System Description of BJTU-NLP MT for NTCIR-10 PatentMT

Peihao Wu, Jinan Xu, Yue Yin and Yujie Zhang
School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 100044, China
{12120465, jaxu, 11120499, yjzhang}@bjtu.edu.cn

ABSTRACT
This paper presents the overview of statistical machine translation systems and example-based machine translation system that BJTU-NLP developed for the NTCIR-10 Patent Machine Translation Task (NTCIR-10 PatentMT). We used Japanese named entity in Japanese word segmentation and found a good result is obtained in EJ subtask. Although we use external chemical dictionary in our Patent SMT of Chinese to English, it does not make a better BLEU score in our experiments.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Machine Translation.*

General Terms
Design, Experimentation

Keywords
Hierarchical Phrase-based Translation Model, Example-based Machine Translation, Chemical Dictionary, Name Entity, NTCIR-10 PatentMT

Team Name
BJTU-NLP

Subtasks/Languages
Chinese to English, English to Japanese, Japanese to English

External Resources Used
Juman7.0, ICTCLAS2011, GIZA++, Moses, SRILM…

1. INTRODUCTION
This year's Patent Machine Translation task at the NTCIR-10 workshop consists of three subtasks.We participate all subtasks and submit two system results for Chinese-English subtask, and one system result for English-Japanese and Japanese-English.

In this paper, we briefly describe our system by different kinds of translation models in all three subtasks in PatentMT Tasks of NTCIR-10. Thus far, we develop hierarchical phrased-based translation model in Subtasks Chinese to English and English to Japanese, and phrased-based translation model in Subtask Japanese to English.

Extra resources was proposed improving the quantity of machine translation. In our experiments, we extract 101,258 Chinese-English parallel chemical dictionary pairs from website. However, Extra resources do not get a better BLEU score in our experiments.

The rest part of this paper is arranged as follows. In section 2 we first describe the main framework of phrased-based translation model, hierarchical phrased-based translation model, then the experimental settings and results of statistical machine translation at NTCIR-10. In section 3 we also describe example-based machine translation on Chinese to English PatentMT Tasks. Finally, we conclude our work and give the future directions in section 4.

2. Statistical Machine Translation Systems

2.1 Phrase-based Translation Model
Phrase-based translation model is distinguished by combining a set of features in a log-linear way. This model expressed the probability of a target-language word sequence (*e*) of a given source language word sequence (*f*) given by:

$$\hat{e} = \arg\max_{e} \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e,f))}{\sum_{e'} \exp(\sum_{m=1}^{M} \lambda_m h_m(e',f))} \quad (1)$$

Where $h_m(e,f)$ is the feature function, such as the translation model or the language model, λ_m is its weight, and M is the number of features. λ_m is tuned by using the Minimum Error Rate Training(MERT) algorithm based on the development set.

2.2 Hierarchical Phrase-based Translation Model
Hierarchical phrase-based translation model can be regarded as an expansion of phrase-based translation. It can extract non-continuous parts from source language, and translate them into sour language. SCFG is used to establish translation model in hierarchical phrase-based translation model, the rules of the form is as follows:

$$X \rightarrow <\gamma, \alpha, \sim> \quad (2)$$

Hierarchical phrase-based translation model uses phrase rules just like phrase-based translation model, and it also imports hierarchical rules to reorder phrases. Log-linear model is also used in hierarchical phrase-based translation model.

2.3 Experiment Settings
We use the open source toolkit Moses[1] to develop hierarchical phrase-based machine translation system and phrase-based

[1] http://www.statmt.org/moses/

translation model machine translation system. Moses is a statistical machine translation system that offers two types of translation models: phrase-based and tree-based.

On one hand, in our experiments, we only use surface as the factors of language we involved, as the following example shows:

例如， 用 具有 广谱抗 微生物 活性 的 聚 腈基 丙烯酸酯 膜 覆盖 皮肤 表面 的 不可 缝合 性 小 伤口 将 会 减弱 伤口 感染 的 可能 。

on the other hand , a cable 324 is connected to the movable plate 321 .

