to enhance the translation in order to solve lexical and structural choice errors. The novel technique uses a similarity metric among each test sentence and each training sentence. First experimental results of this technique are reported in the Arabic and Chinese Basic Traveling Expression Corpus (BTEC) task. Although working in a single domain, there are ambiguities in SMT translation units and slight improvements in BLEU are shown as both tasks (Zh2En and Az2En).

4 EXPERIMENTS

4.1 DATA AND PREPROCESSING

Arabic training, development, test and evaluation sets before the preprocessed (Arabic) and after (Arabic ‘)

• For Arabic, the MADA+TOKAN system was used for disambiguation and tokenization.

Chinese training, development, test and evaluation sets.

• For Chinese, no tokenization was performed.

4.2 OFFICIAL SUBMISSIONS

Primary system: we submitted the MOSES-based system enhanced with the source context information technique. As a contrastive system we submitted the MOSES-based system.

Secondary system: the above MOSES-based system with the following models and feature functions:

– TM(s), direct and inverse phrase/word based TM (10 words as maximum length per phrase).
– Distortion model, which assigns a cost linear to the reordering distance, while the cost is based on the number of source words which are skipped when translating a new source phrase.
– Lexicalized word reordering model.
– Word and phrase penalties, which count the number of words and phrases in the target string.
– Target-side LM (4-gram).

Example of source context information methodology.

4.3 POSTPROCESSING

We used a strategy for restoring punctuation and case information as proposed on the IWSLT 08 web page, using standard SRI LM tools: downbug to restore case information and hidden-ngrams to insert missing punctuation marks.

4.4 EXPERIMENTAL RESULTS

<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>Baseline</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU results for Arabic-English test set.</td>
<td>44.52</td>
<td>44.26</td>
<td>44.94</td>
</tr>
</tbody>
</table>

BLEU results for Chinese-English test set.

<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>Baseline</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU results for Arabic-English evaluation set (case/punctuation)</td>
<td>44.52</td>
<td>44.26</td>
<td>44.94</td>
</tr>
<tr>
<td>BLEU results for Chinese-English evaluation set.</td>
<td>44.52</td>
<td>44.26</td>
<td>44.94</td>
</tr>
</tbody>
</table>

5 CONCLUSIONS

This paper presented a novel technique which allows to introduce source context information into a phrase-based SMT system. The technique is based on using a new concept of translation unit which is composed of a conventional phrase plus its corresponding original source context, which is the context of the source language side of the bilingual sentence pair the phrase was originally extracted from. For simplicity, in this first implementation of the proposed methodology, we will restrict the idea of original source context to the whole source sentence the phrase was extracted from. Notice that, by this definition of translation unit, two identical phrases extracted from different aligned sentence pairs will constitute two different translation units. The similarity metric used as feature function for incorporating the source context information into the translation system is the cosine distance. According to this, the feature is computed for each phrase by considering the cosine distance between the vector models of the input sentence to be translated and the original source sentence the phrase was extracted from. For constructing the vector models, the standard bag of words approach with TFIDF weighting is used. Once the cosine distance is computed for each phrase and each input sentence to be translated, we can add it as feature function (hereinafter, cosine distance feature). Notice that, differently from most of the feature functions commonly implemented by state-of-the-art phrase based systems, the cost of this new feature function depends on the input sentence to be translated, which means that has to be computed during translation time (this, indeed, constitutes a computational overhead that cannot be dealt with beforehand). Because of this, we must keep one translation table for each input sentence to be translated. In the case one phrase table of a specific test sentence contains several identical phrase units with different costs of the cosine distance feature, we keep the one that has the highest cosine distance value.

6 ACKNOWLEDGEMENTS

This work has been partially funded by Barcelona Media Innovation Center and the Spanish Ministry of Education and Science through the Juan de la Cierva research program.