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.
15/F, Tower A, Ocean International Center, 56 Dongsihuan Zhong Rd.
Chaoyang District, Beijing, China
{zhengzhg, genaisheng, mengyao, yu}@cn.fujitsu.com

ABSTRACT

This paper describes the FRDC machine translation system for the NTCIR-9 PatentMT. The FRDC system JIANZHEN is a hierarchical phrase-based (HPB) translation system. We participated in all the three subtasks, i.e., Chinese to English, Japanese to English and English to Japanese. In this paper, we introduce a novel paraphrasing mechanism to handle a certain kind of Chinese sentences whose syntactic component are far separated. The paraphrasing approach based on the manual templates moves far-separated syntactic components closer so that the translation could become more acceptable. In addition, we single parentheses out for special treatment for all the three languages.

General Terms

Experimentation

Keywords

FRDC, PatentMT, HPB model, Paraphrasing, Parentheses

TeamName: [FRDC]

Subtasks/Languages: [PatentMT][Chinese-to-English][English-to-Japanese][Japanese-to-English]

External Resources Used: [Giza++][SRILM][Chasen][Manual templates]

1. INTRODUCTION

FRDC participated in all the NTCIR PatentMT tasks which include

- Chinese to English subtask.
- Japanese to English subtask.
- English to Japanese subtask.

The FRDC statistical machine translation (SMT) system JIANZHEN is totally based on the hierarchical phrase-based (HPB) [3] translation model. The HPB translation model, which employs a synchronous context-free grammar (SCFG), is one of the promising SMT approaches since it has a better ability of reordering and generation than phrase-based model. The HPB model is presented in the form

$$X \longrightarrow < \gamma, \alpha, \sim > \tag{1}$$

where X is a non-terminal, γ and α denote source and target strings, which contain both terminals and non-terminals. $\sim$ is the one-to-one correspondence between terminals and non-terminals in γ and α. The SMT system is built within a log-linear framework [7] denoted as

$$P(e|f) \propto \sum_i \lambda_i h_i(\gamma, \alpha) \tag{2}$$

where e is called English (source language) and f is called foreign language (target language). $h_i(\gamma, \alpha)$ is a feature function and λ_i is the weight of h_i.

There are many factors that may affect the translation quality, i.e., the data processing and the translation model. In the NTCIR-9 PatentMT, we focused on the preprocessing of the training data. After analyzing the translation results, we find that 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. Either parsing or paraphrasing can be used to move the components closer. However, the accuracy of current parsing tools is not satisfactory as many mistakes maybe come to the translation after parsing. Thus, we turn to paraphrasing to rewrite these sentences.

Moreover, there are a lot of parentheses in the patent text. A parenthesis always acts as an interpretation to the main subject of the sentence. However, parenthesis usually disrupts the sentence grammatic structure. For example, the main sentence of "*the distance from the central position (the position of the optical axis) increases*" is "*the distance from the central position increases*" and should be translated as an integral phrase. But the parenthesis "*(the position of the optical axis)*" makes the verb "*increases*" far from the subject "*distance*". Since such long distance structure always harms the translation quality, we use a simple but effective method to extract parentheses in order to restore the main sentence structure and make the sentence shorter. Without the interference of parentheses, the translation may become more acceptable.

