Using Unlabeled Dependency Parsing for Pre-reordering for Chinese-to-Japanese Statistical Machine Translation

Dan Han[1,2] **Pascual Martínez-Gómez**[2,3] **Yusuke Miyao**[1,2]
Katsuhito Sudoh[4] **Masaaki Nagata**[4]
[1]The Graduate University For Advanced Studies
[2]National Institute of Informatics, [3]The University of Tokyo
[4]NTT Communication Science Laboratories, NTT Corporation
{handan,pascual,yusuke}@nii.ac.jp
{sudoh.katsuhito,nagata.masaaki}@lab.ntt.co.jp

Abstract

Chinese and Japanese have a different sentence structure. Reordering methods are effective, but need reliable parsers to extract the syntactic structure of the source sentences. However, Chinese has a loose word order, and Chinese parsers that extract the phrase structure do not perform well. We propose a framework where only POS tags and unlabeled dependency parse trees are necessary, and linguistic knowledge on structural difference can be encoded in the form of reordering rules. We show significant improvements in translation quality of sentences from news domain, when compared to state-of-the-art reordering methods.

1 Introduction

Translation between Chinese and Japanese languages gains interest as their economic and political relationship intensifies. Despite their linguistic influences, these languages have different syntactic structures and phrase-based statistical machine translation (SMT) systems do not perform well. Current word alignment models (Och and Ney, 2003) account for local differences in word order between bilingual sentences, but fail at capturing long distance word alignments. One of the main problems in the search of the best word alignment is the combinatorial explosion of word orders, but linguistically-motivated heuristics can help to guide the search.

This work explores syntax-informed pre-reordering for Chinese; that is, we obtain syntactic structures of Chinese sentences, reorder the words to resemble the Japanese word order, and then translate the reordered sentences using a phrase-based SMT system. However, Chinese parsers have difficulties in extracting reliable syntactic information, mainly because Chinese has a loose word order and few syntactic clues such as inflection and function words.

On one hand, parsers implementing head-driven phrase structure grammars infer a detailed constituent structure, and such a rich syntactic structure can be exploited to design well informed reordering methods. However, inferring abundant syntactic information often implies introducing errors, and reordering methods that heavily rely on detailed information are sensitive to those parsing errors (Han et al., 2012).

On the other hand, dependency parsers are committed to the simpler task of finding dependency relations and dependency labels, which can also be useful to guide reordering (Xu et al., 2009). However, reordering methods that rely on those dependency labels will also be prone to errors, specially in the case of Chinese since it has a richer set of dependency labels when compared to other languages. Since improving parsers for Chinese is challenging, we thus aim at reducing the influence of parsing errors in the reordering procedure.

We present a hybrid approach that boosts the performance of phrase-based SMT systems by pre-reordering the source language using unlabeled parse trees augmented with constituent information derived from Part-of-Speech tags. Specifically, we propose a framework to pre-reorder a Subject-Verb-Object (SVO) language, in order to improve its translation to a Subject-Object-Verb (SOV) language, where the only required syntactic information are POS tags and unlabeled dependency parse trees. We test the performance of our pre-reordering method and compare it to state-of-the-art reordering methods in the news domain for Chinese.

In the next section, we describe similar work on pre-reordering methods for language pairs that in-

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 25–33,
Sofia, Bulgaria, August 8, 2013. ©2013 Association for Computational Linguistics

volve either Chinese or Japanese, and explain how our method builds upon them. From a linguistic perspective, we describe in section 3 our observations of reordering issues between Chinese and Japanese and detail how our framework solves those issues. In section 4 we assess to what extent our pre-reordering method succeeds in reordering words in Chinese sentences to resemble the order of Japanese sentences, and measure its impact on translation quality. The last section is dedicated to discuss our findings and point to future directions.

2 Related Work

Although there are many works on pre-reordering methods for other languages to English translation or inverse (Xia and McCord, 2004; Xu et al., 2009; Habash, 2007; Wang et al., 2007; Li et al., 2007; Wu et al., 2011), reordering method for Chinese-to-Japanese translation, which is a representative of long distance language pairs, has received little attention.

The most related work to ours is in (Han et al., 2012), in which the authors introduced a refined reordering approach by importing an existing reordering method for English proposed in (Isozaki et al., 2010b). These reordering strategies are based on Head-driven phrase structure grammars (HPSG) (Pollard and Sag, 1994), in that the reordering decisions are made based on the head of phrases. Specifically, HPSG parsers (Miyao and Tsujii, 2008; Yu et al., 2011) are used to extract the structure of sentences in the form of binary trees, and head branches are swapped with their dependents according to certain heuristics to resemble the word order of the target language. However, those strategies are sensitive to parsing errors, and the binary structure of their parse trees impose hard constraints in sentences with loose word order. Moreover, as Han et al. (2012) noted, reordering strategies that are derived from the HPSG theory may not perform well when the head definition is inconsistent in the language pair under study. A typical example for the language pair of Chinese and Japanese that illustrates this phenomenon is the adverb “bu4”, which is the dependent of its verb in Chinese but the head in Japanese.

