Syntactic Based Reordering Rules for Chinese-to-Japanese Machine Translation

Han Dan[+] **Katsuhito Sudoh**[*] **Xianchao Wu**[*]
Kevin Duh[*] **Hajime Tsukada**[*] **Masaaki Nagata**[*]
[+]The Graduate University For Advanced Studies
handan@nii.ac.jp
[*]NTT Communication Science Laboratories, NTT Corporation

1 Introduction

In Statistical Machine Translation(SMT), reordering rules have been proved effective in extracting bilingual phrases and in decoding when translating between languages whose word orders are structurally different. Researchers have tackled the reordering problem in multiple ways. One basic idea is pre-ordering (Xia and McCord, 2004; Collins et al., 2005), that is, to pre-order the source sentences following the word order of the target sentences to be used for decoding. For example, making use of a source dependency parser, Xu et al. (2009) manually created dependency-to-string pre-ordering rules for translating English into five SOV(Subject-Object-Verb) languages. Later, dependency tree based pre-ordering rules were automatically extracted by Genzel (2010) from word-aligned parallel sentences.

In this work, we focus on Chinese-to-Japanese translation, motivated by the need of constructing a direct machine translation system without using a pivot language. Chinese and Japanese involve significant differences in syntax, which poses a severe difficulty in SMT. In this direction, we present a detailed syntactic analysis of several reordering issues in Chinese-Japanese translation using the information provided by an HPSG-based deep parser. Then, we introduce novel reordering rules based on head-finalization and linguistically-inspired refinements to make the order of words in Chinese sentences resemble Japanese word order. We empirically show its effectiveness (e.g. 20.70 to 24.23 BLEU improvement).

2 Head Finalization Chinese (HFC) and Chinese Deep Parsing

The structure of languages can be characterized by phrase structures. English is known as a primarily head-initial language, since the head of a phrase can be usually found before its modifiers. On the other hand, Japanese is a typical head-final language because the last word is defined as the head. Isozaki et al. (2010b) proposed Head Finalization (HF) pre-ordering rule to reorder sentences from a head-initial language to resemble the word order in sentences from a head-final language. The essence of this rule is to move the syntactic heads to the end of their constituents by swapping child nodes in a phrase structure tree when the head child appears before the dependent child. Therefore, this reordering rule needs parsed sentences as input. They used *Enju* (Miyao and Tsujii, 2008), an HPSG-based deep parser for English, obtaining strong improvements in English-to-Japanese translation. In this paper, we used *Chinese Enju* (Yu et al., 2011), an HPSG-based parser for Chinese that provides rich syntactic information including phrase structures and syntactic heads.

Since most of the structures of Chinese sentences are head-initial, ideally, HF would reorder Chinese sentences to follow the word order of its Japanese counterpart. Figure 1 shows an example of a head finalized Chinese sentence based on the output of Chinese Enju. Notice that the exception rule described in (Isozaki et al., 2010b) is also implemented. The exception rule says that child nodes are not swapped if the node is a coordination or punctuation. As it can be seen in the example of Figure 1, the nodes of **c3**, **c6** and **c8** were not swapped with their dependencies. In this account, only the verb "去" had been moved to the end of the sentence.

3 Syntax-based Reordering Rules

Although a simple adaptation of HF can improve the word order of Chinese sentences to resemble its Japanese counterpart, we found that HFC has prob-

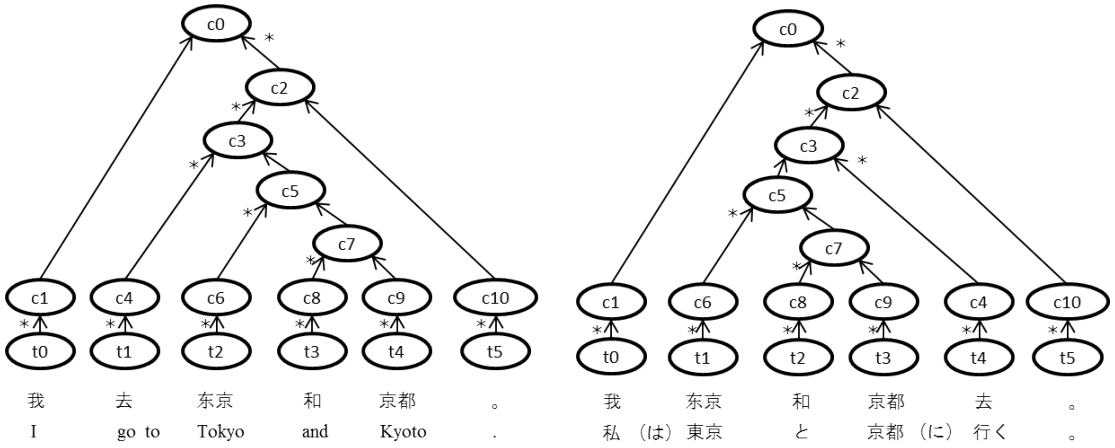

Figure 1: Simple example for HFC. The left figure shows the parsing tree of the original sentence and its English translation. The right figure shows the reordered sentence along with its Japanese translation. ("*" indicate the syntactic head).

lems due to peculiarities in Chinese syntax. In this section we analyze several distinctive cases of the problem in detail. Following this analysis, Section 3.2 proposes a couple of exception rules for pre-reordering, as a refinement of the original HFC.

