Integrating empty category detection into Preordering Machine Translation

Shunsuke Takeno[†] **Masaaki Nagata**[‡]**, Kazuhide Yamamoto**[†]
[†]Nagaoka University of Technology,
1603-1 Kamitomioka, Nagaoka, Niigata, 940-2188 Japan
{takeno, yamamoto}@jnlp.org
[‡]NTT Communication Science Laboratories, NTT Corporation,
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
nagata.masaaki@labs.ntt.co.jp

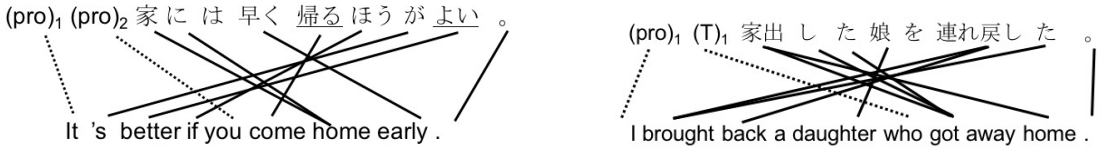
Abstract

We propose a method for integrating Japanese empty category detection into the preordering process of Japanese-to-English statistical machine translation. First, we apply machine-learning-based empty category detection to estimate the position and the type of empty categories in the constituent tree of the source sentence. Then, we apply discriminative preordering to the augmented constituent tree in which empty categories are treated as if they are normal lexical symbols. We find that it is effective to filter empty categories based on the confidence of estimation. Our experiments show that, for the IWSLT dataset consisting of short travel conversations, the insertion of empty categories alone improves the BLEU score from 33.2 to 34.3 and the RIBES score from 76.3 to 78.7, which imply that reordering has improved For the KFTT dataset consisting of Wikipedia sentences, the proposed preordering method considering empty categories improves the BLEU score from 19.9 to 20.2 and the RIBES score from 66.2 to 66.3, which shows both translation and reordering have improved slightly.

1 Introduction

Empty categories are phonetically null elements that are used for representing dropped pronouns ("pro" or "small pro"), controlled elements ("PRO" or "big pro") and traces of movement ("T" or "trace"). Dropped pronouns are one of the major problems caused on machine translation from the pro-drop language such as Japanese to the non-pro-drop language such as English because it is difficult to produce the correct pronouns on the target side when the pronoun is missing on the source side.

The effects of empty categories in machine translation have previously been examined (Chung and Gildea, 2010; Taira et al., 2012; Xiang et al., 2013; Kudo et al., 2014; Wang et al., 2016). In this paper, we address two new problems that were not fully discussed in previous work. The first problem is that, even if empty categories are correctly recovered, it is difficult to automatically obtain the correct word alignment for languages with a completely different word order such as Japanese and English. The second problem is that it is not only difficult to translate non-existent pronouns but also relative clauses because relative pronouns do not exist in Japanese. In theory, we can safely ignore PRO in control structures for translation because they are absent from both Japanese and English.



(a) Noun clause (b) Relative clause

Figure 1: Recovering empty categories makes word alignment more difficult

Fig. 1 shows examples for which there are two empty categories in the source Japanese sentence, which results in complicated word alignments. In Fig. 1(a), the first *pro* should be aligned to "it"

Proceedings of the 3rd Workshop on Asian Translation,
pages 157–165, Osaka, Japan, December 11-17 2016.

because it is the subject of a matrix clause of which the verb is “よい (good)”, whereas the second *pro* should be aligned to “you” because is it the subject of a noun clause of which the verb is “帰る (come home)”. In Fig. 1(b), the first *pro* should be aligned to “I” because it is the subject of matrix clause whose verb is “連れ戻す (bring back)”, while the second *T* could arguably be aligned to the relative pronoun “who” because it is the subject of the relative clause of which the verb is “家出した (ran away from home)”.