The work in (Xu et al., 2009) used an English dependency parser and formulated handcrafted reordering rules with dependency labels, POS tags and weights as triplets and implemented them recursively into sentences. This design, however, limited the extensibility of their method. Our approach follows the idea of using dependency tree structures and POS tags, but we discard the information on dependency labels since we did not find them informative to guide our reordering strategies in our preliminary experiments, partly due to Chinese showing less dependencies and a larger label variability (Chang et al., 2009).

3 Methodology

In Subject-Verb-Object (SVO) languages, objects usually follow their verbs, while in Subject-Object-Verb (SOV) languages, objects precede them. Our objective is to reorder words in Chinese sentences (SVO) to resemble the word order of Japanese sentences (SOV). For that purpose, our method consists in moving verbs to the right-hand side of their objects. However, it is challenging to correctly identify the appropriate verbs and objects that trigger a reordering, and this section will be dedicated to that end.

More specifically, the first step of our method consists in identifying the appropriate verb (and certain words close to it) that need to be moved to the right-hand side of its object argument. Verbs (and those accompanying words) will move as a block, preserving the relative order among them. We will refer to them as *verbal blocks* (Vbs). The second step will consist in identifying the right-most argument object of the verb under consideration, and moving the verbal block to the right-hand side of it. Finally, certain invariable grammatical particles in the original vicinity of the verb will also be reordered, but their positions will be decided relative to their verb.

In what follows, we describe in detail how to identify verbal blocks, their objects and the invariable grammatical particles that will play a role in our reordering method. As mentioned earlier, the only information that will be used to perform this task will be the POS tags of the words and their unlabeled dependency structures.

3.1 Identifying verbal blocks (Vbs)

Verbal blocks are composed of a head (Vb-H) and possibly accompanying dependents (Vb-D). In the Chinese sentence “wo3 (I) chi1 le5 (ate) li2 (pear).”[1], “chi1” refers to the English verb “eat”

[1] In this paper, we represent a Chinese character by using Pinyin plus a tone number (there are 5 tones in Chinese). In the example, “chi1(eat)” is a verb and “le5(-ed)” is an aspect particle that adds preterit tense to the verb.

Vb-H	VV VE VC VA P
Vb-D	AD AS SP MSP CC VV VE VC VA
BEI	LB SB
RM-D	NN NR NT PN OD CD M FW CC ETC LC DEV DT JJ SP IJ ON
Oth-DEP	LB SB CS

Table 1: Lists of POS tags in Chinese used to identify blocks of words to reorder (Vb-H, Vb-D, BEI lists), the POS tags of their dependents (RM-D lists) which indicate the reordering position, and invariable grammatical particles (Oth-DEP) that need to be reordered.

and the aspect particle "le5" adds a preterit tense to the verb. The words "chi1 le5" are an example of verbal block that should be reordered as a block without altering its inner word order, i.e. "wo3 (I) li2 (pear) chi1 le5 (ate).", which matches the Japanese SOV order.

Possible heads of verbal blocks (Vb-H) are verbs (words with POS tags VV, VE, VC and VA), or prepositions (words with POS tag P). The Vb-H entry of Table 1 contains the list of POS tags for heads of verbal blocks. We use prepositions for Vb-H identification since they behave similarly to verbs in Chinese and should be moved to the rightmost position in a prepositional phrase to resemble the Japanese word order. There are three conditions that a word should meet to be considered as a Vb-H:

i) Its POS tag is in the set of Vb-H in Table 1.
ii) It is a dependency head, which indicates that it may have an object as a dependent.
iii) It has no dependent whose POS tag is in the set of BEI in Table 1. BEI particles indicate that the verb is in passive voice and should not be reordered since it already resembles the Japanese order.

Chinese language does not have inflection, conjugation, or case markers (Li and Thompson, 1989). For that reason, some adverbs (AD), aspect particles (AS) or sentence-final particles (SP) are used to signal modality, indicate grammatical tense or add aspectual value to verbs. Words in this category preserve the order when translating to Japanese, and they will be candidates to be part of the verbal block (Vb-D) and accompany the verb when it is reordered. Other words in this category are coordinating conjunctions (CC) that connect multiple verbs, and both resultative "de5" (DER) and manner "de5" (DEV). The full list of POS tags used to identify Vb-Ds can be found in Table 1. To be a Vb-D, there are three necessary conditions as well:

i) Its POS tag is in the Vb-D entry in Table 1.
ii) It is a dependent of a word that is already in the Vb.
iii) It is next to its dependency head or only a coordination conjunction is in between.

