Source Phrase Segmentation and Translation for Japanese-English Translation Using Dependency Structure

Junki Matsuo
Graduate School of System Design, Tokyo Metropolitan University, Japan
matsuo-junki@ed.tmu.ac.jp

Kenichi Ohwada
Graduate School of System Design, Tokyo Metropolitan University, Japan
ohwada-kenichi@ed.tmu.ac.jp

Mamaru Komachi
Graduate School of System Design, Tokyo Metropolitan University, Japan
komachi@tmu.ac.jp

Abstract

There are various approaches to statistical machine translation (SMT). In particular, phrase-based SMT (PBSMT) is used as a de facto standard for many language pairs because it works robustly across languages and it is easy to implement. However, the results of PBSMT can include ungrammatical sentences, since it typically does not take syntactic structure into account. To overcome this problem, we propose a linguistically motivated approach based on segmenting a source phrase using a dependency structure and translating each phrase with PBSMT. This paper presents the results of our method on Japanese-English translation and discusses potential improvements.

1 Introduction

It is difficult for statistical machine translation (SMT) to perform translation between languages such as Japanese and English, which have a systematic difference in their word orders: typically, Japanese is a subject-object-verb (SOV) language, whereas English is a subject-object-verb (SVO) language.

Although PBSMT is used as a de facto standard for many language pairs because it works robustly across languages and it is easy to implement; it typically does not take syntactic structure into account. It is difficult to recognize syntactic information for phrase-based SMT therefore it cannot handle long-distance reordering that frequently occurs in these language pairs.

To incorporate syntactic information into the PBSMT framework, we attempt to identify the SOV of the source language (Japanese) and then correctly produce the SVO of the target language (English). Concretely, we devise a dependency-based method that extracts a sentence's frame (hareafter "basic frame"), consisting of the predicate and its direct children (hereafter "anchor words"), and its dependent phrases consisting of the anchor words and their all descendants. After extracting these words and phrases, our method translates them separately and then combines their translation.

We conducted an experiment with the proposed method on a Japanese-to-English task at the Second Workshop on Asian Translation (Nakazawa et al., 2015). Although the results of our method are not positive, we discuss potential improvements.

The rest of this paper is organized as follows. In the next section, we discuss related work. In Section 3, the details of our method are explained. Then, we describe our experiments and analyze the results.

2 Related Work

A substantial, systematic difference in word orders creates difficulty for SMT, especially, PBSMT, which is not based on syntactic phrases. Good translation can be achieved in such situations by segmenting the input sentences into portion for simpler and adequate scale inputs.

For translating a long and complex sentence composed of several clauses in English into Japanese translation, Sudoh et al. (2010) proposed segmenting the sentence into clauses that include non-terminals as placeholders corresponding to embedded clauses using an HPSG parser, translating the clauses, and then replacing the non-terminals with the corresponding clause's translations. By representing an embedded clause with

Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 99–104,
Kyoto, Japan, 16th October 2015.

