A Phrase Combination Approach to Patent SMT

Junguo Zhu[1] Muyun Yang[1] Tiejun Zhao[1] Sheng Li[1] Haoliang Qi[2]

[1]School of Computer Science and Technology, Harbin Institute of Technology
{jgzhu; ymy; tjzhao; lish}@mtlab.hit.edu.cn
[2]Dept. of Computer, Heilongjiang Institute of Technology
haoliang.qi@gmail.com

Abstract

This paper presents a phrase combination approach to patent SMT (Statistical Machine Translation) for Japanese to English. To minimize the segmentation problems caused by the rich OOV (out-of-vocabulary) words in the patent texts, the character based translation phrases are first introduced to avoid the segmentation errors in translation modeling. Then the word based translation phrases, which are established to utilize the dependent word level information, are combined with character translation table by linearly integrating their probability. Our experiments on NTCIR corpus indicate that the proposed method significantly out-performed the originally word based approach.

Keywords: statistical machine translation, patent, phrase combination, word segmentation

1. Introduction

With the booming of patents at home and abroad recently, the patent translation technology becomes an important issue for the circulation of patent information in the world. Plenty of patent documents are demanded to be automatically translated between different languages. SMT, as the most promising tool in the computational linguistics research, is naturally adopted as the solution to patent translation. And its Japanese-English translation has been consistently researched by NTCIR. From NTCIR-3 to NTCIR-6, the technology of searching and retrieval of large-scale patent data publicly available are carried in both English and Japanese. Finally, the patent translation is decided as an independent task newly added into the patent information processing at NTCIR-7[1].

Traditionally, most researches of SMT are based on word level because the word is usually considered to be the smallest operable language unit. However, word segmentation is a difficult task for languages like Japanese and Chinese, in which word boundaries are not orthographically marked. However, few papers with detailed analysis of the effects of the word segmentation on SMT have been published except [2], which points out that recognition errors of OOV words clearly decreases the translation performance. In the patent translation, the impact of word segmentation on translation performance is non-trivial because the OOV words are prevailing in the texts, which, in turn, causes increasing errors in the text after word segmentation.

This paper focus on the STM for patent translation, investigating how to resolve the OOV words in the patents and conqueror the negative impacts of word segmentation on quality of patent SMT. In this paper we propose an approach to patent SMT by a phrase combination approach. We first establish the translation

phrases at the character level to decrease the negative impacts of word segmentation. Then we integrate the character and word information by linearly combine the two translation tables. This approach is proved positive by a favorable BLEU and NIST [3] score on NTCIR-7 data.

The rest of paper is arranged as following. Section 2 presents a phrase combination approach to patent SMT. Section 3 offers the experimental results of Japanese-English translation of patent translation data in the subtask of NTCIR-7. And the conclusion is given in section 4.

2. A Phrase Combination Approach to Patent SMT

2.1. Phrase Based SMT

SMT as a research issue was first proposed in early 1990s[4], and met its blooming era in the beginning decade of 21st century. Most current statistical translation models treat the aligned sentences in the corpus as sequences of words and punctuations[5,6,7]. Then the translation probability for translating a foreign sentence F into English E is formulated as the following:

$$\begin{aligned} &\arg\max_{E} \Pr(E \mid F) \\ &= \arg\max_{E} \Pr(E) \bullet \Pr(F \mid E) \end{aligned} \quad (1)$$

where *Pr (E)* is the occurrence probability of the target language which allows for language model; *Pr (F|E)* is probability of translation from target language to source language which allows for translation model.

Phrase-based statistical machine translation systems are usually modeled through a log-linear framework[8], which is generally described as:

$$\begin{aligned} &\Pr(E \mid F) = P_{\lambda_M^1}(E \mid F) \\ &= \frac{\exp(\sum_{m=1}^{M} h_m f_m(F,E))}{\sum_{E'_1{}^{I}} \exp(\sum_{m=1}^{M} h_m f_m(F,E))} \end{aligned} \quad (2)$$

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

where $f_m(F, E)$ is the logarithmic value of the m-*th* feature, and h_m is the weight of the m-*th* feature. The candidate target sentence $\hat{\lambda}_1^M$ is the solution of the equation.

