Patent SMT Based on Combined Phrases for NTCIR-7

Zhu Junguo[1], Qi Haoliang[2], Yang Muyun[1,2], Li Jufeng[1], Li Sheng[1,2]
1 School of Computer Science and Technology, Harbin Institute of Technology
2 HIT2 Joint Lab of NLP, Dept. of Computer, Heilongjiang Institute of Technology
{jgzhu;ymy;jfli}mtlab.hit.edu.cn, haoliang.qi@gmail.com, lisheng@hit.edu.cn
Harbin, Heilongjiang, 150001

Abstract

In this paper, we describe a combined phrase approach to the Statistical Machine Translation of Japanese patents into English. To resolve the segmentation errors caused by the rich OOV (out-of-vocabulary) words in the patent texts, the character based translation phrases are first employed. Then the word based translation phrases are established to utilize the dependable word level information. Finally the two translation phrases tables are linearly combined to capture both character and word level translation correspondences. Preliminary experiments on NTCIR-7 corpus indicate that the BLEU scores of the proposed method significantly out-perform the usual word based approach.

Keywords: *Word Segmentation, Phrase, Patent Translation, Statistical Machine Translation.*

1. Introduction

In this paper, we describe a combined phrase table approach to Statistical Machine Translation (SMT) for the Japanese-English Patent Translation Task. SMT is a hot issue in current computational linguistics. The patent translation technologies are of substantial contribution the circulation of patent information in the world, and the Japanese-English translation has been consistently investigated by NTCIR.

Traditionally, in the phrase-based SMT system, the word is considered to be the smallest information unit. But in such language without clear word boundary as Japanese and Chinese, there is no ready solution for automatic word identification. Usually, the Japanese corpus cannot be directly used by most SMT systems without preprocess the corpus by various Word Segmentation systems. Therefore, the impact of Word Segmentation systems on quality of SMT is non-trivial and deserves further investigation. In the Patent Translation Task, the impacts might be increased due to the rich OOV words, which are the main cause for word segmentation errors. Therefore, one of the aims of this work is how to conqueror the negative impacts of word segmentation on quality of patent SMT.

Obviously, a lot of OOV words exist in the Patent Translation Task data. Most of word segmentation systems can recognize some OOV words at the cost of certain errors. In this case, the systems maybe unpredictably produce wrong segmented words. And different segmentation systems maybe produce different words in the same corpus. Therefore, different segmentation results will contain some errors and disagreements inescapably. In this paper we propose an approach to decrease the negative impacts of word segmentation. We combine the character and word information into phrase table, compensating OOV words by translation phrase extraction upon character level. This approach is proved positive by a favorable BLEU[1] score on NTCIR-7' Patent Translation Task.

The rest part of this paper is arranged as follows. Section 2 presents a SMT approach based on combined phrase in Patent Translation Task. Section 3 describes the experimental results of Japanese-English translation of Patent Translation Task data in the Patent Translation Task of NTCIR-7. And Section 4 concludes the paper.

2. Patent SMT Based on combined phrase

2.1. Phrase Based SMT

SMT method was first proposed in the late 1990s. But it was not widely accepted and thoroughly discussed until this century. Basically speaking SMT would formulate the translation probability from a foreign sentence F into English E as the following:

$$\arg\max_{E} \Pr(E \mid F) = \arg\max_{E} \Pr(E)\Pr(F \mid E)$$

where *Pr (E)* is generally called as the language model, representing the occurrence probability of the target language sentence; *Pr (F|E)* is the translation probability evaluating how likely the target language resembles to the source language.

Many researchers have come up with a lot of methods to estimate and implement this equation in practical MT system. In contrast to previous noisy-channel model based IBM methods, the most popular approach today is characterized as phrase-based model[2]. The phrase translation model is distinguished by its log-linear form, which facilitates to integrate multiple features:

$$\Pr(E \mid F) = P_{\lambda_1^M}(E \mid F)$$

$$= \frac{\exp(\sum_{m=1}^{M} h_m f_m(F, E))}{\sum_{E'^I_1} \exp(\sum_{m=1}^{M} h_m f_m(F, E))}$$

$$\hat{\lambda}_1^M = \arg\max(\sum_{s=1}^{S} \log P_{\lambda_1^M}(E_s \mid F_s))$$

where $f_m(F, E)$ is the logarithmic value of the feature m, and h_m is its corresponding weight. The candidate target translation λ_1^M is the solution of the equation.

2.2. Combined Phrase Tables of Different Segmentation

Usually, the phrase table is extracted from the word aligned parallel corpus, which means that the sentences as in Japanese or Chinese need to be segmented for the words at first. For the Patent Translation Task, we can simply follow this procedure, which suffers from the segmentation errors and disagreements. And it is naturally that the word boundary errors would significantly impact the performance of patent SMT.

