Reducing Morpho-Syntactic Difference for Chinese-Japanese Translation

Chooi-Ling Goh and Eiichiro Sumita
Multilingual Translation Laboratory, MASTAR Project
National Institute of Information and Communications Technology, 619-0289 Kyoto
{chooiling.goh,eiichiro.sumita}@nict.go.jp

1 Introduction

Translation between different word order languages has been a critical problem in phrase-based statistical machine translation (SMT). This is because the distance for word reordering from the source language to the target language becomes larger. For example, translation from subject-verb-object (SVO) languages (e.g. Chinese and English), to subject-object-verb (SOV) languages (e.g. Japanese and Korean). There are researches working on English to Japanese or English to Korean translation by pre-ordering the English sentence word order into Japanese/Korean-like prior to translation and the results showed that pre-ordering does help a lot to improve the translation. Furthermore, Japanese and Korean are free word order languages where the function of arguments is determined by the function particles, whereas English and Chinese are quite strict in the word order to decide the functions of the arguments. However, while Chinese has similar word order to English, it has also similarities to Japanese due to the use of Hans character especially for terms in scientific domains. Therefore, pre-ordering Chinese to Japanese-like order should be easier. Besides, pseudo function particles can be inserted into the Chinese sentence in order to show the semantic functions of the words. We propose a method of pre-ordering by using a dependency parser to move the head to the final position and insert pseudo function particles based on the dependency relations, in order to improve Chinese-Japanese translation.

2 Previous Research

There are a few research that pre-reorder the English text into SOV-like languages. [3] and [10] used Korean as the target language, and applied some hand-crafted rules based on a dependency parser. [3] also inserted some pseudo words (function words) into the source sentence which usually exist in Korean but do not have their corresponding words in English. This helps to reduce the null alignment for English-Korean. [10] also further verified on other SOV-type target languages such as Japanese, Hindi, Urdu and Turkish. Besides the phrase-based SMT model, they also showed that the pre-ordering rules can be applied to a Hierarchical model.

[5] introduced head finalization rule for English-Japanese pre-ordering based on a syntactic parser. The basic idea is that Japanese is a head-final language, therefore, a syntactic head word always comes after its dependent word(s). Only one reordering rule is used: move the syntactic heads to the end of the corresponding syntactic constituents. [9] further improved the head-finalization rule by extracting pre-ordering rules from predicate-argument structures automatically. Besides, three types of pseudo particles were added into the English text: *wa* (topic marker), *ga* (nominative case marker), and *wo* (objective case marker).

All the methods proposed above require a good parser to analyze the source sentence. It is easy to get a good parser for English but not for other languages. In this work, we want to reproduce the work similar to [3] and [5] for Chinese-Japanese translation, using a Chinese dependency parser.

3 Proposed Method

There are a few major differences between Chinese and Japanese morpho-syntactic structures. Japanese uses inflections to indicate tense, aspect, modality and etc, but Chinese does not have any inflection. Japanese uses case markers to show the grammatical function of the marked words but Chinese does not use any case marker. Japanese is a head-final language where the head is always at the end of the phrases (SOV) while Chinese is not (SVO). During word alignment for machine translation, missing corresponding words will cause null alignment. Furthermore, it is difficult to guess the translation in the target if the information is missing in the source text. In a standard phrase-based SMT

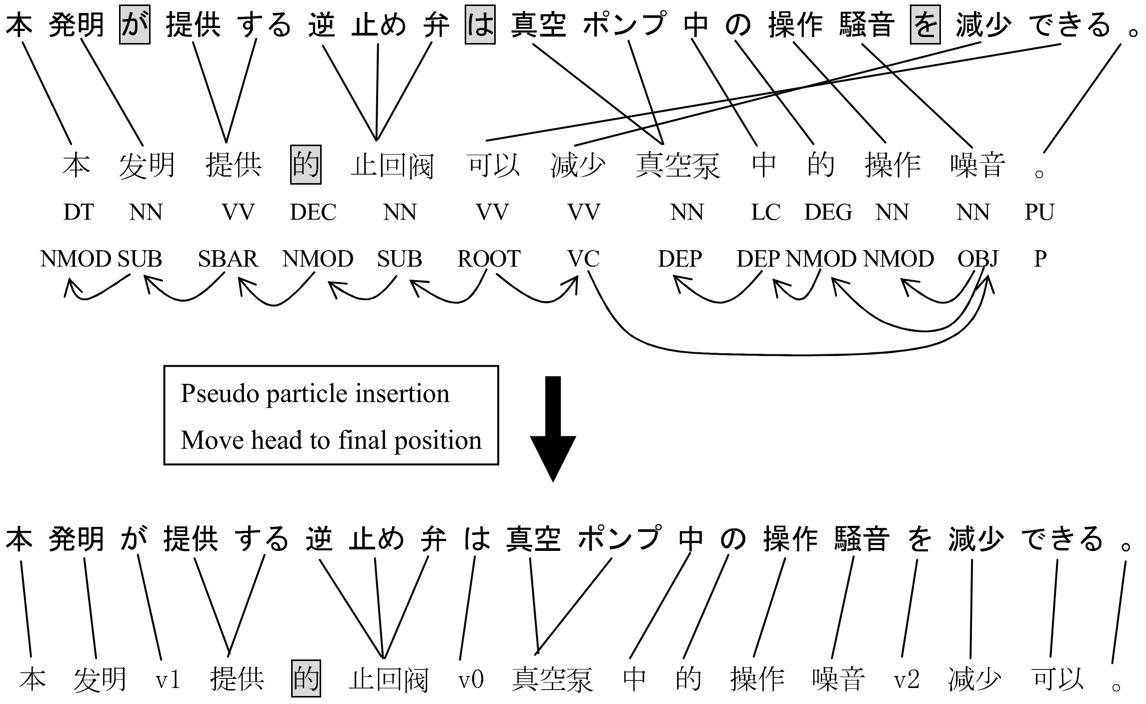
本 発明 が 提供 する 逆 止め 弁 は 真空 ポンプ 中 の 操作 騒音 を 減少 できる 。
本 发明 提供 的 止回阀 可以 减少 真空泵 中 的 操作 噪音 。
DT NN VV DEC NN VV VV NN LC DEG NN NN PU
NMOD SUB SBAR NMOD SUB ROOT VC DEP DEP NMOD NMOD OBJ P
Pseudo particle insertion
Move head to final position
本 発明 が 提供 する 逆 止め 弁 は 真空 ポンプ 中 の 操作 騒音 を 減少 できる 。
本 发明 v1 提供 的 止回阀 v0 真空泵 中 的 操作 噪音 v2 减少 可以 。