2.2. Phrase Combination Integrating Character and Word Information

Usually, the phrase table in statistical translation model is established on the word aligned parallel corpus. Therefore, the sentences of language without word boundary, such as the Japanese as well as Chinese, need to be first segmented into the word sequences. For the patent translation task, we can simply follow this classic procedure. However, this approach suffers from the word segmentation noises. And the noises will take a significant impact on performance of patent SMT.

It is natural here to wonder if one can minimizing the word segmentation noises for phrase-based SMT. It is well known that the key to good translation performance is having a good phrase translation table, and phrases in SMT never assumes to be either semantically or grammatically sufficient unit. In this sense, both the word and the character are qualified for translation phrase extraction.

Hence, we can process Japanese of the parallel training data by two approaches, word segmentation and character segmentation which means each character is assumed as later processing unit. Then the phrase-based models can be easily processed current SMT toolkit like Moses after word alignment via GIZA++. The translation model, lexicalized word reor-

dering 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.

Although the extracted translation phrases based on characters may be linguistically meaningless, they are still helpful to SMT owing to the tiling effect[8]. In another aspect, the character based translation table may have some additional advantages. First, the Japanese segmentation by characters is so simple that it is less prone to errors than Japanese word segmentation. Second, each character as one unit provides a more reliable and flexible granular of information. But, of course, the character brings more spaces compared to the word for alignment, which is an additional burden for this pre-phase of translation table extraction.

To make best use of the words and characters information, a translation strategy based on combined translation phrase table for SMT is formulated as: 1) process the Japanese data by word segmentation and build the word-based translation phrase table W; 2) process the Japanese data by character segmentation and obtain the character-based translation phrase table C. 3) combine the two phrase tables W and C by Equation as follows:

$$PH = \sum_{i=1}^{N} ph_i \bullet \alpha_i \qquad \sum_{i=1}^{N} \alpha_i = 1 \tag{3}$$

where N is the number of phrase tables, and N=2 in our experiments. ph_i represents feature in i-th phrase table. represents weight of i-th phrase table. PH represents feature in combined phrase table. The following features are linear integrated: phrase translation probability p(e|f), phrase inverse probability p(f|e), lexical probability lex(e|f, a) and lexical inverse probability lex(f|e, a). We use linear search method to find the best weight for the two phrases. The step is set to 0.1. The method will take a long time to train in MER (minimum error rate) to find the best weight for the two.Do not add any headings or footers to your document.

There are two extreme approximations to the combination of the translation tables. One is to augment the character based phrase table C with those form word based phrase W. And the other is to add character based phrases into word based translation table. Although these two approaches are fast to implement, they suffer from the defects of inconsistency in probability space.

3. Experiment

3.1. Data Setting

NTCIR-7 provides sentence-aligned Japanese-English parallel patent data PSD-1 which can be used for training and development of MT systems based on parallel corpora. PSD-1 contains four Japanese-English bilingual parallel corpus (Table 1).

Table 1: Statistics of Used Corpora

Data name	# of Sents	average length
Training corpus	428, 287	18.0 (En) 31.0(Jp by char) 21.6(Jp by word)
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)

The training corpus and corpus of language model come from the same corpus which has 1,798,571 sentences. But the training corpus is selected by filtering out long sentences with a pre-defined 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 of word segmentation and character segmentation. The SRILM toolkit is used to build language model[9]. The BLEU4 and NIST5 are adopted to measure the translation quality. A word segmentation tool developed by our lab is used to Japanese word segmentation[10].

3.2. All papers must be sent in Word and PDF format.

We compared various phrase combinations, and re-weighted them in the MER training, then decode the two test corpus in PSD-1 through Moses decoder. The BLEU score of them are listed in Table 2.

