Analyzing the Influence of Parsing Errors on Pre-reordering Performance for SMT

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

1 Introduction

Word alignment for long distance language pairs is problematic in state-of-the-art phrase-based statistical machine translation. Linguistically motivated reordering models have been widely studied to conquer this challenge. One of the most popular and effective methods is called pre-reordering, where words in sentences from the source language are re-arranged with the objective to resemble the word order of the target language. There are mainly two ways to formulate re-arranging rules. One is to learn automatically from the data (Xia and McCord, 2004; Genzel, 2010); while another one is to handcraft reordering rules based on linguistic studies (Isozaki et al., 2010; Han et al., 2012; Han et al., 2013). In both methods, syntactic information are obtained by using automatic parsers. However, although these parsers produce parsing errors, current reordering methods do not include any parsing error identification or correction mechanism. In order to improve the robustness of pre-reordering method, it is useful to explore the relationship between parsing and reordering.

In this work, we use both empirical and descriptive approaches to analyze the effects of parsing errors on pre-reordering performance for Chinese-to-Japanese statistical machine translation. We examine the impact of parsing errors that are produced by a dependency parser called Corbit[1] (Hatori et al., 2011), on a pre-reordering framework called unlabeled dependency parsing based pre-reordering for Chinese (DPC) (Han et al., 2013). We not only quantify the distribution of general parsing errors along with reordering quality, but also examine the influence of concrete parsing errors on reordering.

2 Gold Data

In order to reveal how and which parsing errors influence reordering, we contrast the reordering based on error-free parse trees, which are considered as Gold-Trees, with the reordering based on parse trees that are generated by parser, which are referred as Auto-Trees. We select sentences from Chinese Penn Treebank ver.7.0 (CTB-7) to build up gold data. CTB-7 is a human annotated treebank that comprises sentences from five genres: broadcast news (BN), broadcast conversations (BC), news magazine (NM), newswire (NS), and newsgroup weblogs (NW). Since our parser was trained on this corpus, in order to maximize the analysis accuracy, we obtain sentences from the development set (Wang et al., 2011).

We first randomly sampled 517 unique sentences (set-1) from all five genres. But sentences in BC and NW tend to include faults like repetitions, incomplete sentences, corrections, and so on. We thus randomly selected another $2,126$ unique sentences (set-2) from other three genres: NS, NM, and BN. Table 1 lists the statistics of the selected sentences in five genres.

To obtain Japanese references, professional human translators translated all the sentences in set-1 and set-2. Based on the Japanese references, a bilingual speaker reordered Chinese

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

	BN	BC	NM	NS	NW	Total
set-1	100	100	100	117	100	517
set-2	797	-	578	751	-	$2,126$
Total	897	100	678	868	100	$2,643$
AL.	29.8	20.0	33.5	28.4	25.9	29.8
Voc.	$5.5K$	690	$5K$	$5.1K$	972	$9.5K$

Table 1: Statistics of Selected Sentences in Five Genres of CTB-7. AL. stands for the average length of sentences, while Voc. for vocabulary.

sentences in set-1 to follow the word order of their Japanese counterparts. After constructing these two data sets, we use Penn2Malt[2] (Nivre, 2006) to convert CTB-7 sentences to dependency Gold-Trees. We added three new head rules for Penn2Malt converting as follows: FLR (Fillers) and DFL (Disfluency) head on right-hand branch; INC (Incomplete sentences) follows the same head rule as FRAG (Fragment).

3 Analysis Method

We first use an empirical approach to provide a general idea of the sensitiveness of DPC pre-reordering method on parsing errors, and also test an upper bound of the reordering performance. Then, we dig more on the effects of concrete parsing errors from the aspect of Part-of-Speech tag (POS tag) to reveal what types of parsing errors influence the most. In both stages, we calculate Kendall's tau (τ) rank correlation coefficient (Kendall, 1938) to evaluate the reordering quality (Isozaki et al., 2010).

