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

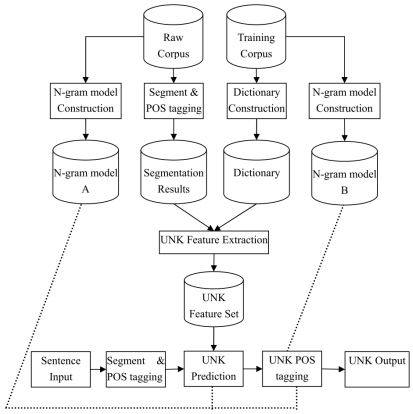
Junjie Jiang, Jinan Xu, Youfang Lin and Yujie Zhang
School of Computer and Information Technology, Beijing Jiaotong University

Introduction

BJTU-NLP participated in two PatentMT subtasks at NTCIR-9: Chinese to English and English to Japanese.
We developed phrase-based translation model and factored translation model SMT system, and compared the differences between them.
The results showed that phrase-based translation model systems gave better performance.

Chinese Unknown Word Prediction Using SVM

According to the characteristics of the patent documents, we used a SVM based method that predicted the unknown words for the result of word segmentation and tagging by ICTCLAS2011. First of all, we manually marked unknown words boundary to construct training corpus. Then, we used SVM tool to train the model. And finally, we predicted the unknown words of input sentence.



Example

Training Data:

+1 优选/v/O 的/ude1/O 亚/b/B-S 乙烯/n/B-I 基芳/n/B-I 族/ng/B-E 单体/n/O 包括/v/O
-1 优选/v/O 的/ude1/B-S 亚/b/B-I 乙烯/n/B-I 基芳/n/B-I 族/ng/B-E 单体/n/O 包括/v/O
-1 优选/v/O 的/ude1/O 亚/b/B-S 乙烯/n/B-I 基芳/n/B-I 族/ng/B-I 单体/n/B-E 包括/v/O

Test Data:

0 自/p/O C1-C10/x/O 亚/b/B-S 烷基/n/B-E ， /wd/O C6-C12/x/O
0 自/p/O C1-C10/x/B-S 亚/b/B-I 烷基/n/B-E ， /wd/O C6-C12/x/O
0 自/p/O C1-C10/x/O 亚/b/B-S 烷基/n/B-I ， /wd/B-E C6-C12/x/O

Results:

0.86 自/p/O C1-C10/x/O 亚/b/B-S 烷基/n/B-E ， /wd/O C6-C12/x/O
-0.75 自/p/O C1-C10/x/B-S 亚/b/B-I 烷基/n/B-E ， /wd/O C6-C12/x/O
-0.59 自/p/O C1-C10/x/O 亚/b/B-S 烷基/n/B-I ， /wd/B-E C6-C12/x/O

Experiments

For factored translation model system, we used 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

Preprocessing:

- English sentence: `tokenizer.perl`, `lowercase.perl`, Stanford POS Tagger.
- Chinese Sentence: ICTCLAS2011.
- Japanese sentence: Mecab.
- Before building the translation model, long sentences with more than 90 words are removed.

Training:

- The GIZA++ is applied to align words.
- Parameter of phrase alignment heuristic is “grow-diag-final”.
- Parameter of reordering model is “msd-bidirectional-fe”.
- The SRILM toolkit is used to build trigram models with Kneser-Ney smoothing.

Tuning:

- MERT

Decode:

- Moses

Post-Processing:

- Japanese output: Remove the spaces.
- English outputs: `detokenizer.perl`, recaser in Moses toolkit.

Data

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

Result & Analytics

Table 2. BLEU score of using different translation models

Subtasks	Translation models	BLEU/Adequacy	
		Dev	Test
C-E	Phrase-based model	0.3092	0.2808
	Factored model	0.3121	0.2779/3.1133
E-J	Phrase-based model	0.2681	0.2705/1.7933
	Factored model	0.2556	0.2584

As illustrated in table 2, factored translation model only gets a higher BLEU on dev for CE subtask. In other case, phrase-based translation model gives a better performance. The reasons may be:

•Factored model that uses surface and POS factors has lower phrase table size. We need richer factors.
•The accuracy of POS tagger toolkit can’t achieve 100%.

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. We reported the results that phrase-based translation model gave better performance than factored translation model.

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

About BJTU-NLP

Founded in 2010 by professor Yujie Zhang and associate professor Jinan Xu, the Natural Language Processing Research Group at Beijing Jiaotong University conducts research on algorithms that allow computers to process and understand human languages. Our work covers areas such as word segmentation, parsing, WSD, MT, ASR, IR, Sentimental Analysis (SA), etc. Currently, we focus on IR, SA and SMT.

****

School of Computer and Information Technology, Beijing Jiaotong University
No.3 ShangYuanCun, HaiDian District Beijing, Post-Code 100044 China
E-mail: jaxu@bjtu.edu.cn
Tel: +86-010-51688451
http://nlp.bjtu.edu.cn

****