Rule-based Syntactic Preprocessing for Syntax-based Machine Translation

Nara Institute of Science and Technology, Japan
- Yuto Hatakoshi
- Graham Neubig
- Sakriani Sakti
- Tomoki Toda
- Satoshi Nakamura

25 October 2014, Doha, Qatar
Phrase-based Machine Translation (PBMT)

- Doesn’t incorporate syntactic information
- Difficulty estimating the probability of long distance reordering

Translation Model

Reordering Model

Preprocessing using syntactic information
Lexical Processing for PBMT

Change the words in the sentence

Example: Reduce errors in verb conjugation and noun case agreement

[Avramidis and Koehn, 2008]

Annotation

[we: nominative] resolved the [issue: dative] of … or [relations: dative] with Serbia

Source language (Tagged)

Target language

PBMT
Pre-ordering

Rearrange source sentence into target language word order

Preordering method
hand-crafted or trained rules

Translation Model

Reordering Model

[John hit a ball] [Xia and McCord, 2004]
Motivation

- Rule-based syntactic preprocessing is useful for PBMT
- Few attempts have been made for Syntax-based SMT
- Examine whether it also can contribute to Syntax-based SMT
Head Finalization:
A Syntactic Preprocessing Method for PBMT
Head Finalization

- Syntactic preprocessing method for \textit{English to Japanese PBMT}
- Show significant improvements through 2 steps

1. \textbf{Reordering}
   Convert English sentence into Japanese word order

2. \textbf{Lexical Processing}
   Generate more Japanese-like sentences

[Isozaki et al., 2010]
Reordering

Move head words to the end of the corresponding syntactic constituents

Before:

```
S
  NP
    VBD
    NP
      NN
      hit
    NN
      DT
      a
      NN
      ball
```

After:

```
S
  NP
    VP
      NN
      DT
      a
    NN
      NN
      ball
  VBD
```
Reordering

Move head words to the end of the corresponding syntactic constituents

Japanese word order
Lexical Processing

Generate sentences closer to Japanese

Determinar elimination / Singularization

John \(\times\) a ball hit

Pseudo-particle insertion

John \(\_wa\_\) ball \(\_wo\_\) hit

- \(\_wa\_\) : Subject particle of the main verb
- \(\_ga\_\) : Subject particle of other verb
- \(\_wo\_\) : Object particle of any verb
Syntactic Preprocessing for T2S
Tree-to-String machine translation (T2S)

[Liu et al., 2006]

- Use parsing results of the source sentence
  - Possible to generate translations that are more accurate
  - Possible to handle long distance reordering

```
John hit a ball
[ha wo utta]
```

```
S
  | NP0
  |   VP
  |     VBD
  |       NN
  |       John

x0:NP0
ha
x1:NP1
wo utta
```
Potential Effect of Preprocessing on T2S

Reordering
- Improve word alignment
- Identify good translation patterns

Lexical Processing
- Improve translation quality of words
Proposed method

Apply three methods to T2S:

1. **Reordering**
   - Convert English sentence into Japanese word order

2. **Lexical Processing**
   - Determiner elimination / Singularization / Particle insertion

3. **HF-feature**
   - Apply reordering information to T2S as soft constraints
Reordering for T2S

- Convert English sentence into Japanese word order
- Output the reordered tree

```
John NP
hit VBD
a DT
ball NN

S NP
VP

S NP
VP
```

Improve word alignment
Lexical Processing for T2S

- Determinar elimination / Singularization / Particle insertion
- Transform not strings, but trees

Improve translation performance of word
Reordering Information as Soft Constraints

- Some translation patterns do not obey head final order due to bad alignment.
- Sometimes head final order is not applicable in Japanese grammar.

