An Improved Statistical Machine Translation Method for United Chinese-Japanese Word Segmentation

Xiaowei Wang[1] , Jinke Wang[2, *]

[1]Department of Foreign Languages, Harbin University of Science and Technology, Rongcheng, China

[2]Software Engineering Department, Harbin University of Science and Technology, Rongcheng, China

[*] jkwang@hitwh.edu.cn

Keywords. machine translation; segmentation granularity; Kanji-Hanzi; Chinese-Japanese;

Abstract. As Chinese and Japanese word segmentation is processed with different tagging system and semantic performance, the granularity of word segmentation results should be readjusted to improve the performance of Statistical Machine Translation (SMT). This paper proposes an approach to adjust the word segmentation granularity for improving the performance of SMT, which combines Hanzi-Kanji comparison table and Japanese-Chinese dictionary. Experimental results express that the proposed method could adjust the granularity between Chinese and Japanese effectively and improve the performance of SMT.

Introduction

Since the different kinds of languages' vocabulary, grammar and Semantics are mostly non – isomorphic, then it's difficult to achieve one relationship between words [1-4].Therefore, the effect of word segmentation granularity for Chinese and Japanese bilingual statistical machine, to be further studied. Wang et al.'s [5] experiments show that the result of fine word segmentation granularity, which can improve the performance of Statistical Machine Translation (SMT). Ma et al. [6] propose the adaptive method of training corpus, using the Align Construction Trusted word lattice to adjust the Chinese side, in order to enhance the ability to adapt to the field of segmentation. Dyer [7] and Zhang [3] et al. base on Multi Strategy Chinese-English word, to optimize it for statistical machine translation decoding process. Chu et al. [8] use the Chinese-Japanese characters to correspond information, to tune the result through the Japanese side at Chinese side.

This paper proposes a strategy, which can improve the bilingual word segmentation adjustment of Chinese and Japanese statistical machine translation system performance. The first section of this article discusses and elaborates the dealing method of Chinese-Japanese characters' control construct tables and dictionary; The second section describes to use Chinese-Japanese characters chart and dictionary for the adjustment strategy of word segmentation granularity, and analysis the differences from bilingual Chinese-Japanese language; The final section introduces this paper's experiment method, experiment's result and analysis.

Character Chart Construction and Processing dictionary

Construction of Chinese and Japanese Chart. The correspondence of Japanese characters and Chinese characters is very complex. Chu et al. [8] use Open Source Resources to construct Kanji, Traditional Chinese characters, Simplified Chinese characters table. 1) Character pattern changes dictionary. In this paper, taking variants of Japanese kanji font change, if there is a link between the two characters through variants, then the two characters can be transformed into each other. 2) Chinese-Japanese kanji dictionary. This paper use Kanconvit 2 in Chinese and Japanese kanji conversion table as a Chinese-Japanese kanji dictionary. The dictionary contains a total of 1,159 words table variants of different characters on the information. 3) Traditional and simplified

Chinese dictionary. This paper use Chinese Encoding Converter 3 in traditional and simplified Chinese traditional and simplified Chinese translation table as a dictionary.

Japanese-Chinese Dictionary. In the EDR dictionary, the same semantic value does not correspond. Therefore, this article uses two steps to integrate the dictionary:1) using practical Chinese and Japanese Kanji lookup tables to Japanese kanji characters into Chinese.2) If the presence of the same word in two rows dictionary information, it believes that the two lines are all words in the dictionary synonyms and the combined data of the two lines. Through the above two steps, the final control of Japanese and Chinese dictionaries is obtained.

Chinese and Japanese bilingual word granularity adjustment

Bilingual granularity difference extraction. The process of extraction mainly includes the following two aspects: 1) The same word pairs are extracted from the word list: if a word through a Chinese-Japanese character table for Chinese character conversion, and the result is exactly the same as the sequence of consecutive words at the other end, then the word table information for the word pair is called the same. 2) Extracting the same word pairs for dictionary information: If word table information is different, basing on the dictionary information, the phrase is extracted as a word, and the other language end is a word sequence, and the word pairs which exist in the dictionary.

Chinese Fine Grain Analysis. The following two sections will analyze the granularity of bilingual word segmentation from Chinese fine-grained analysis, and Japanese fine-grained analysis. Chinese-side words are divided into fine-grained reasons mainly : The Japanese proper nouns appearing in Chinese can't be correctly segmented. Mainly including the Japanese named in the specific entity, That is, names, place names, organization names.