This means that inserting empty categories into source sentences could worsen automatic word alignment and result in less accurate machine translation outputs. We solve this problem by integrating empty category detection into preordering-based statistical machine translation. We first insert empty categories into the source sentence, and then reorder them such that the word order is similar to that of the target sentence. We find it is effective to filter out unreliable empty category candidates to improve the accuracy of machine translation. In the following sections, we first briefly describe related works. We then describe empty category detection method (Takeno et al., 2015) and discriminative preordering method (Hoshino et al., 2015) used in the proposed method. We then report experiment results of Japanese-to-English translation on both spoken (IWSLT dataset) and written (KFTT dataset) languages.

2 Related works

Conventional approaches to recover zero pronouns in Japanese are to frame it as zero anaphora resolution, which is a sub-problem of predicate argument structure analysis (Nakaiwa and Ikehara, 1995; Iida et al., 2007; Hangyo et al., 2013). Zero anaphora resolution consists of two procedures: zero pronoun detection and anaphora resolution.

It is difficult to integrate zero anaphora resolution (or predicate-argument structure analysis) into SMT for two reasons. The first is that anaphora resolution requires context analysis, which complicates the translation method. The second is that predicate argument structure analysis provides semantic relations, not syntactic structure. This makes it difficult to use the information of recovered zero pronouns in SMT, because there is no position information for the zero pronouns in the word sequence (for phrase-based translation) or syntactic tree (for tree-based translation).

Only a few studies on the recovery of zero pronouns for Japanese-to-English statistical machine translation have been reported. Taira et al.(2012) reported that recovering zero pronouns in source Japanese sentence, both by human and by simple rule-based methods, improved the accuracy of generating correct personal pronouns in target English sentence. However, they also reported that the BLEU scores remained unchanged in both cases. Kudo et al. (2014) showed that generating zero subjects in Japanese improved the BLEU score in preordering-based translation by about 0.5 points. They designed a specific probabilistic model for dependency-based preordering to generate the subject when it was omitted from the source Japanese sentence.

Chinese also has zero pronoun problems. Based on Chinese Penn Treebank, recovering zero pronouns in Chinese is framed as a sub-problem of empty category detection, and some previous work on applying empty category detection in Chinese-to-English statistical machine translation has been published.

Chung and Gildea (2010) reported that the automatic insertion of empty categories improved the accuracy of phrased-based machine translation. Xiang et al. (2013) proposed a log-linear model for the empty category detection as a post-processor of the constituent parser, and combined it with Hiero and a tree-to-string system. Wang et al. (2016) proposed NN-based unsupervised empty category detection and its integration into phrase-based SMT. Their method succeeded in dialogue corpus in which the difference in the word-order problems between Chinese and English are alleviated compared to written language corpus because both Chinese and English have an SVO grammatical structure in shorter sentence.

Our approach is very close to the Xiang et al., (2013)’s method. We used an empty category detector (Takeno et al., 2015) implemented as a post-processor of a Japanese constituent parser, and combined it with preordering-based translation system (Hoshino et al., 2015). Yet, there are some differences between our work and theirs. We used preordering as a way to improve the word alignment accuracy after empty categories are recovered. We examined the effect of recovering *T* (trace) for the translation of a relative clause.

