Continuous space LM:
- Tries to tackle the data sparseness problem [Y. Bengio, NIPS’01]
- Idea: projection of the word indices onto a continuous space
- α-gram probability estimation in this continuous space
- → Better generalization to unseen α-grams can be expected
- Implementation using a 3-layer neural network
- Backpropagation training to learn the continuous representation of the words and the α-gram LM probabilities
- Several tricks to tackle the high complexity

Dev Data:
- Dev4 and Dev5 seem to be very similar
- Dev6 is mainly close to the BTEC training corpus.
- Analysis of the Arabic source: Test08 seems to be close to Dev4/5
- → All tuning is done on Dev5
- Results on Dev6 and Test08

Post-processing:
- Translations of Test08 data contain only few punctuation marks
- This is in contrast to Dev5 and Dev6
- → Negative impact on our system
- We were unable to analyze the Arabic source
- Simple post-processing to restore end-of-sentence punctuation

EXPERIMENTAL EVALUATION

Language Modeling:
- English part of BTEC train and Dev1-4 (all English references)
- GALE part of the 2006 NIST test set (1.1M words).
-– Contains WEB logs (travel-related?)
-– We realized after the evaluation that this data was only distributed to participants of the NIST MT eval
- 4-gram back-off LMs 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>Dev5</th>
<th>Dev6</th>
<th>Dev4</th>
<th>Dev5</th>
<th>Dev6</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>NIST</td>
<td>44.19</td>
<td>47.52</td>
<td>38.26</td>
<td>47.62</td>
<td>45.80</td>
</tr>
<tr>
<td>BTEC + Dev1-4</td>
<td>NIST</td>
<td>47.16</td>
<td>50.76</td>
<td>31.63</td>
<td>44.15</td>
<td>50.08</td>
</tr>
<tr>
<td>+ Gale</td>
<td>NIST</td>
<td>47.16</td>
<td>50.76</td>
<td>31.63</td>
<td>44.15</td>
<td>50.08</td>
</tr>
<tr>
<td>+ Gale</td>
<td>BTEC</td>
<td>49.39</td>
<td>47.92</td>
<td>43.94</td>
<td>46.72</td>
<td>44.19</td>
</tr>
<tr>
<td>+ Gale</td>
<td>BTEC + Gale</td>
<td>51.20</td>
<td>52.10</td>
<td>48.09</td>
<td>52.42</td>
<td>47.52</td>
</tr>
<tr>
<td>BTEC + Dev1-4</td>
<td>BTEC + Dev1-6</td>
<td>51.20</td>
<td>52.10</td>
<td>48.09</td>
<td>52.42</td>
<td>47.52</td>
</tr>
<tr>
<td>+ Gale</td>
<td>BTEC + Dev1-6</td>
<td>51.20</td>
<td>52.10</td>
<td>48.09</td>
<td>52.42</td>
<td>47.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Improved tokenization:</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC + Dev1-4 + Gale</td>
</tr>
<tr>
<td>BTEC + Dev1-6 + Gale</td>
</tr>
</tbody>
</table>

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 Buckwalter transliterator and 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 step uses lexical knowledge and a global analysis of the sentence.
- Integration of linguistic knowledge, but difficult to apply onto a word lattice from ASR
- Result analysis:
- Substantial improvements in the BLEU score: Dev6: 47.62 ↔ 52.10, Test08: 44.19 ↔ 48.09
- Gale bitexts are not useful any more
- The morphological decomposition seems to achieve better translations than adding additional bilingual out-of-domain data.
- Relation to SPE:
- word-based system: SMT performs the full translation task
- SPE: SMT only corrects the output of rule-based system
- SYSTRAN’s tokenisation + SMT: somewhere in the continuum between both.

INTERFACE WITH SPEECH RECOGNITION
- Simple 1-best coupling
- → Bad performance on ASR transcriptions

CONCLUSION AND PERSPECTIVES
- Based on Moses decoder
- Two extensions achieved significant improvements:
- morphological word decomposition based on SYSTRAN’s rule-based translation system
- n-best list rescoring with a continuous space language model
- No gain with additional bitexts
- Small improvements with additional LM data

Ongoing work
- Explore unsupervised training of translation model
- Comparison of SYSTRAN’s morphological decomposition with MADA/TOKAN and other standard tools

ACKNOWLEDGMENTS

This work has been partially funded by the French Government under the project INSTAR (ANR JCJC06_143038).