3.1 Discrepancies in Head Definition

In Chinese, there has been much debate on the definition of syntactic head, possibly because Chinese has less surface syntactic features when compared to other languages. This causes some discrepancies between the definitions of the head in Chinese and Japanese, leading to unexpected and undesirable reorderings. These differences are described below.

Aspect Particle: Although Chinese has no syntactic tense marker, three aspect particles following verbs can be used to identify the tense semantically. They are "了" (did), "着" (doing), "过" (done) and their Japanese counterparts are "た", "ている", "た", respectively. The third word represent past perfect.

The Chinese parser treated aspect particles as dependents of verbs, whereas their Japanese counterparts are identified as the head. For example in Table 1[1], "去" (go) and "过" (done) aligned to "い" and "った", respectively. However, since "过" is treated as a dependent of "去", the sentence will be reordered like HFC in Table 1 which is not following the word order of the Japanese (Ja) translation. On the contrary, the reordered sentence from refined-HFC (R-HFC) can be translated monotonically.

Adverbial Modifier '不': Both in Chinese and Japanese, verb phrase modifiers typically occur in preverbal position, especially when the modifiers are adverbs which are identified as dependents in both languages. For this reason, head finalization works perfectly for them. However, "不" is an exceptional adverb, which is usually translated into an auxiliary verb "ない" in Japanese and thus is the head. For example in Table 2, "不" is the dependent of the word "看"(watch), but "ない", which is aligned to "不", is the head. Therefore, the HFC is not in the same order but the reordered sentence by R-HFC obtained the same order as the Japanese translation.

En	I have been to Tokyo.
Ch	我去 过 东京.
HFC	我东京过 去.
R-HFC	我东京去 过.
Ja	私 (は) 東京 (に) いった.

Table 1: An example for Aspect Particle.

En	I do not watch TV.
Ch	我不 看 电视.
HFC	我电视不 看.
R-HFC	我电视看 不.
Ja	私 (は) テレビ (を) 見ない.

Table 2: An example for Adverbial Modifier `bu4`.

Sentence-final Particle: Sentence-final particles often appear at the end of a sentence to express speaker's attitude: e.g. "吧, 啊" in Chinese, and "なぁ, ね" in Japanese. Although they are in the same position in two languages, they are identified as the dependent and head respectively. In Table 3, "啊" had been reordered to the beginning of the sentence

[1]English translation (En); Chinese original sentence (Ch); reordered Chinese by Head-Final Chinese (HFC); Refined Head-Final Chinese (R-HFC) sentence; Japanese translation (Ja).

as dependent while its Japanese translation “ね” is at the end of the sentence as head. Likewise, by refining the HFC, we can improve the word alignment.

En	It is good weather.
Ch	天气真好啊.
HFC	啊 天气真好.
R-HFC	天气真好啊.
Ja	いい天気ですね.

Table 3: An example for Sentence-final Particle.

Et cetera: In Chinese, “等”, “等等” are used to represent the meaning of “and other things”, and they are identified as dependent, while “など” is always the head in Japanese since it appears as the right-most word in a noun phrase. Table 4 shows an example.

En	Fruits include apples, etc.
Ch	水果包括苹果等.
HFC	水果等 苹果包括.
R-HFC	水果苹果等 包括.
Ja	果物 (は) りんごなど (を) 含んでいる.

Table 4: An example for Et cetera.

3.2 Refinement of HFC

In the preceding subsection, we have discussed syntactic constructions that cause wrong application of Head Finalization to Chinese sentences. Following the observations, we proposed a method to improve the original Head Finalization reordering rule to obtain better alignment with Japanese.

For the refined-HFC, we defined a list of POSs to use as exceptions to the application of the HF reordering rule. Table 5 shows the list of POSs[2] that we defined in the current implementation. While interjection is not discussed in detail, it is obvious that we should not apply reordering to interjection because they are position-independent. PU and CC are basically equivalent to the exception rule that we mentioned in section 2.

4 Experiments

The corpus we used as training data was obtained from China Workshop on Machine Translation (CWMT). It is a Japanese-Chinese parallel corpus in the news domain containing $281,322$ sentence pairs. We also collected another Japanese-Chinese parallel corpus of the same domain containing $529,769$ sentences and merged it with CWMT corpus. We refer to this combination as “CWMT ext.”. For development and test, we used 1,000 sentence pairs, respectively.

[2]The definition of POSs are following Penn Chinese Treebank.