Table 2: Comparison of Various Combinations

	DEVTEST		TEST	
	BLEU%	NIST	BLEU%	NIST
Char	22.9	6.704	23.9	6.854
Word	24.3	6.946	24.9	7.045
Word-into-Char	23.1	6.925	24.6	7.115
Char-into-Word	23.8	7.021	**25.2**	**7.207**
w9c1	24.0	7.013	25.1	7.180
w8c2	24.0	7.007	25.1	7.163
w7c3	23.8	7.005	24.9	7.190
w6c4	24.2	7.024	25.2	7.180
w5c5	24.1	7.001	25.0	7.161
w4c6	24.3	7.032	**25.4**	7.197
w3c7	**24.4**	**7.041**	25.2	**7.202**
w2c8	23.9	6.974	25.1	7.146
w1c9	24.3	6.911	25.0	7.050

In the table, the “Char” and “Word” respectively represent character based phrase and word based phrase. “Word-into-Char” and “Char-into-Word” respectively represent the former mentioned extreme approximation of phrase combinations. And such string of “w9c1” represents a combination with word and character based phrases weight by 9:1.

As is illustrated in Table 2, we compared word segmentation and character segmentation approaches of Japanese segmentation. Obviously, word segmentation has a little better on SMT performance than character segmentation. This is because the word is indeed able to reduce the complexity of the word alignment. So it can get a better result of word alignment than result of character alignment. But the score shows that the two have not significant discrepancies on BLEU scores. This is indicates that character information alone is a good source for translation modeling, and noises by segmenting OOV word do hurt translation quality.

In the respect of weight, the best weight of character is 0.6-0.7, and the best weight of word is 0.3-0.4. So, it presents that character based phrase can describe the patent text better and include much more information of OOV words than word based phrase.

It also can be inferred from Table 2 that we can get a much better translation performance by combined phrase table via Char-into-Word. The BLEU score reaches to 25.2%, and NIST score reaches to 7.207. But the SMT performance of Word-into-Character table is somewhat inferior. There are several possible reasons for such results. At the first, the phrase table derived word segmentation and the phrase table 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 high recognition rate in OOV word. That can

reduce the noise in the word alignment. What is more, because the phrase come from the phrase table based on words takes a more important role, the phrase come from the phrase table derived from character segmentation makes a subsidiary function. So, Word-into-Character is better than Word-into-Character on performance of SMT.

At the same time, we used the phrase combinations with linear integration. The results are hardly significant difference. But, the intuitive phrase combination is earlier.

4. Conclusion

This paper focuses on patent phrase-based SMT of from Japanese to English. Through a brief analysis, we discover that the word segmentation cannot take much higher improvement on patent SMT performance than the character segmentation. The reason of this case is that word segmentation produce too many noises in segmenting the patent Japanese corpus with rich OOVs. And we also discover that the combined phrase table is formulated to enhance the performance of phrase-base SMT.

The results of the experiments indicate that the phrase combination Char-into-Word can provides a better precise, a easier approach than the others, while the character based phrase can decrease the noises which are produced by word segmentation.

5. Acknowledgment

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

6. References

[1] F. Atsushi, U. Takehito, Y. Mikio. Definition of Patent Translation Task at NTCIR-7: Formal Run. *http://if-lab.slis.tsukuba.ac.jp/fujii/ ntc7patmt*, 2008.

[2] R. Zhang, K. Yasuda, and E. Sumita. Chinese Word Segmentation and Statistical Machine Translation. *ACM Transactions on Speech and Language Processing*, 2008,1.

[3] 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.Use "**References**" style here.

[4] 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.

[5] P. Brown, S. Pietra, V. Pietra, and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguis-tics*, pp. 263–311, 1993.

[6] S. Vogel, H. Ney, and C. Tillmann. HMM-based word alignment in statis-tical translation. *In Proceedings of COLING*, pp. 836–841, 1996.

[7] Y. Deng and W. Byrne. 2005. HMM word and phrase alignment for statistical machine translation. *In Proceedings of HLT-EMNLP*, pp. 169–176, 2005.

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

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

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