System Description: Dependency-based Pre-ordering for Japanese-Chinese Machine Translation

Jingsheng Cai Yujie Zhang Hua Shan Jinan Xu
School of Computer and Information Technology, Beijing Jiaotong University
{12120397, yjzhang, 13120422, jaxu}@bjtu.edu.cn

Abstract

This paper describes the Beijing Jiaotong University Japanese-Chinese machine translation system which participated in the 1st Workshop on Asian Translation (WAT 2014). We propose a pre-ordering approach based on dependency parsing for Japanese-Chinese statistical machine translation (SMT). Our system achieves a BLEU of 24.12 and a RIBES of 79.48 on the Japanese-Chinese translation task in the official evaluation.

1 Introduction

Difference in word order between source language and target language will cause troubles in word alignment and further affect the quality of statistical machine translation (SMT). Therefore, it becomes an issue in SMT, especially in the language pairs where there are great difference in word order between source language and target language.

Syntax-based pre-ordering has demonstrated effectiveness in previous research. These kinds of approaches first parse the sentences in the side of source language. Then pre-ordering rules are applied to the created parse trees, in order to obtain source language sentences which have similar word order with target language sentences. Finally, the reordered source language sentences are used in the SMT system, not only in training, but also in tuning and translating. In other words, all of the sentences in the training set, the development set and the test set should be reordered while applying these kinds of approaches.

Since parsing is required in pre-ordering approaches, two popular parsing are often considered, i.e. constituent parsing and dependency parsing. Constituent parsing has been employed in pre-ordering for the translation of English-French (Xia and McCord, 2004), Germany-English (Collins et al., 2005), Chinese-English (Wang et al., 2007), etc. and shown its effectiveness. In recent years, more and more pre-ordering research used dependency parsing for the translation of Arabic-English (Habash, 2007), English-SOV languages (Xu et al., 2009) and Chinese-English (Cai et al., 2014), because the accuracy of dependency parsing is greatly improved.

This paper introduces a Japanese-Chinese SMT system which employs a pre-ordering approach based on dependency parsing. Since Japanese is a language with a fairly free word order, dependency parsing can describe the relation between two Japanese cases in a sentence better than constituent parsing. Therefore, we adopt dependency parsing for pre-ordering. Experimental results show that our approach can improve the BLEU score on the test set by 0.18, compared with the baseline system (without pre-ordering).

Section 2 describes some issues in Japanese-Chinese translation and proposes our dependency-based pre-ordering approach. Section 3 reports on our experiment results on a Japanese-Chinese phrase-based SMT (PBSMT) system. Section 4 concludes this paper.

2 Dependency-based Pre-ordering Approach

Japanese is a kind of SOV language and Chinese is a kind of SVO language. Therefore, great difference exists in their word order. Moreover, both of them have very free word order, i.e. a sentence may have several expressions by only changing its word order. This fact causes troubles while translating Japanese into Chinese.

This section describes some issues in the translation of Japanese-Chinese and then proposes our dependency-based pre-ordering approach to tackle these problems.

Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 39–43,
Tokyo, Japan, 4th October 2014.

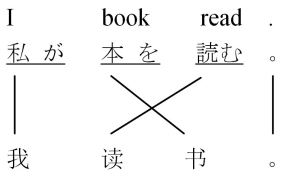


Figure 1: An example Japanese sentence and its Chinese counterpart, along with their word alignment.

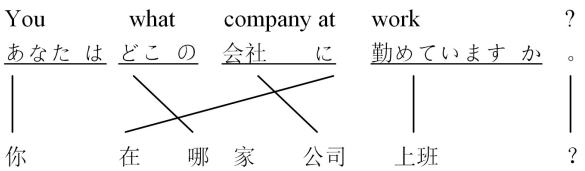


Figure 2: An example with a cross word alignment, owing to the existence of a Japanese particle “に”.

2.1 Issues in Japanese-Chinese Translation

As is known to all, Japanese is a kind of SOV language, in which the verb (V) occurs after the object (O), and Chinese is a kind of SVO language, in which the verb (V) occurs before the object (O). This is the most common difference between Japanese and Chinese. Figure 1 shows an example of a Japanese sentence and its corresponding Chinese sentence with their word alignment. These sentences means “I read book.”. As described in above, the verb “read” occurs after the object “book” in Japanese but before the object in Chinese. This kind of phenomena is often observed in Japanese-Chinese parallel corpus, which unavoidably brings about wrong results word alignment and further affect the quality of machine translation.

Japanese is a kind of agglutinative language but Chinese is not. A Japanese sentence consists of phrases which are usually ended with a particle, indicating the case of the phrase. Although all Japanese particles are not content words, the particles are often incorrectly aligned with Chinese content words in conventional word alignment. To relieve this problem, we consider alignment to connect the particles with Chinese prepositions. Figure 2 shows such an example, of which the Japanese sentence consisting of four phrases - “wa” case, “no” case, “ni” case and “bunmatu” case. It means “What company do you work for?” in English. The tokens within one phrase are displayed by one underline. The Japanese particle “に” is aligned to the Chinese preposition “在”, which is a cross word alignment within the “ni” case (“どこの会社に”). Erroneous word alignments often appear in such kinds of sentences.