Fine Grain Analysis of Japanese. Japanese-side words are divided into fine-grained reasons mainly as follows: 1) Numerals, time words. Chinese word segmentation will be the number of words and related follow-up words are combined, the Japanese side is handled separately. 2) Chinese proper nouns. Including the Chinese in the proper nouns such as the name "Ding Meiyuan", "one year old" and other Japanese cannot be correctly cut.

Chinese-Japanese Bilingual Particle Size Adjustment. In this paper, only the extracted word pairs are considered, and there is a word at one end. For word pairs with the same dictionary information, we take different approaches. If the word is the same for the dictionary information, any of the ends of the word pair are combined into one word processing. we can use the following two ways to deal with. 1) A word in a word is segmented into a sequence of words in accordance with another verbal sequence. 2) The word segmentation results for one end of the word sequence, combined into one word for processing.

Experiment and analysis

Experimental data and tools. All the experiments are carried out using Moses translation model training and decoding, the use of GIZA++ as an alignment tool, Srilm build language model. The Chinese-Japanese language model uses the 5-gram model; Moses uses grow-diag-final-and to optimize the alignment results. BLEU and NIST were used as the test results.

Bilingual Granularity Fusion Experiment. In order to verify the effect of bilingual granularity fusion and statistical machine when the bilingual particle size is different, we use the method described in section III D to extract the word pairs with different bilingual granularity, and process the training corpus as follows: 1) The use of word segmentation tool for the baseline results of the word (baseline); 2) Bilingual word segmentation of different particle pairs, the Chinese word sequence merged into the word (cn-mix); 3) In the word pairs with different particle sizes, the Japanese word sequences are combined into words (ja-mix); 4) Bilingual word segmentation of different particle pairs, the bilingual term sequence merged into the word (both-mix).

In this paper we extracts 23,274 sentence pairs with the granularity adjustment, and 80,000 sentence pairs are randomly selected from the remaining pairs to increase the ratio of the granularity

adjustment corpus to all the corpus, and then carry out a set of experiment in Chinese-Japanese .The results of the four groups of experiments are shown in Table I.

TABLE I. DIFFERENT CHINESE GRANULARITY AND DATA SIZE OF CHINESE-JAPANESE STATISTICAL MACHINE TRANSLATION PERFORMANCE

	Corpus scale		baselin	n-mix	ja-mix	both-mix	cn-split	ja-split	both-split
Chinese-Japanese	282,476	BLEU	0.1673	0.1666	0.1705	0.1704	0.155	0.1602	0.1593
		NIST	4.4573	4.4307	4.5042	4.4599	4.3605	4.4148	4.4093
	103,274	BLEU	0.1461	0.1474	0.1468	0.1609	0.1562	0.1644	0.1571
		NIST	4.1968	4.1732	4.1968	4.349	4.2885	4.4135	4.3068
Japanese-Chinese	282,476	BLEU	0.1413	0.1381	0.1391	0.1415	0.141	0.139	0.1362
		NIST	4.2623	4.1999	4.2405	4.2614	4.2726	4.1775	4.1588
	103,274	BLEU	0.1299	0.1305	0.1299	0.131	0.1288	0.1256	0.1239
		NIST	3.973	3.9423	3.9771	3.9557	3.9685	3.8977	3.8058

Experiment result and discussion. Through the experimental results in Section III B we can get the following conclusions: 1)Through adjusting the granularity of bilingual word segmentation, it can improve the performance of Chinese-Japanese bilingual statistical machine translation system; 2)Not all granularity adjustments can improve the performance of statistical machine translation systems. In this paper, we propose a method to measure the difference of granularity between pairs of parallel sentences in Bilingual Linguistics.