3 Preordering with empty categories

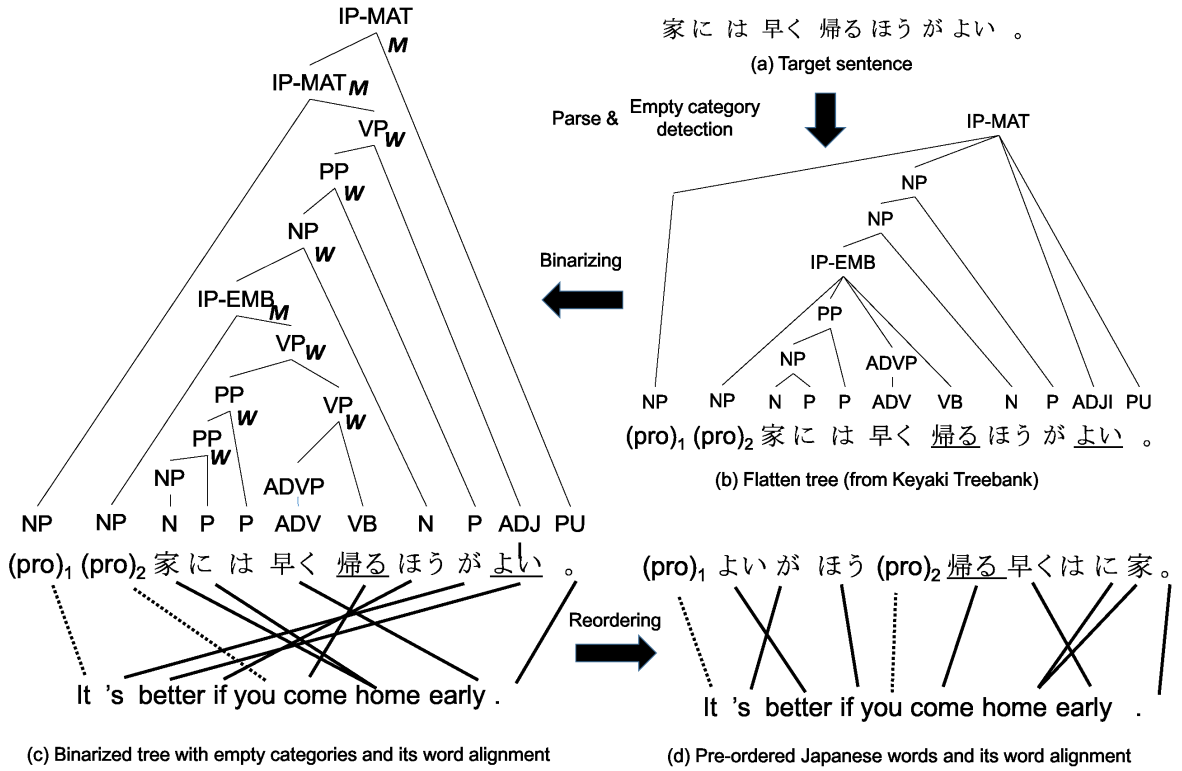


Figure 2: Progress to integrate empty category detection into machine translation. In Fig.2(b), we annotate reordering mark.W indicates that branches are to be swapped M indicates monotone

Fig. 2 shows the process of preordering with empty category detection for Japanese-to-English translation. We first parse the source Japanese sentence to obtain a constituent parse tree and apply empty category detection to recover empty categories. We then binarize the augmented tree and apply the discriminative preordering model to the binary tree to decide whether the children of each node should be swapped (W=swap) or not (M=monotone). We then obtain reordered Japanese sentence as the yield of the reordered tree. We provide details of each step as follows.

The remainder of this section contains further details of each step.

Japanese constituent parsing and empty category detection

We used a spinal tree-based shift-reduce parser (Hayashi et al., 2016) to obtain a constituent parse tree for the source Japanese sentence. It is trained on the Keyaki Treebank and outputs flat phrase structure as shown in Fig. 2(b). As this parser ignores empty categories, we used a log-linear model-based empty category detector (Takeno et al., 2015) to recover empty categories. The parser can detect two empty category types: dropped pronouns *pro* and the trace of the movement of the noun phrase (NP) *T*. Although the original Keyaki Treebank has sub-categories of *pro*, such as *speaker* and *hearer*, we unified them into *pro* as was done in our previous work of Takeno et al., (2015).

We used a rule-based tree binarization tool [1] provided with the Keyaki Treebank to convert a flat tree as shown in Fig. 2(b) to a binary tree as shown in Fig. 2(c).

