QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic Spoken Language Translation

Hassan Sajjad, Francisco Guzmán, Preslav Nakov, Ahmed Abdelali, Kenton Murray, Fahad Al Obaidli, Stephan Vogel
Qatar Computing Research Institute

1. Baseline System
- Train: TED 2013 training data;
- Dev: dev2010;
- Dev test: tst2010;
- Maximum sentence length: 100 tokens;
- English truecasing: For AR-EN only;
- Word alignments: IBM4 + grow-diag-final-and;
- Maximum phrase length: 7 tokens;
- Language model: 5-gram;
- Reordering: msd-bidirectional-fe, mono-punct;
- Tuning: PRO.

2. Adaptation
- Phrase table combination (TED+UN)
  - Three additional features
    - F1 if a phrase pair came from TED
    - F2 if a phrase pair came from UN
    - F3 if a phrase pair came from both TED and UN
  - Preferring TED data performs best
  - +0.6 BLEU points
- Backoff phrase tables (TED, UN)
  - n-gram order 6 or less
  - +0.6 BLEU points
- Modified Moore-Lewis filtering on UN
  - -0.3 BLEU points (UN filtered combined with TED)

3. Arabic Segmentation
- Arabic segmentation schemes
  - D0, D1, D2, D3, S2, ATB (using MADA)

4. System Combination
- 1. Decoder settings
  - OSM, MBR, 100 translations per input phrase
- 2. Arabic segmentations
- 3. Adaptation
- 4. Decoders
  - Moses, cdec, Jane

System combination: +0.6 BLEU points over best individual system

5. QCRI Normalizer for Arabic Output and References
- Translating into Arabic:
  - Spelling inconsistencies (Ta Marbuta, Alef)
  - Punctuation symbols (Arabic & English mixed)
  - Digits (Arabic & Indian mixed)
  - Diacritics (with, without or wrong)
- Evaluation unfairly penalizes the translation output

6. Arabic – English

7. Official Scores

<table>
<thead>
<tr>
<th></th>
<th>tst2011</th>
<th>tst2012</th>
<th>tst2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR-EN</td>
<td>Primary</td>
<td>27.8</td>
<td>30.3</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>26.9</td>
<td>28.7</td>
</tr>
<tr>
<td>EN-AR</td>
<td>Primary</td>
<td>15.5</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>15.2</td>
<td>15.7</td>
</tr>
<tr>
<td>EN-AR (SLT)</td>
<td>Primary</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Primary: system combination. Secondary: best individual system

8. Conclusion & Future Work

+3.4 BLEU points over the baseline AR-EN system

What helped most
- System combination
- Interpolated language model
- Adaptation using full UN data
- Operation sequence model
- PRO with fixed BLEU+1

Future work
- Why less improvement for EN-AR than for AR-EN?

* The system uses OSM and MBR with baseline settings