一方、 可動 プレート 321 に は ケーブル 324 が 接続 さ れ て いる 。

In order to improve the translation quantity of chemical name entity in Chinese to English translation, we use extra Chinese-English chemical dictionary which was extracted from websites. We use this resource when we trained the translation model and language model. We totally extracted 101,258 pairs of chemical dictionary. Here are the examples of chemical dictionary.

Table 1. Examples of Chemical Dictionary

Chinese	English
酒石酸氢钾	potassium bitartrate
2,6-二氯苯甲酸	2,6-dichlorobenzoic acid
4,5-二(羟甲基)-2-苯基-1H-咪唑	4,5-bis(hydroxymethyl)-2-phenyl-1H-imidazole

Experiments are carried out on all subtasks' sentence-aligned parallel patent data provide by NTCIR-10. English sentences are tokenized and lowercased by using `tokenizer.perl` and `lowercase.perl`, which provided by WMT2008 organizers. We segment the Chinese sentences by using ICTCLAS2011[2]. Japanese sentences are segmented by using the open source Japanese morphological analyzer Juman[3] and we use more than 30,000 name entities to improve Japanese segmentation. All sentences are encoded in UTF-8 and we convert full-width word to half-width word.

Before building the translation model, long sentences with more than 90 words are removed by using the script `clean-corpus-n.perl`. Both translation model and language model are generated from the resulting bilingual sentences pairs. The dataset were used are in table 2.

Table 2. Statistics of datasets used in training

Subtask	Datasets	#of sentences
C-E	Training	849,012
	Dev	2,000
E-J & J-E	Training	2,522,589
	Dev	2,000

The GIZA++[4] is applied to align words. Parameter of phrase alignment heuristic is "grow-diag-final-and", and in JE subtask, the parameter of reordering model is "msd-bidirectional-fe". The SRILM toolkit[6] is used to build trigram models with Knerser-Ney smoothing in phrase-based translation model system. In factored translation model system, surface language models are trigram model with Knerser-Ney smoothing.

The decoder is Moses. The BLEU4 metric is adopted to measure the translation quality. For Japanese outputs, we remove the spaces. For English outputs, detokenization is done by the script `detokenizer.perl`. To recover the case information, we used the recaser in Moses toolkit which is based on heuristic rules and HMM models.

2.4 Results and Analysis

Part of the experimental results are shown in table 3.

Table 3. BLEU score in partial subtasks

Subtasks	Translation models	Task Name	BLEU
C-E	Hierarchical phrase-based model	IE	0.2637
		ChE	0.2490
		ME	0.1576
E-J	Hierarchical phrase-based model	IE	0.3445
		ChE	0.3377
J-E	Phrase-based model	ME	0.2090

As illustrated in table 3, the BLEU scores in our subtask CE are not as expected, and the extra resource does not improve the quantity of machine translation. In my opinion, the main reason is that the chemical dictionary has some pairs which do not have data correlation with training data. When a chemical substring was used in different chemical words, it may translate into different words, which may add more noise while training the translation model. So when we use some extra resources, we should take data correlation into consideration.

Compared with NTCIR-9, we try different ways to improve the translation quantity of SMT in Chinese-English subtask, however, both of them do not improve the quantity of machine translation. At the same time, the work in English-Japanese proves efficient, and we get a relatively high BLEU score at NTCIR-10 ,and it implies that details determine success or failure at machine translation.

3. Example-based Machine Translation (EBMT) System

EBMT is proposed by Makoto Nagao [11]. Three key problems in EBMT are selection of suitable example, extraction of translation example and generation of translation from translation example [12]. In resolving the former two problems, we propose to use dependency structures on both source and target side, since syntactic information is useful in machine translation [13]. For the selection of suitable example, instead of extracting 2^{N-1}-1 subtrees in a dependency structure of a sentence with N words, we only extract such two kinds of subtrees as defined below: all subtrees in any father-its son two lays and subtrees of any father-its all descendants.