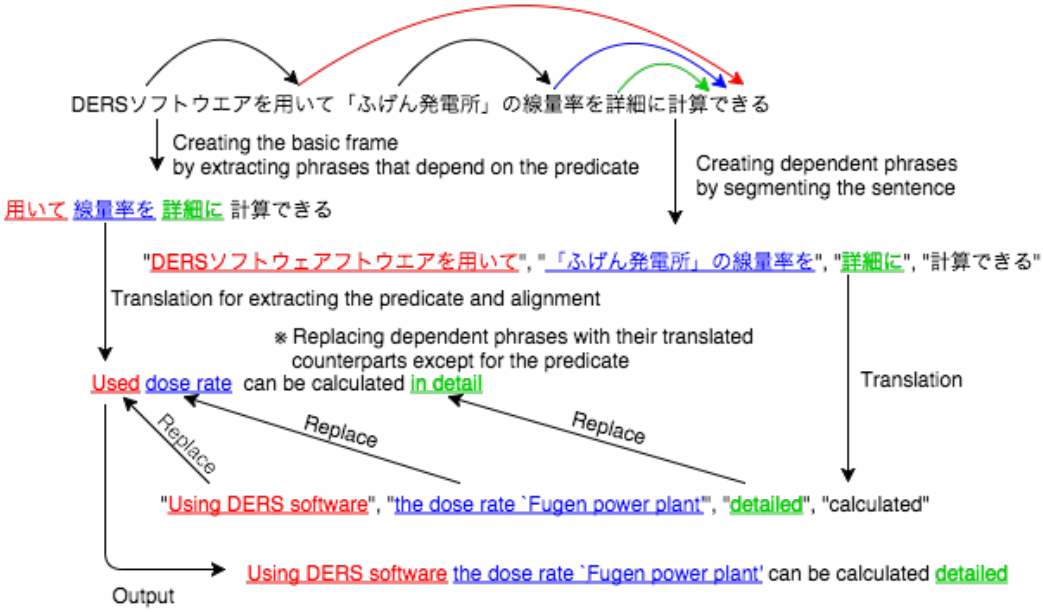

Figure 1 : Illustrative example of our method.

a non-terminal as a placeholder, they reduced the problem of reordering a complex sentence to a simple problem of word-level reordering. For each clause translation, non-terminals are treated as words and the clause including them is translated using the model trained on a clause-level aligned corpus that they developed. Their method uses a high-quality HPSG parser to segment an English sentence, but such a rich parser is not publicly available for Japanese. Thus, we opt to use dependency analysis to segment Japanese sentences. In addition, they segmented a sentence into clauses, whereas we segment a sentence into linguistically-motivated phrases, and in contrast to their approach, we do not arrange the corpus to be suitable for segmentation unit.

There are other segmentation methods in machine translation. Roh et al. (2001) proposed a method that recognizes the range of sub-sentences such as a relative and a conjunctive clause using sentence pattern information to overcome the problem of a syntactic ambiguity in a long sentence. Doi et al. (2003) split an input sentence into some smaller units to deal with long sentences in speech translation. Their method does not split sentences in a pre-processing phase or a parsing phase. It uses partial translation results and some criteria that judge the results to determine the best split positions. Other than that, Lee et al. (2012) proposed training a phrase segmentation model using a PBSMT decoder. The model is incorporated into the log-linear model of PBSMT, and the phrase segmenter based on the decoder annotates the source language phrase boundaries. The annotated data are used to train a new phrase segmentation model, which is then reused by the decoder. This process is performed iteratively, improving the phrase segmentation model.

Our approach also involves pre-ordering, one of the means of coping with the reordering problem. It reorders the word order of a source language sentence in the pre-processing phase to bring the sequence of words closer to the word order of the target language. Previous work addressing pre-ordering in SOV/SVO language pairs such as Japanese and English includes Isozaki et al. (2010), Komachi et al. (2006), Katz-Brown and Collins (2008), Xu et al. (2009), and Hoshino et al. (2013). These methods use some source language information to reorder the words of source language words with manual rules: morphological analysis (Katz-Brown and Collins, 2008), dependency analysis (Katz-Brown and Collins, 2008), and predicate argument structure analysis (Komachi et al., 2006; Hoshino et al., 2013). Our method also uses dependency analysis for pre-processing, but reordering is not performed. We use a dependency parser only to extract the basic frame and dependent phrases.

