ISTIC Statistical Machine Translation System for PatentMT in NTCIR-10

Yanqing He, Chongde Shi, Huilin Wang
Institute of Scientific and Technical Information of China
No 15, Fuxing Road, Haidian District,Beijing, China, 100038
{heyq,shicd,wanghl}@istic.ac.cn

ABSTRACT

This paper describes statistical machine translation system of ISTIC used in the evaluation campaign of the patent machine translation task at NTCIR-10. In this year's evaluation, we participated in patent machine translation tasks for Chinese-English, Japanese-English and English-Japanese. Here we mainly describe the overview of the system, the primary modules, the key techniques and the evaluation results.

Keywords

Machine translation; System combination; Patent machine translation.

Team Name: ISTIC

Subtasks/Languages: PatentMT from Chinese to English, from Japanese to English and from English to Japanese.

External Resources Used: No

1. INTRODUCTION

This paper describes the statistical machine translation system of ISTIC (Institute of Scientific and Technical Information of China), which is used for the evaluation campaign of patent machine translation task at NTCIR-10 [1]. We participated in patent translation tasks for Chinese-English, Japanese-English and English-Japanese. The main improvement is that we tried to implement all the tasks of patent machine translation. We use different language models to train phrase-based statistical machine translation (SMT) model: Moses decoder [2] to get multiple translation results. In some tasks system combination based on word and phrase are implemented on the multiple output results of Moses to obtain the final translation result. In other tasks we choose the best output from multiple translation results.

This paper is structured as follows: Section 2 presents the overview of ISTIC translation system. In Section 3, the experimental results of our system are reported and the details on analyses of the results are given. Section 4 gives the conclusions

2. SYSTEM OVERVIEW

Figure 1 depicts our system architecture. After the test data are preprocessed, they are passed into multiple translation systems respectively to produce an N-Best translation list, and then all the N-Best translations in the list are combined to obtain 1-Best translation. We post-process the best translation to get the final translation results. We will detail each module as follows:

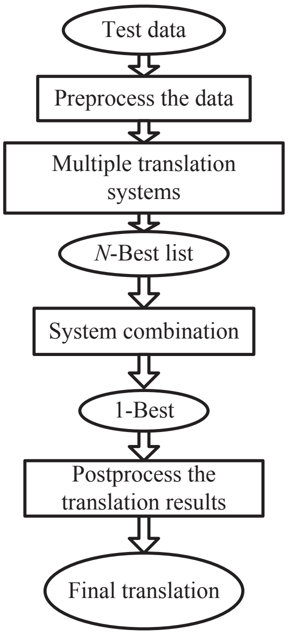

Figure 1. Our system architecture.

2.1 Preprocessing

For the Chinese part of the training data, two types of preprocessing are performed:

- Segmenting the Chinese characters into Chinese words using the free software toolkit ICTCLAS3.0[1];
- Transforming the SBC case into DBC case.

For the English part of the training data, also two types of preprocessing are performed:

- Tokenization of the English words which separates the punctuations with the English words;
- Transforming the uppercase into lowercase.

For the Japanese part of the training data, also two types of preprocessing are performed:

- Segmenting the Japanese characters into Japanese words using the free software toolkit Mecab[2];

[1] http://www.nlp.org.cn

[2] http://sourceforge.net/projects/mecab/files/

- Transforming the uppercase into lowercase

2.2 Multiple translation systems

We use phrase-based machine translation system and hierarchical phrase-based machine translation system in the open source Moses package[3]. They both are run by the default parameters. We only train 4-gram language model [3] and extract phrase pairs no more than 7 words. We filter all language files distributed based on the English size of the training data. The criterion is: the ratio of the words in the sentence falling into the vocabulary of English training data. We use different ratio to get different size of language model, which are employed to train Moses to get multiple translation results.

