# Predicate-Argument Structure-based Preordering for Japanese-English Statistical Machine Translation of Scientific Papers

**Kenichi Ohwada, Ryosuke Miyazaki** and **Mamoru Komachi**
{ohwada-kenichi, miyazaki-ryosuke}@ed.tmu.ac.jp, komachi@tmu.ac.jp
Graduate School of System Design
Tokyo Metropolitan University, Japan

## Abstract

Translating Japanese to English is difficult because they belong to different language families. Naïve phrase-based statistical machine translation (SMT) often fails to address syntactic difference between Japanese and English. Preordering methods are one of the simple but effective approaches that can model reordering in a long distance, which is crucial in translating Japanese and English. Thus, we apply a predicate-argument structure-based preordering method to the Japanese-English statistical machine translation task of scientific papers. Our method is based on the method described in (Hoshino et al., 2013), and extends their rules to handle abbreviation and passivization frequently found in scientific papers. Experimental results show that our proposed method improves performance of both (Hoshino et al., 2013)'s system and our phrase-based SMT baseline without preordering.

## 1 Introduction

Preordering method is one of the popular techniques in statistical machine translation. Preordering the word order of source language in advance can enhance alignments on a pair of languages with a large difference in syntax like japanese and English, and thus improve performance of machine translation system.

One of the advantages of preordering is that it can incorporate rich linguistic information on the source side, whilst off-the-shelf SMT toolkit can be plugged in without any modification. Preordering methods employ various kinds of linguistic information to achieve better alignment between source and target languages. Specifically, previous work in the literature uses morphological analysis (Katz-Brown and Collins, 2008), dependency structure (Katz-Brown and Collins, 2008) and predicate-argument structure (Komachi et al., 2006; Hoshino et al., 2013) for preordering in Japanese-English statistical machine translation.

However, these preordering methods are tested on limited domains: travel (Komachi et al., 2006) and patent (Katz-Brown and Collins, 2008; Hoshino et al., 2013) corpora. Translating Japanese to English in a different domain such as scientific papers is still a big challenge for preordering-based approach. For example, academic writing in English traditionally relies on passive voice to give an objective impression, but one can use either passive construction or a zero-pronoun in the Japanese translation of passive construction on the English side. It is not clear whether existing preordering rules are applicable to scientific domain due to such stylistic difference.

Predicate-argument structure-based preordering is one of the promising approaches that can solve syntactic and stylistic difference between a language pair. Predicate-argument structure analysis identifies who does what to whom and generalizes grammatical relations such as active and passive construction. Following (Hoshino et al., 2013), we perform predicate-argument structure analysis on the Japanese side to preorder Japanese sentences to form an SVO-like word order. We propose three modifications to the preordering rules to extend their model to better handle translation of scientific papers.

The main contribution of this work is as follows:

- We propose an extension to (Hoshino et al., 2013) in order to deal with abbreviation and passivization frequently found in scientific papers.

Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 44–49,
Tokyo, Japan, 4th October 2014.

## 2 Previous work

There are several related work that take preordering approaches to Japanese-English statistical machine translation.

First, Komachi et al. (2006) suggested a preordering approach for Japanese-English speech translation in travel domain based on predicate-argument structure. They used an in-house predicate-argument structure analyzer and reordered Japanese sentences using heuristic rules. They only performed preordering at sentence-level, whereas other Japanese-English preordering methods we will describe below perform preordering at both sentence- and phrase-level[1].

Second, Katz-Brown and Collins (2008) presented two preordering methods for Japanese-English patent translation based on morphological analysis and dependency structure, respectively. Morphological analysis-based method splits sentences into segments by punctuation and a topic marker ("は"), and then reverses the segments. Dependency analysis-based method reorders segments into a head-initial sentence, and moves verbs to make an SVO-like structure. Unlike (Komachi et al., 2006), they also reverse all words in each phrase.

Third, Hoshino et al. (2013) proposed predicate-argument structure-based preordering rules in two-level for the Japanese-English patent translation task. The first is sentence-level and the second is phrase-level. Furthermore, sentence-level preordering rules are divided into three parts. In total, sentences are reordered sequentially by four rules. Since this method is the one we re-implemented in this paper, we will describe their method in detail below.