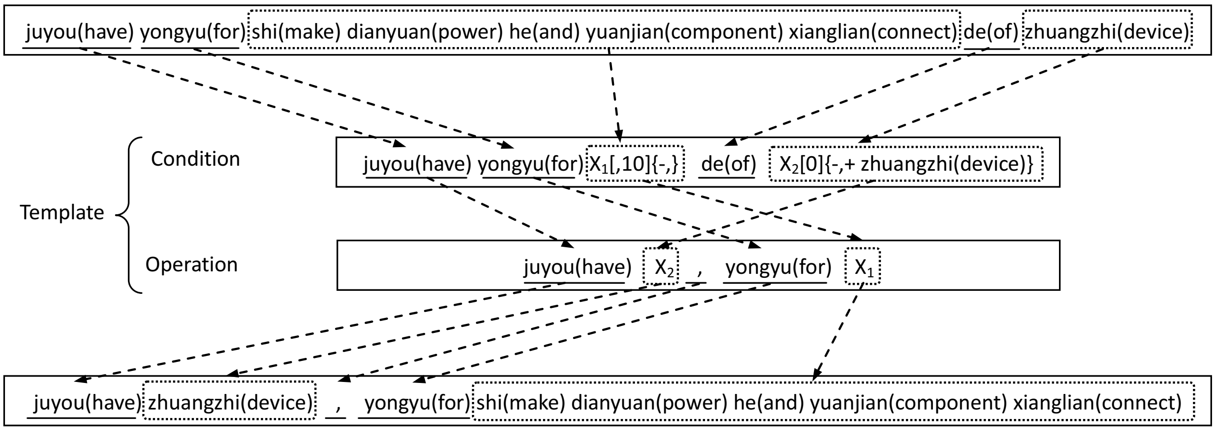
juyou(have) yongyu(for) shi(make) dianyuan(power) he(and) yuanjian(component) xianglian(connect) de(of) zhuangzhi(device)
Template
Condition
juyou(have) yongyu(for) X1[,10]{-,} de(of) X2[0]{-,+ zhuangzhi(device)}
Operation
juyou(have) X2 , yongyu(for) X1
juyou(have) zhuangzhi(device) , yongyu(for) shi(make) dianyuan(power) he(and) yuanjian(component) xianglian(connect)

Figure 1: Example of sentence rewriting.

The rest of the paper is organized as follows. Section 2 introduces the preprocessing methods. Experiment is shown in Section 3. Conclusion and Future Work is presented in Section 4.

2. PREPROCESSING METHODS

2.1 Chinese Sentence paraphrasing

As we know, 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:

$$X_i\,[m,n]\,\{+/-w\}?|||Operation \qquad (3)$$

The left part is the condition and the right part is the operation. The signs in the template mean:

- X_i: the generalized variation in the sentence. The subscript i denotes the index started from 1. X_i in the condition corresponds to the X_i in the operation.
- $[m,n]$: The *character number* covered by X_i. There are several variations of this condition.

  $[m,n]$: $m \leq$ *character number* $\leq n$

  $[m,]$: $m \leq$ *character number*

  $[,n]$: $0 \leq$ *character number* $\leq n$

  $[0]$: No limitation of *character number*.
- $\{+/-$ *character*$\}$: the variation X_i must have or must not have some certain characters in the brace.

  $\{+$*character*$\}$: X_i must have some certain characters.

  $\{-$*character*$\}$: X_i must not have certain characters.

  $\{0\}$: no limitation to the characters in the X_i.
- ?: means to choose the first match if the generalized variation X_i can match many parts of the sentence. If there is no ?, it means to choose the last match.
- *target pattern*: the paraphrasing result.

We also introduce "$$" to describe the characters in the condition part. If the character is between two $, it means the character is the result of word segmentation. Otherwise, it is just the string of character.

For example, "$*shi(is)*$$X_1$" excludes the sentences containing "*danshi(but)*" because "*danshi(but)*" and "*shi(is)*" are different results of word segmentation. However, "*shi(is)*X_1" includes the sentences containing "*danshi(but)*" because "*shi(is)*" is a substring of "*danshi(but)*".

One sentence may match more than one templates. If two templates overlap, we use the leftmost one. If the templates are nested, we just preserve the topmost one. An example of matching a Chinese sentence with a human rule is shown in Figure 1. The source sentence means "*have a device for the connection to the component of power supply*". In the sentence, structure "*juyou(have) zhuangzhi(device)*" is separated far away and makes it difficult for HPB model to translate. However, after paraphrasing the structure is changed into a form whose syntactic components are closer, which is easier to translate. Meanwhile the result could become more readable by human.

2.2 Handling the Parentheses

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

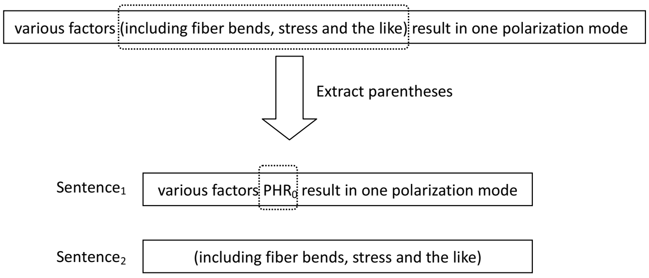
various factors (including fiber bends, stress and the like) result in one polarization mode
Extract parentheses
Sentence₁
various factors PHR₀ result in one polarization mode
Sentence₂
(including fiber bends, stress and the like)

Figure 2: Example of handling parenthesis.