Figure 1: An illustration of pseudo particle insertion and re-ordering based on dependency parsing

system, there is a limitation on the reordering (distortion limit). If the syntactic structures between the source language and the target language are very different, it is necessary to increase the distortion limit in order to allow long distance reordering. However, this will increase the decoding time and yet the reordering may not be correct. Therefore, in order to reduce the distance (i.e. morpho-syntactic differences) between Chinese and Japanese, we propose five steps to re-phrase the Chinese text into Japanese-like.

1. Split the sentence into small clauses
2. Parse the clauses individually by a dependency parser
3. Insert pseudo function particles
4. Re-order the head into final position
5. Join the clauses

Figure 1 shows an example where after pre-ordering and pseudo particle insertion, the cross and null alignment have been reduced. In this case, the translation can be done monotonously.

3.1 Sentence Splitting

[2] proposed to use some hand-crafted rules to split long sentences into small clauses for translation. This is to avoid the reordering between the clauses after translation as the material before and after the split could be translated almost independently. We applied the similar rules for both parsing and translation. Sentences are first split into clauses before parsing. Therefore, pre-ordering can only be done within the clauses. During decoding, the split is marked by the "wall" constraint in a phrase-based SMT system [7, 6]. We used the Penn Chinese Treebank POS tagset in our rules for splitting. If any of the POS tags shown in Table 1 is found before (tail position) or after (head position) a comma, then the comma will be a split position.

POS tag	Description
Head Position	
AD	adverb
BA	*BA* in ba-construction
CC	coordination conjunction
CS	subordinating conjunction
DT	determiner
LB	*BEI* in long bei-construction
P	preposition excluding *BEI* and *BA*
PN	pronoun
VV	other verb
Tail Position	
LC	localizer

Table 1: POS tags used for wall constraint

3.2 Chinese Parser

We used the CNP Parser provided by the ALAGIN Forum[1]. This parser generates dependency parsing with high accuracy [1]. Based on the parsing results, we pre-order the sentence using the methods as described in the following sections.

3.3 Pseudo Particle Insertion

Japanese language uses case markers to indicate the functionality of the words, whereas Chinese does not have this type of marker. Therefore, there is a gap between them for word alignment. In order to reduce the gap, we insert pseudo particles to the Chinese text. Based on the dependency relation type, three pseudo particles were defined as in [5].

1. The subject (SUB) of the ROOT is assigned with the particle v0 (acts as は-*wa* in Japanese)
2. The other subject (SUB) in the sentence is assigned with the particle v1 (acts as が-*ga* in Japanese)
3. The object (OBJ) is assigned with the particle v2 (acts as を-*wo* in Japanese)

3.4 Pre-ordering Rules

Similar to [3] and [5], the head is moved to the end of the phrases. In general, we only have one rule: move the head to the end of the last dependent. The dependents are remained in the original positions. However, there are a few exceptions on the rule:

- Punctuation and sentence-final particle are excluded as dependents
- Aspect marker is moved together with the head
- Negation marker is placed after the head and moved together with the head
- The countable noun is moved together with the head if it is a determiner

3.5 Joining Clauses

After reordering, the clauses are joined sequentially. For the test data, a wall constraint marker (<wall/>) is inserted for decoding. This indicates a position where the translation reordering cannot go across the wall boundary.