- Log-linear model

\[
\hat{e} = \arg\max_e \ w^T \cdot h(f, e)
\]

- \(f\): source sentence
- \(e\): target sentence
- \(h(\cdot)\): feature function
Procedure of HF-feature addition

Source

VP

VBD
hit

NP
x0:NP

Target

x0 wo utta

Target

VP

NP
x0:NP

VBD
hit

Reordering
Procedure of HF-feature addition

Source

Target

Reordering

Non-crossing \( h_{HF}(f, e) = 1 \)

Crossing \( h_{HF}(f, e) = 0 \)
Experiment and Result
Experimental Environment

Translation Task

- English → Japanese
- NTCIR-7 (train: 3.08M, dev: 0.82k, test: 1.38k sentences)

Translation Method

- PBMT ( default settings of moses )
- T2S ( default settings of travatar )

Evaluation

- BLEU, RIBES
Translation quality

- Translation quality is improved by Lexical Processing
- Reordering is not effective for T2S

**BLEU**

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBMT</td>
<td>32.11</td>
</tr>
<tr>
<td>PBMT + Reordering + Lexical</td>
<td>37.77</td>
</tr>
<tr>
<td>T2S</td>
<td>38.94</td>
</tr>
<tr>
<td>T2S + Lexical</td>
<td>39.51</td>
</tr>
<tr>
<td>T2S + Reordering</td>
<td>38.44</td>
</tr>
<tr>
<td>T2S + Reordering + Lexical</td>
<td>39.60</td>
</tr>
</tbody>
</table>
Translation quality

- Translation quality is not improved by HF-feature
  - Reordering quality achieved by T2S was already high

<table>
<thead>
<tr>
<th>BLEU</th>
<th>T2S</th>
<th>T2S</th>
<th>T2S</th>
<th>T2S</th>
<th>T2S</th>
<th>T2S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>38.94</td>
<td>38.74</td>
<td>38.44</td>
<td>38.48</td>
<td>39.60</td>
<td>39.38</td>
</tr>
</tbody>
</table>

+ HF-feature + Reordering + Reordering + Reordering + Reordering + Lexical + Lexical + HF-feature
Conclusion and Future Work

- Applied rule-based syntactic preprocessing designed for PBMT to T2S

- Reordering +

- Lexical Processing +

- Reordering feature →

- Examine other language pairs / Apply preprocessing to F2S
Improvement by Lexical Processing

- Lexical Processing

図に示すように、電気絶縁性のハウジング97に一列に並ぶ複数の雄型コンタクト98とから構成されている。

\[
\text{s ( x0:np vp ( pn (, ("","" ) ) vp (vx (vbz ("comprises") ) x1:np ) ) )} \\
\rightarrow x0 \times x1 \text{ "と" "から" "構成" "さ" "れ" "で" "い"
}
\]

+ Lexical Processing

図に示すように、電気絶縁性のハウジング97に一列に並ぶ複数の雄型コンタクト98を有して構成される。

\[
\text{np ( np (x0:nx) va (_va2 "_va2") ) } \rightarrow x0 \text{ "を"
}
\]
Optimized weight of HF-feature

- HF-feature led to confusion in MERT optimization
- There is no consistent pattern of learning weights

<table>
<thead>
<tr>
<th>HF-feature</th>
<th>Reordering</th>
<th>Lexical</th>
<th>Weight of HF-feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.00707078</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>0.00524676</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>0.156724</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-0.121326</td>
</tr>
</tbody>
</table>
# Translation Quality

<table>
<thead>
<tr>
<th>HF-feature</th>
<th>Reordering</th>
<th>Lexical</th>
<th>PBMT BLEU</th>
<th>PBMT RIBES</th>
<th>T2S BLEU</th>
<th>T2S RIBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>32.11</td>
<td>69.06</td>
<td>38.94</td>
<td>78.48</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>33.16</td>
<td>70.19</td>
<td>39.51</td>
<td>79.47</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>37.62</td>
<td>77.56</td>
<td>38.44</td>
<td>78.48</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>37.77</td>
<td>77.71</td>
<td>39.60</td>
<td>79.26</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>38.74</td>
<td>78.33</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>39.29</td>
<td>79.23</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>38.48</td>
<td>78.44</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>39.38</td>
<td>79.21</td>
</tr>
</tbody>
</table>