3.1 General analysis

In the first stage, we set up two analysis scenarios since we have one gold reordered Chinese data set and one Japanese references set, which can be used as two benchmarks. In scenario 1, manually reordered Chinese sentences are used as the benchmark. We measure the word order similarities which reveal the reordering quality in two sentence pairs. One pair is composed of the benchmark sentences and the Gold-Tree based reordered sentences, while another pair is composed of the benchmark sentences and the Auto-Tree based reordered sentences. By comparing the measurement results, we quantify the extent of parsing errors that influence reordering. Meantime, the former measurement gives a general idea of the upper bound of the reordering method. In scenario 2, instead of using handcrafted reordered Chinese, Japanese references are used as the benchmark and we calculate the word order similarities in the same way as in scenario 1.

In both scenarios, we carry out reordering method DPC. Accordingly, there are two automatically reordered data sets that are produced by two reordering systems: Gold-Tree based reordering system (Gold-DPC) and Auto-Tree based reordering system (Auto-DPC). Auto-Trees are automatically generated by Corbit. The baseline system uses unreordered Chinese sentences. Due to the fact that the reordering method is identical but the Auto-Trees may contain errors, we will be able to observe reordering differences directly caused by parsing errors.

Scenario 1 Although there are totally 517 sentences in set-1, 26 sentences were failed during the converting from CTB-7 parsed text to parse trees. For comparison, 491 available (Gold- and Auto-) dependency trees are used to reorder sentences. Our first observation on the effects of parsing errors to reordering performance is to examine word order similarities between manually reordered Chinese sentences and automatically reordered Chinese sentences. Figure 1 shows the distribution of τ values of the 491 sentences. Comparing to baseline, both Gold-DPC and Auto-DPC show higher average τ values which imply that DPC have positively reordered the Chinese sentences and improved the word alignment between Chinese and Japanese. However, comparing with Gold-DPC, Auto-DPC has reordered more sentences with lower τ value and less sentences with high τ, which shows that DPC is sensitive to parsing errors. Since the sentence number of set-1 is limited, in order to enhance the conclusions, we increased the test data by adding set-2 for further experiment in scenario 2.

Scenario 2 We merge set-1 and set-2, then use Japanese references as benchmark in this scenario. In order to calculate the value of Kendall's tau between Chinese sentences and their Japanese counterparts, we use

[2] http://stp.lingfil.uu.se/ nivre/research/Penn2Malt.html

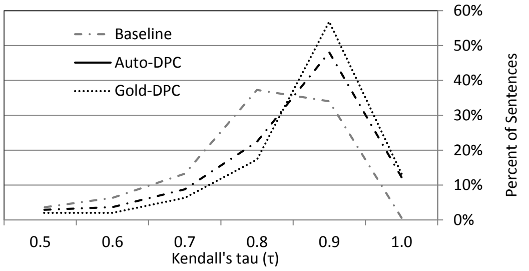


Figure 1: Distribution of τ for 491 sentences.