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. Here we define parenthesis as the contents between parentheses, curve, square bracket, angle bracket and brace.

We use an indexed symbol "PHR_{index}" to denote the place where the parenthesis is located. Figure 2 shows an example of our method. If we do not extract the parenthesis, the result may be translated from "*factors include ...*" rather than "*factors result in ...*". After substituting the parenthesis, the source sentence is changed to the form "*factors* PHR_0 *result in ...*", which is more feasible for SMT system.

Note that if there are nested parentheses, we just extract the topmost one, and We do not handle the constituent if a bracket drops its counter part, i.e., "*Figure 1)*".

3. EXPERIMENT

3.1 Data Set

We used the corpora provided by the NTCIR-9 workshop[2, 10] in the three subtasks. Table 1 lists all the corpora we used. No external corpora were applied in our experiment.

We extracted parentheses of the training set and then trained our translation model. Language model was trained on the target portion of the training set in each subtask. For Chinese to English subtask, manual templates were obtained according to the development corpus. Those templates were employed in the final evaluation.

3.2 Primary System

Our primary system JIANZHEN is a hierarchical phrase-based translation system [3, 4]. Eight features are used in our system:

- Language model.
- Word penalty.
- Weight of normal rules.
- Weight of glue rules.
- English to foreign language translation weights $P(\gamma|\alpha)$.
- Foreign language to English translation weights $P(\alpha|\gamma)$.

Table 1: Information of our data sets.

Subtasks	Training Set	Development Set
Chinese to English	1 million	2 thousand
Japanese to English English to Japanese	3.2 million	2 thousand

Table 2: Information of our data sets.

Subtask	Corpus	Refine	Paraphrasing	parentheses
CE	Training	no	no	yes
	Dev	no	yes	yes
	Test	no	yes	yes
JE	Training	yes	no	yes
	Dev	yes	no	yes
	Test	yes	no	yes
EJ	Training	yes	no	yes
	Dev	yes	no	yes
	Test	yes	no	yes

- English to foreign language lexical weights $P_w(\gamma|\alpha)$.
- Foreign language to English lexical weights $P_w(\alpha|\gamma)$.

We used in-house toolkits for Chinese word segmentation and English tokenization. *Chasen*[1] was adopted for Japanese segmentation. GIZA++ [8] was run in both translation directions to obtain the word alignment, and the alignment result was refined by "*grow-diag-final*" method [12].

For the language model, we used the SRI Language Modeling Toolkit (SRILM) [1] to train 4-gram language models on the target portion of each training set.

We used the minimum error rate training algorithm (MERT) [6] for tuning the feature weights of the log-linear model, and adopted BLEU [9] as our evaluation metric.

3.3 Results of Dry Run

In the dry run phase, in order to obtain the test set, we divided the official development corpus into two sub corpora equally for each subtask. Besides the two preprocessing methods described before, we also adopted some basic methods to handle the segmented (tokenized) corpora, i.e., separated numerics "*0 . 6*" → "*0.6*", abbreviation with a period "*Fig . 1*" → "*Fig. 1*", escape characters "*&#x3bc;*" → "μ". We call this procedure $Refine$. After that, we processed the parentheses in all the corpora. The paraphrasing method was applied only on the Chinese development set and test set. Table 2 depicts the detailed information of our preprocessing procedure for each subtask.

We conducted several groups of experiments to verify our methods. The results are listed in Table 3.

From the results we can see that the *Refine* method is quite effective on JE and EJ tasks, because there are lots of numerics and escape characters in Japanese and English corpora.

[1] http://chasen.naist.jp

Table 3: Dry run results.

System	CE	EJ	JE
Baseline	32.60%	26.44%	26.20%
Refine	–	28.25%	27.51%
Paraphrasing	32.70%	–	–
Parentheses	32.92%	28.56%	28.09%
Paraphrasing + Parentheses	32.92%	–	–

We did not use *Refine* method to process Chinese since this procedure was implemented together with the word segmentation.