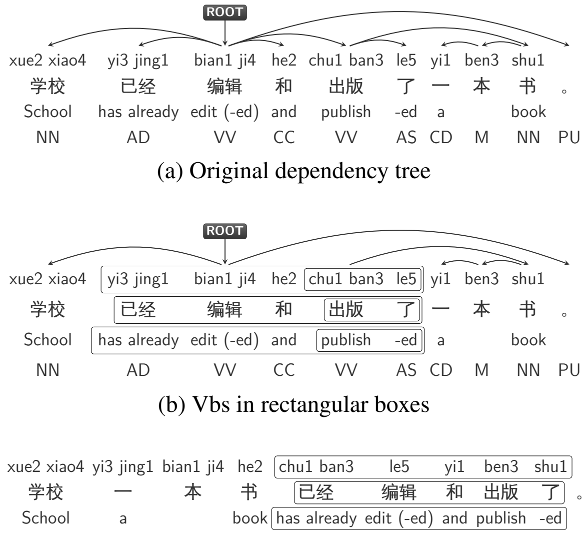
To summarize, to build verbal blocks (Vbs) we first find the words that meet the three Vb-H conditions. Then, we test the Vb-D conditions on the words adjacent to the Vb-Hs and extend the verbal blocks to them if they meet the conditions. This process is iteratively applied to the adjacent words of a block until no more words can be added to the verbal block, possibly nesting other verbal blocks if necessary.

Figure 1a [2] shows an example of a dependency tree of a Chinese sentence that will be used to illustrate Vb identification. By observing the POS tags of the words in the sentence, only the words "bian1 ji4 (edit)" and "chu1 ban3 (publish)" have a POS tag (i.e. VV) in the Vb-H entry of Table 1. Moreover, both words are dependency heads and do not have any dependent whose POS tag is in the BEI entry of Table 1. Thus, "bian1 ji4 (edit)" and "chu1 ban3 (publish)" will be selected as Vb-Hs and form, by themselves, two separate incipient Vbs. We arbitrarily start building the Vb from the word "chu1 ban3 (publish)", by analyzing its adjacent words that are its dependents.

We observe that only "le5 (-ed)" is adjacent to "chu1 ban3 (publish)", it is its dependent, and its POS tag is in the Vb-D list. Since "le5 (-ed)" meets all three conditions stated above, "le5 (-ed)" will be included in the Vb originated by "chu1 ban3 (publish)". The current Vb thus consists of the sequence of tokens "chu1 ban3 (publish)" and "le5 (-ed)", and the three conditions for Vb-D are tested on the adjacent words of this block. Since the adjacent words (or words separated by a coordinating conjunction) do not meet the conditions, the block is not further extended. Figure 1b shows the dependency tree where the Vb block that consists of the words "chu1 ban3 (publish)" and "le5 (-ed)" is represented by a rectangular box.

By checking in the same way, there are three dependents that meet the requirements of being

[2] For all the dependency parsing trees in this paper, arrows are pointing from heads to their dependents.


ROOT
xue2 xiao4 yi3 jing1 bian1 ji4 he2 chu1 ban3 le5 yi1 ben3 shu1
学校 已经 编辑 和 出版 了 一 本 书 。
School has already edit (-ed) and publish -ed a book
NN AD VV CC VV AS CD M NN PU
xue2 xiao4 yi3 jing1 bian1 ji4 he2 chu1 ban3 le5 yi1 ben3 shu1
学校 一 本 书 已经 编辑 和 出版 了 。
School a book has already edit (-ed) and publish -ed

(a) Original dependency tree

(b) Vbs in rectangular boxes

(c) Merged and reordered Vb

Figure 1: An example that shows how to detect and reorder a Verbal block (Vb) in a sentence. In the first two figures 1a and 1b, Chinese Pinyin, Chinese tokens, word-to-word English translations, and POS tags of each Chinese token are listed in four lines. In Figure 1c, there are Chinese Pinyin, reordered Chinese sentence and its word-to-word English counterpart.

Vb-Ds for “bian1 ji4 (edit)”: “yi3 jing1 (has already)”, “he2 (and)” and “chu1 ban4 (publish)” and hence this Vb consists of three tokens and one Vb. The outer rectangular box in Figure 1b shows that the Vb “bian1 ji4 (edit)” as the Vb-H. Figure 1c shows an image of how this Vb will be reordered while the inner orders are kept. Note that the order of building Vbs from which Vb-Hs, “chu1 ban3 (publish)” or “bian1 ji4 (edit)” will not affect any change of the final result.

