Morpho-Syntax in Statistical Machine Translation

Young-Suk Lee
IBM T. J. Watson Research Center

OpenLab 2006
March 30 – April 1, 2006
Reordering Rules: Motivations

de transportes especialmente peligrosos
of extremely dangerous transport

los procedimientos administrativos complejos
∅ complex administrative procedures
Outline

• Baseline Phrase Translation System
  o Block Acquisition & Decoding

• Acquisition of Reordering Rules
  o Base Reordering Rules
  o Lexicalized Reordering Rules

• Experimental Results

• Related and Ongoing Work
Baseline Block Acquisition

Block (b): a phrase translation pair consisting of source $f$ & target $\bar{e}$ phrase

Tillmann 2003, EMNLP Proceedings
Extended Block Acquisition Algorithm

- Expansion word list: A list of target words typically aligned to null source words (e.g. I, we, are)
- Extend the target phrase to include an expansion word if it occurs in the neighborhood of a block
Decoding

• Phrase translation models
  
  • Direct model: \[ p(\bar{e} | \bar{f}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{e}'} \text{count}(\bar{e}', \bar{f})} \]

  • Source channel model: \[ p(\bar{f} | \bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}'} \text{count}(\bar{f}', \bar{e})} \]

  • Block unigram model: \[ p(b) = \frac{\text{count}(b)}{\sum_{b'} \text{count}(b')} \] , \( b = (\bar{e}, \bar{f}) \)
Decoding Cont’d …

• IBM Model 1 cost per phrase in both directions

\[ \sum_{j=1}^{m} -\log_{10} \max_{i} p(f_j | e_i) \text{, } 1 \leq i \leq n \]

• Word & part-of-speech tag trigram language models

• Word-level distortion models applied to blocks
  • Al-Onaizan 2004, DARPA MT Evaluation Workshop

• Word & block count penalty
  • Zens and Ney 2004, HLT Proceedings
Acquisition of Base Reordering Rules

- Viterbi-align
  - Part-of-speech tagged source language corpus
  - Un-tagged target language corpus

- Identify the source language part-of-speech tag sequence (monotone increasing)
  - whose corresponding target word sequence is not monotone increasing

- Compute the reordering probabilities of each part-of-speech tag sequence
Reordering Probability Computation

\[ p(\text{reorder}_i \mid \text{tag}_k) = \frac{\text{count}(\text{reorder}_i, \text{tag}_k)}{\sum_{\text{reorder}'} \text{count}(\text{reorder}', \text{tag}_k)} \]

<table>
<thead>
<tr>
<th>DTS₁</th>
<th>NNS₂</th>
<th>JJS₃</th>
<th>JJS₄</th>
<th>IN₁</th>
<th>NNS₂</th>
<th>RB₃</th>
<th>JJS₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{reorder}'</td>
<td>[ p(\text{reorder}_i \mid \text{tag}_k) ]</td>
<td>\text{reorder}'</td>
<td>[ p(\text{reorder}_i \mid \text{tag}_k) ]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2 3 4</td>
<td>0.107</td>
<td>1 2 3 4</td>
<td>0.179</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2 4 3</td>
<td>0.059</td>
<td>1 2 4 3</td>
<td>0.047</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 3 2 4</td>
<td>0.239</td>
<td>1 3 2 4</td>
<td>0.094</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 3 4 2</td>
<td>0.109</td>
<td>1 3 4 2</td>
<td>0.560</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 4 2 3</td>
<td>0.111</td>
<td>1 4 2 3</td>
<td>0.072</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 4 3 2</td>
<td>0.375</td>
<td>1 4 3 2</td>
<td>0.048</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
One Best Reordering Rules

\[ p(\text{reorder}_f | \text{tag}) > p(\text{reorder}_s | \text{tag}) + \alpha \]
Lexicalization of Exceptions

El apoyo operativo de la Secretaría General del Consejo
the operational support of the Secretary General of the Council

DT₁ NN₂ JJ₃[~General] IN₄ → DT₁ JJ₃ NN₂ IN₄

El Fondo, por supuesto, debe continuar cumpliendo con su misión de investigación sobre la búsqueda de variedades más adaptadas a la demanda y lo menos nocivas posible,
The Fund must of course continue to serve its purpose and pursue research into varieties more suited to demand and causing as little harm as possible.

