Word Order Does NOT Differ Significantly Between Chinese and Japanese

Chenchen Ding[•†‡] **Masao Utiyama**[○†] **Eiichiro Sumita**[○†] **Mikio Yamamoto**[•‡]
[†]Multilingual Translation Laboratory,
National Institute of Information and Communications Technology
3-5 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0289, Japan
[‡]Department of Computer Science, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
•{tei@mibel., myama@}cs.tsukuba.ac.jp
○{mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

We propose a pre-reordering approach for Japanese-to-Chinese statistical machine translation (SMT). The approach uses dependency structure and manually designed reordering rules to arrange morphemes of Japanese sentences into Chinese-like word order, before a baseline phrase-based (PB) SMT system applied. Experimental results on the ASPEC-JC data show that the improvement of the proposed pre-reordering approach is slight on BLEU and mediocre on RIBES, compared with the organizer's baseline PB SMT system. The approach also shows improvement in human evaluation. We observe the word order does not differ much in the two languages, though Japanese is a subject-object-verb (SOV) language and Chinese is an SVO language.

1 Introduction

The state-of-the-art techniques of statistical machine translation (SMT) (Koehn et al., 2003; Koehn et al., 2007) demonstrate good performance on translation of languages with relatively similar word orders (Koehn, 2005). However, word reordering is a problematic issue for language pairs with significantly different word orders, such as the translation between a subject-verb-object (SVO) language and a subject-object-verb (SOV) language (Isozaki et al., 2012).

To resolve the word reordering problem in SMT, a line of research handles the word reordering as a separate pre-process, which is referred as *pre-reordering*. In pre-reordering, the word order on source-side is arranged into the target-side word order, before a standard SMT system is applied, on both training and decoding phases.

An effective rule-based approach, *head finalization* has been proposed for English-to-Japanese translation (Isozaki et al., 2012). The approach takes advantage of the *head final* property of Japanese on the target-side. It designs a head finalization rule to move the head word based on the parsing result by a head-driven phrase structure grammar (HPSG) parser. Generally, the idea can be applied to other SVO-to-Japanese translation tasks, such as its application in Chinese-to-Japanese translation (Dan et al., 2012).

However, the head finalization cannot be applied on the reverse translation task, i.e. Japanese-to-SVO translation, which becomes a more difficult task. Specifically, Japanese-to-English translation has been studied and several rule-based pre-reordering approaches have been proposed, taking advantage of the characters of Japanese and English (Komachi et al., 2006; Katz-Brown and Collins, 2008; Sudoh et al., 2011; Hoshino et al., 2013; Ding et al., 2014). A comparison of these approaches is reported in Ding et al. (2014).

Because both Chinese and English are SVO languages, to transfer approaches of Japanese-to-English to Japanese-to-Chinese translation is a natural idea. Based on the framework of Ding et al. (2014), we propose dependency-based pre-reordering rules for Japanese-to-Chinese translation in this paper. Contrary to our expectations, from the experimental results on ASPEC-JC data, we discover that the rule-based pre-reordering cannot improve the Japanese-to-Chinese significantly as in the case of Japanese-to-English translation. In a further investigation, we find a basic reason is that the word order is actually similar between Chinese and Japanese. Chinese and English, though both of them are SVO languages, have very different properties.

Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 77–82,
Tokyo, Japan, 4th October 2014.

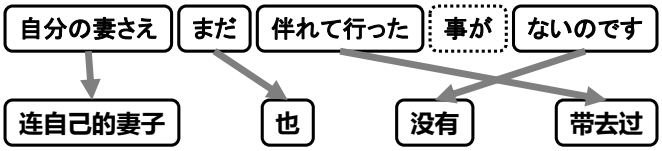

Figure 1: *Japanese tends to use formal noun (in dash-line box) to "wrap" a long clause.*

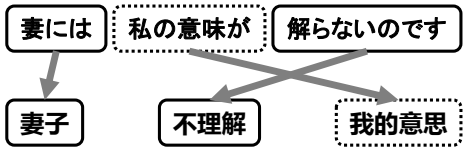

Figure 2: *A nominative phrase in Japanese turns to be an accusative phrase in Chinese.*

2 Character of Chinese and Japanese

In Ding et al. (2014), they proposed three kinds of verb, noun, and copula rules to arrange the order of Japanese chunks[1], with further two rules of morphemes to achieve a more correct pre-reordering. However, their system can not directly applied on Japanese-to-Chinese. Generally, Chinese shares much more similar characters with Japanese than English does, as listed in follows.

- Chinese is head-final in noun phrases, just as Japanese.
- Chinese has no clear boundary between verbs and prepositions. The coverb and serial verb constructions are common in Chinese, which somewhat like Japanese.