3.2 Identifying objects

In the most general form, objects are dependents of verbal blocks[3] that act as their arguments. While the simplest objects are nouns (N) or pronouns (PN), they can also be comprised of noun phrases or clauses (Downing and Locke, 2006) such as nominal groups, finite clauses (e.g. *that* clauses, *wh*-clauses) or non-finite clauses (e.g. -*ing* clauses), among others.

For every Vb in a verb phrase, clause, or sentence, we define the right-most object dependent (RM-D) as the word that:

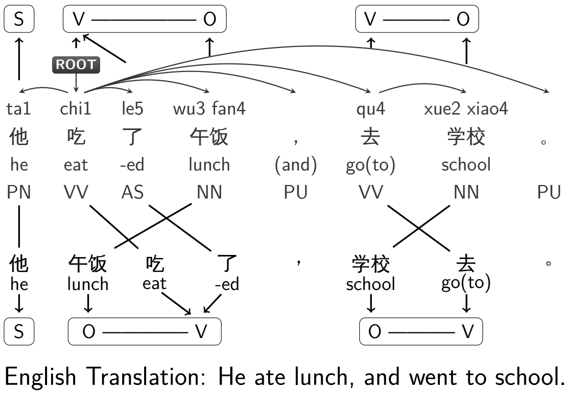

S
V O
V O
ROOT
ta1 chi1 le5 wu3 fan4 qu4 xue2 xiao4
他 吃 了 午饭 ， 去 学校 。
he eat -ed lunch (and) go(to) school
PN VV AS NN PU VV NN PU
他 午饭 吃 了 ， 学校 去 。
he lunch eat -ed school go(to)
S
O V
O V
English Translation: He ate lunch, and went to school.

Figure 2: An example of a Chinese sentence with a coordination of verb phrases as predicate. Subject(S), verbs(V), and objects(O) are displayed for both verb phrases. Lines between the original Chinese sentence and the reordered Chinese sentence indicate the reordering trace of Verbal blocks(Vb).

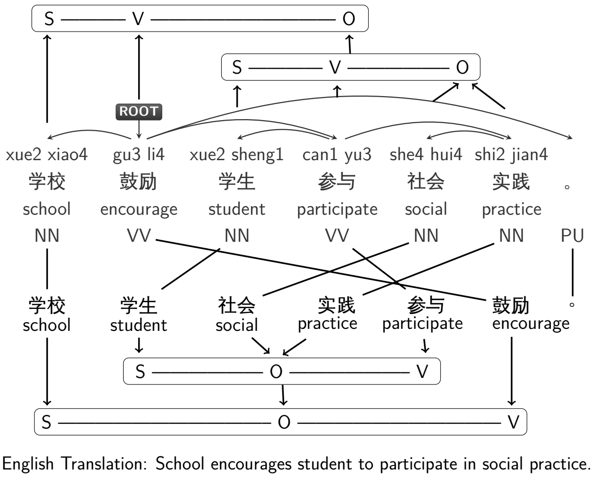
i) its POS tag is in the RM-D entry of Table 1,
ii) its dependency head is inside of the verbal block, and
iii) is the right-most object among all objects of the verbal block.

All verbal blocks in the phrase, clause, or sentence will move to the right-hand side of their correspondent RM-Ds recursively. Figure 1b and Figure 1c show a basic example of object identification. The Chinese word corresponding to “shu1 (book)” is a dependent of a word within the verbal block and its POS tag is within the RM-D entry list of Table 1 (i.e. NN). For this reason, “shu1 (book)” is identified as the right-most dependent of the verbal block (Vb), and the Vb will move to the right-hand side of it to resemble the Japanese word order.

A slightly more complex example can be found in Figure 2. In this example, there is a coordination structure of verb phrases, and the dependency tree shows that the first verb, “chi1 (eat)”, appears as the dependency head of the second verb, “qu4 (go)”. The direct right-most object dependent (RM-D) of the first verb, “chi1 (eat)”, is the word “wu3 fan4 (lunch)”, and the verb “chi1 (eat)” will be moved to the right-hand side of its object dependent.

There are cases, however, where there is no coordination structure of verb phrases but a similar dependency relation occurs between two verbs. Figure 3 illustrates one of these cases, where the main verb “gu3 li4 (encourage)” has no direct de-

[3] Dependents of verbal blocks are dependents of any word within the verbal block.



