IBM Arabic-to-English Translation
for IWSLT 2006

Young-Suk Lee

IWSLT 2006
November 28, 2006, Kyoto, Japan
Outline

• Baseline Decoding
• Technical Challenges and Solutions
• IWSLT 2006 Evaluation Results
• Conclusions
Baseline Decoding

Input sentence

Pre-processing

DP-based phrase decoder

Translation output

**Phrase translation models**
- Direct / source channel / unigram models
  - [Tillmann 2003], [Lee et al. 2006]
**Modified IBM Model 1 cost** [Lee et al. 2006]
**Word trigram language models**
**Distortion models** [Al-Onaizan & Papineni 2006]
**Word/block count penalty** [Zens & Ney 2004]

*We are open from seven p.m. to midnight.*
IWSLT 2006 Challenge: High OOV Rate

Correct Recognition Result

<table>
<thead>
<tr>
<th></th>
<th>Token Cnt</th>
<th>OOV Cnt</th>
<th>Avg. seg length</th>
<th>OOV Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval06</td>
<td>5,229</td>
<td>609</td>
<td>10.5</td>
<td>11.65 %</td>
</tr>
<tr>
<td>Dev06</td>
<td>4,763</td>
<td>489</td>
<td>10.1</td>
<td>10.27 %</td>
</tr>
<tr>
<td>Eval05</td>
<td>3,164</td>
<td>157</td>
<td>6.3</td>
<td>4.96 %</td>
</tr>
</tbody>
</table>
Word Segmentation & Morphological Analysis

Word Segmentation  [Lee et al. 2003], Morphological Analysis  [Lee 2004]
OOV Rate Reduction

- **Eval06**
  - BTEC baseline: 12%
  - BTEC segmentation: 3%
  - BTEC analysis: 4%

- **Dev06**
  - BTEC baseline: 10%
  - BTEC segmentation: 2%
  - BTEC analysis: 3%

- **Eval05**
  - BTEC baseline: 8%
  - BTEC segmentation: 2%
  - BTEC analysis: 3%
Translation Quality Improvement

![Graph showing BLEU scores for Eval06, Dev06, and Eval05 with BTEC baseline, segmentation, and analysis models.]

- Eval06: BTEC baseline (0.22), BTEC segmentation (0.24), BTEC analysis (0.25)
- Dev06: BTEC baseline (0.29), BTEC segmentation (0.25), BTEC analysis (0.29)
- Eval05: BTEC baseline (0.51), BTEC segmentation (0.56), BTEC analysis (0.56)
## Out-of-Domain Corpora: Newswires

<table>
<thead>
<tr>
<th>Source</th>
<th># AR words</th>
<th># EN words</th>
<th># sent pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>159,213</td>
<td>189,239</td>
<td>20,000</td>
</tr>
<tr>
<td>LDC2003T18</td>
<td>26,146</td>
<td>33,869</td>
<td>1,043</td>
</tr>
<tr>
<td>LDC2003E05</td>
<td>103,717</td>
<td>129,181</td>
<td>4,235</td>
</tr>
<tr>
<td>LDC2003E09</td>
<td>123,505</td>
<td>150,865</td>
<td>5,003</td>
</tr>
<tr>
<td>LDC2004E07</td>
<td>520,971</td>
<td>681,613</td>
<td>20,358</td>
</tr>
<tr>
<td>LDC2004E11</td>
<td>227,792</td>
<td>310,079</td>
<td>8,576</td>
</tr>
<tr>
<td>LDC2004E08</td>
<td>1,771,893</td>
<td>2,207,934</td>
<td>52,042</td>
</tr>
<tr>
<td>LDC2005E46</td>
<td>616,879</td>
<td>819,354</td>
<td>24,874</td>
</tr>
<tr>
<td>LDC2001T55</td>
<td>70,183</td>
<td>80,354</td>
<td>2,346</td>
</tr>
<tr>
<td>FBIS</td>
<td>86,614</td>
<td>117,420</td>
<td>2,624</td>
</tr>
<tr>
<td><strong>OOD Total</strong></td>
<td><strong>3,547,700</strong></td>
<td><strong>4,530,669</strong></td>
<td><strong>121,119</strong></td>
</tr>
</tbody>
</table>
OOV Rate Reduction

- **Eval06**: Baseline - 12%, BTEC+OOD baseline - 4%, BTEC+OOD segmentation - 2%, BTEC+OOD analysis - 1%
- **Dev06**: Baseline - 10%, BTEC+OOD baseline - 4%
- **Eval05**: Baseline - 8%, BTEC+OOD baseline - 2%
Eval05 Translation Quality Improvement

BTEC ; OOD Corpora Ratio

BLEU16n4c

baseline
segmentation
analysis
Dev06 Translation Quality Improvement

BLEU7n4c

BTEC ; OOD Corpora Ratio
Eval06 Translation Quality Improvement

Correct Recognition

ASR Output

BTEC ; OOD Corpora Ratio = 4 ; 1
# IWSLT 2006 Open Data Track Results

<table>
<thead>
<tr>
<th>Correct Recognition Result</th>
<th>BLEU4</th>
<th>NIST</th>
<th>METEOR</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official</td>
<td>0.2549</td>
<td>6.3769</td>
<td>0.5316</td>
<td>0.5668</td>
<td>0.4825</td>
</tr>
<tr>
<td>Additional</td>
<td>0.2773</td>
<td>7.1681</td>
<td>0.5314</td>
<td>0.5593</td>
<td>0.4480</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ASR Output</th>
<th>BLEU4</th>
<th>NIST</th>
<th>METEOR</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official</td>
<td>0.2274</td>
<td>5.8466</td>
<td>0.4845</td>
<td>0.6049</td>
<td>0.5198</td>
</tr>
<tr>
<td>Additional</td>
<td>0.2428</td>
<td>6.4867</td>
<td>0.4842</td>
<td>0.6035</td>
<td>0.4958</td>
</tr>
</tbody>
</table>

Scores in BLUE indicate the best scores under the given condition.
Conclusions

• Techniques for improving translation quality & increasing vocabulary coverage
  • Word segmentation & morphological analysis
  • Proper combination of domain-specific and out-of-domain corpora for model training

• Effectiveness of the techniques
  • Demonstrated in the IBM Arabic-to-English translation system performances in the Open Data Track