ISTIC Statistical Machine Translation System for Patent machine translation in NTCIR-9

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-9. In this year's evaluation, we participated in patent machine translation task for Chinese-English. 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

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-9 [1]. We participated in patent translation task for Chinese-English. We use different language models to train phrase-based statistical machine translation (SMT) model: Moses decoder [2] to get multiple translation results. Then system combination based on word and phrase is implemented on the multiple output results of Moses to obtain the final translation result.

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:

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.

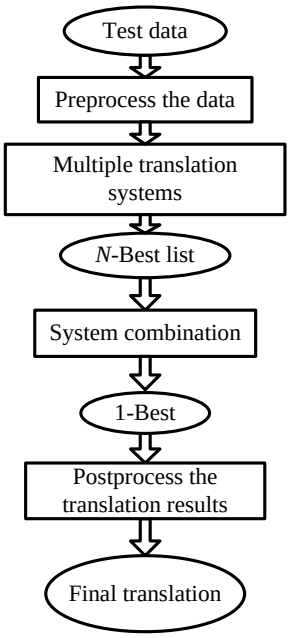

Figure 1. Our system architecture.

2.2 Multiple translation systems
We use phrase-based machine translation systems, Moses system (MOSES). The MOSES decoder provided in the open source Moses package[2] is run by the default parameters. We only train 4-gram language model and extract phrase pairs no more than 7 words.

In the evaluation we didn't use other translation system, such as hierarchical phrase-based machine translation system and syntax-

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

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

based machine translation system. 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 [3] to N-Best list from multiple translation systems. The overall framework of system combination is shown in Figure 2.

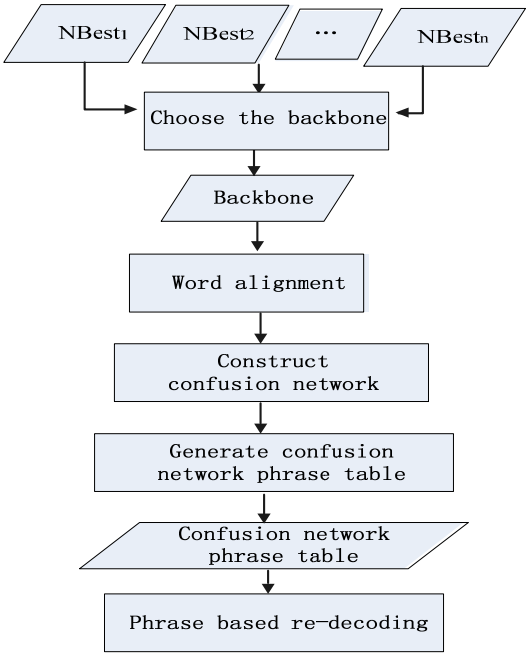

Figure 2. System combination architecture.

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 GIZA++ [4] is used to train word alignment. We pair the backbone with each other translation hypothesis to obtain the parallel sentences. It is to be noted that the parallel sentences here are in mono-language. In general the word alignment produced by GIZA++ is sensitive to the size of parallel sentences. The more the parallel sentences are, the better the word alignment is. In order to solve the problem, we extract each word in all the translation results and add parallel sentence by pairing it with itself. In this way the identical words will be aligned by GIZA++. Grow-Diag-Final heuristics are used to extend the result of GIZA++ to get the final 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. For the Chinese-to-English translation task, our source sentence is in Chinese and the backbone is in English. The probability of each target phrase is calculated according to its frequency. Each target phrase's frequency in phrase pair is calculated as its posterior probability by voting. Identical target phrase is counted repeatedly.

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 c^* with highest probability:

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

where $h_k(b, c)$ 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.

Here the language model feature is calculated as [5]. The phrase reordering feature is easily modeled as a distance-based probability: $P_{d(a_k - b_{k-1})} = |a_k - b_{k-1} - 1|$ where a_k is the starting position of the source phrase for the k^{th} target phrase and b_{k-1} is the ending position of the source phrase for the k-1^{th} target phrase. The word penalty feature is the size of the words in the sentence. 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.

Our combination strategy is different from word level system combination in 1) We extract a phrase table from the word alignment 2) Our decoding algorithm can use more features than confusion network decoding although the features used in this paper may be not more than those in confusion network decoding. Different from phrase level system combination, the source side and the target side of our phrase table is in the same language. In such way we can guarantee the maximum possibility of some target positions' candidate translations.

2.4 Post-processing

The post-processing for the output result mainly includes:

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

3. Experiments

Experiments were carried out on the Chinese-English translation task for NTCIR-9. We will describe each step in detail and give our analysis on the experiment results.

3.1 Corpus

Besides the training data for Chinese-English Patent Translation provided by NTCIR9, we didn't use any other data. Table 1 gives the detail statistics of our data. 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

Data		Sentence	Vocabulary	Average Sentence Length.
Training set	**C**	818,643	101,398	71.6
	E	818,643	329,612	40.7
Development set	**C**	2,103	5,576	38.3
	E	2,103	7,967	37.3
Test set	**C**	2,000	4,645	29.0
	E	2,000	6,273	29.1

3.2 Experiment results

For the Chinese-to-English translation track we participated in, we give the experiment results on development set shown in Table 2. Table 3 gives the experimental result on the test set. Here "PB-*" represents MOSES results which are different from the language model used. "PB-3" uses the English side of the training data to train language model. "PB-2" adds some English sentence filtered from a large-scale monolingual patent corpus in English distributed by NTCIR-9 based on the English size of the training data. "PB-1" adds more English sentences. The rate of the former is 80% and the latter is 60%. "COM" means system combination. "COM-1" combines all the three "PB-*" results. "COM-2" combines the first two "PB-*" results. Both the two kinds of system combination we uses "PB-1" as the backbone.

Table 2: Results of development set

System	BLEU	NIST
PB-1	0.3290	8.5049
PB-2	0.3244	8.4629
PB-3	0.3195	8.3664
COM-1[3]	0.3327	8.4797
COM-2	0.3283	8.4291

From the translation results on development set, we find that the more sentences the language model uses, the better the translation performance will have. But the improvement is not very significant. The systems combination improves the performance of each single translation results on both the development set and the test set. Our strategy of system combination is that we use the best translation result as the backbone and employ the other words in the translation hypotheses to improve the words of the backbone. It is a conservative method which has fewer paths than the confusion network decoding but the performance can almost be better than the best results

[3] The score of COM-1 and COM-2 on test set is different from the score released. All the score in our paper is obtained by mteval-v13a.pl on the corresponding reference set. The difference may lie in that our reference set is lowercased.

Table 3: Results of test set

System	BLEU	NIST
PB-1	0.3038	8.1055
PB-2	0.3013	8.0772
PB-3	0.2977	7.9877
COM-1	0.3087	8.0928
COM-2	0.3057	8.0773

4. Conclusions

In summary, this paper presents our statistical machine translation system in NTCIR-9 Patent machine translation evaluation campaign. Our system combines the output results of multiple machine translation systems to get our final translation outputs.

The translation result proves that the combination module is effective in the SMT system. But there are much more space for us to ameliorate. In our experiment the word alignment is too simple. We can try other methods of word alignment, such as WER[7], TER[8,9], INHMM[10], INCHMM[11] and so on. 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.

5. Acknowledge

This research has been partially supported by ISTIC research foundation projects XK2011-6,ZD2011-3-3 ,YY-201122 and YY -201126.

6. References

[1] 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, NTCIR-9, 2011.

[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] 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.

[4] Franz Josef Och, Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

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

[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] 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*.

[11] 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.