HPB SMT of FRDC Assisted by Paraphrasing for the NTCIR-9 PatentMT

Zhongguang Zheng Naisheng Ge Yao Meng Hao Yu

Fujitsu R&D Center CO., LTD
{zhengzhg,genaisheng, mengyao, yu}@cn.fujitsu.com

◆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).

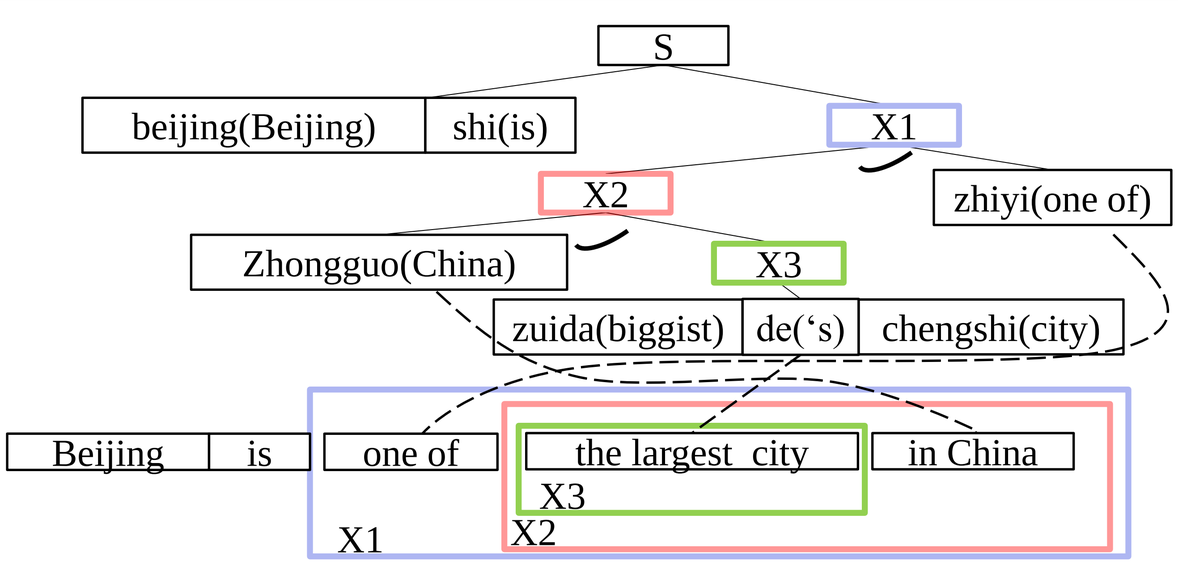

Figure 1. Hierarchical phrase-based model.

◆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 wording are highly repetitive so that it is very likely to paraphrase the sentences by some descriptive templates.

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

$$X_i[m,n]\{+/- w\}?|||Operation$$

The left part is the condition and the right part is the operation. X_i 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.

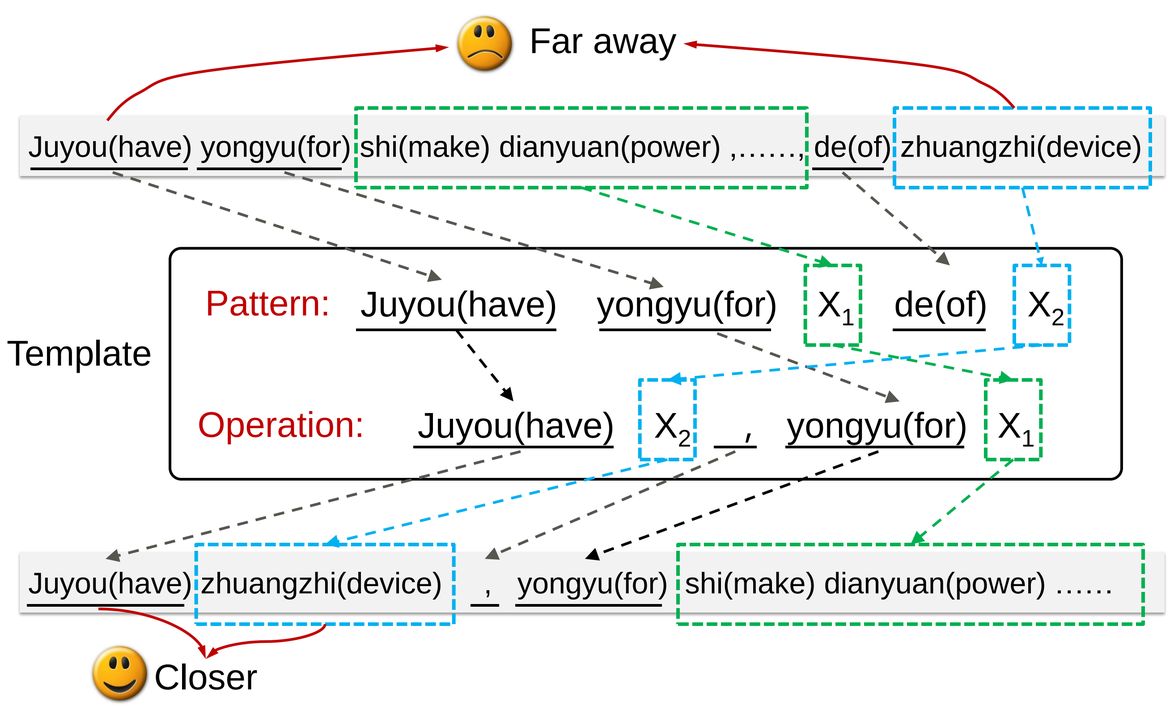

Figure 2. A sample of paraphrasing.