2.3 System combination

We implement system combination based on word and phrase to N-Best list from multiple translation systems. The overall framework of system combination is same as [4]. The only difference lies in the word alignment. We collect the N-Best list translation hypotheses from each translation results in Section 2.2, and find a hypothesis as the alignment reference. In order to guarantee a more robust combination result we choose 1-Best translation from the system with the best performance on the development set as the backbone. After getting all the translation results, the word alignment between the backbone and all the other translation hypotheses is implemented. Here an incremental HMM Alignment method [5] is used to train word alignment. When constructing the confusion network we don't use the null word to extend the network in order to reduce unreliable words in the decoding. We consider all the words in the backbone as nodes in the confusion network. Then the words in other translation hypotheses which are aligned to each node according to the word alignment are collected to obtain a word bag. Each node will have a word bag where there are one or many candidate words. We extract a phrase table from the confusion network. A phrase table is transformed from the confusion network in this way: the word index of backbone translation is looked as source phrase and each word in the word bag as the target phrase. Here our phrase table is actually a dictionary where all the source phrases only have one word and so do the target phrases. We use a phrase-based re-decoding to get the final translation which is similarly to a phrase level system combination. Here the source sentence is the backbone. A log linear model is executed to search for the target translation $f*$ with highest probability:

$$f* = \sum_{k=1}^{K} \lambda_k h_k(b, f)$$

where $h_k(b, f)$ is the feature function, λ_k is the corresponding weight. The features are listed as follows:

- Posterior probability of phrases;
- Language model;
- Distance-based phrase reordering model;
- Word penalty.

A beam searching is implemented to find the 1-Best translation for combination output. We perform the maximum BLEU training [6] on a development set to train the feature weighs.

2.4 Post-processing

The post-processing for the English output result mainly includes:

- Case restoration in English words;
- Recombination of the separated punctuations with its left closest English words;

There are no post-processing for the Japanese output results.

3. Experiments

Experiments were carried out on the Chinese-English, English-Japanese and Japanese-English translation tasks for NTCIR-10. We will describe experiments results and give our analysis on the experiment results.

3.1 Corpus

Besides the training data for all the three tasks of Patent Translation provided by NTCIR-10, there is no any other data used. Table 1-3 gives the detailed statistics of our data for each task. Here we extract the bilingual sentence pair from the training data whose sentence length is not larger than 100.

Table 1. The statistics of our corpus for Chinese-English task

Data		Sentence	Vocabulary	Average Sentence Length.
Training set	C	100,000	435,326	37.65
	E	100,000	456,168	42.34
Development set	C	2,000	7,525	36.65
	E	2,000	7,898	38.75

Table 2. The statistics of our corpus for Japanese-English task

Data		Sentence	Vocabulary	Average Sentence Length.
Training set	J	3,192,352	178,646	37.37
	E	3,192,352	280,268	33.92
Development set	J	4,000	6,915	37.74
	E	4,000	7,921	34.50

Table 3. The statistics of our corpus for English-Japanese task

Data		Sentence	Vocabulary	Average Sentence Length.
Training set	J	3,194,352	178,759	37.37
	E	3,194,352	280,420	33.92
Development set	J	2,000	5,156	38.12
	E	2,000	5,836	35.31

3.2 Experiment results

For the four tasks in Chinese-to-English translation track, we use the same training corpus and the same development corpus. So do Japanese-English and English-Japanese tracks. But different systems are trained to output final results in Chinese-to-English track. Since the corpus for Japanese-to English and English-to-Japanese are too large, only phrase-based statistical machine translation systems are implemented. Table 4 gives the system types for each task.

Table 4. The system types for each task

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

Task	System description
CE-chr	Hierarchical phrase-based machine translation using two language models.
CE-exa	Hierarchical phrase-based machine translation using two language models.
CE-int	An incremental HMM Alignment method was used to system combine the four statistical machine translation results, such as two phrase-based SMT, two hierarchical phrase-based SMT.
CE-mul	Hierarchical phrase-based machine translation using two language models.
JE-exa	Modified version of the Moses phrase-based MT system. Two language models are used.
JE-int	Modified version of the Moses phrase-based MT system. Two language models are used.
JE-chrmul	Modified version of the Moses phrase-based MT system. Two language models are used.
EJ-chr	Modified version of the Moses phrase-based MT system. Two language models are used.
EJ-int	Modified version of the Moses phrase-based MT system. Two language models are used.