S — V — O
S — V — O
ROOT
xue2 xiao4 学校 school NN
gu3 li4 鼓励 encourage VV
xue2 sheng1 学生 student NN
can1 yu3 参与 participate VV
she4 hui4 社会 social NN
shi2 jian4 实践 practice NN
。 PU
学校 school
学生 student
社会 social
实践 practice
参与 participate
鼓励 encourage
。
S — O — V
S — O — V
English Translation: School encourages student to participate in social practice.

Figure 3: An example of a Chinese sentence in which an embedded clause appears as the object of the main verb. Subjects (S), verbs (V), and objects (O) are displayed for both the sentence and the clause. Lines between the original Chinese sentence and the reordered Chinese sentence indicate the reordering trace of Verbal blocks (Vb).

pendent that can be considered as an object since no direct dependent has a POS tag in the RM-D entry of Table 1. Instead, an embedded clause (SVO) appears as the object argument of the main verb, and the main verb "gu3 li4 (encourage)" appears as the dependency head of the verb "can1 yu2 (participate)".

In the news domain, reported speech is a frequent example that follows this pattern. In our method, if the main verb of the sentence (labeled as ROOT) has dependents but none of them is a direct object, we move the main verb to the end of the sentence. As for the embedded clause "xue2 sheng1 (student) can1 yu2 (participate) she4 hui4 (social) shi2 jian4 (practice)", the verbal block of the clause is the word "can1 yu2 (participate)" and its object is "shi2 jian4 (practice)". Applying our reordering method, the clause order results in "xue2 sheng1 (student) she4 hui4 (social) shi2 jian4 (practice) can1 yu2 (participate)". The result is an SOV sentence with an SOV clause, which resembles the Japanese word order.

3.3 Identifying invariable grammatical particles

In Chinese, certain invariable grammatical particles that accompany verbal heads have a different word order relative to their heads, when compared to Japanese. Those particles are typically "bei4" particle (POS tags LB and SB) and subordinating conjunctions (POS tag CS). Those particles appear on the left-hand side of their dependency heads in Chinese, and they should be moved to the right-hand side of their dependency heads for them to resemble the Japanese word order. Reordering invariable grammatical particles in our framework can be summarized as:

i) Find dependents of a verbal head (Vb-H) whose POS tags are in the Oth-DEP entry of Table 1.
ii) Move those particles to the right-hand side of their (possibly reordered) heads.
iii) If there is more than one such particle, move them keeping the relative order among them.

3.4 Summary of the reordering framework

Based on the definitions above, our dependency parsing based pre-reordering framework can be summarized in the following steps:

1. Obtain POS tags and an unlabeled dependency tree of a Chinese sentence.
2. Obtain reordering candidates: Vbs.
3. Obtain the object (RM-D) of each Vb.
4. Reorder each Vb in two exclusive cases by following the order:
   (a) If RM-D exists, reorder Vb to be the right-hand side of RM-D.
   (b) If Vb-H is ROOT and its RM-D does not exist, reorder Vb to the end of the sentence.
   (c) If none of above two conditions is met, no reordering happens.
5. Reorder grammatical particles (Oth-DEPs) to the right-hand side of their corresponding Vbs.

Note that, unlike other works in reordering distant languages (Isozaki et al., 2010b; Han et al., 2012; Xu et al., 2009), we do not prevent chunks from crossing punctuations or coordination structures. Thus, our method allows to achieve an authentic global reordering in reported speech, which is an important reordering issue in news domains.

In order to illustrate our method, a more complicated Chinese sentence example is given in Figure 4, which includes the unlabeled dependency