Although it seems that the pre-reordering for Japanese-to-Chinese may be easier than the case of Japanese-to-English, there are factors turning the case more complex, as listed in follows.

- Chinese has a strong tendency to avoid long attributes before nouns, while long attributes are acceptable in Japanese, especially for the *formal nouns* in Japanese (Fig. 1).
- Between Chinese and Japanese, the transitive verbs and intransitive verbs are not in accordance in some cases. Essentially, case frame of verbs are not marched well between the two languages (Fig. 2).

According to the mentioned issues, the pre-reordering cannot be conducted by only seeing specific Japanese functional morphemes, such as various case-markers, which are often used in many Japanese-to-English pre-reordering approaches. We design three new inter-chunk rules and modified the intra/extra-rules based on Ding et al. (2014).

[1] i.e. *bunsetsu*

3 Proposed Approach

We use three inter-chunk rules, classified by the corresponding head morpheme in a head chunk. They are *formal-noun rule*, *stative-verb rule* and *dynamic-verb rule*.

Formal-noun rule

Head-initialization is conducted. We only apply the rule to chunks with the formal noun *koto* as its head morpheme.

Stative-verb rule

The modifier chunk with nominative case-marker *ga* is moved after the head chunk. We only apply the rule to chunks with a verb head among *aru*, *iru*, and *dekiru* as its head morpheme.

Dynamic-verb rule

The following modifier chunks are moved after the head chunk.

- with an accusative case-marker *wo*
- with a quotation particle *to*
- with a formal noun as a head morpheme

We apply the rule to chunks with a verb except the three verbs used in the stative-verb rule, as its head morpheme.

We also use extra/intra-chunk rules to arrange certain functional morphemes.

Intra-chunk rule

Auxiliary verbs, except *ta* and copula morphemes, are moved before head morpheme within a chunk.

Extra-chunk rule

Functional morphemes attached to nouns, except genitive case-marker *no* and parallel markers, are moved before the governing range of the chunk.

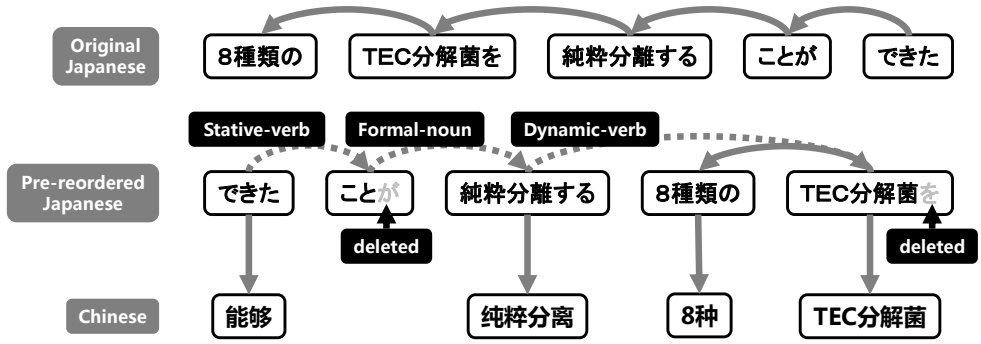

Figure 3: *Example of three inter-chunk rules: formal-noun rule, stative-verb rule and dynamic-verb rule. The alignment between pre-reordered Japanese and Chinese is manually aligned. By the three rules, Morphemes of Japanese are arranged in a Chinese-like order.*

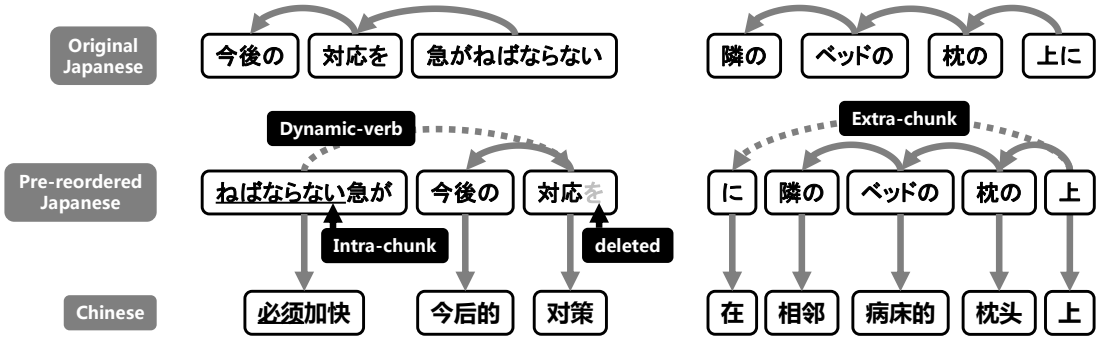

Figure 4: *Examples of Intra- and Extra-chunk rules. The left example shows an intra-chunk move within a verb chunk, leading to a more correct reordering of Japanese functional morphemes, where the underlined parts are corresponding translation. The right example shows an extra-chunk move. A Japanese post-positioned case-marker is arranged to the left-most position of the range it governs. This move makes the Japanese postposition phrase have an identical order to the Chinese preposition phrase.*

We further delete several Japanese functional morphemes which do not have exact Chinese translations. They are as follows.

