Edinburgh System Description for
2005 IWSLT Speech Translation Evaluation

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne,
Chris Callison-Burch, Miles Osborne, David Talbot

School of Informatics
University of Edinburgh
Outline

- SMT at the University of Edinburgh
- Adaptations to the IWSLT task
  - optimizing word alignment
  - optimizing lexicalized reordering
  - optimizing reordering limit
- Results
SMT at the University of Edinburgh

- Currently involved: 2 faculty, 6 graduate students
- Systems for
  - DARPA/NIST (GALE): Arabic-English, Chinese-English
  - Euromatrix: European Union languages
  - IWSLT: first attempt at speech domain
- Available software
  - Pharaoh (incl. training code)
  - New open source decoder (with Linear-B)
Phrase-Based Translation

- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered
Log-Linear Model

- Several model components combined in log-linear model

\[
\hat{e} = \arg \max_e p(e|f) = \arg \max_e \sum_{m=1}^{M} \lambda_m h_m(e, f)
\]

- Tuning with minimum error rate training [Och, 2003]
Word Alignment based on IBM Models

- Intersection of GIZA++ bidirectional alignments
- Growing additional alignment points
Extract Phrase Pairs

- Consistent with the word alignment :=
  phrase alignment has to contain all alignment points for all covered words

\[(\bar{e}, \bar{f}) \in BP \iff \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f}\]

AND \[\forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e}\]
Probability Distribution over Phrase Pairs

- We need a probability distribution $\phi(f|\bar{e})$ over the collected phrase pairs

$\Rightarrow$ Different scoring methods

- relative frequency of collected phrases: $\phi(f|\bar{e}) = \frac{\text{count}(f,\bar{e})}{\sum_{\bar{f}} \text{count}(f,\bar{e})}$
- conversely $\phi(\bar{e}|f)$
- use lexical translation probabilities, also both directions
- word penalty
- phrase penalty
IWSLT Task vs. DARPA/NIST Task

- Participated in supplied track for all language pairs
- Much less training data
  - 20,000 sentences vs. millions
- Different text type
  - shorter sentences
  - questions and answers
  - travel domain vs. news
- Faster Training
  - 15 minutes vs. many days
Adaptations for IWSLT Task

- Optimizing word alignment
  - small corpus $\rightarrow$ high alignment error rate
  - high precision alignments may be better than high recall

- Optimizing lexicalized reordering
  - new model similar to block-orientation model [Tillmann, 2004]
  - tuning of model variants

- Optimizing reordering limit
  - so far, used reordering limit of maximum movement of 4 words
  - better reordering model may allow more reordering
Optimizing Word Alignment

- Heuristics with increasing number of alignment points:
  - intersection of bidirectional GIZA++ alignments (intersection)
  - add union alignment points, if directly (block) neighboring (grow)
  - above + also allow diagonal neighborhood (grow-diag)
  - above + also add points that connect two unaligned words (final-and)
  - above + also add points that connect one unaligned words (final)

- Effect on phrase table size (Japanese–English):

<table>
<thead>
<tr>
<th></th>
<th>final</th>
<th>final-and</th>
<th>grow-diag</th>
<th>grow</th>
<th>intersect</th>
</tr>
</thead>
<tbody>
<tr>
<td>English words</td>
<td>187,843</td>
<td>187,843</td>
<td>187,843</td>
<td>187,843</td>
<td>187,843</td>
</tr>
<tr>
<td>Alignment points</td>
<td>282,110</td>
<td>234,027</td>
<td>220,318</td>
<td>185,714</td>
<td>79,200</td>
</tr>
<tr>
<td>Distinct phrase pairs</td>
<td>61,168</td>
<td>270,654</td>
<td>447,550</td>
<td>854,680</td>
<td>2,561,715</td>
</tr>
</tbody>
</table>
Optimizing Word Alignment

- Fewer alignment points, larger phrase table often better
  - intersection: +5% BLEU for Japanese, Chinese

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>final (default)</th>
<th>final-and</th>
<th>grow-diag</th>
<th>grow</th>
<th>intersect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic-English</td>
<td>48.8</td>
<td>48.5</td>
<td>49.9</td>
<td>39.9</td>
<td>47.5</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>40.4</td>
<td>39.9</td>
<td>39.0</td>
<td>39.1</td>
<td>45.1</td>
</tr>
<tr>
<td>Korean-English</td>
<td>33.9</td>
<td>35.7</td>
<td>27.7</td>
<td>13.5</td>
<td>35.4</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>28.9</td>
<td>32.4</td>
<td>31.7</td>
<td>32.8</td>
<td>34.6</td>
</tr>
<tr>
<td>English-Chinese</td>
<td><strong>15.4</strong></td>
<td>9.6</td>
<td>8.1</td>
<td>15.4</td>
<td>15.2</td>
</tr>
</tbody>
</table>
New Lexicalized Reordering Model