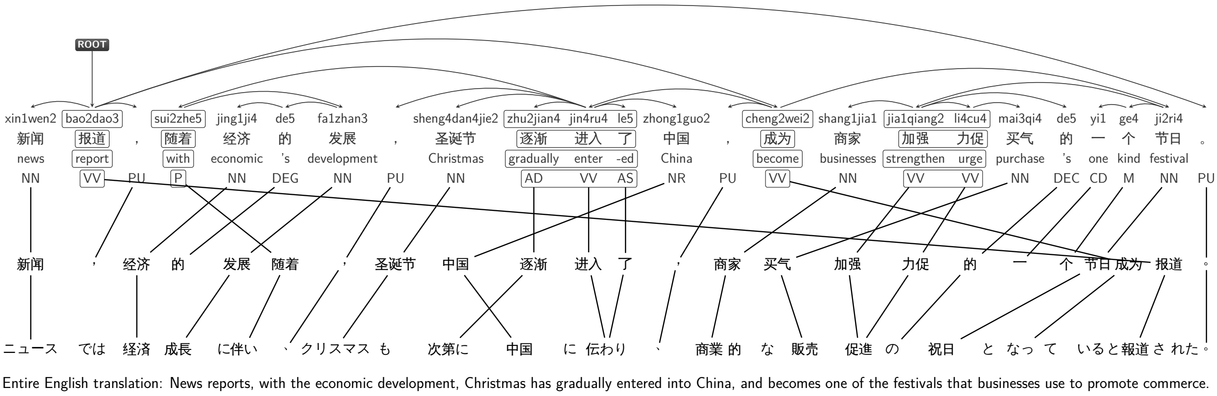

ROOT
xin1wen2 bao2dao3 sui2zhe5 jing1ji4 de5 fa1zhan3 sheng4dan4jie2 zhu2jian4 jin4ru4 le5 zhong1guo2 cheng2wei2 shang1jia1 jia1qiang2 li4cu4 mai3qi4 de5 yi1 ge4 ji2ri4
新闻 报道 ， 随着 经济 的 发展 ， 圣诞节 逐渐 进入 了 中国 ， 成为 商家 加强 力促 买气 的 一 个 节日 。
news report with economic 's development Christmas gradually enter -ed China become businesses strengthen urge purchase 's one kind festival
NN VV PU P NN DEG NN PU NN AD VV AS NR PU VV NN VV VV NN DEC CD M NN PU
新闻 ， 经济 的 发展 随着 ， 圣诞节 中国 逐渐 进入 了 ， 商家 买气 加强 力促 的 一 个 节日 成为 报道 。
ニュース では 经済 成長 に伴い 、 クリスマス も 次第に 中国 に 伝わり 、 商業 的 な 販売 促進 の 祝日 と なっ て いると報道 さ れた 。
Entire English translation: News reports, with the economic development, Christmas has gradually entered into China, and becomes one of the festivals that businesses use to promote commerce.

Figure 4: Dependency parse tree of a complex Chinese sentence example, and word alignments for reordered sentence with its Japanese counterpart. The first four lines are Chinese Pinyin, tokens, word-to-word English translations, and the POS tags of each Chinese token. The fifth line shows the reordered Chinese sentence while the sixth line is the segmented Japanese translation. The entire English translation for the sentence is showed in the last line.

parsing tree of the original Chinese sentence, and the word alignment between reordered Chinese sentence and its Japanese counterpart, etc.

Based on both POS tags and the unlabeled dependency tree, first step of our method is to obtain all Vbs. For all heads in the tree, according to the definition of Vb introduced in Section 3.1, there are six tokens which will be recognized as the candidates of Vb-Hs, that is "bao4 dao3 (report)", "sui2 zhe5 (with)", "jin4 ru4 (enter)", "cheng2 wei2 (become)", "jia1 qiang2 (strengthen)", and "li4 cu4 (urge)". Then, for each of the candidate, its direct dependents will be checked if they are Vb-Ds. For instance, for the verb of "jin4 ru4 (enter)", its dependents of "zhu2 jian4 (gradually)" and "le5 (-ed)" will be considered as the Vb-Ds. For the case of "jia1 qiang2 (strengthen)", instead of being a Vb-H, it will be recognized as Vb-D of the Vb "li4 cu4 (urge)" since it is one of the direct dependents of "li4 cu4 (urge)" with a qualified POS tag for Vb-D. Therefore, there are five Vbs in total, which are "bao4 dao3 (report)", "sui2 zhe5 (with)", "zhu2 jian4 (gradually) jin4 ru4 (enter) le5 (-ed)", "cheng2 wei2 (become)", and "jia1 qiang2 (strengthen) li4 cu4 (urge)".

The next step is to identify RM-D for each Vb, if there is one. By checking all conditions, four Vbs have their RM-Ds: "fa1 zhan3 (development)" is the RM-D of the Vb "sui2 zhe5 (with)"; "zhong1 guo2 (China)" is the RM-D of the Vb "zhu2 jian4 (gradually) jin4 ru4 (enter) le5 (-ed)"; "jie2 ri4 (festival)" is the RM-D of the Vb "cheng2 wei2 (become)"; "mai3 qi4 (purchase)" is the RM-D of the Vb "jia1 qiang2 (strengthen) li4 cu4 (urge)".

After obtaining all RM-Ds, we find those Vbs that have RM-Ds and move them to right of their RM-Ds. As for the case of "bao4 dao3 (report)", since it is the root and does not have any matched RM-D, it will be moved to the end of the sentence, before any final punctuation. Finally, since there is no any invariable grammatical particle in the sentence that need to be reordered, reordering has been finished. From the alignments between the reordered Chinese and its Japanese translation showed in the figure, an almost monotonic word alignment has been achieved.