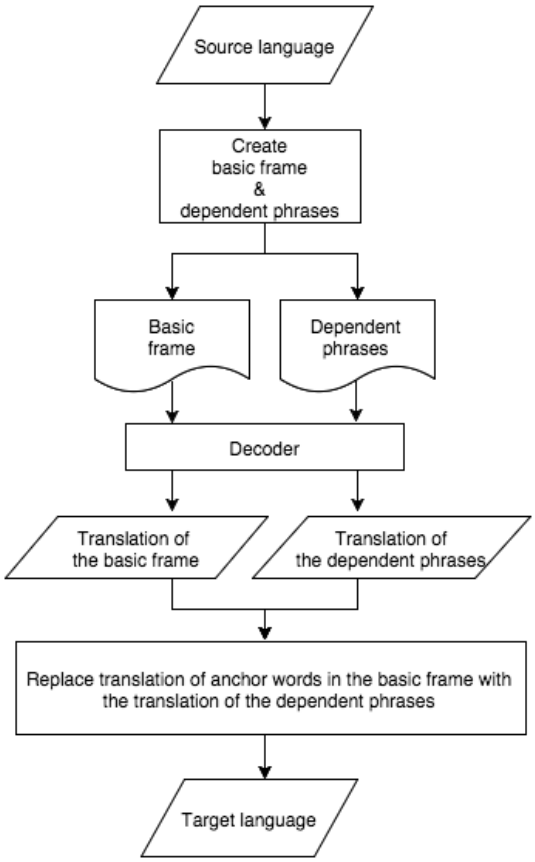

Figure 2: Flowchart of our method.

3 Source Phrase Segmentation and Translation Using Dependency Structure

Because of its ignorance of syntax, PBSMT may output ungrammatical sentences. Thus, we propose a source-phrase segmentation and translation approach to compensate for the lack of linguistically motivated information in the source language. We propose two methods. First method is "segmentation and translation", and second method is "segmentation and translation without preposition". "Segmentation and translation" is the method described in Subsection 3.1, and "segmentation and translation without preposition" is described in Subsection 3.2. In Subsection 3.3, we describe an additional rule for our method (for a sentence that includes a substantive verb).

Figures 1 and 2 show an example and outline of our method, respectively. Although our method is similar to that of Sudoh et al. (2010)'s work, we segment a sentence into its basic frame consisting of a predicate, anchor words and the dependent phrases that are the phrases consisting the subtrees rooted at the anchor words (Figure 3) by

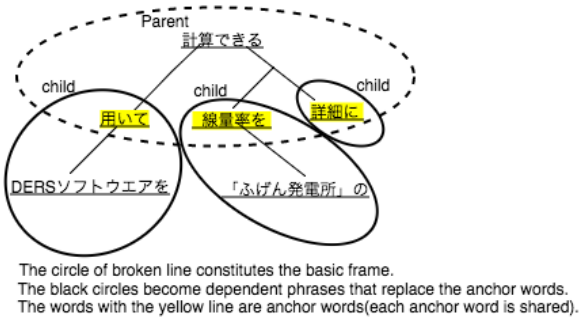

Figure 3 : Image of a basic frame, anchor words, and dependent phrases.

segmenting a sentence. The basic frame and dependent phrases are translated using a decoder. In the translation of a basic frame, each clause except for the predicate is subject to replacement. We replace the anchor words in the basic frame with the translations of the corresponding dependent phrases.

3.1 Segmentation and Translation

First, a basic frame is created by extracting the predicate and its direct children. Then, dependent phrases are created by segmenting the phrases that depend on the predicate. Second, the translation of the basic frame and those of the dependent phrases are obtained.

Finally, the output is created by replacing the translations of anchor words in the basic frame with those of the corresponding dependent phrases except for the translation of the predicate in the basic frame.

3.2 Segmentation and Translation without Preposition

We propose another but similar method that does not replace the preposition in the translation of the basic frame with one in the translation of dependent phrases. Because the translation of dependent phrases in our first method uses the language and translation models optimized for a sentence, our first method might not be able to use a model optimized for translating phrases.