AS	Aspect particle
SP	Sentence-final particle
ETC	*et cetera* (i.e. `deng3` and `deng3 deng3`)
IJ	Interjection
PU	Punctuation
CC	Coordinating conjunction

Table 5: The list of POSs for exception reordering rules

For decoding, we used the MT toolkit Moses in its default configuration. Phrase pairs were extracted from symmetrized word alignments and distortions generated by GIZA++ using the combination of heuristics “grow-diag-final-and” and “msd-bidirectional-fe”. We used the SRILM toolkit to generate a 5-gram language model. The weights of the log-linear combination of feature functions were estimated using MERT. The effectiveness of the reordering proposed in Section 3.2 was assessed by using two precision metrics, BLEU and RIBES (Isozaki et al., 2010a), and two error metrics, TER and WER. Table 6 shows the assessment of translation quality.

As it can be observed in Table 6, the translation quality was consistently and significantly increased when using the HFC reordering rule and further significant improvements were obtained when using the refinement proposed in this work. Specifically, the BLEU score increased from 19.94 to 20.79 when using the CWMT corpus, and from 23.17 to 24.14 when using the CWMT extended corpus.

5 Error Analysis

In Section 3 we analyzed the definition of syntactic head differences between Chinese and Japanese which led to the design of an effective refinement. A manual error analysis of the Refined-HFC results evidenced that some more reordering issues are left and, although they are not side-effects of our proposed rule, they are worth to be mentioned separately.

Serial Verb Construction: This construction is a phenomenon occurring both in Chinese and Japanese, where several verbs are progressively or parallelly put together as one unit without any conjunction. Apparently, HFC are not fit for this construction.

Complementizer: In Chinese, other types of words can act as complementizers (e.g. verb, adjective, quantifiers, etc.) and they are identified as the dependent of the verb that modify. In Japanese, how-

	CWMT				CWMT ext.			
	BLEU	RIBES	TER	WER	BLEU	RIBES	TER	WER
Baseline	16.74	71.24	70.86	77.45	20.70	74.21	66.10	72.36
HFC	19.94	73.49	65.19	71.39	23.17	75.35	61.38	67.74
Refined-HFC	**20.79**	**75.09**	**64.91**	**70.39**	**24.14**	**77.17**	**59.67**	**65.31**

Table 6: Evaluation of translation quality when using CWMT and CWMT extended corpus for training. Results are given in terms of BLEU, RIBES, TER and WER for baseline, Head Finalization Chinese and proposed refined-HFC reordering rules.

ever, they are considered as heads, which represents another head-definition issue.

Adverbial Modifier: Unlike the adverb "不" we discussed in Section 3.1, the ordinary adverbial modifier comes directly before the verb it modifies both in Chinese and Japanese. Nevertheless, according to the principle of identifying the head for Chinese, the adverb will be treated as the dependent and thus the alignment between adverbs and verbs is non-monotonic after reordering.

Verbal Nominalization and Nominal Verbalization: As Guo (2009) discussed when comparing to English and Japanese, Chinese has little inflectional morphology, namely no inflection about tense, case, etc. Thus, words are extremely flexible, making verbal nominalization and nounal verbalization to appear frequently and commonly without any conjugation or declension. As a result, it is difficult to do disambiguation during POS tagging and parsing.

POS tagging and Parsing Errors: These two errors are not caused solely by differences in syntactic structures and they are difficult to remedy during reordering. They are also hard to avoid since reordering rules are highly dependent on the tagger and parser.

6 Conclusions and Future Work

In the present work we have proposed novel Chinese-to-Japanese reordering rules inspired in (Isozaki et al., 2010b) based on linguistic analysis on Chinese HPSG. Although a straight implementation of HF on reordering Chinese sentences performs well, further substantial improvements on translation quality were achieved by including linguistic knowledge into the refinement of the reordering rule.

In Section 5, we have found more patterns on reordering issues when reordering Chinese sentences to resemble Japanese word order. The extraction of those patterns and their effective implementation may lead to further improvements on translation quality and we are planning to explore this possibility in future works. We also believe that using semantic information can further increase the expressive power of reordering rules. With this objective, Chinese Enju can be used since it provides the semantic head of nodes and can interpret sentences with their semantic dependency.

Acknowledgments

This work was mainly developed during an internship at NTT Communication Science Laboratories. We would like to thank Prof. Yusuke Miyao for his invaluable support on this work.

References

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for Statistical Machine Translation. In *Proc. of ACL*, pages 531–540.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proc. of COLING*, pages 376–384.

Yuqing Guo. 2009. *Treebank-based acquisition of Chinese LFG resources for parsing and generation*. Ph.D. thesis, Dublin City University.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010a. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP*.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010b. Head finalization: A simple reordering rule for SOV languages. In *Proc. of WMTMetricsMATR*, pages 244–251.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34:35–80.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proc. of COLING*.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve SMT for Subject-Object-Verb languages. In *Proc. of Human Language Technologies: NAACL*, pages 245–253.

Kun Yu, Yusuke Miyao, Takuya Matsuzaki, Xiangli Wang, and Junichi Tsujii. 2011. Analysis of the difficulties in Chinese deep parsing. In *Proc. of the 12th International Conference on Parsing Technologies*, pages 48–57.