For comparison purposes, particle seed words had been inserted into the reordered sentences in the same way as the Refined-HFC method, which is using the information of predicate argument structure output by Chinese Enju (Yu et al., 2011). We therefore can not entirely disclaim the use of the HPSG parser at the present stage in our method. However, we believe that dependency parser can provide enough information for inserting particles.

4 Experiments

We conducted experiments to assess how our proposed dependency-based pre-reordering for Chinese (DPC) impacts on translation quality, and compared it to a baseline phrase-based system and a Refined-HFC pre-reordering for Chinese to Japanese translation.

We used two Chinese-Japanese training data

	News		CWMT+News	
	BLEU	RIBES	BLEU	RIBES
Baseline	39.26	84.83	38.96	85.01
Ref-HFC	39.22	84.88	39.26	84.68
DPC	**39.93**	**85.23**	**39.94**	**85.22**

Table 3: Evaluation of translation quality of two test sets when CWMT, News and the combination of both corpora were used for training.

sets of parallel sentences, namely an in-house-collected Chinese-Japanese news corpus (News), and the News corpus augmented with the CWMT (Zhao et al., 2011) corpus. We extracted disjoint development and test sets from News corpus, containing $1,000$ and $2,000$ sentences respectively. Table 2 shows the corpora statistics.

We used MeCab [4] (Kudo and Matsumoto, 2000) and the Stanford Chinese segmenter [5] (Chang et al., 2008) to segment Japanese and Chinese sentences. POS tags of Chinese sentences were obtained using the Berkeley parser [6] (Petrov et al., 2006), while dependency trees were extracted using Corbit [7] (Hatori et al., 2011). Following the work in (Han et al., 2012), we re-implemented the Refined-HFC using the Chinese Enju to obtain HPSG parsing trees. For comparison purposes with the work in (Isozaki et al., 2010b), particle seed words were inserted at a preprocessing stage for Refined-HFC and our DPC method.

DPC and Refined-HFC pre-reordering strategies were followed in the pipeline by a standard Moses-based baseline system (Koehn et al., 2007), using a default distance reordering model and a lexicalized reordering model "msd-bidirectional-fe". A 5-gram language model was built using SRILM (Stolcke, 2002) on the target side of the corresponding training corpus. Word alignments were extracted using MGIZA++ (Gao and Vogel, 2008) and the parameters of the log-linear combination were tuned using MERT (Och, 2003).

Table 3 summarizes the results of the Baseline system (no pre-reordering nor particle word insertion), the Refined-HFC (Ref-HFC) and our DPC method, using the well-known BLEU score (Papineni et al., 2002) and a word order sensitive metric named RIBES (Isozaki et al., 2010a).

As it can be observed, our DPC method obtains around 0.7 BLEU points of improvement when compared to the second best system in both corpora. When measuring the translation quality in terms of RIBES, our method obtains an improvement of 0.3 and 0.2 points when compared to the second best system in News and CWMT + News corpora, respectively. We suspect that corpus diversity might be one of the reasons for Refined-HFC not to show any advantage in this setting.

We tested the significance of BLEU improvement for Refined-HFC and DPC when compared to the baseline phrase-based system. Refined-HFC tests obtained p-values 0.355 and 0.135 on News and CWMT + News corpora, while our proposed DPC method obtained p-values 0.002 and 0.0, which indicates significant improvements over the phrase-based system.

5 Conclusions

In the present paper, we have analyzed the differences in word order between Chinese and Japanese sentences. We captured the regularities of ordering differences between Chinese and Japanese sentences, and proposed a framework to reorder Chinese sentences to resemble the word order of Japanese.

Our framework consists in three steps. First, we identify verbal blocks, which consist of Chinese words that will move all together as a block without altering their relative inner order. Second, we identify the right-most object of the verbal block, and move the verbal block to the right of it. Finally, we identify invariable grammatical particles in the original vicinity of the verbal block and move them relative to their dependency heads.

Our framework only uses the unlabeled dependency structure of sentences and POS tag information of words. We compared our system to a baseline phrase-based SMT system and a refined head-finalization system. Our method obtained a Chinese word order that is more similar to Japanese word order, and we showed its positive impact on translation quality.