- Traditionally, only a distance-based reordering model
- Lexicalized reordering: conditioned on translated phrase
- Orientation types: monotone, swap, discontinuous
- Extension of work at IBM [Tillmann, 2004]
Orientation Types

monotone (m), swap (s), discontinous (d)
Lexicalized Reordering Model Variants

- Distinguish between monotone, swap, and discontinuous... or just check monotonicity?
- Condition on foreign phrase... or on both foreign and English phrase?
- Model reordering in respect to previously translated phrase... or also in respect to following phrase?
Different Methods for Language Pairs

- Optimized alignment heuristic and lexicalized reordering model variant
  - baseline: only distance-based reordering model
  - often intersection gives best results
  - mixed picture for lexicalized reordering model

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Reordering</th>
<th>Word Alignment</th>
<th>Baseline</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic-English</td>
<td>orientation-bi-fe</td>
<td>final-and</td>
<td>49.9</td>
<td>50.9</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>orientation-fe</td>
<td>intersect</td>
<td>45.1</td>
<td>47.6</td>
</tr>
<tr>
<td>Korean-English</td>
<td>orientation-fe</td>
<td>intersect</td>
<td>35.7</td>
<td>42.3</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>monotonicity-fe</td>
<td>intersect</td>
<td>34.6</td>
<td>38.6</td>
</tr>
<tr>
<td>English-Chinese</td>
<td>monotonicity-bi-fe</td>
<td>grow-diag</td>
<td>15.2</td>
<td>16.6</td>
</tr>
</tbody>
</table>
Optimizing Reordering Limit

- Traditionally, we used a reordering limit 4 words
  - no reordering worse and faster
  - unlimited reordering worse and slower

- Lexicalized reordering model allows more reordering

<table>
<thead>
<tr>
<th>Reordering Limit</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic-English</td>
<td>50.3</td>
<td>50.4</td>
<td>50.1</td>
<td><strong>50.6</strong></td>
<td>50.0</td>
<td>50.1</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>46.4</td>
<td>48.3</td>
<td>48.8</td>
<td>49.1</td>
<td>49.0</td>
<td><strong>49.9</strong></td>
</tr>
<tr>
<td>Korean-English</td>
<td>37.8</td>
<td>41.8</td>
<td>42.0</td>
<td>44.1</td>
<td>44.1</td>
<td><strong>45.2</strong></td>
</tr>
<tr>
<td>Chinese-English</td>
<td>36.8</td>
<td>36.7</td>
<td>37.2</td>
<td><strong>37.5</strong></td>
<td>36.9</td>
<td>37.2</td>
</tr>
<tr>
<td>English-Chinese</td>
<td>16.6</td>
<td>16.8</td>
<td>16.0</td>
<td>16.4</td>
<td><strong>17.2</strong></td>
<td>17.1</td>
</tr>
</tbody>
</table>
Also Tried, but no Success

- Optimizing GIZA++ parameters
  - no consistent gains
  - not given up yet — this needs to be revisited

- Manual reordering rules for Japanese
  - move verb in front of sentence
  - move markers in front of nouns
  - ... but we have no Japanese POS tagger, parser
## Results

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>BLEU</th>
<th>NIST</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic-English</td>
<td>0.5105 (0.93)</td>
<td>7.6382 (0.70)</td>
<td>5th of 8</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>0.3778 (0.81)</td>
<td>4.0784 (0.41)</td>
<td>4th of 7</td>
</tr>
<tr>
<td>Korean-English</td>
<td>0.3672 (0.88)</td>
<td>5.6172 (0.60)</td>
<td>1st of 4</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>0.4650 (0.90)</td>
<td>6.4922 (0.62)</td>
<td>3rd of 10</td>
</tr>
<tr>
<td>English-Chinese</td>
<td>0.2127 (0.94)</td>
<td>5.1807 (0.98)</td>
<td>1st of 2</td>
</tr>
</tbody>
</table>

- Comparably good results
- Why so low NIST scores?
  - tuned for shortest reference length
  → relatively short output (length penalty in parenthesis)
## Tuning for *Average* Reference Length

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic-English</td>
<td>0.5180 (0.98)</td>
<td>9.7749 (0.94)</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>0.3941 (0.95)</td>
<td>8.1209 (0.91)</td>
</tr>
<tr>
<td>Korean-English</td>
<td>0.3859 (1.00)</td>
<td>8.4455 (0.99)</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>0.4364 (1.00)</td>
<td>9.0834 (0.99)</td>
</tr>
<tr>
<td>English-Chinese</td>
<td>0.2230 (0.91)</td>
<td>5.2391 (0.97)</td>
</tr>
</tbody>
</table>

- **Much higher NIST scores**

- **BLEU not changed much**
  - 4 out of 5 higher, but not by much
Conclusion

- Successful participation, competitive results
- System could be easily applied
- Lessons from adaptation
  - high precision word alignment heuristics better
  - lexicalized reordering model validated
  - longer reordering windows