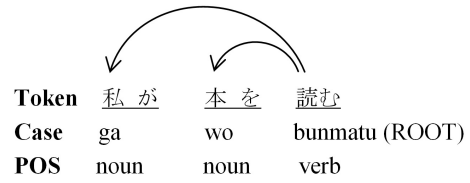


Figure 3: Dependency parse tree for the Japanese sentence in Figure 1.

2.2 Dependency-based Pre-ordering

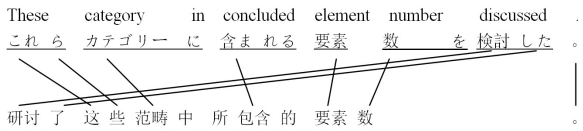
After investigating the issues in Japanese-Chinese translation, we attempt to tackle them by introducing a pre-ordering approach into the PBSMT system. Considering that current Japanese dependency parsing is of high accuracy and provides case information, we think these conditions are suitable for both global and local pre-ordering operations. We therefore propose a dependency-based pre-ordering approach.

Figure 3 shows the dependency parse tree of the Japanese sentence in Figure 1. There are three cases in this sentence. Both of the “ga” case and the “wo” case depend on the last phrase called as “bunmatu” case, i.e. the ROOT. According to the dependency tree, we can easily move the “wo” case after the “bunmatu” to obtain an SVO word order sentence. After operating this pre-ordering, the Japanese sentence can be translated into “我/I 读/read 书/book 。/.” in Chinese, whose word order is the same as its Chinese counterpart.

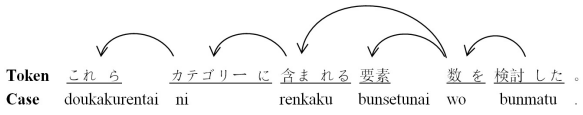
We investigate the phenomena of word order exchanging, such as “wo” case and “ni” case, on the ASPEC Japanese-Chinese paper excerpt corpus[1]. We take statistics of the various cases and observe the samples of the cases with high frequency. Based on the observed results, we build pre-ordering rules, which consist of global pre-ordering and local pre-ordering.

Inspired by the work of Xu et al. (2009), we obtain the global pre-ordering rules in the following way. We focus on verbs and classify the cases that directly depend on verbs into several groups. We then set sequence numbers to the groups by considering the word orders of their Chinese translations in a Chinese sentence, as we observe from

[1]http://lotus.kuee.kyoto-u.ac.jp/ASPEC/



(a) A Japanese sentence and its Chinese counterpart, along with their word alingment.



(b) Dependency parse tree for the Japanese sentence.

Figure 4: An example of pre-ordering operation on a Japanese sentence.

the corpus. Table 1 lists the groups and their sequence numbers.

We first define the subtree of a case as the phrase ended with the particle and those phrases that directly or indirectly depend on this phrase. Then we design pre-ordering operation. For a given Japanese dependency tree, if its root node contains a verb, conduct pre-ordering as the following steps.

Step 1. Global Pre-ordering: Change the order of the subtrees of the cases directly depend on the "bunmatu" case and the "bunmatu" case itself, according to the sequence number of the groups that they belong to, as listed in Table 1. The cases of Group 1 should occur before those of Group 2, and so forth. Note that if any two cases belong to the same group, we keep their relative order unchanged.

Step 2. Local Pre-ordering: For "wo" case, "de" case, "ni" case, "he" case, "kara" case and "made" case, move the particles to the front of the whole subtree.

Figure 4 gives an example of conducting pre-ordering operation. Figure 4 (a) displays the Japanese sentence to be reordered, the corresponding Chinese translation and the word alignment between them. The English meaning is "(We) discussed the number of the elements that are contained in these categories.". The dependency tree of the Japanese sentence is shown in Figure 4 (b). First, global pre-ordering is conducted. The subtree of the "wo" case, i.e. all of the tokens occurring before "を", is moved to the position after the "bunmatu" case. Then, local pre-ordering is conducted. The particle "を" and "に" are moved to the front of the subtree of their cases, respectively. The reordered Japanese sentence can be translated into "研讨了/discussed 这些/these 范畴/category 包含/conclude 要素/element 数/number 。/." in Chinese, whose word order is more similar to its Chinese counterpart.

Group	Case
1	ga/が wa/は
2	renyou/conjunction
3	de/で ni/に he/へ kara/から made/まで yori/より to/と
4	bunmatu/文末
5	wo/を

Table 1: Groups of the cases and their sequence numbers.

3 Experiments

Section 2 describes our dependency-based pre-ordering approach for the translation of Japanese-Chinese. This section reports on our experiments and evaluation results.

We use MOSES PBSMT system (Koehn et al., 2007) in our experiments and use BLEU scores (Papineni et al., 2002) and RIBES score (Isozaki et al., 2010) for evaluation.