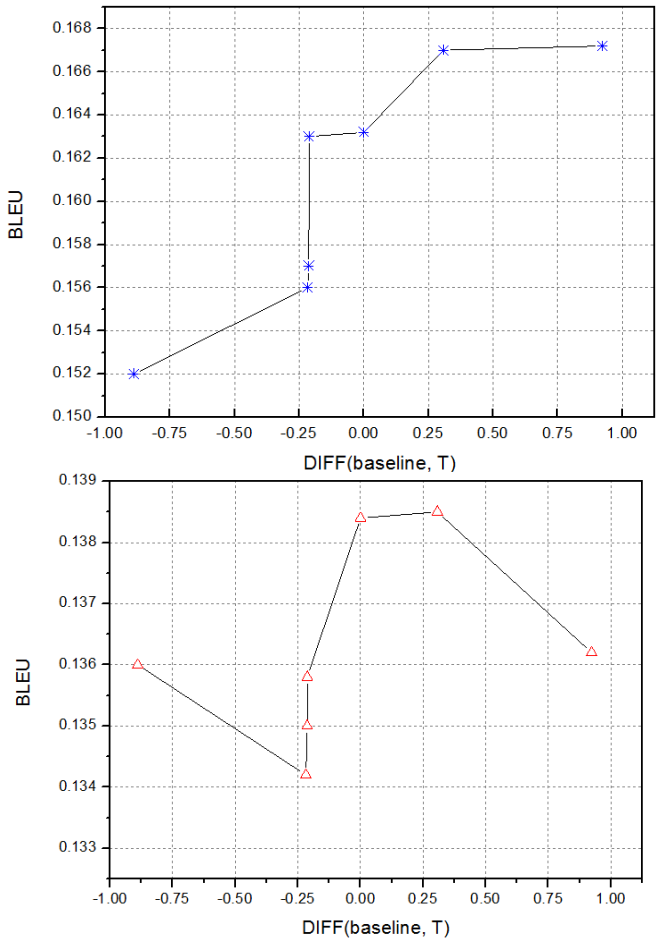

Figure 1. The effect of the relative granularity different for BLEU in Chinese and Japnese SMT

The relationship between the BLEU value and the DIFF value in statistical machine translation is shown in Fig. 1, where the horizontal axis is the machine translation performance evaluation index BLEU with vertical axis T. Figure 1 shows that there is a positive correlation between the

performance of Chinese-Japanese statistical machine translation and the relative granularity difference under the experimental conditions in this paper, that is, the smaller the relative granularity of the training corpus word, the smaller the statistical machine translation The better the performance of the system. The rest of the results in Fig. 1 show that there is also the same nature as the Chinese-Japanese statistical machine translation in Japanese-Chinese statistical machine translation. According to the results of Fig. 1, it is deduced that there is a positive correlation between the granularity difference and the performance of statistical machine translation when the absolute granularity difference of bilingual corpus is within a certain range.

Conclusion

In this paper, we propose a method to adjust the granularity of bilingual word segmentation by using the Chinese-Japanese character table and dictionary information. The experimental results show that the method proposed in this paper can effectively adjust the granularity of bilingual word segmentation and improve the performance of Chinese-Japanese bilingual statistical machine translation system.

Acknowledgements

This work was supported by Philosophy and Social Sciences Research Program of Heilongjiang Province (No.16WWC02), Youth Fund of Harbin Institute of Technology Rongcheng College (No.2014YF018) and Educational and Scientific Research of Higher Education Institute of Heilongjiang Province (No.16Q080).

References

[1] Chu C, Nakazawa T, Kawahara D, et al. Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation[C]//EAMT 2012, Proceedings of the 16th Annual Conference of the European Association for Machine Translation. Trento. 2012: 35-42.

[2] Chang P C, Galley M, Manning C D. Optimizing Chinese word segmentation for machine translation performance[C]//Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008: 224-232.

[3] Zhang R, Yasuda K, Sumita E. Improved statistical machine translation by multiple Chinese word segmentation[C]//Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008: 216-223.

[4] Xu J, Zens R, Ney H. Do we need Chinese word segmentation for statistical machine translation[C]//Proceedings of the Third SIGHAN Workshop on Chinese Language Learning. 2004: 122-128.

[5] Wang Y, Kazama J, Tsuruoka Y, et al. Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data[C]//Proceedings of 5th International Joint Conference on Natural Language Processing. 2011:309-317.

[6] Ma Y, Way A. Bilingually motivated domain-adapted word segmentation for statistical machine translation[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 549-557.

[7] Dyer C, Muresan S, Resnik P. Generalizing word lattice translation[R]. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2008.

[8] Chu C, Nakazawa T, Kurohashi S. Japanese-chinese phrase alignment using common chinese characters information[C]//Proceedings of MT Summit. 2011, 13: 475-482.