6 Discussion and future work

In the literature, there are mainly two types of parsers that have been used to extract sentence structure and guide reordering. The first type corresponds to parsers that extract phrase structures (i.e. HPSG parsers). These parsers infer a rich

[4] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[5] http://nlp.stanford.edu/software/segmenter.shtml

[6] http://nlp.cs.berkeley.edu/Software.shtml

[7] http://triplet.cc/software/corbit

		News		CWMT+News	
		Chinese	Japanese	Chinese	Japanese
Training	Sentences	342,050		621,610	
	Running words	7,414,749	9,361,867	9,822,535	12,499,112
	Vocabulary	145,133	73,909	214,085	98,333
News Devel.	Sentences	1,000		–	
	Running words	46,042	56,748	–	–
	Out of Vocab.	255	54	–	–
News Test	Sentences	2,000		–	
	Running words	51,534	65,721	–	–
	Out of Vocab.	529	286	–	–

Table 2: Basic statistics of our corpora. News Devel. and News Test were used to tune and test the systems trained with both training corpora. Data statistics were collected after tokenizing and filtering out sentences longer than 64 tokens.

annotation of the sentence in terms of semantic structure or phrase heads. Other reordering strategies use a different type of parsers, namely dependency parsers. These parsers extract dependency information among words in the sentence, often consisting in the dependency relation between two words and the type of relation (dependency label).

Reordering strategies that use syntactic information have proved successful, but they are likely to magnify parsing errors if their reordering rules heavily rely on abundant parse information. This is aggravated when reordering Chinese sentences, due to its loose word order and large variety of possible dependency labels.

In this work, we based our study of ordering differences between Chinese and Japanese solely on dependency relations and POS tags. This contrasts with the work in (Han et al., 2012) that requires phrase structures, phrase-head information and POS tags, and the work in (Xu et al., 2009) that requires dependency relations, dependency labels and POS tags.

In spite of the fact that our method uses less syntactic information, it succeeds at reordering sentences with reported speech even in presence of punctuation symbols. It is worth saying that reported speech is very common in the news domain, which might be one of the reasons of the superior translation quality achieved by our reordering method. Our method also accounted for ordering differences in serial verb constructions, complementizers and adverbial modifiers, which would have required an increase in the complexity of the reordering logic in other methods.

To the best of our knowledge, dependency parsers are more common than HPSG parsers across languages, and our method can potentially be applied to translate under-resourced languages into other languages with a very different sentence structure, as long as they count with dependency parsers and reliable POS taggers.

Implementing our method for other languages would first require a linguistic study on the reordering differences between the two distant language pairs. However, some word ordering differences might be consistent across SVO and SOV language pairs (such as verbs going before or after their objects), but other ordering differences may need special treatment for the language pair under consideration (i.e. Chinese “bei” particles).

There are two possible directions to extend the present work. The first one would be to refine the current method to reduce its sensitivity to POS tagging or dependency parse errors, and to extend our linguistic study on ordering differences between Chinese and Japanese languages. The second direction would be to manually or automatically find common patterns of ordering differences between SVO and SOV languages. The objective would be then to create a one-for-all reordering method that induces monotonic word alignments between sentences from distant language pairs, and that could also be easily extended to account for the unique characteristics of the source language of interest.

Acknowledgments

We would like to thank Dr. Takuya Matsuzaki for his precious advice on this work and Dr. Jun Hatori for his support on using Corbit.

References

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proc. of the 3rd Workshop on SMT*, pages 224–232.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D Manning. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proc. of the Third Workshop on Syntax and Structure in Statistical Translation*, pages 51–59.

Angela Downing and Philip Locke. 2006. *English grammar: a university course*. Routledge.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proc. of Machine Translation Summit XI*, pages 215–222.

Dan Han, Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2012. Head finalization reordering for Chinese-to-Japanese machine translation. In *Proc. of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 57–66.

Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Junichi Tsujii. 2011. Incremental joint POS tagging and dependency parsing in Chinese. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 1216–1224.

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

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proc. of ACL '07, Demonstration Sessions*, pages 177–180.

Taku Kudo and Yuji Matsumoto. 2000. Japanese dependency structure analysis based on support vector machines. In *Proc. of the EMNLP/VLC-2000*, pages 18–25.

Charles N Li and Sandra Annear Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proc. of ACL*, page 720.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. of the 21st COLING and the 44th ACL*, pages 433–440.

Carl Jesse Pollard and Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. The University of Chicago Press and CSLI Publications.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of the 7th international conference on Spoken Language Processing, 2002*, pages 901–904.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proc. of the 2007 Joint Conference on EMNLP-CoNLL*, pages 737–745.

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2011. Extracting pre-ordering rules from predicate-argument structures. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 29–37.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proc. of the 20th international conference on Computational Linguistics*.

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 HLT: NA-ACL 2009*, 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.

Hong-Mei Zhao, Ya-Juan Lv, Guo-Sheng Ben, Yun Huang, and Qun Liu. 2011. Evaluation report for the 7th China workshop on machine translation (CWMT2011). *The 7th China Workshop on Machine Translation (CWMT2011)*.