- topic marker *wa*
- nominative case-marker *ga*
- accusative case-marker *wo*
- conjunctive particle *te*

We illustrate examples of our pre-reordering approach in Figs. 3 and 4.

4 Experiment

We tested our approach on the ASPEC-JC data (Nakazawa et al., 2014). For the source side Japanese sentences, we used MeCab (IPA dictionary)[2] for morpheme analysis, CaboCha[3] (Kudo and Matsumoto, 2002) for chunking and dependency parsing. We used the Stanford Chinese Word Segmenter[4] (Tseng et al., 2005) with the *Chinese Penn Treebank standard* (CTB) to seg-

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[3] https://code.google.com/p/cabocha/
[4] http://www-nlp.stanford.edu/software/segmenter.shtml

ment each Chinese sentence. We used the phrase-based (PB) translation system in Moses[5] (Koehn et al., 2007) as a baseline SMT system. Word alignment was automatically generated by GIZA++[6] (Och and Ney, 2003) with the default setting of Moses, and symmetrized by the *grow-diag-final-and* heuristics (Koehn et al., 2003). In phrase extraction, the *max-phrase-length* was 7 with *GoodTuring* option in scoring. The language model used in decoding is an interpolated modified Kneser-Ney discounted 5-gram model, trained on the English side of the training corpus by SRILM[7] (Stolcke, 2002). In decoding, the *distortion-limit* was 9. The MERT (Och, 2003) was used to tune the feature weights on the development set and the translation performance was evaluated on the test set with the tuned weights. We used identical decoding settings on development and test sets.

Our approach reached a test set BLEU of 28.18 with CTB segmentation in the final evaluation, which had a slight improvement compared with the 28.01 of the organizer's PB SMT baseline. As to the reordering measure RIBES, our approach reached a score of 0.8087, which had a mediocre improvement compared with the organizer's 0.7926. In the human evaluation, our approach also had an improvement of 6.5 percent according to the organizer's evaluation score.

5 Discussion

In a further investigation, we used the Kendall's τ (Isozaki et al., 2012) on training data as an automatic measure, to investigate the difference in word order between Chinese and Japanese and the effect of our pre-reordering approach.

We discovered that the baseline Japanese-to-Chinese word alignment already had an average τ of 0.847 and our approach only improved the score to a slightly higher one of 0.865, which explained why our pre-reordering approach cannot lead to a significant improvement. In Isozaki et al. (2012), their pre-reordering approach (head finalization) improves the average τ from 0.43 to 0.75 in English-to-Japanese translation and In Ding et al. (2014), their pre-reordering approach improves the average τ from 0.49 to 0.67 in Japanese-to-English translation. However, the τ between Chinese and Japanese is very high even without pre-reordering, which suggests Japanese has a much more similar word order with Chinese than English. Then, we conducted more comparisons. We calculated the average τ of French and English, which are usually referred as a language pair with similar word order, on Europarl V7 data[8] (Koehn, 2005). As a result, the score is 0.898, which is not significantly higher than the 0.847 between Chinese and Japanese in our experiment. We further calculated an average τ between Chinese and Japanese on a parallel corpus with 400,000 sentences crawled from Internet[9] and the score decreased to 0.764, which, however, is still relatively high. So we conclude that Chinese and Japanese can be classified as a language pair with *fairly similar* word order but not as a language pair with *significantly different* word order.

As to the phenomenon, we consider a main reason is that Chinese is not a *typical* SVO language. Chinese can even have a SOV word order with the help of a specific particle *ba* preceding the object. As a typical *topic-prominent language*, the word order in Chinese is affected more by semantic constraints than by syntactic ones. We show an example in Fig. 5 to illustrate the strengths and weaknesses of our approach.

From the above discussion, we conclude that Japanese-to-Chinese is a relatively easier translation task than Japanese-to-English, and thus a state-of-the-art PB SMT system can achieve an acceptable performance. On the other hand, rather than the word order problem, which we have found not such a serious issue, we consider the errors in word alignment around Japanese functional morphemes is a critical issue in our experiment. Japanese has a sophisticated system of case-markers, particles, and auxiliary verbs, which are deeply tangled with its syntax and semantics. However, these functional morphemes usually have not exactly corresponding translations in Chinese, and an unsupervised automatic word aligner tends to scatter the alignment of them. We assume that using more explicit syntax information in training and decoding phrases of an SMT system may improve the alignment of these functional morphemes and lead to a better performance in translation quality.