The data sets are from the ASPEC Japanese-Chinese paper excerpt corpus. The training data contains 672,315 sentences, the development data contains 2,090 sentences and the test data contains 2,107 sentences.

We use a bilingual-based approach proposed in Su et al. (2013) for Chinese word segmentation, in which n-gram feature from the raw Chinese part and word alignment from the parallel corpus are introduced to augment the conventional model based on annotation data.

We employ the KNP parser[2] for Japanese dependency parsing. The KNP parser can create dependency tree for a Japanese sentence and provide the part-of-speech (POS) for each word, both of which are used in our dependency-based pre-ordering approach.

After Japanese dependency trees are obtained, we conduct pre-ordering on the training set, the development set and the test set by applying the pre-ordering rules described in Section 2.2. The reordered data sets are then used in training, tuning and test in the Japanese-Chinese PBSMT system.

On the other hand, data sets without pre-

[2]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

System	BLEU on Dev	BLEU on Tst	RIBES on Tst
Baseline	21.55	20.71	77.34
Pre-ordering	21.71	20.89	77.67

Table 2: Comparisons of the pre-ordering system and the baseline system.

Category	Positive	Unchanged	Negative
Global	115	49	36
Local	63	108	29

Table 3: Human evaluation for 200 reordered sentences, respectively.

ordering are used in our baseline system for comparison.

Table 2 presents the evaluation results of the pre-ordering system and the baseline system, including the BLEU scores on the development set, the BLEU scores and the RIBES scores on the test set. The results show that the system employing our dependency-based pre-ordering approach outperforms the baseline system slightly, both on the development set and the test set. Note that these results are different with those in the official evaluation because we should keep the consistency in Chinese word segmentation, while the official evaluation uses other word segmenters (kytea and Stanford).

To access the accuracy of our pre-ordering approach, we also conduct human evaluation for the reordered sentences. 200 global reordered sentences and 200 local reordered sentences are extracted from the training data, along with their corresponding Chinese sentences. For global pre-ordering, we align the Japanese cases with the Chinese phrases. For local pre-ordering, we pick the reordered Japanese cases and align the words in them with the Chinese words. For both categories of pre-ordering, we count the number of cross alignment and compare it with the one of the original sentence pairs. Table 3 shows the results of our human evaluation. Here, "Positive" means that the number of the sentences in which the cross alignment decreases after applying the pre-ordering, "Negative" means that the number of the sentences in which the cross alignment increases after applying the pre-ordering, and "Unchanged" means that the number of the sentences in which the cross alignment stays the same after applying the pre-ordering. As shown in Table 3, global pre-ordering helps improve the word alignment in 57.5% sentences and 24.5% stays the same. Local pre-ordering helps improve the word alignment in 31.5% sentences and 54% stays the same. The results access the accuracy and effectiveness of our approach. Note that the "Unchanged" local pre-ordering is 108, which is over half of all. The reason for this is that in fact many Japanese particles have no alignment relation with any Chinese word, pre-ordering of these particles will not change the number of cross alignment at all.

4 Conclusion

This paper describes the Beijing Jiaotong University Japanese-Chinese PBSMT system participated in WAT 2014. The system employs a dependency-based pre-ordering approach. The approach conducts two categories of pre-ordering: 1) global pre-ordering reorders the cases directly depending on the verb of "bunmatu" case; 2) local pre-ordering moves the Japanese particles to the front of the case they belong to. Experimental results show that our pre-ordering approach improves the BLEU score by 0.18 and the RIBES score by 0.34 on the test set of the ASPEC Japanese-Chinese corpus. The accuracy and effectiveness of this approach are also accessed by human evaluation.

Note also that some cases which are not listed in Table 1 may also be parsed to depend on a verb of "bunmatu" case. We do not operate any pre-ordering in such situation currently because their frequency is very low and it is difficult to summarize their reordering phenomena.

Our research builds up a framework for Japanese-Chinese pre-ordering. However, the improvement for the PBSMT system is not as great as our expectation. As shown in Table 3, some negative results are brought into the word alignment after pre-ordering. This implies that the there

are some problems in current pre-ordering rules. More features from dependency tree may be introduced as constrains to the pre-ordering approach in the future.

Acknowledgments

This work is supported by Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China and International Science & Technology Cooperation Program of China (Grant No. 2014DFA11350).

References

Jingsheng Cai, Masao Utiyama, Eiichiro, Sumita, and Yujie Zhang. 2014. Dependency-based Pre-ordering for Chinese-English Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 155-160.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531-540.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of the 11th Machine Translation Summit (MT-Summit)*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944-952.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177-180.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318.

Chen Su, Yujie Zhang, Zhen Guo and Jinan Xu. 2013. *Exploring Multiple Chinese Word Segmentation Results Based on Linear Model*. Natural Language Processing and Chinese Computing, 400:5059.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 737-745.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508-514.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz J. Och. 2009. Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of HLT-NAACL*, pages 245-253.