Building preordering model with empty categories

We extended Hoshino et al., (2015)'s preordering method to process empty categories in the source Japanese sentence. According to Hoshino et al., (2015), they build a classifier for each node in the

[1] http://www.compling.jp/haruniwa/#create_stripped

source tree to decide whether its children need to be swapped. The oracle is decided to maximize the Kendall's τ between the reordered source sentence and the target sentences based on the word alignment between the source and target sentences.

We used two methods to process empty categories in Hoshino et al., (2015)'s preordering method, namely REORDERING(H) and REORDERING(C). The former of these methods trains the preordering model using sentences with a manually constructed word alignment. As the currently available manual word alignment examples do not have alignment information on empty categories, the trained preordering model is agnostic with regard to empty categories. If the input parse tree has an empty category, it is treated as an NP with an unknown lexical symbol.

The latter of these two methods trains the preordering model using sentences with automatic word alignment, which is obtained by using unsupervised word alignment tool GIZA++ (Och, 2007) for the source Japanese sentences with empty categories recovered and the target sentences. If the input parse tree has an empty category, it is treated as a noun phrases with a known lexical symbol.

It is noteworthy that the preordering procedure involves training the translation model on the reordered source sentence as shown in Fig. 2(d). We can expect the word alignment for empty categories is improved by this preordering. This is different from previous approaches such as Xiang et al., (2013), where word alignment is automatically obtained from original source sentences with empty categories recovered.

Filtering out unreliable empty categories

As we report with the experiment, the accuracy of Japanese empty category detection is relatively low, even if we were to use the state-of-the-art Takeno et al., (2015)'s method. Therefore, we modified this method to filter out unreliable empty categories.

Let $T = t_1 t_2 \cdots t_n$ be the sequence of nodes produced by the post-order traversal of the parse tree from its root node, and e_i be the empty category tag associated with t_i. In Takeno et al., (2015), the empty category tag for each node is decided by the following log-linear model:

$$\hat{e_i} = \arg\max_{e \in \mathcal{E}} P(e|e_1^{i-1}, T) = \arg\max_{e \in \mathcal{E}} \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(e, e_1^{i-1}, T))}{Z(e_1^i, T)}$$

where $\mathcal{E}$ represents the set of all empty category types to be detected including NULL label (in our case, either *pro*, *T*, or NULL).

The above equation means that an empty category is inserted if its probability is larger than that of the NULL label. We modified the decision function so that, for a given threshold θ, we remove empty categories if its probability is less than θ:

$$\hat{e_i} = \begin{cases} NULL & if \quad \arg\max_{e \in \mathcal{E}} P(e|e_1^{i-1}, T) < \theta \\ \arg\max_{e \in \mathcal{E}} P(e|e_1^{i-1}, T) & otherwise \end{cases}$$

The threshold θ is decided using development set on experiment.

4 Experiments

4.1 Empty category detection (before filtering)

We trained and evaluated the empty category detection model, following Takeno et al., (2015) settings.

We used the Keyaki Treebank as of November 15, 2015, which included 30,872 annotated sentences. We used 1,000 sentences as the development set, and 1,003 sentences as the test set. These sentences were taken from the files blog_KNB.psd (blog), spoken_CIAIR.psd (transcript), newswire_MAINICHI-1995.psd (newswire) to balance the domain. The remaining 20,646 sentences were used for training. We used GloVe as word embedding, and Wikipedia articles in Japanese as of January 18, 2015, were used for training, which amounted to 660 million words and 23.4 million sentences. We used the development set to decide the dimension of word embedding and the window size for co-occurrence counts as 200 and 10, respectively.

We performed the tests under two conditions: gold parse and system parse. Under the gold parse condition, we used trees from Keyaki Treebank without empty categories as input to the systems. Under the system parse condition, we used the output of the spinal tree-based shift-reduce parser (Hayashi et al., 2016).