Based on the extracted subtrees of source side and word alignment obtained by GIZA++, we extract subtrees of target side for the extraction translation examples using the following method:

[2] http://ictclas.org/

[3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[4] http://code.google.com/p/giza-pp/

[5] http://www.speech.sri.com/projects/srilm

(1) Collecting target words corresponding to the words of source subtree.

(2) If a subtree is formed by the target words in the target side, pair it with the source subtree as a translation example. Otherwise, add other words in the target side into the collection to form a subtree. The added word should satisfy the following conditions: 1) the word has direct modified/modifier relation with one word of the collection; and 2) the word does not have corresponding word in the source sentence.

In translation generation, we search translation examples based on LCS (Longest Common Subsequence). The generation process contains two types of operations: 1) Delete operation: when the source subtree of the searched translation example contains the subtree of the input sentence, delete the redundant part of the translation example and keep the reminder part being a subtree; 2) Replace operation with condition: the corresponding word in the target subtree will be replaced with the translation of the word in the input subtree only if the father of the word of input subtree and the word of the source subtree are identical word, and the dependency relations between the father and the word on both input subtree and source subtree are same.

For implementation, we use Gparser [14] and Stanford parser [15] for dependency parsing. At first we extracted translation examples from the CE training data and then conducted experiment on the test data. The experimental result is shown in Table 4. The performance of the current system is still low. Many improvements on the selection of suitable example and the strategy of translation generation should be conducted.

Table 4. BLEU score of our EBMT

Subtasks	Translation models	Task Name	BLEU
C-E	EBMT based on Dependency Structure	IE	0.1076

4. CONCLUSION & FUTURE WORK

This paper describes our experiments for NTCIR-10 PatentMT, which used hierarchical phrase-based translation model in Chinese to English translation and English to Japanese translation, example-based translation model in Chinese to English translation, and phrased-based translation model in Japanese to English translation. While extra chemical dictionary adds extra alignment information to phrases, it does not get a higher BLEU score than original hierarchical phrase-based translation model. Besides, training translation model with extra chemical dictionary is time-consuming. Therefore, attention should first be given on another way to improve the quantity of machine translation.

In future work, we will continue to do research about the effect of hierarchical phrase-based model and syntax-based model, and analyze the effect of extra resources in machine translation.

5. REFERENCES

[1] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita and Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In Proceeding of the NTCIR-10 Workshop.

[2] Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, 901-904.

[3] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 263-270.

[4] David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201-228.

[5] Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, 160-167.

[6] Franz Joseph Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

[7] Marcello Federico, Nicola Bertoldi, Mauro Cettolo. 2008 IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech 2008*, 1618-1621.

[8] Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, 48-54.

[9] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, 177-180.

[10] Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. In *Technical Report TR-10-98, Center for Research in Computing Technology (Harvard University)*.

[11] Zhu Junguo, Qi Haoliang, Yang Muyun, Li Jufeng, Li Sheng. 2008. Patent SMT Based on Combined Phrases for NTCIR-7. In *Proceeding of the NTCIR-7 Workshop Meeting*, 471-474.

[12] Nago M. A Framework of a Mechanical Translation between Japanese English by Analogy Principle, In: Proceedings of the International NATO Symposium on Artificial and Human Intelligence, 1984, 173~180.

[13] Somers H. Review Article: Example-based Machine Translation. Machine Translation, 1999, 14(2):113~157.

[14] WANG Haifeng, LIU Zhanyi, and WU Hua, Semi-Structured Example Based Machine Translation. CNCCL-2007.

[15] MA Jinshan, Research on Chinese Dependency Parsing Based on Statistical Methods.

[16] Klein, Dan. And Manning, Christopher.(2003) Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.