System Description of BJTU-NLP SMT for NTCIR-9 PatentMT

Junjie Jiang, Jinan Xu, Youfang Lin and Yujie Zhang
School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 100044, China
{10120469, jaxu, yflin, yjzhang}@bjtu.edu.cn

ABSTRACT
This paper presents the overview of statistical machine translation systems that BJTU-NLP developed for the NTCIR-9 Patent Machine Translation Task (NTCIR-9 PatentMT). We compared the performance between phrase-based translation model and factored translation model in our Patent SMT of Chinese to English and English to Japanese. Factored translation model was proposed as an extended phrase-based statistical machine translation model. Many languages have shown off it to good effect. However, factored translation model didn't get a better BLEU score than phrase-based translation model in our experiments.

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

General Terms
Design, Experimentation

Keywords
phrase-based translation model, factored translation model, Moses, NTCIR-9 PatentMT

Team Name: [BJTU-NLP]

Subtasks/Languages: [Chinese to English, English to Japanese]

External Resources Used: [mecab, ICTCLAS2011, GIZA++, moses, SRILM…]

1. INTRODUCTION
In this paper, we briefly describe our system by different kinds of translation models in the Chinese to English and English to Japanese PatentMT Tasks at NTCIR-9. Thus far, we develop phrase-based translation model and factored translation model, and compare the different performance between them.

Factored translation model was proposed as an extended phrase-based translation model. Phrase-based model uses word sequences as a phrase, while factored translation model uses factor sequences. A factor is different from representation of word, such as surface form, lemma and part-of-speech etc. Kohen et al.[8] reported that factored model has better performance in syntactically complex language. However, factored translation model didn't get a better BLEU score than phrased-based model in our experiments. This result is consistent with Takahiro Oda's study [11].

The rest part of this paper is arranged as follows. Section 2 presents the main framework of phrased-based translation model and factored translation model. In section 3 we describe the experimental settings and results of C-E and E-J PatentMT Tasks at NTCIR-9. Finally, we conclude our work and give the future directions in section 4.

2. TRANSLATION MODELS
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:

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

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 Factored Translation Model
The current start-of-the-art approach to statistical machine translation, namely phrase-based model, is limited to the mapping of small text chunks without any explicit use of linguistic information like morphological, syntactic, or semantic. A word in factored translation model is not anymore only a token any more, but a vector of factors that represents different levels of annotation. In such a model, we would want to translate factors separately, and combine this information on the output side and ultimately generate the output surface words. An illustration of factored translation model is shown in figure 1.

In order to train a factored model, three steps are processed:

1. factorize training data
2. train translation model
3. train generation model

In this paper, we didn't use generation model. Thus the factor in output is only surface.

When we train a factored translation model from factorized corpus, the test data must be factorized too.

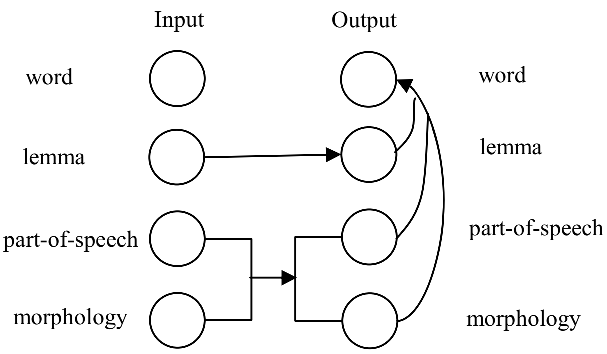

Figure 1. An illustration of factored model.

3. EXPERIMENTS

We use the open source toolkit Moses[1] to develop a phrase-based SMT system and a factored translation model SMT system. Moses is a statistical machine translation system that offers two types of translation models: phrase-based and tree-based. Besides, Moses features factored translation models, which enables the integration linguistic and other information at the word level.