IN₁ NNS₂ RB₃ JJS₄[~adaptadas] → IN₁ RB₃ JJS₃ NNS₂
Lexicalized Reordering Rules

- Identify the key part-of-speech tag in the base reordering rules

- Replace the key part-of-speech tag with the corresponding word
  - DT NN JJ IN $\rightarrow$ DT NN General IN

- Compute reordering probabilities of lexicalized part-of-speech tag sequences

- Exception word list
  - If the reordering pattern with the highest probability is monotone increasing, select the word in the pattern as an exception
## Lexicalized Reordering Probabilities

| $DT_1$ | $NN_2$ | $General_3$ | $IN_4$ | $p(reorder_i | tag_k)$ |
|--------|--------|-------------|--------|----------------------|
| 1      | 2      | 3           | 4      | 0.454                |
| 1      | 2      | 4           | 3      | 0.021                |
| 1      | 3      | 2           | 4      | 0.201                |
| 1      | 3      | 4           | 2      | 0.012                |
| 1      | 4      | 2           | 3      | 0.194                |
| 1      | 4      | 3           | 2      | 0.098                |
| 4      | 1      | 2           | 3      | 0.007                |
| 4      | 1      | 3           | 2      | 0.013                |
Performance Evaluations

• Translation model training corpus
  • \( \sim 1.3 \text{ M sentence pairs from EPPS distributed by RWTH} \)

• Language model training corpus
  • EPPS English corpus: \( \sim 35 \text{ M words} \)
  • UN parallel corpus English (LDC94T4A): \( \sim 45 \text{ M words} \)
  • English gigaword second edition (LDC2005T12): \( \sim 2.5 \text{ B words} \)
## Evaluation Corpus Statistics

<table>
<thead>
<tr>
<th>Data Sets</th>
<th># of Segments</th>
<th>Avg. Segment Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPPS Dev06 FTE</td>
<td>699</td>
<td>35 words/segment</td>
</tr>
<tr>
<td>EPPS Dev06 VHT</td>
<td>792</td>
<td>31 words/segment</td>
</tr>
<tr>
<td>CORTES Dev06 FTE</td>
<td>753</td>
<td>37 words/segment</td>
</tr>
<tr>
<td>CORTES Dev06 VHT</td>
<td>920</td>
<td>31 words/segment</td>
</tr>
</tbody>
</table>
Lexicalized Reordering Rules: Impact

BLEUr2n4c

- EPPS Dev06 FTE: 0.5322
- EPPS Dev06 VHT: 0.5204
- CORTES Dev06 FTE: 0.4439
- CORTES Dev06 VHT: 0.4186
Base vs. Lexicalized Reordering Rules

![Bar chart comparing BLEU scores between Base and Lexicalized Reordering Rules for different datasets: EPPS Dev06 FTE, EPPS Dev06 VHT, CORTES Dev06 FTE, CORTES Dev06 VHT. The scores range from 0.4 to 0.55.](image)
Related Work

• N-best Reordering in Arabic-to-English Translation
  o Statistically significant performance improvement by applying local reordering to noun phrase parsed Arabic
  o *IBM Site Report: DARPA MT Evaluation Workshop 2004*

• Morphological Analysis for Statistical Machine Translation
  o Identify one to one word correspondences between Arabic and English to improve word to word translation qualities
  o *Companion Volume of HLT-NAACL 2004, pages 57–60*

• Local Reordering for Spanish-English Translations
  o Presentation at TC-STAR 2005 Evaluation Workshop
  o *April 21-22, 2005, Trento, Italy*
Ongoing Work

• Non-local reordering models
  • [Se ha puesto a prueba]_{VP} [su voluntad]_{NP} → [Its will]_{NP} [has been put to the test]_{VP}

• Todas sus Señorías firmaron [con los electores]_{PP} [un contrato]_{NP} → All your ladies and gentlemen signed [a contract]_{NP} [with the electors]_{PP}

• Integration of reordering models into the decoder