It is therefore reasonable here to wonder if these word segmentation errors could be minimized in phrase-based SMT. And in fact, this is quite feasible within the phrase modeling in SMT. It is well known that the key to good performance of phrase SMT is to have a good phrase translation table. If word segmentation is not correct, we may turn to the character segmentation because character in itself is also qualified as the unit for phrase extraction. Of course the whole procedure is still conforms to the assumption well-founded in [3], which can be fully operated with word alignment obtained from GIZA++ and Moses decoder[4]. The translation model, lexicalized word reordering model are trained using the tools provided in the open source Moses package. So we could obtain two translation models derived from word segmentation and character segmentation without any addition programming.

To make best use of the words and characters information, we adopt a translation strategy based on combined translation phrase table for patent SMT in NTCIR-7, consisting the following steps: 1) process the Japanese data by word segmentation, and train it to obtain the word-based translation phrase table W; 2) process the Japanese data by character segmentation and train it to obtain the character-based translation phrase table C. 3) combine the two phrase tables W and C.

To combine this two phrase tables, we can treat them as two features of translation models and weighted them via the MER (minimum error rate training). However the MER training is extremely time-consuming and a simple way to approximate this process is to linearly combine these two translation tables into one. Here we just chose two extreme conditions of this idea as an approximation: AWTC (add word phrase table to character phrase table) and ACTW (add character phrase table to word phrase table). AWTC is the phrase table with all of phrases in C and augmented with those phrases form W, which cannot be found in C. Reversely, we can get ACTW.

It should be noted here that although the extracted translation phrases form character may be linguistically meaningless, they are still helpful to SMT owing to titling effect. In addition, the extracted translation table derived from character segmentation may bring some advantages to SMT. First, the Japanese segmentation by character is simple and lest prone to errors and disagreements than the word segmentation. Secondly, each character as one unit provides a smaller and more reliable granularity of information, although it is too flexible in terms of alignment modeling.

3. Experiment

3.1. Data Setting

Experiments are carried out sentence-aligned Japanese-English parallel patent data provided by NTCIR-7. The corpus are used for training and development of MT systems based on parallel corpora. The data contain four Japanese-English bilingual parallel corpus (Table 1).

Table 1: Statistics of Corpora Used in Experiments

Data name	#of Sentences	average length
Training corpus	428,287	18.0 (En) 31.0(Jp by char)
Language Model corpus	1,798,571	32.1(En)

Dev corpus	915	32.5 (En)
		65.3(Jp by char)
		42.9(Jp by word)

Devtest corpus	927	31. 8(En)
		64.0(Jp by char)
		42.9(Jp by word)

Test corpus	899	31.7 (En)
		64.0(Jp by char)
		42.9(Jp by word)

All the experiment data come from NTCIR-7's PSD-1. Our training corpus is extracted from train.txt in the PSD-1, which has 1,798,571 sentences. But the training corpus is the remnant of the corpus, which has 428,287 sentences, after filtrating out long sentences. And the maximization count of the characters is limit by 40.

The Moses toolkit is selected to build the phrase-based SMT system and the tool Giza++ is applied to align English words to both Japanese words and characters. The SRILM toolkit is used to build language model[5], trained only on the texts from train.txt in PSD-1. The BLEU4 metric is adopted to measure the translation quality by using the tools named mteval-v11b.pl. A word segmentation tool developed by our lab is used to Japanese word segmentation[6], whose dictionary has 215,125 words. We should have use other segmentation tools to compare the differences, but the time is so limit that the compare experiences are in progress.

3.2. The Influence of Combining phrase table on Phrase Based SMT

We respectively use four different phrase tables in the four approaches, the sizes of whom are list in Table 2,

Table 2: The size of Phrase tables

Phrase table	Size
character segmentation	825331
word segmentation	869053
AWTC	31968347
ACTW	31968347

and weight them in the MER. Then we decode the devtext.txt and test.txt in PSD-1through Moses decoder. We list the BLEU score of four approaches in Table 3.

Table 3: BLEU score of using different phrase tables

	Devtest	Test
character segmentation	0.2285	0.2386
word segmentation	0.2430	0.2499
AWTC	0.2309	0.2455
ACTW	0.2376	0.2516

which means that the phrase table based on words takes a more important role than what derived from character segmentation.

In the table, character segmentation and word segmentation means two different ways of getting phrase tables. AWTC and ACTW means two different phrase tables which are obtained on the training parallel corpus followed the approaches in section 2. Later we will describe the differences of the two approaches.