For example, the adverbial phrase "詳細に (in detail)" in Figure 1 must be translated as an adverbial phrase "in detail". Our first method translates the phrase into "detailed", because the baseline decoder assumes a sentence as input. The baseline decoder correctly translates the phrase into "in detail" since it recognizes the verb "計算できる (can be calculated)" next to the phrase. In order

	energy	fluctuation	is	shown	.
エネルギー	■				
揺らぎ		■			
を			■		
示した				■	
。					■

Table 1: Alignment example to which the rule applies. The English word “is” is aligned to “を” included in the bunsetsu-chunk “揺らぎを”. Hence, without the rule, “is” is replaced by the translation of “揺らぎを” and deleted.

	BLEU	RIBES
no-seg&trans	18.32	0.641456
seg&trans	15.85	0.628897
seg&trans (w/o prep)	15.72	0.628463

Table 2: BLEU and RIBES of the baseline and our methods.

error types	frequency
Dependency parsing	3
Translation of a basic frame	18
Translation of dependent phrases	46
Total (Each error may overlap)	57

Table 3: Types of errors in the first 100 sentences of the test-set.

not to replace the correct prepositions, our second method preserves the preposition coming from the basic frame.

3.3 Additional Rule for Predicate Alignment

We observe that there is a problem in aligning a Japanese predicate to an English counterpart. Here, if a word governed by VP in an English sentence is not aligned to one of the words included in the corresponding predicate bunsetsu-chunk of the Japanese sentence, the word of the English side will be deleted by our replacement.

For example, when translating “エネルギー揺らぎを示した (energy fluctuation is shown)” in Table 1, the predicate of Japanese sentence “示した” is not aligned to the English words “is shown” but to the word “shown”. In this case, the English word “is” is not aligned to the word in the Japanese predicate bunsetsu-chunk “示した”; hence, the word will be deleted by the dependent phrases.

To avoid this unwanted occurrence, we add a rule to take effect that if there is a substantive verb in the translation of a basic frame, the replacement will not be applied to the substantive verb, since this accident occurs only for the substantive verb in our observation. As a result of this rule, “is shown” remains in the resulting sentence.

Our method uses this rule for both “segmentation and translation” and “segmentation and translation without preposition”.

4 Experiments

4.1 Experimental Settings

We use three million parallel sentences from the Asian Scientific Paper Excerpt Corpus. We use JUMAN (version 7.0) for segmentation, and GIZA++ (version 1.0.7) for alignment. We use Moses (version 2.1.1)’s default configurations: monotone, swap, and discontinuous. The language and translation models of Moses are trained with the ASPEC. In translating the basic frame and dependent phrases, we use the same language and translation models. MERT is performed on the full dev-set. We follow the split of dev-set and test-set provided by the organizer of the workshop.

To preprocess the input sentences, basic frames and dependent phrases are extracted by a dependency parser CaboCha (version 0.68). [1]

We use Moses as a baseline. Moses’s settings are the same as the above settings. Our method is evaluated using Bilingual Evaluation Understudy (BLEU) and Rank-based Intuitive Bilingual Evaluation Score (RIBES).

4.2 Experimental Results

Table 2 reports our official evaluation results for the WAT 2015 and an additional experiment after the official evaluation campaign. Both BLEU and RIBES deviated from the baseline.

5 Discussion

In Figure 1, we present an example for which our first method fails but our second method succeeds to translate. Our first method creates “DERS ソフトウェアを用いて”, “「ふげん発電所」の線量率を”, “詳細に”, and

[1] We do not use CaboCha for segmentation but only to create dependency phrases.

“計算できる” as the dependent phrases. These dependent phrases are translated into “using DERS software”, “the dose rate ‘Fugen power plant”’, “detailed”, and “calculated”, respectively. Our first method combines them with a translation of the basic frame to create output “using DERS software the dose rate ‘Fugen power plant’ can be calculated detailed”, but this is ungrammatical.

This problem is caused by lack of contextual information in the the dependent phrases. For instance, the reason why the translation of the dependent phrase “詳細に” lacks any preposition is that the proper preposition corresponding to the particle “に” cannot be identified only by the information obtained from the input “詳細に”. Our second method can solve this problem. It does not replace the preposition following the predicate to preserve the plausible preposition that seems to be translated properly in the translation of the basic frame.