It is a little disappointing that *Paraphrasing* method produces only a slight improvement. We have overall 300 templates and 165 sentences are paraphrased in the test set. After analyzing the results, we find that the *Paraphrasing* may not contribute much to the BLEU score. Because only a few sentences are paraphrased (165/1000) and the translation errors still remain though the sentence structure is improved. Here is an actual example, a source sentence

mouxie(some) shuju(data) bingbu(not) zongshi(always) keyong(available) , qie(and) keneng(may) cunzai(be) yunxu(allow) zhezhong(this) shuju(data) lianxu(continuously) xianshi(display) de(of) shuju(data) texing(characteristic)

means

Some data is not always available, and there may be a data characteristic that allows such data to be continuously displayed

We can see that the structure "*cunzai(be) A de(of) B*" should be translated into a attributive clause "*be B that A*". However, the baseline translation is

Some data is not always available, and possible have allows the data shown in succession data characteristics

which fails to translate the structure and not understandable. Thus we use the following template to paraphrase the sentence

$$cunzai(be)X_1\,[10,]\,\{-,\}?de(of)X_2\,[0]\,\{-,\}\;|||\;cunzai(be)X_2, X_1 \tag{4}$$

Then the source sentence becomes

mouxie(some) shuju(data) bingbu(not) zongshi(always) keyong(available) , qie(and) keneng(may) cunzai(be) shuju(data) texing(characteristic), yunxu(allow) zhezhong(this) shuju(data) lianxu(continuously) xianshi(display)

The structure "*cunzai(be) A de(of) B*" is changed to "*cunzai(be) B, A*". This time the translation is

Certain data are not always available, and may have a characteristic data which allows this data continuously displayed

Table 4: Evaluation results. "AA" denotes average adequacy and "PC" denotes pairwise comparison

Subtasks	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	0.471666667
Baseline2_EJ	31.9%	2.476666667	0.456333333

which is more understandable but not improves the BLEU score due to the translation error of the noun phrase, i.e., "*shuju(data) texing(characteristic)*" is translated into "*characteristic data*" which should be "*data characteristic*".

3.4 Results of Formal Run

The system settings remained the same as in the dry run phase except that we used the integral development corpora. In the NTCIR-9 PatentMT automatic evaluation and human evaluation are both adopted as measurement metrics. Table 4 lists the official results of our system. We refer to BLEU as the automatic evaluation metric. The Average Adequacy and Pairwise Comparison are described in [11]. Baseline systems are SMT systems provided by the NTCIR-9 workshop [11].

From the results we can see that we beat the baseline systems on the Chinese-to-English subtask. However our results are not as good as the baseline on the other two subtasks. One possible reason is that the word order between Japanese and English is quite different. Sophisticated techniques should be adopted to solve the reordering problem. Without high quality language analysis, we could not get satisfactory results.

4. CONCLUSION AND FUTURE WORK

This paper describes FRDC SMT system for the NTCIR-9 patent machine translation subtask. FRDC SMT system JIANZHEN is a hierarchical phrase-based translation system. 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.

In the future work, we will improve our paraphrasing method by utilizing POS messages to develop more generalized template to cover more sentences. Japanese language analysis is also an essential work. Word order between Japanese and Chinese or English should be studied.

5. REFERENCES

[1] Andreas Stolcke. SRIM - An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pages 901-904. 2002.

[2] Bin Lu, Benjamin K. Tsou, Tao Jiang, Oi Yee Kwong and Jingbo Zhu. Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. In *Proceedings of the 1st CIPS-SIGHAN Jiont Conference on Chinese Language Processing (CLP-2010)*, 2010.

[3] 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.

[4] David Chiang. Hierarchical phrase-based translation. In *Computational Linguistics*, pages 201-228. 2007.

[5] 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.

[6] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of Computational Linguistics (ACL)*, pages 160-167. 2002.

[7] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of Computational Linguistics (ACL)*, pages 295-302. 2002.

[8] Franz Josef Och and Hermann Ney. A system comparison of various statistical alignment models. *Computational Linguistics*, pages 19-51. 2003.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhou. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, pages 311-318. 2002.

[10] Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. *MT Summit XI.*

[11] 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.

[12] Philipp Koehn, Franz Josef Och and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 40th Annual Meeting of HLT-NAACL 2003*, pages 127-133. 2003.