**Pseudo head-initialization** Since Japanese is a head-final language but English is a head-initial language, this rule transforms a Japanese dependency tree as to become a head-initial phrase sequence. Concretely, we move the last phrase, which is a predicate of a Japanese sentence in almost all cases, to the beginning of the sentence. We then order each phrase as their children located immediately after them.

**Inter-chunk preordering** We move a predicate of a sentence to an adequate place. If a subject exists in a sentence[2], the predicate is moved immediately after the subject. If a subject is not present but an object, the predicate is moved just before the object. If there is neither a subject nor an object, the predicate is moved just before the rightmost phrase in the predicate's children. Thanks to this rule, even when a subject and a object do not exist, we avoid having a predicate at the beginning of a sentence.

**Inter-chunk normalization** We restore the order of coordinate expressions which are reversed by the first rule. Also, since a full stop is moved along with the predicate, we restore it back to the end of the sentence.

**Intra-chunk preordering** We apply the phrase-level rule, which swaps function words and content words in a phrase. It will improve alignments because function words in Japanese (e.g. postposition) appear after content words while those in English (e.g. preposition) appear before content words.

## 3 Extension to (Hoshino et al., 2013)

Our proposed preordering model is based on (Hoshino et al., 2013) with three extensions to better handle academic writing in scientific papers.

### 3.1 Parenthesis preordering

Scientific papers often include parenthetical expressions. The training data (1,000,000 parallel sentences, hereafter referred to as 1M training corpus) contains 168,336 (16.8%) parentheses on Japanese side, and 187,400 parentheses on English side. However, Japanese dependency analyzer fails to parse parenthetical expressions appropriately. In particular, if a parenthesis is used to show apposition (e.g. abbreviation and acronym), the pseudo head-initialization described in the last section swaps an acronym and its full spelling, which is not desirable. Therefore, we modify the pseudo head-initialization rule to restore the position of parenthetical expressions.

Figures 1a and 1b illustrate how parenthesis preordering transforms original sentences. Parenthesis preordering rule handles not only single phrase parenthetical expressions but also multiple phrase parenthetical expressions.

### 3.2 Passive voice preordering

In scientific papers, zero-pronouns in Japanese are often translated into passive construction in English. The number of passive construction in the

[1] In this paper, "phrase" means "bunsetsu".

[2] A pro-drop language like Japanese often omits subjects.

Figure 1: Examples of extended preordering rules.

(a) Parenthesis preordering (single phrase)

Japanese: 一般病棟における | コンサルテーション型の | 緩和ケアチーム | （ＰＣＴ）が | 注目されている。
English literal: in the general ward | of the consultation type | palliative care team | (PCT)$_{NOM}$ | notice$_{PRESENT|PASSIVE}$.
English translation: “Palliative care team (PCT) of the consultation type in the general ward is noticed.”

Pseudo head-initialization: 注目されている。| （ＰＣＴ）が | 緩和ケアチーム | コンサルテーション型の | 一般病棟における
Parenthesis preordering: 注目されている。| 緩和ケアチーム | （ＰＣＴ）が | コンサルテーション型の | 一般病棟における

(b) Parenthesis preordering (multiple phrases)

Japanese: ＰＡＬチャートを | 使った | 適切な | ガラス部材（庇を | 含めた）を | 選択する | 手法を | 紹介した。
English literal: PAL chart$_{ACC}$ | using | appropriate | glass member ( a hood$_{ACC}$ | including )$_{ACC}$ | selecting | method$_{ACC}$ | present$_{PAST}$.
English translation: “This paper presents a technique for selecting an appropriate glass member ( including a hood ) using a PAL chart.”