In our experiments, we only use surface and part-of-speech as the factors of language we involved, as the following example:

例如|v ， |wd 用|p 具有|v 广谱抗|n 微生物|n 活性|b 的|ude1 聚腈基丙烯酸酯|n 膜覆盖|n 皮肤|n 表面|n 的|ude1 不可|v 缝合|v 性|ng 小|a 伤口|n 将|d 会|v 减弱|v 伤口|n 感染|v 的|ude1 可能|n 。|wj

on|IN the|DT other|JJ hand|NN ,|, a|DT cable|NN 324|CD is|VBZ connected|VBN to|TO the|DT movable|JJ plate|NN 321|CD .|.

一方|接続詞 、|特殊 可動|名詞 プレート|名詞 321|名詞 に|助詞 は|助詞 ケーブル|名詞 324|名詞 が|助詞 接続|名詞 さ|動詞 れて|接尾辞 いる|接尾辞 。|特殊

3.1 Experiment Settings

Experiments are carried out on the sentence-aligned Chinese-English and English-Japanese parallel patent data provide by NTCIR-9. English sentences are tokenized and lowercased by using `tokenizer.perl` and `lowercase.perl`, which provided by WMT2008 organizers. The POS tagger tool we used is Stanford POS Tagger[2]. For Chinese sentences, we segment the sentences and tag the word's POS by using ICTCLAS2011[3]. As for Japanese sentences, they are segmented and tagged POS by using the open source Japanese morphological analyzer Mecab[4].

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 1.

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

[2] http://nlp.stanford.edu/software/tagger.shtml

[3] http://ictclas.org/

[4] http://mecab.sourceforge.net

The GIZA++[5] is applied to align words. Parameter of phrase alignment heuristic is "grow-diag-final", and 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, while POS language models are trigram model with Witten-Bell 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.

Table 1. Statistics of datasets used in experiments

Subtask	Datasets	#of sentences
C-E	Training	747,754
	Dev	2,000
	Test	2,000
E-J	Training	2,522,589
	Dev	2,000
	Test	2,000

3.2 Results and Analysis

Experimental results of CE and EJ subtasks are shown in table 2.

As illustrated in table 2, factored translation model only gets a high BLEU on dev for CE subtask. In other case, phrase-based translation model has a better performance. We analyze one of the reason is that surface and part-of-speech factors are not enough for factored model. Therefore we need richer factors. The other reason may be that the accuracy of POS tagger toolkit can't achieve 100%. As Takahiro Oda's study [11] shows, surface-surface factored model, ie phrase-based translation model has the best performance among different factored translation models.

Table 2. BLEU score of using different translation models

Subtasks	Translation models	BLEU	
		Dev	Test
C-E	Phrase-based model	0.3092	**0.2808**
	Factored model	**0.3121**	0.2779
E-J	Phrase-based model	**0.2681**	**0.2705**
	Factored model	0.2556	0.2584

4. CONCLUSION & FUTURE WORK

This paper describes our experiments for NTCIR-9 PatentMT, which compared the different performance between phrase-based translation model and factored translation model in Chinese to English translation and English to Japanese translation. While factored translation model added more information to words, it didn't get a high BLEU score than phrase-based translation model. Besides, factorizing the corpus is time-consuming. Therefore, attention should first be given on phrase-based translation model.

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

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

In the future work, we will do a research about the effect of hierarchical phrase-based model and syntax-based model, and analyze the feature and advantages of these translation models.

5. ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities (2009JBM027) (2010JBZ2007) and by the RenCai Funding of Beijing Jiaotong University(2011RC034).

6. REFERENCES

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

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] Philipp Koehn and Hieu Hong. 2007. Factored Translation Models. In *Proceeding of Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 868-876.

[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] Takahiro Oda, Tomoyosi Akiba. 2008. Analyzing Effects of Factored Translation Models in English to Japanese Statistical Machine Translation. In *Proceeding of NTCIR-8 Workshop Meeting*, 411-414.

[12] 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.