We evaluated these conditions using the word-position-level identification metrics described in Xiang et al.,(2013). This approach projects the predicted empty category tags to the surface level. An empty node is regarded as correctly predicted surface position in the sentence if and only if type (*T* or *pro*) and function (SBJ, OB1 and so on) matches with the reference.

The results are presented in Table 1. For *pro* and *T*, the detector achieved 74.9%, 91.9% in F scores, respectively, under the gold parse condition. However, the performance of detector is reduced considerably under the system parse condition. In particular, the decline in the accuracy of *T* is remarkable. These tendencies are the same as described in Takeno et al., (2015).

types	INPUT	P	R	F
pro	GOLDEN	74.3	75.6	74.9
T	GOLDEN	89.0	95.0	91.9
pro	SYSTEM	60.9	66.2	63.4
T	SYSTEM	50.0	42.2	45.8

Table 1: Empty category detection results[%]

4.2 Effects of empty categories on Machine Translation

Datasets and Tools

We tested the proposed method on two Japanese-to-English translation tasks; one of which involved the IWSLT dataset, which was provided during the International Workshop on Spoken Language Translation in 2005 (Eck and Hori, 2005). The dataset contains 19,972 sentences for training, 500 sentences for tuning, and 1,000 sentences for testing. Although this dataset is small, it is appropriate for evaluating the effectiveness of the proposed method since a spoken language corpus generally has many empty categories. In particular, *pro* appears very often. The other dataset is the Kyoto Free Translation Task corpus, the so called KFTT (Neubig, 2011). The KFTT is made from the "Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles", which is created by manually translating Japanese Wikipedia articles related to Kyoto City into English. The dataset consists of 440,000, 1,235, and 1,160 sentences for training, tuning, and testing, respectively.

We built the preordering model by applying the empty category detection method to source Japanese sentences to obtain syntactic trees with empty categories, as described in the previous section. We achieved this by first tokenizing Japanese sentences by using a CRF-based tokenizing and chunking software (Uchimoto and Den, 2008) to obtain the long unit words required by the Japanese parser (Hayashi et al., 2016). We then achieved word alignment by using short unit words in Japanese obtained by using the MeCab morphological analyzer with the UniDic dictionary[2].

For the Japanese-to-English translation experiment, we used a phrase-based translation model (Koehn et al., 2007). For all systems we compared, the language model is a 5-gram KenLM (Heafield, 2011), which uses modified Kneser-Ney smoothing and tuning is performed to maximize the BLEU score using minimum error rate training (Och, 2007). Other configurable setting of all tool use default values unless otherwise stated.

We compared three translation methods, each with and without empty category detection. BASELINE is a phrase-based machine translation system (Moses) (Koehn et al., 2007) which consists of training data comprising a bilingual dataset without preordering. REORDERING(H) and REORDERING(C) are described in the previous section. For REORDERING(H), 5,319 sentences with manual word alignment

