Introduction

FRDC participated in all the NTCIR PatentMT tasks.
- Chinese to English
- Japanese to English
- English to Japanese

The FRDC statistical machine translation (SMT) system JIANZHEN is totally based on the hierarchical phrase-based (HPB) translation model (Fig. 1).

Method 1. Chinese Sentence Paraphrasing

Some low quality translations occur in the sentences whose syntactic components are far separated. If we could move the far-separated components closer, the translation quality could be better.

Paraphrasing is a highly mental process which is hard to be programmed for the computer. However, in a specific domain, especially for the patent documents, some sentence patterns or programmed for the computer. However, in a specific domain, especially for the patent documents, some sentence patterns or

The templates are developed by regular expression, which consist of characters, generalized variation and word segmentation results. Here is the expression of template:

$$X[m,n][+/-]\;w?$$ Operation

The left part is the condition and the right part is the operation. X denotes the generalized variation in the sentence. [m,n] presents the character number covered by X. [+/- w] means that the variation X must have or must not have some certain characters in the brace. A sample is shown in Figure 2. The actual template is shown in Figure 3.

Method 2. Handling parentheses.

Parentheses are very common in the patent corpus. Long parentheses always break the main structure of the sentence and result in translation errors.

Considering that parentheses are independent of the main content of a sentence, we just extract them out from the sentence and translate the parentheses and the main sentence separately, then combine the translations. Figure 4 shows an example.

Results

Experiment Settings (for all the tasks)
In-house Chinese word segmentation toolkit and English tokenization script. Chosen for Japanese word segmentation.

Word Alignment: GIZA++ "grow-dia-final";

Language Model: SRILM toolkit, 4-gram;

Training algorithm: MERT;

Post process: Remove unknown words;

<table>
<thead>
<tr>
<th>Subtask</th>
<th>BLEU</th>
<th>AA</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRDC_CE</td>
<td>31.46%</td>
<td>3.34</td>
<td>0.495277778</td>
</tr>
<tr>
<td>Baseline1_CE</td>
<td>30.72</td>
<td>3.29</td>
<td>0.475833333</td>
</tr>
<tr>
<td>Baseline2_CE</td>
<td>29.32</td>
<td>2.893333333</td>
<td>--</td>
</tr>
<tr>
<td>FRDC_JE</td>
<td>27.76%</td>
<td>2.516666666</td>
<td>0.448079623</td>
</tr>
<tr>
<td>Baseline1_JE</td>
<td>28.96</td>
<td>2.618666666</td>
<td>0.473974359</td>
</tr>
<tr>
<td>Baseline2_JE</td>
<td>28.61</td>
<td>2.426666666</td>
<td>0.446794872</td>
</tr>
<tr>
<td>FRDC_EJ</td>
<td>27.81</td>
<td>2.346666666</td>
<td>--</td>
</tr>
<tr>
<td>Baseline1_EJ</td>
<td>31.66</td>
<td>2.603333333</td>
<td>0.475833333</td>
</tr>
<tr>
<td>Baseline2_EJ</td>
<td>31.9</td>
<td>2.476666666</td>
<td>0.456333333</td>
</tr>
</tbody>
</table>

Table 1. Official Results.

Conclusion

This paper describes FRDC HPB SMT system for the NTCIR-9 patent machine translation subtask. We focused on the preprocessing of the training data. A regular expression based paraphrasing method was applied to simplify the structure of Chinese sentences. We also specially handled the parentheses in the sentence. Experimental results showed that our methods are effective for improving the translation quality by both human judge and BLEU score.

References