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

[6] http://code.google.com/p/giza-pp/

[7] http://www.speech.sri.com/projects/srilm/

[8] http://www.statmt.org/europarl/

[9] http://japanese.donga.com/

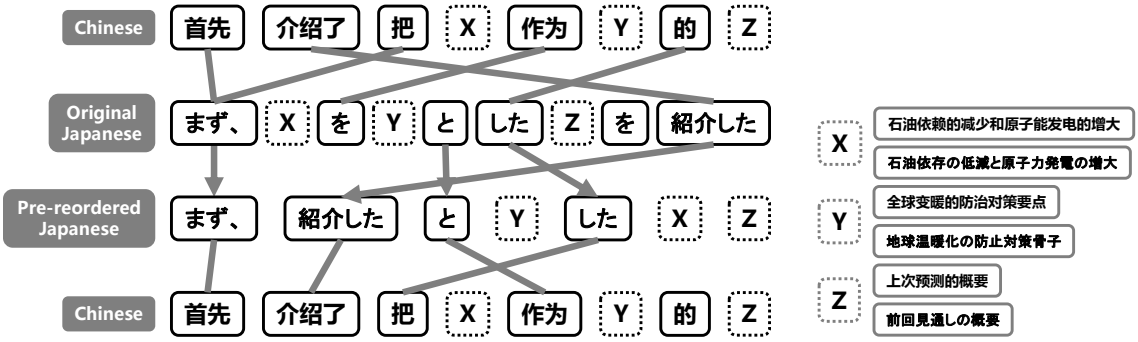

Figure 5: *Example of pre-reordering for a Japanese sentence and the automatically generated word alignment of its Chinese translation. The upper row and the lower row are identical Chinese sentence. The two rows in middle are original and pre-reordered Japanese sentences. X, Y, Z are monotonic noun phrases not affecting the sentence structures. The pre-reordering approach can move the final-positioned verb properly and lead to a more correct alignment of functional words. However, the original Japanese sentence has already had a similar word order as the Chinese sentence except the final verb, while the pre-reordering conducted a rigid reordering of swapping X and Y, which leads to an excess move.*

6 Conclusion and Future Work

We propose a rule-based pre-reordering approach for Japanese-to-Chinese SMT, using the dependency parsing of source-side Japanese sentences. The approach can bring a mediocre improvement in automatic and human evaluation metrics. We investigate the experimental results and discover that Chinese and Japanese actually have relatively similar word orders. It seems that the proposed approach is too rigid to handle the word reordering of Japanese-to-Chinese translation task. We plan to conduct deeper investigation of the two languages and design more flexible pre-reordering rules.

Acknowledgment

We thank Mitsuo Yoshida for providing parallel data on Japanese and Chinese crawled form Internet, which helps us to investigate the characters of the two languages.

References

Han Dan, Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2012. Head finalization reordering for Chinese-to-Japanese machine translation. In *Proc. of SSST*, pages 57–66.

Chenchen Ding, Keisuke Sakanushi, Hirona Touji, and Mikio Yamamoto. 2014. Dependency tree-based pre-reordering rules for statistical Japanese-to-English machine translation. In *Proc. of ANLP*, pages 963–966. (In Japanese).

Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. 2013. Two-stage pre-ordering for Japanese-to-English statistical machine translation. In *Proc. of IJCNLP*, pages 1062–1066.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2012. HPSG-based preprocessing for English-to-Japanese translation. *ACM Transactions on Asian Language Information Processing*, 11(3). Article 8.

Jason Katz-Brown and Michael Collins. 2008. Syntactic reordering in preprocessing for Japanese English translation: MIT system description for NTCIR-7 patent translation task. In *Proc. of NTCIR*, pages 409–414.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of HTL-NAACL*, pages 48–54.

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 *Proc. of ACL*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT summit*, pages 79–86.

Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. 2006. Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proc. of IWSLT*, pages 77–82.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proc. of CoNLL*, pages 63–69.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st Workshop on Asian Translation. In *Proc. of WAT*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167.

Andreas Stolcke. 2002. SRILM–an extensible language modeling toolkit. In *Proc. of ICSLP 2002*, pages 901–904.

Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Masaaki Nagata, Xianchao Wu, Takuya Matsuzaki, and Jun'ichi Tsujii. 2011. NTT-UT statistical machine translation in NTCIR-9 PatentMT. In *Proc. of NTCIR*, pages 585–592.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *SIGHAN 2005, Workshop on Chinese Language Processing*, volume 171.