[2]http://taku910.github.io/mecab/unidic

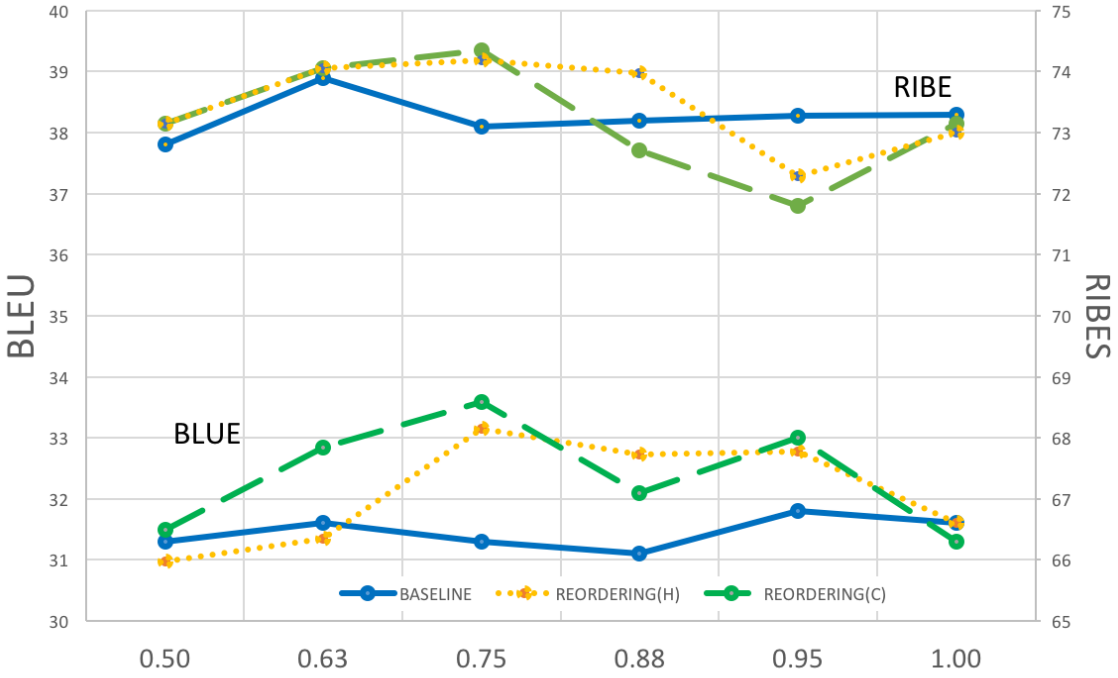


Figure 3: Characteristic of machine translation evaluation scores to empty categories filtered for development set of the IWSLT dataset

is used. These systems are equivalent to Hoshino et al., (2015)'s method. They are taken from both the spoken language (CSJ) and written (KTC) language corpus.

As for evaluation measures, we use the standard BLEU (Papineni et al., 2002) as well as RIBES (Isozaki et al., 2010), which is a rank correlation based metric that has been shown to be highly correlated with human evaluations of machine translation systems between languages with a very different word order such as Japanese and English.

Result of filtering empty categories

In this experiment, we search for the best threshold value to filter out empty categories in Sec 3. Changing the threshold values θ from 0.50 to 1.0, we measure both BLEU and RIBES, where $\theta = 1.0$ corresponds to the result produces by machine translation systems trained from a dataset without empty categories. When decoding the text into English, we set the distortion limit to 20 in all systems.

The result of the IWSLT dataset is shown in Fig. 3. It indicates that the threshold that are used to filter out empty categories affect the result of machine translation accuracy and that setting the threshold appropriately improves the result. For REORDERING(C), we achieved 33.6 in BLEU and 74.4 in RIBES for a threshold value of $\theta = 0.75$.

Fig. 3 shows that the BLEU score generally decreases as the threshold value is lowered. In particular for REORDERING(H), the BLEU score drops dramatically for lower threshold value. A decrease in the threshold value signifies an increase in the number of empty categories inserted into source languages and REORDERING(H) does not consider empty categories explicitly on reordering. Therefore, we suspect that REORDERING(H) tends to locate empty categories in unfavorable places.

The tendency displayed by BLEU and RIBES are differs for lower threshold values; BLEU decreases dramatically, whereas RIBES is reduced moderately. We consider the difference to be caused by their definitions: BLEU is sensitive to the word choice while RIBES is sensitive to the word order.

Inserting an element in the source sentence could result in inserting some words in the target sentence. The change directly could affect word-sensitive metrics such as BLEU, but it does not necessarily affect order-sensitive metrics such as RIBES, since RIBES changes only when the same word appears in both

the decoded sentence and the reference sentence.