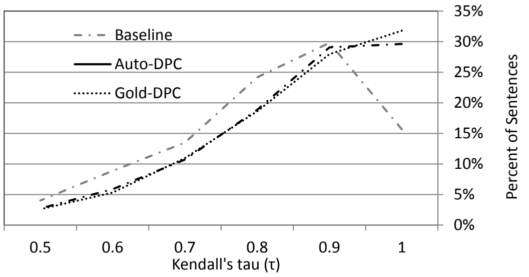
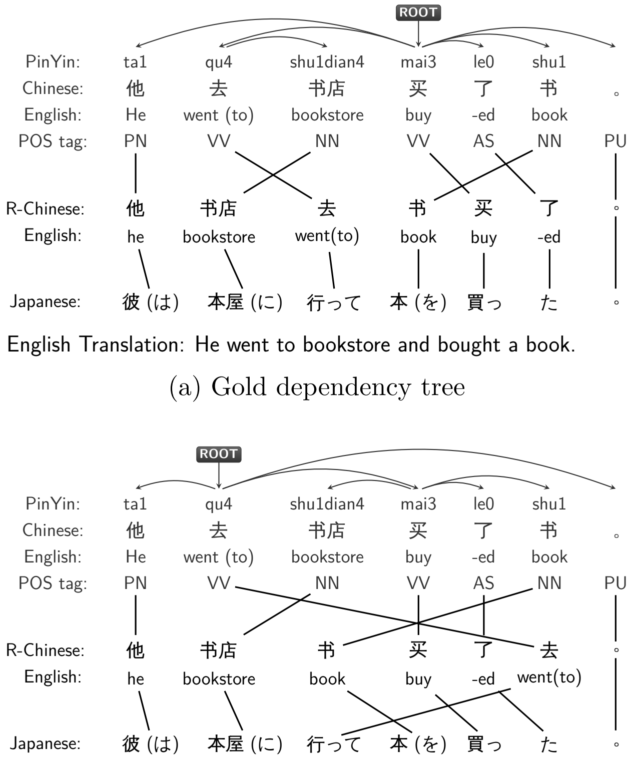


Figure 2: Distribution of τ for 2, 164 sentences.


ROOT
PinYin: ta1 qu4 shu1dian4 mai3 le0 shu1
Chinese: 他 去 书店 买 了 书 。
English: He went (to) bookstore buy -ed book
POS tag: PN VV NN VV AS NN PU
R-Chinese: 他 书店 去 书 买 了 。
English: he bookstore went(to) book buy -ed
Japanese: 彼 (は) 本屋 (に) 行って 本 (を) 買っ た 。
English Translation: He went to bookstore and bought a book.
ROOT
PinYin: ta1 qu4 shu1dian4 mai3 le0 shu1
Chinese: 他 去 书店 买 了 书 。
English: He went (to) bookstore buy -ed book
POS tag: PN VV NN VV AS NN PU
R-Chinese: 他 书店 书 买 了 去 。
English: he bookstore book buy -ed went(to)
Japanese: 彼 (は) 本屋 (に) 行って 本 (を) 買っ た 。

(a) Gold dependency tree

(b) Possible wrong dependency parse tree

Figure 3: Example for calculating parsing errors in terms of POS tag.

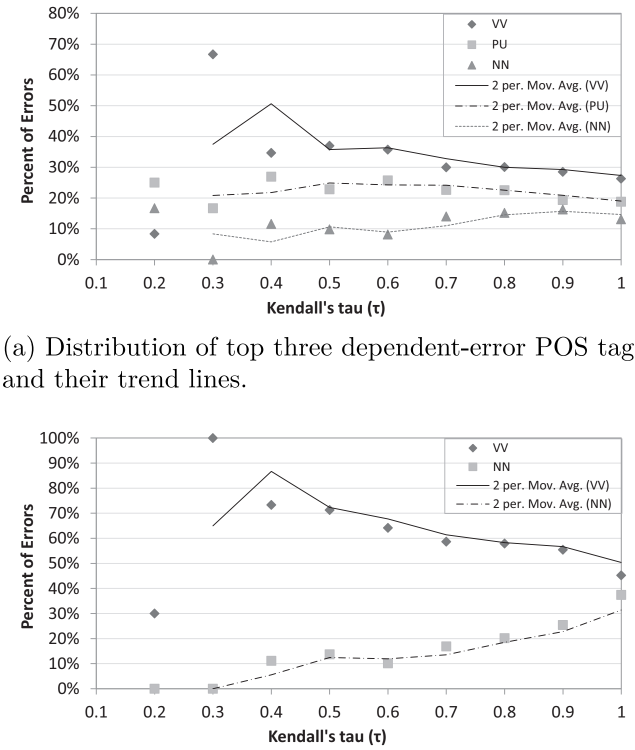
MGIZA++ to obtain the alignment file, *ch-ja.A3.final*. Therefore, the comparison implies how monotonically the Chinese sentences have been reordered to align with their Japanese references. There are totally 2, 164 available (Gold- and Auto-) trees. Figure 2 shows the distribution of τ values and baseline system was using unreordered Chinese sentences.