As is illustrated in Table 3, we compared word segmentation and character segmentation approaches of Japanese segmentation. Obviously, word segmentation produces a little better translation performance than character segmentation. This is because the word is able to reduce the complexity of the word alignment, producing more correct translation equivalences than pure character alignment. But the score shows that the two have not significant discrepancies on BLEU scores. We guess this may come from form our hypothesis that word segmentation tool cannot deal with the OOV words in the patent texts and make a lot of noises in the process of segmenting. But it might as well comes from that our Japanese word segmentation module is not well developed. And in fact, the study of the later issue involving a number of well-known Japanese word segmentation toolkits is still in progress.

The method we proposed has two approaches to combining the former two phrase tables W and C. The one is AWTC that is the phrase table, that contains all of phrases in C and then if given a phrase form W, we cannot find in C, add it to AWTC. This means that if one phrase exists in the phrase table based on segmentation by characters, the phrase in the phrase table based on segmentation by words will not add in the combined phrase table. The other one is just reverse, we can get ACTW. So we obtain two new phrase tables, and introduce them into phrase-based SMT. We tested the two phrase tables using the same approach and found that the BLEU score has a significant improvement. To better understand and compare the influence of the two combining phrase tables, the results of the experiment are also displayed in Table 3.

It also can be inferred from Table 3 that we can get a much better translation performance by combined phrase table ACTW. The BLEU score reaches to 0.2515. But the SMT performance of combined phrase table AWTC is between the phrase table based on characters and the phrase table based on words. There are several possible reasons for such results. At the first, the phrase table derived word segmentation and that derived from character segmentation can affect each other. The former can provide a lower complexion and higher precision in word alignment. But the latter can provide a good coverage of OOV translation without pre-set noise in the word alignment. Note that ACTW is better than AWTC on performance of SMT, which means that the phrase table based on words takes a more important role than what derived from character segmentation.

It should not be neglected that the results we submitted are produced by the system based on character segmentation alone. And that is because whole work has not been completed when the submission of Patent Translation Task is due. In fact, we do believe we will get a much better performance than current results (ID is 'C' as mentioned in [11] with single RUN) if the translation table are finished with proper combination.

4. Conclusion and Further Work

This paper describes the efforts on Japanese patent SMT into English by combining character information to conquer OOV influence during translation modeling. Through a contrastive experiment, we discover that the word segmentation cannot make much improvement in patent SMT performance than the character segmentation. The reason of this case is assumed as that word segmentation produce too many errors and disagreements in segmenting the patent Japanese corpus with rich OOVs. And we also formulate the method of phrase table combination to enhance the performance of phrase-based SMT.

The results of experiments on NTCIR-7 indicate that the phrase come from ACTW can provides a better precise, while the phrase come from the phrase table based on characters can provides additional translation phrase coverage as an counter-effect of word segmentation errors in OOV issues.

In the further, we will investigate the impact the weight of the phrase derived different approaches on the performance of the patent SMT.

Acknowledgment

This work is supported by Natural Science Foundation China (grant No. 60773066 & 60736044).

References

[1] K. Papineni, S. Roukos, T. Ward, and W. Zhu, BLEU: A method for automatic evaluation of machine translation. *In Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL),* pp. 311–318,2002.

[2] P. Koehn, F. Och, D. Marcu. Statistical phrase-based translation. *In Proceedings of the HLT-NAAC,* pp.127–133, 2003

[3] Y. Xue. Research on Some Key Aspects of Statistical Machine Translation. *PhD thesis of Harbin Institute of Technology,* 19~37. 2006.

[4] F. Och and H. Ney. A systematic comparison of various statistical alignment mode*ls. Computat. Linguist,* pp. 19–51. 2003.

[5] A. Stolcke. SRILM -- an extensible language modeling toolkit. *Proceeding of International Conference on Spoken Language Processing,* 2002.

[6] Wang Jing. Japanese Morphological Analysis and Its Application in CLIR. *Master Thesis, Harbin institute of Technology*, 2008.

[7] P. Brown, J. Cocke, S. Pietra, V. Pietra, F. Jelinek, J. Lafferty, R. Mercer and P. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics,* 1990.

[8] P. Koehn, F. Och, D. Marcu. Statistical phrase-based translation. *In Proceedings of the HLT-NAACL* .pp. 127–133., 2003

[9] M. PAUL. Overview of the IWSLT 2006 Evaluation Campaign. *In Proceedings of the IWSLT ,* pp. 1–15, 2006.

[10] F. Och. Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, 2003.

[11] G. Foster and R. Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the 2ndWorkshop on Statistical Machine,* pp. 128-135, 2007.

[12] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access,* 2008.A.B. Smith, C.D. Jones, and E.F. Roberts. Article Title. *Journal*, Publisher, Tokyo, January 1999