Pseudo head-initialization: 紹介した。| 手法を | 選択する | 含めた）を | ガラス部材（庇を | 使った | ＰＡＬチャートを | 適切な
Parenthesis preordering: 紹介した。| 手法を | 選択する | ガラス部材（庇を | 含めた）を | 使った | ＰＡＬチャートを | 適切な

(c) Passive voice preordering

Japanese: 研究班の | 概要を | 紹介した。
English literal: research team$_{GEN}$ | outline$_{ACC}$ | introduce$_{PAST}$
English translation: “The outline of the research team is introduced.”

Pseudo head-initialization: 紹介した。| 概要を | 研究班の
Passive voice preordering: 概要を | 研究班の | 紹介した。

(d) Subject preordering

Japanese: ＣＯＰＤ患者の | 換気増加は | 日常活動を | 制限する | 重要な | 因子である。
English literal: COPD patient$_{GEN}$ | ventilation increase$_{TOP}$ | daily activity$_{ACC}$ | limit$_{PRES}$ | important | be$_{PRES}$ factor.
English translation: “Ventilation increase of the COPD patients is the important factor which limits the daily activity.”

Pseudo head-initialization: 因子である。| 換気増加は | ＣＯＰＤ患者の | 制限する | 日常活動を | 重要な
Inter-chunk preordering: 換気増加は | 因子である。| ＣＯＰＤ患者の | 制限する | 日常活動を | 重要な
Modified inter-chunk preordering: 換気増加は | ＣＯＰＤ患者の | 因子である。| 制限する | 日常活動を | 重要な

Final preordering result: は換気増加 | のＣＯＰＤ患者 | ある因子で | 制限する | を日常活動 | な重要。

1M training corpus is 166,057 (17%), whereas the number of active construction starting with "They ..." and "It is ..." are 4,700 and 17,104 (2%), respectively. Hence, we move a predicate to the end of the sentence if there exists no subject in active voice.

Figure 1c describes how this rule transforms a Japanese sentence with a zero-pronoun. Even though the Japanese side is in active voice, English translation is expressed in passive voice. Note that a Japanese sentence in active voice may be translated into different expressions even in the same passive construction (e.g. "... を説明した (explained ...)" can be either "... was explained" or "It was explained that ...".).

### 3.3 Subject preordering

Hoshino et al. (2013) proposed to move a predicate after the subject (*inter-chunk preordering*). However, if a subject is modified by other phrases, this rule moves the predicate to the middle of a subjective phrase composed of multiple phrases. Thus, we move a predicate to the end of the subjective phrase.

Table 1d depicts how subject preordering moves a predicate in a sentence. As we can see, this rule prevents subjective phrase "ＣＯＰＤ患者の | 換気増加 (Ventilation increase of the COPD patients)" to be split by the predicate movement.

## 4 Experiments

We compared translation performance using a standard phrase-based statistical machine translation technique with three kinds of data:

- original data (baseline),
- preordered data by our re-implementation of (Hoshino et al., 2013), and
- preordered data by our proposed methods.

We analyzed predicate-argument structure of only the last predicate for each sentence, regardless of the number of predicates in a sentence. Also, following (Hoshino et al., 2013), we did not consider event nouns as predicates.

### 4.1 Experimental settings

We used 1M Japanese-English parallel sentences extracted from scientific papers (`train-1.txt`) from the Asian Scientific Paper Excerpt Corpus (ASPEC) [3]. We varied the size of the training corpus and used the best size determined by preliminary experiments.

We identified predicate-argument structure in Japanese by SynCha[4] 0.3. It uses MeCab[5] 0.996 with IPADic 2.7.0 for morphological analysis and CaboCha[6] 0.68 for dependency parsing.

We used SRILM[7] 1.7.0 for language model, GIZA++[8] 1.0.7 for word alignment, and Moses[9] 2.1.1 for decoding. We set distortion limits to default value 6 for all systems[10].

Translation quality is evaluated in terms of BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010), as determined by the workshop organizers (Nakazawa et al., 2014).

We performed minimum error rate training (Och, 2003) optimized for BLEU using the development set (`dev.txt`) of the ASPEC corpus. We conducted all the experiments using the scripts distributed at KFTT Moses Baseline v1.4 [11].