4 Experiment Results

Our experiment were carried out using a Chinese-Japanese Patent corpus. The Chinese text data were purchased from the CNIPR (China Intellectual Property Net) and translated into Japanese by a translation agency. This corpus has about 270K sentence pairs. However, we removed sentences that are more than 100 words. This is because when the sentence is too long, it is difficult to parse the sentence and will cause more parsing errors. Secondly, the word alignment result will also be deteriorated when there are long sentences in the corpus. Finally, we used only 240K sentence pairs for training the SMT model, 1K for MERT tuning and 2K is used as test data. We used Moses [7] in our experiment with the following settings.

- alignment with grow-diag-final-and heuristic
- 5-gram language model, interpolated Kneser-Ney discounting
- msd-bidirectional-fe lexicalized reordering
- distortion-limit = 10

Our preliminary experiment showed that distortion limit of 10 is better than 6 as the word orders between Chinese and Japanese are quite different. Even after pre-ordering, the distortion limit should have no influence to the translation but in reality, the pre-ordering was not perfect and we still need to set the distortion limit as 10 for reordering.

The Chinese text were first split using the wall constraint as describes in Section 3.1 before parsing. Table 2 shows the number of sentences over the number of clauses before and after the split. Splitting the sentences into small clauses not only fasten up the parsing time but also the decoding time.

	before	after
train	240,217	404,666
dev	1,000	2,972
test	2,000	5,841
total	243,217	413,479

Table 2: Number of sentences/clauses before and after split

Table 3 shows the experiment results. We use BLEU [8] and RIBES [4] as the automatic evaluation metrics. While BLEU compares only the n-grams, RIBES considers also the overall word order. We obtained slight improvements for both metrics.

	BLEU	RIBES
Baseline	41.08	0.817046
Pre-ordering	41.54	0.828761

Table 3: Evaluation results using BLEU and RIBES

Table 4 shows the human evaluation results using 5-rank metric. While S is the best, D is the worst.

[1] http://www.alagin.jp/

We randomly selected 100 sentences for evaluation. Similar to automatic evaluation, we only obtained slight improvements over the baseline.

Rank	S	+A	+B	+C	+D
Baseline	0	4	28	74	100
Pre-ordering	0	3	30	80	100

Table 4: Human Assessment Results

Although previous research showed that pre-ordering for English-SOV language translation generates excellent improvements, in our experiment for Chinese-Japanese Patent translation, the improvement is small. The main reason could be the parsing errors. Patent text is always long and complex, therefore, the parsing accuracy is also low. Although we have tried to reduce the parsing errors by splitting the long sentence into small clauses before parsing, there still remain some other problems. Since Chinese language does not use spaces to indicate the word boundaries, so word segmentation has be to done during morphological analysis. If the segmentation is wrong, then there will be errors for POS tagging and further parsing will be incorrect as well. In this case, there is a possibility that the pre-ordering will move the incorrect head word to the end, and cause the distance to be even larger. While English dependency parser achieved 92.5% accuracy, Chinese dependency parser achieved only 89.4% accuracy even when the gold standard segmentation and POS tags were used [1]. In the future, we want to try to reduce the complexity of Patent texts in order to improve the parsing accuracy. Besides, our pre-ordering rules are quite shallow now where only general cases are considered. Although we have split the sentences into small clauses in order to reduce parsing errors, however, we did not consider cases of coordination expressions as pointed out by [5], which may cause errors to pre-ordering.

5 Conclusion

We have done the experiment on pre-ordering on Chinese text for Chinese-Japanese Patent translation. While previous research showed significant improvements on translation from English to other SOV languages, we only obtained slight translation improvements. However, there are still rooms for improvement as our parsing result is not perfect and the pre-ordering rules could be furthered refined.

References

[1] Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Improving Dependency Parsing with Subtrees from Auto-Parsed Data. In *Proceedings of EMNLP*, pages 570–579.

[2] Chooi-Ling Goh, Takashi Onishi, and Eiichiro Sumita. 2011. Rule-based Reordering Constraints for Phrase-based SMT. In *Proceedings of the 15th Conference of the EAMT*, pages 113–120.

[3] Gumwon Hong, Seung-Wook Lee, and Hae-Chang Rim. 2009. Bridging Morpho-Syntactic Gap between Source and Target Sentences for English-Korean Statistical Machine Translation. In *Proceedings of ACL-IJCNLP*, pages 233–236.

[4] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010a. 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.

[5] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010b. Head Finalization: A Simple Reordering Rule for SOV Languages. In *Proceedings of the Joint Workshop on SMT and MetricsMATR*, pages 244–251.

[6] Philipp Koehn and Barry Haddow. 2009. Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 160–164.

[7] 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 ACL, demo and poster session*, pages 177–180.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL*, pages 311–318.

[9] Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2011. Extracting Pre-ordering Rules from Predicate-Argument Structures. In *Proceedings of IJCNLP*, pages 29–37.

[10] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a Dependency Parser to Improve SMT for Subject-Object-Verb Languages. In *Proceedings of HLT-NAACL*, pages 245–253.