Head Finalization Reordering for Chinese-to-Japanese Machine Translation

Dan Han *, Katsuhito Sudoh †, Xianchao Wu $, Kevin Duh §, Hajime Tsukada †, Masaaki Nagata †

* National Institute of Informatics
† NTT Communication Science Laboratories
§ Baidu Japan, Inc.
§ Nara Institute of Science and Technology(NAIST)
Outline

- Introduction
- Motivation & Objective
- Syntactic-Based Reordering Rules
- Other Reordering Issues
- Conclusion
Introduction

- English and Chinese are head-initial languages, while Japanese is a typical head-final language.

- Head Finalization Reordering Rule for English-Japanese MT (Isozaki et al. 2010).
  - Using the parsed result of Enju, an HPSG parser (Miyao and Tsujii, 2008) that outputs the syntactic heads.

Move syntactic heads to the end of the corresponding syntactic constituents
● Head-Final English (HFE)
Motivation

- HF works well for Japanese-English, but not for other language pairs.
- Attempt to remedy it at the level of syntactic-head analysis.
- Discrepancies in Head Definition among Chinese, Japanese, and English cause reordering issues while implementing HF.
- In this work, example of analysis and solution for Chinese-to-Japanese.
Problem!

- Head-Final Chinese (HFC)
Perfect!

- Refined-Head-Final Chinese (Refined-HFC)
Objective

- Present detailed syntactic analysis of reordering issues.
- Define novel reordering rules based on HF and linguistically inspired refinements.
Syntactic-based Reordering Rules

- Aspect Particle
- Adverbial Modifier 'bu4'
- Sentence-final Particle
- Et cetera
- Punctuation
- Coordination
Aspect Particle

Example for Aspect Particle

Ch 我去过东京。

En I have been to Tokyo.

<table>
<thead>
<tr>
<th>HFC</th>
<th>我 (I) 东京(Tokyo) 过(have) 去(been to)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja</td>
<td>私(は) 東京(に) いった</td>
</tr>
<tr>
<td>R-HFC</td>
<td>我 东京 去 过</td>
</tr>
</tbody>
</table>
Adverbial Modifier 'bu4'

Example for Adverbial Modifier bu4

| Ch   | 我不看电视。 |
| En   | I do not watch TV. |
| HFC  | 我(l) 电视(TV) 不(do not) 看(watch) |
| Ja   | 私(は) テレビ(を) 見 ない |
| R-HFC| 我 电视 看 不 |
天气是真好啊。
It is good weather.

<table>
<thead>
<tr>
<th>HFC</th>
<th>Ja</th>
<th>R-HFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>啊</td>
<td>いい</td>
<td>天気</td>
</tr>
<tr>
<td>天气(weather)</td>
<td>真好(good)</td>
<td>是(is)</td>
</tr>
<tr>
<td>ね</td>
<td>です</td>
<td>啊</td>
</tr>
</tbody>
</table>

Example for Sentence-final Particle
Et cetera

Example for Et cetera.

Ch   水果包括苹果等。

En   Fruits include apples, etc.

<table>
<thead>
<tr>
<th>HFC</th>
<th>水果(fruits)</th>
<th>等(etc.)</th>
<th>苹果(apples)</th>
<th>包括(include)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja</td>
<td>果物(は)</td>
<td>リンゴ</td>
<td>など(を)</td>
<td>含む</td>
</tr>
<tr>
<td>R-HFC</td>
<td>水果</td>
<td>苹果</td>
<td>等</td>
<td>包括</td>
</tr>
</tbody>
</table>
Experiments

- Training data: CWMT data set (News domain, 282K sentences)
- Additional training data: XINHUA parallel data sets (News domain, 593K sentences) is used to compare the results.
- Dev & Test data: CWMT data set (1K sentences)
- Word alignment: GIZA++
- Decoder: Moses
Experiments (cont.)

BLEU and RIBES scores while CWMT corpus was used for training.
BLEU and RIBES scores while CWMT ext. corpus was used for training.
Other Reordering Issues

- Serial Verb Construction
- Complementizer
- Verbal Nominalization and Nounal Verbalization
- Adverbial Modifier
- POS tagging and Parsing Errors
Example for Adverbial Modifier

Ch  严格处罚违法行为
En  Severely penalize unlawful act

Example for Serial Verb Construction

Ch  维持深化中日关系
En  Maintain and deepen the Japan-China relations

Example for Complementizer

Ch  忙完了
En  Have finished

Example for Verbal Nominalization

Ch  健全安定发展的促进
En  sound, stably and increasingly promote
POS Tagging & Parsing Errors

- POS Tagging Errors
  - 「伊朗」(イラン, Iran)
    - POS = "VV" or "JJ" ➔ POS="NR"
  - 「胡主席」(フー・チンタオ, Hu Jintao)
    - POS = "VV" ➔ POS="NR"
  - 「実施」(実施する, Implement)
    - POS = "NN" ➔ POS="VV"
POS Tagging & Parsing Errors

- Parsing Errors

Invest 2 billion US dollars

According to “TONAN NEWS”
POS Tagging & Parsing Errors

- Parsing Errors

Invest 2 billion US dollars

According to “TONAN NEWS”
Conclusion

- Basic Head Finalization reordering rules improved the Chinese-to-Japanese machine translation quality.
- However, the refined-HFC substantially achieved further improvement.
  - Due to more monotonic word alignment.
Thank you for your attention!
Suggestions & Questions
The effect frequency of each exception rule during reordering on CWMT extended corpus.
<table>
<thead>
<tr>
<th></th>
<th>CWMT</th>
<th></th>
<th>CWMT ext.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>RIBES</td>
<td>TER</td>
<td>WER</td>
</tr>
<tr>
<td>baseline</td>
<td>16.74</td>
<td>71.24</td>
<td>70.86</td>
<td>77.45</td>
</tr>
<tr>
<td>HFC</td>
<td>19.94</td>
<td>73.49</td>
<td>65.19</td>
<td>71.39</td>
</tr>
<tr>
<td>refined HFC</td>
<td><strong>20.79</strong></td>
<td><strong>75.09</strong></td>
<td><strong>64.91</strong></td>
<td><strong>70.39</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>AS</th>
<th>SP</th>
<th>ETC</th>
<th>IJ</th>
<th>PU</th>
<th>COOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>3.7790</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td>0.7946</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETC</td>
<td>1.3359</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IJ</td>
<td>0.0016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU</td>
<td>21.0191</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COOD</td>
<td>38.2641</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>