Result of machine translation of empty categories

	IWSLT-2005 JE dataset				KFTT JE dataset			
	BLEU		RIBES		BLEU		RIBES	
METHODS	dl=6	dl=20	dl=6	dl=20	dl=6	dl=20	dl=6	dl=20
BASELINE w/o EC	29.6	33.1	73.6	74.2	17.9	18.5	62.4	66.4
BASELINE w/ EC	29.2	33.6	74.1	75.7	18.1	18.6	62.5	65.4
REORDERING(C) w/o EC	29.4	33.2	74.6	76.3	19.2	19.3	64.8	65.7
REORDERING(C) w/ EC	29.6	**34.3**	**75.8**	**78.8**	19.4	19.8	65.2	66.0
REORDERING(H) w/o EC	**29.7**	33.8	74.1	76.8	19.3	19.9	65.2	66.2
REORDERING(H) w/ EC	29.3	34.1	75.6	78.6	**19.5**	**20.2**	**65.5**	**66.3**

Table 2: machine translation results with empty categories. dl means the distortion limit. EC indicates empty categories are detected in dataset

We compared the machine translation accuracies between the baseline systems and the proposed systems integrated with empty category detection. In all experiments, we set the threshold value to 0.75 to remove unreliable empty categories by filtering. Table 2 shows the results for IWSLT dataset and KFTT datasets.

In the result for the IWSLT dataset, we find that empty category detection improves both of the metrics RIBES and BLEU in each system when the distortion limit is set to 20. Empty category detection increases the BLEU score by +0.5, +1.1, and +0.3 points for BASELINE, REORDERING(C) (empty categories are preordered as known words) and REORDERING(H) (empty categories are preordered as unknown words), respectively. As for the RIBES metrics, it increases +1.5 points, +2.5 points and +1.8 points respectively. The best result we achieved was 34.3 in the BLUE score and 78.8 for the RIBES score when REORDERING(C) with empty categories was used.

The result for the KFTT dataset showed that integration of empty category detection into the preordering model slightly improves both of the metrics RIBES and BLEU in each system when the distortion limit is set to 6. Empty category detection has slightly bad effect on the BLEU score when the distortion limit is set to lower value. The differences are +0.1 point, +0.2 point and 0.2 point for BASELINE, REORDERING(C) and REORDERING(H) respectively. The RIBES metrics increase by +0.1 point, +0.4 point and +0.3 point respectively. The best result we achieved was 20.2 in BLEU score and 66.3 for the RIBES score when REORDERING(H) with empty categories was used.

Empty category detection considerably improves the IWSLT dataset, which is a spoken language corpus, whereas it moderately improves the KFTT dataset, which is a written language corpus. Although the improvement resulting from inserting empty categories into REORDERING(H), of which the empty categories are regarded as unknown words, is +0.3 points in BLEU and 1.8 points in RIBES for the IWSLT dataset, the improvement of inserting empty categories in REORDERING(C) is +0.5 points in BLEU and +0.3 point RIBES. This shows that a preordering method which considers empty categories has a slightly better.

Finally, we include several translation samples in Table 3 to illustrate the translation errors caused by empty categories. The insertion of empty categories enables us to improve the translation if there are missing elements on the source side. The first and second sample showed that we can obtain additional grammatical output by making null elements explicit.

Some problems remain to be solved on the translation of empty categories. One of them is the excessive insertion of empty categories as we mentioned in our experiment. Filtering unreliable empty category candidates enables us to alleviate the problem. However, we expect to improve the translation accuracy by using both source and target contexts for filtering. Another major problem is the inference of the attribute of empty categories such as the person, gender, and number. The last example in Table 3 necessity of inferring the person information of *pro*.