However, the BLEU and RIBES scores of our second method are lower than those of our first method. Especially, the BLEU scores of our methods differ more than the RIBES scores. We will discuss the reason below.

We count the number of the pairs of input and reference that have different voices from the sample data, which comprises 100 sentences that we selected as the top 100 sentences from the test-set. Thirty-five percent of the translation outputs differ in their voice from the corresponding reference, and we suppose that this is the cause of the degradation in BLEU. For example, suppose a reference is in the active voice and the output of our methods is in the passive voice. Since BLEU penalizes incorrect translation based on n-gram precision, mis-ordering affect BLEU more than RIBIES.

In addition, the proposed methods have problems in the translation and language models. Since dependent phrases are typically noun phrases, the baseline decoder trained on parallel sentences might not produce appropriate translations. Moses tends to translate a noun phrase into a sentence if the dependent phrase contains a verb in a relative clause construction. For example, the dependent phrases “水の運動 (the water movement) の基本である (a basis of) 水素結合ネットワークの変化を (the change of hydrogen bond network)” were translated into “the change of hydrogen bond network is a basis of the water movement”. From this perspective, we need to perform translation with different models for a basic frame and for dependent phrases.

In order to perform error analysis, we evaluate our methods quantitatively. Table 3 shows the number of errors in using a dependency parser (CaboCha in Table 3), translating a basic frame and dependent phrases (Translation of a basic frame and Translation of dependent phrases in Table 3). The errors resulting from CaboCha are very few; there are only three instances out of 100.

The most prominent errors originate in phrase translation. We can solve the problem by creating optimized language and translation models for noun phrases. Because the language and translation models must create a noun phrase when dependent phrase is noun phrase, they must be optimized for noun phrases.

The next most frequent errors are translations of basic frames. We count this type of error when the translation of a basic frame is not a sentence. In contrast to the case of dependent phrases, the translation of a basic frame must output a sentence with a predicate. Alternatively, we can re-rank the output to remove ungrammatical translations of a basic frame.

6 Conclusion

This paper proposed segmentation and translation methods for Japanese to English translation and presented its evaluation. We discussed the sample data consisting of the top 100 sentences that we had selected from the test-set. As a result, we found that the output has three problems: dependency parsing, translation of a basic frame, and translation of dependent phrases. In the future, we plan to optimize the language and the translation models suitable for phrase translation.

References

Takao Doi and Eiichiro Sumita. 2003. Input Sentence Splitting and Translating. In *Proceedings of the HLT-the North American Chapter of the Association for Computational Linguistics 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3*, pages 104–110.

Sho Hoshino, Yusuke Miyao Katsuhito Sudoh, and Masaaki Nagata. 2013. Two-Stage Pre-ordering for Japanese-to-English Statistical Machine Translation. In *Proceedings of The 6th International Joint*

Conference on Natural Language Processing 2013, pages 1062–1066.

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

Jason Katz-Brown and Michael Collins. 2008. Syntactic Reordering in Preprocessing for Japanese→English Translation: MIT System Discriptioin for NTCIR-7 Patent Translation Task. In *Proceedings of the NII Testbeds and Community for Information access Research-7 Workshop Meeting*, pages 409–414.

Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. 2006. Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 77–82.

Hyoung-Gyu Lee and Hae-Chang Rim. 2012. Decoder-based Discriminative Training of Phrase Segmentation for Statistical Machine Translation. In *Proceedings of International Conference on Computational Linguistics*, pages 611–620.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*.

Yoon-Hyung Roh, Young-Ae Seo, Ki-Young Lee, and ung Kwon Choi. 2001. Long Sentence Partitioning using Structure Analysis for Machine Translation. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 646–652.

Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, and Masaaki Nagata. 2010. Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 418–427.

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 the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253.