In Figure 2, disregarding the alignment accuracy of MGIZA++, both Gold-DPC and Auto-DPC produced more sentences with higher τ values against the baseline system, which indicate the same conclusion that DPC improved word alignment as shown in scenario 1. However, comparing with Gold-DPC, Auto-DPC has about 2% drop in $0.9 < \tau <= 1$, which shows that reordering method DPC is sensitive to parsing errors. Furthermore, the performance of reordering system Gold-DPC sketches the figure of the upper bound of DPC.

3.2 Dependency errors by POS tag

In this stage, we define two types of dependency parsing errors given by POS tags, namely dependent-error and head-error. As a dependent-error, a token points to a wrong head, whereas a head-error means that a token is wrongly recognized as a head. We collect all the POS tags of the tokens which are involved in either parsing error in sentences, and then calculate the proportion of different POS tags in terms of the parsing error types. Figure 3 shows an example of a gold parse tree (Fig. 3a) with an error-containing tree (Fig. 3b). By comparing these two parse trees, tokens of “他 (he, PN)”, “去 (went, VV)”, “书店 (bookstore, NN)”, “买 (buy, VV)”, and “。(., PU)” are dependent-errors, since they point to wrong heads. “去 (went, VV)” and “买 (buy, VV)” are head-errors since they are wrongly recognized as heads. Therefore, in this sentence, dependent-error includes twice of VV and once of PN, NN, PU; head-error includes twice of VV.

With the objective of observing the effects of parsing errors on reordering, we group sentences based on their τ values according to Figure 2, and plot the distributions of POS tags that are involved in parsing errors for dependent-error and head-error separately. Figure 4a gives the distribution of three most frequent POS tags that point to wrong heads, and Figure 4b shows



(a) Distribution of top three dependent-error POS tags and their trend lines.

(b) Distribution of top two head-error POS tags and their trend lines.

Figure 4: Distribution of top influential error POS tags for two types of errors.

the distribution of two most frequent POS tags that are wrongly recognized as heads. VV represents most of verbs except a few exceptions, such as predicative adjective and copula. NN represents most of nouns except proper noun or temporal noun, etc., and PU is punctuation.

Both figures exhibit that verbal tokens take the largest proportion in both dependent-error and head-error than other types of tokens in low reordering performance sentences. Therefore, the parsing errors on verbal tokens influence more on the reordering quality. On the other hand, parsing errors on nouns do not influence severely on reordering. Moreover, since the proportion of PU involved in dependent-error is consistent, we can not say whether parsing errors on punctuations are influential or not.

4 Conclusion

We carried out linguistically motivated analysis methods by using empirical and descriptive approaches to examine the effects of different parsing errors on pre-reordering performance. We achieved three objectives that we quantified effects of parsing errors on reordering, estimated upper bounds in performance of the pre-reordering method, and examined the effects of specific parsing errors on reordering from the aspect of POS tags. The current study on exploring influential parsing errors is not exhaustive. There are more analysis possibilities, such as parsing errors that are produced by different parsers and their effects on different pre-reordering models. We believe that such relationship analysis between two types of techniques can benefit the improvement of both.

References

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

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 SSST-6*, pages 57–66.

Dan Han, Pascual Martínez-Gómez, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. 2013. Using unlabeled dependency parsing for pre-reordering for Chinese-to-Japanese statistical machine translation. In *Proc. of HyTra*, pages 25–33.

Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun' ichi Tsujii. 2011. Incremental joint POS tagging and dependency parsing in Chinese. In *Proc. of IJCNLP*, pages 1216–1224.

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

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Joakim Nivre. 2006. *Inductive dependency parsing.* Springer.

Yiou Wang, Jun' ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. 2011. Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *Proc. of IJCNLP-5*, pages 309–317.

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