Success translation	
Reference	i 'm in a hurry .
Source(EC)	*pro* 急い で いる ん です 。
NO EC	are in a hurry .
ECs	i 'm in a hurry .
Reference	how much to rent it for three days ?
Source	*pro* 三 日間 借りる と いくら に なり ます か 。
Reordered Source	*pro* いくら に ます なり と 借りる 三 日間 か 。
NO ECs	i have a three days and how much will it be ?
ECs	i have a three days and how much will it be ?
Pre-ordered w/o EC	what would you like to hire and three days .
Pre-ordered w/ EC	how much will it cost to three days ?
Failed translation	
Reference	do you have any fruits or plants ?
Source	*pro* 果物 や 植物 を 持っ て い ます か 。
Reordered Source	*pro* い て 持っ ます 果物 や 植物 を か 。
NO ECs	i have a carrying any plants and fruits ?
ECs	i have fruit or plant ?
Pre-ordered w/o EC	do you have some fruit or plants ?
Pre-ordered w/ EC	i have a carrying any plants and fruits ?

Table 3: Translation Examples

5 Conclusion

In this paper, we propose a method to integrate empty category detection into preordering-based machine translation system.

We examined the effect of empty category detection on both the IWSLT and KFTT datasets. We showed by experiments that empty category detection results in an improvement in machine translation, in particular for the IWSLT dataset, which is a spoken language corpus. We also showed that, by using preordering with empty categories, we were able to achieve consistent improvement in translation accuracy for the KFTT dataset.

In future, we would like to improve the filtering strategy for empty categories. The integration of empty categories into machine translation is problematic in that empty categories are inserted excessively. There are some empty categories that are not aligned to any words in the target language. In this work, we simply filtered these empty categories based on the probability to alleviate the problem. However, for addressing this problem more appropriately, we should consider both source language context and target language context. We expect the corpus-based approach such as Wang et al., (2016) address this problem.

References

Tagyoung Chung and Daniel Gildea. 2010. Effects of Empty Categories on Machine Translation. In *Proceedings of the EMNLP2010*, pages 636–645.

Matthias Eck and Chiori Hori. 2005. Overview of the iwslt 2005 evaluation campaign. In *Proceedings of the IWSLT2005*, pages 1–22.

Masatugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2013. Japanese Zero Reference Resolution Considering Zero Exophora and Author/Reader Mentions. In *Proceedings of the EMNLP2013*, pages 924–934.

Katsuhiko Hayashi, Jun Suzuki, and Masaaki Nagata. 2016. Shift-reduce spinal tag parsing with dynamic programming. *Transactions of the Japanese Society for Artificial Intelligence*, 31(2):1–8.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, Katsuhiko Hayashi, and Masaaki Nagata. 2015. Discriminative preordering meets kendall 's tau maximization. In *Proceedings of the ACL-IJCNLP2015*, pages 139–144.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing*, 6(4):1–22.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the EMNLP2010*, pages 944–952.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL2007*, pages 177–180.

Taku Kudo, Hiroshi Ichikawa, and Hideto Kazawa. 2014. A joint inference of deep case analysis and zero subject generation for Japanese-to-English statistical machine translation. In *Proceedings of the ACL2014*, pages 557–562.

Hiromi Nakaiwa and Satoru Ikehara. 1995. Intrasentential Resolution of Japanese Zero Pronouns using Pragmatic and Semantic Constraints. In *Proceeedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 96–105.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Franz Josef Och. 2007. Minimum error rate training in statistical machine translation. In *Proceedings of the ACL2007*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the ACL2002*, pages 311–318.

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero pronoun resolution can improve the quality of je translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 111–118.

Shunsuke Takeno, Nagata Masaaki, and Kazuhide Yamamoto. 2015. Empty Category Detection using Path Features and Distributed Case Frames. In *Proceedings of the EMNLP2015*, pages 1335–1340.

Kiyotaka Uchimoto and Yasuharu Den. 2008. Word-level dependency-structure annotation to corpus of spontaneous japanese and its application. In *Proceedings of the LREC2008*, pages 3118–3122.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A Novel Approach to Dropped Pronoun Translation. In *Proceedings of the NAACL-HLT2016*, pages 983–993.

Bing Xiang, Xiaoqiang Luo, and Bowen Zhou. 2013. Enlisting the ghost: Modeling empty categories for machine translation. In *Proceedings of the ACL2013*, pages 822–831.