This paper describes the system developed by the LIUM laboratory for the 2008 IWSLT evaluation. We only participated in the Arabic/English BTEC task. We developed a statistical phrase-based system using the Moses toolkit and SYSTRAN's rule-based translation system to perform a morphological decomposition of the Arabic words. A continuous space language model was deployed to improve the modeling of the target language. Both approaches achieved significant improvements in the BLEU score. The system achieves a score of 49.4 on the test set of the 2008 IWSLT evaluation.

**INTRODUCTION**

- Only Arabic/English BTEC task (mainly text)
- Similar architecture than Ar/En NIST or Fr/En WMT system
- Only BTEC bitests
- Small improvements using additional LM data (Gigaword)
- Two different tokenizations of the Arabic source text:
  - full word mode
  - morphological decomposition kindly provided by SYSTRAN
- No system combination

**SYSTEM ARCHITECTURE**

- Statistical phrase-based system using Moses and own tools
- Two pass approach:
  - Decode with Moses and generate 1000-best lists
  - Rescore 1-best lists with continuous space LM
- Maximum BLEU tuning on rescoring 1-best lists using public CONDOR tool
- All models are case sensitive models
- Punctuation markers are considered as normal words

**EXPERIMENTAL EVALUATION**

**Language Modeling**

- English part of BTEC train and Dev1-4 (all English references)
- LDC Gigaword (3.3 billion words)
- GALE part of the 2006 NIST test set (1.1M words).
  - contains WEB blogs (travel related?)
- We realized after the evaluation that this data was only distributed to participants of the NIST MT eval
- 4-gram back-off LM with Modified Kneser-Ney smoothing
- Individual LMs are interpolated together

**Baseline experiment with NIST Arabic/English system**

<table>
<thead>
<tr>
<th>Translation model</th>
<th>Language model</th>
<th>Dev1</th>
<th>Dev5</th>
<th>Dev6</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>BTEC</td>
<td>21.35</td>
<td>47.09</td>
<td>43.45</td>
</tr>
<tr>
<td>BTEC + Dev1-4</td>
<td>BTEC + Dev1-4</td>
<td>22.90</td>
<td>45.16</td>
<td>42.98</td>
</tr>
<tr>
<td>BTEC + Dev1+4</td>
<td>BTEC + Dev1+4</td>
<td>23.18</td>
<td>44.15</td>
<td>43.62</td>
</tr>
<tr>
<td>BTEC + Dev1+4</td>
<td>BTEC + Dev1-4+Ga</td>
<td>25.10</td>
<td>47.52</td>
<td>46.52</td>
</tr>
<tr>
<td>BTEC + Dev1+4+Ga</td>
<td>BTEC + Dev1+4+Ga</td>
<td>28.39</td>
<td>47.62</td>
<td>49.19</td>
</tr>
</tbody>
</table>

**Improved tokenization**

- The large LM with the Gigaword data has only a small impact on the BLEU scores, despite a good gain in perplexity
- Gale bitext seem to be useful

**IMPROVED TOKENIZATION**

- It is known that a morphological decomposition of the Arabic words can improve the word coverage and by these means the translation quality
- Particularly true for under-resourced tasks like BTEC
- Usually the Backus-Naur and also the MADA and TOKAN tools from Columbia University are used

Using SYSTRAN's sentence analysis

- Sentence analysis represents a large share of the computation in a rule-based system
- Apply first decomposition rules coupled with a word dictionary
- For words that are not known in the dictionary, the most likely decomposition is guessed

In general, all possible decompositions of each word are generated and then filtered in the context of the sentence.

This steps uses lexical knowledge and a global analysis of the sentence.

- Integration of linguistic knowledge, but difficult to apply onto a word lattice from ASR

**ACKNOWLEDGMENTS**

This paper has been partially funded by the French Government under the project INSTAR (ANR JCJCo6_143038).