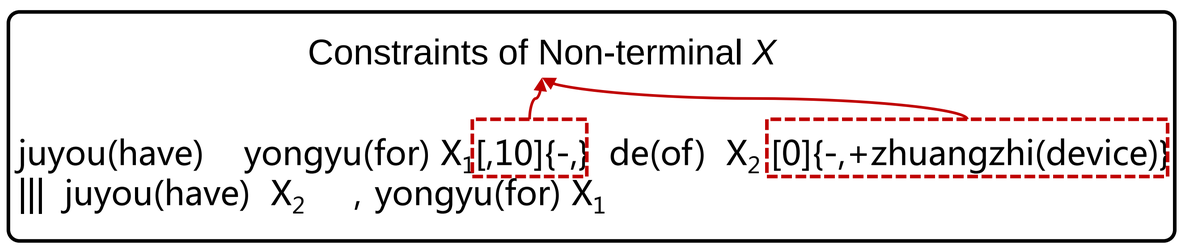

Figure 3. Actual form of template.

◆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.

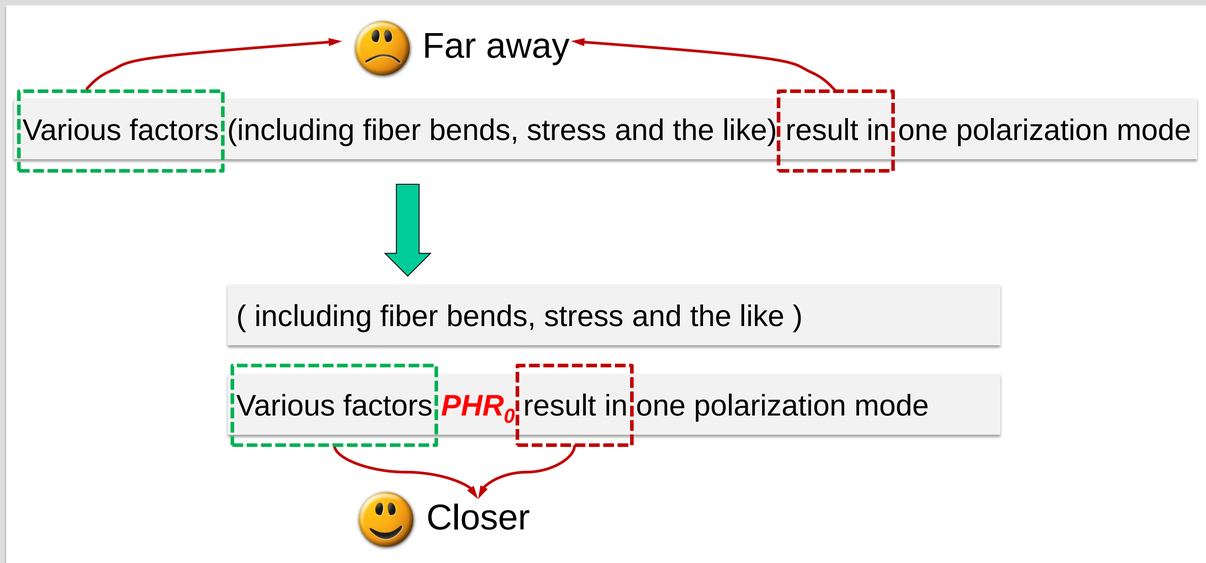

Figure 4. A sample of handling parentheses.

◆Results

Experiment Settings (for all the tasks)
In-house Chinese word segmentation toolkit and English tokenization script. Chasen for Japanese word segmentation.
Word Alignment: GIZA++, “grow-diag-final”;
Language Model: SRILM toolkit, 4-gram;
Training algorithm: MERT;
Post process: Remove unknown words;

Subtask	BLEU	AA	PC
FRDC_CE	31.46%	3.34	0.495277778
Baseline1_CE	30.72	3.29	0.475833333
Baseline2_CE	29.32	2.893333333	--
FRDC_JE	27.76%	2.516666667	0.448076923
Baseline1_JE	28.95	2.616666667	0.473974359
Baseline2_JE	28.61	2.426666667	0.446794872
FRDC_EJ	27.81	2.346666667	--
Baseline1_EJ	31.66	2.603333333	0.475833333
Baseline2_EJ	31.9	2.476666667	0.456333333

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

[1] David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*, pages 263-270. 2005.
[2] Fu Lei, Lv Yajuan and Liu Qun. A Translation Method Integrating Sentence Structure Templates with Statistical Machine Translation. *9th Chinese National Conference on Computational Linguistics, (CNCCL)*, 2007.
[3] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. 2011.

15/F, Tower A, Ocean International Center, No. 56, Dongsihuan Zhong Rd, Chaoyang District, Beijing, China
Tel. (010) 5969 1000