### 4.2 Experimental results

Table 1 shows the experimental results. In terms of BLEU, our re-implementation of (Hoshino et al., 2013) is below the baseline method while our proposed methods better than the baseline. In terms of RIBES, all preordering methods outperform the baseline, and our proposed method archieve the highest score.

All methods including parenthesis preordering outperform the baseline method, and when we subtract three modifications one by one from proposed method, the parenthesis rule has the largest impact on the translation quality.

### 4.3 Discussion

Some of the errors found in a translation result are due to the errors in predicate-argument structure analysis. We found that it is hard for predicate-argument structure analyzer trained on a newswire

[3] `http://lotus.kuee.kyoto-u.ac.jp/ASPEC/`

[4] `http://www.cl.cs.titech.ac.jp/~ryu-i/syncha/`

[5] `http://mecab.googlecode.com/`

[6] `http://cabocha.googlecode.com/`

[7] `http://www.speech.sri.com/projects/srilm/`

[8] `https://code.google.com/p/giza-pp/`

[9] `http://www.statmt.org/moses/`

[10] We confirm in another experiment that the highest performance of the proposed system is archieved by the distortion limit around 15.

[11] `http://www.phontron.com/kftt/`

Table 1: Comparison of the preordering methods. All the preordering models using (Hoshino et al., 2013) are our re-implementation of their paper.

| Method | BLEU | RIBES |
|---|---|---|
| Phrase-based SMT baseline (w/o preordering) | 15.74 | 0.620162 |
| (Hoshino et al., 2013) (preordering baseline) | 15.45 | 0.645954 |
| Proposed method − parenthesis preordering | 15.73 | 0.652461 |
| Proposed method − passive voice preordering | 15.93 | 0.654454 |
| Proposed method − subject preordering | 15.88 | 0.650964 |
| Proposed method | 16.02 | 0.654600 |
| Phrase-based SMT baseline (ORGANIZER) | 18.45 | 0.645137 |

Figure 2: Error analysis of predicate-argument structure-based preordering.

(a) Reordering error with "It is . . ." construction.

Japanese: ＬＡＮの | 仕組みに関する | 解説記事である。
English literal: LAN$_{GEN}$ | on mechanism | be$_{PRES}$ explanation article.

Final reordering result: に関する仕組み | のＬＡＮ | ある解説記事で。
English translation: It is the explanation article on the mechanism of the LAN.

(b) Reordering error with "They . . ." construction.

Japanese: 新しい | 規格の | 項目と | 構成を | 示した。
English literal: new | standard$_{GEN}$ | item and | composition$_{ACC}$ | show$_{PAST}$.

Final reordering result: と項目 | を構成 | の規格 | 新しい | た示し。
English translation: They showed item and composition of the new standard.

corpus to parse scientific papers. It may be necessary to perform domain adaptation at some level.

Apart from apparent errors in predicate-argument structure analysis, there exist errors in preordering rules. Figure 2 shows errors in preordering. In both examples, the passive voice preordering rule moves the predicate to the end of a sentence, but the English counterpart uses active construction instead of passive construction. It would be necessary to not only perform predicate-argument structure analysis on the source side but also on the target side to correctly align predicate-argument structures between a language pair.

## 5 Issues for Context-aware MT

In this paper, we did not consider any inter-sentential context, even though the off-the-shelf predicate-argument structure analyzer is able to perform co-reference and zero-anaphora resolution (Iida and Poesio, 2011). It is only because the training corpus at hand does not come with inter-sentential information. If we have access to the whole article, we may perform zero-anaphora resolution to better handle passivization in Japanese-English translation.

## 6 Conclusion

In this paper, we propose to modify several rules to (Hoshino et al., 2013) in order to address the stylistic differences for translating scientific papers. Experimental results show that all preordering methods combined improve the system performance.

In future work, we investigate the effectiveness of these rules in different domains.

## References

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 (IJCNLP2013)*, pages 1062–1066.

Ryu Iida and Massimo Poesio. 2011. A Cross-Lingual ILP Solution to Zero Anaphora Resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 804–813.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 944–952.

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

Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 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.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st workshop on asian translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 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 Annual Meeting of the ACL*, pages 311–318.