Table 5 gives some experimental result released on the test set. Due to some error in JE-int submission, the score is not listed in the table.

Table 5: Results of test set

System	BLEU	NIST
CE-mul	0.1993	6.4249
JE-chrmul	0.2531	7.194
EJ-chr	0.3071	7.9265
EJ-int	0.3143	8.0971

From the evaluation results on test set, we find that there are still large gap between our result and the best result in each task of the evaluation. In order to improve our translation systems, we summarize the following shortages:

- For the pre-processing, we only use some common pre-processing strategies to do with the input part for training, tuning and testing. Patent texts have unique characteristics, such as lots of patent entities and long sentences, which need entities recognition and translation.

- Our system combination can improve the translation results from multiple translation engines, but the improvement is not significant statistically. There are much more space for us to ameliorate. In our experiment the choice of backbone is too simple. More word alignment method need to be tried, such as [7,8,9,10]We only choose the best result as our backbone. We can use MBR decoding to choose the backbone to get better performance. The features in the re-decoding are not enough to guarantee the combination performance. We will add some syntactic features in the future.

- For the post-processing, we only use some common post-processing strategies to do with the English output part for testing. Our Japanese output is not processed due to lack of experiences in Japanese language processing.

- The corpus between Japanese and English is too large for us to implement more complicated translation models, such as hierarchical phrase-based statistical machine translation or other systax-based statistical machine translation. More strategies should be employed to strengthen machine translation training on large corpus.

4. Conclusions

In summary, this paper presents our statistical machine translation system in NTCIR-10 Patent machine translation evaluation campaign. We participate in all the tasks in all the three tracks. In some tasks we combine the output results of multiple machine translation systems to get our final translation outputs. In other tasks we only run a modified version of the Moses phrase-based MT system based on two language models.

By participating in the patent translation evaluation in NTCIR-10, we have accumulated experiences and lessons for processing Japanese. There are still a lot of things for us to do to meet the need of patent machine translation. We believe our translation system will be better in the next year.

5. Acknowledgements

This research has been partially supported by ISTIC research foundation projects XK2012-6, ZD2012-3-3, YY201220.

6. References

[1] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop, *Proceedings of NTCIR-10,* 2013.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, “ Moses: Open Source Toolkit for Statistical Machine Translation”, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL),* Poster Session, pp. 177-180, Prague, Czech Republic, June 2007.

[3] Andreas Stolcke, 2002. SRILM-An extensible language modeling toolkit. In *Proceedings of International Conference on spoken language processing*, volumn 2, pages 901-904.

[4] Yanqing He, Junsheng Zhang and Huilin Wang. Combining Multiple Translations based on Words and Phrases. *Journal of the China Society for Scientific and Technical Information*, in press.

[5] Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore, Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems. In *Proceedings of EMNLP 2008.*

[6] Ashish Venugopal, Stephan Vogel. Considerations in Maximum Mutual Information and Minimum Classification Error training for Statistical Machine Translation. In *the Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*, Budapest, Hungary May 30-31, 2005.

[7] Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proc. ASRU*, pages 351–354.

[8] Antti-Veikko I.Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, and Richard Schwartz, Bonnie J.Dorr. Combining Outputs from Multiple Machine Translation Systems. In *Proceedings of NAACL HLT*, pages 228-235, Rochester,NY, April 2007.

[9] K.C. Sim, W. Byrne, M. Gales, H. Sahbi and P. Woodland. Consensus Network Decoding For Statistical Machine Translation System [A]. In: *ICASSP*, 2007.

[10] Chi-Ho Li, Xiaodong He, Yupeng Liu and Ning Xi, Incremental HMM Alignment for MT System Combination, In *Proceedings of the 4th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, page 949-957, Suntec, Singapore, 2-7 August 2009.