# OkaPU's Japanese-to-English Translator for NTCIR-10 PatentMT

Hideki Isozaki
Okayama Prefectural University
isozaki@cse.oka-pu.ac.jp

## ABSTRACT

This paper describes Okayama Prefectural University's system for NTCIR-10 PatentMT JE task. It is a variant of the REV method proposed by Katz-Brown and Collins [KBC08] which obtained the best human evaluation score among Statistical Machine Translation systems at NTCIR-7 [FUYU08]. Their REV method preorders Japanese sentences without syntactic parsing. They split each Japanese sentence into segments at punctuations and the Japanese topic marker "wa". Then, they reversed words in each segment and concatenated the reversed segments into one. For NTCIR-10, we tried to improve the REV method by keeping Japanese word order in noun phrases and coordinations.

## Keywords

Japanese-to-English Translation, Statistical Machine Translation, preordering

## Team Name

OKAPU

## Subtasks

Japanese to English

## 1. INTRODUCTION

Our simple preordering method, "Head Finalization", worked well for English-to-Japanese SMT [ISTD12]. NTT-UT's English-to-Japanese translator for NTCIR-9 [SDT+11] based on Head Finalization was better than RBMT systems in terms of human judgement score. This was the first time that an SMT system outperformed RBMT systems in the NTCIR PatentMT history. [GLC+11] It divided English-to-Japanese translation into two steps: English-to-HFE (Head Final English) translation and HFE-to-Japanese translation. We can implement the first step easily by using an HPSG parser, Enju [MT08][1]. The second step is almost monotone and we can use a conventional phrase-based SMT system.

For Japanese-to-English translation, we considered feasibility of "Head Initialization", because English is a "head-initial" language. However, English has some "head-final" expressions such as noun phrases. It is not easy to find a set of simple reordering rules for Japanese-to-English translation.

Sudoh et al. [SWD+11] proposed a simple solution for this problem. They divided J-to-E translation into two steps: J-to-HFE translation and HFE-to-E translation. Each translation was solved by conventional SMT methods. Goto et al. [GUS12] refined this approach.

Here, we searched a direct preordering method for Japanese-to-English SMT. In NTCIR-7, Katz-Brown and Collins [KBC08] proposed two preordering methods: REV preorder and CaboCha preorder.

- The REV preorder uses MeCab[2], one of the most popular Japanese morphological analyzers.
- The CaboCha preorder uses CaboCha[3], the de facto standard Japanese dependency analyzer.

According to the PATMT overview paper [FUYU08], their REV method obtained the best human evaluation score among Statistical Machine Translation systems.

## 2. METHODOLOGY

For our formal run submission, we tried to improve their REV method by keeping Japanese word order in each base noun phrase, because English base noun phrase also follows head-final word order just like Japanese.

For English-to-Japanese translation, we introduced the "Coordination Exception" rule to keep order of elements in coordinations [ISTD12]. We also need the "Coordination Exception" rule for Japanese-to-English translation.

We implemented these "keep Japanese order" rules by Part-of-Speech tag check. We used MeCab-0.994 for morphological analysis and treated the following Part-of-Speech tags as "**Japanese word order keepers**": *alphabets, parallel case markers, conjunctions, noun prefixes, nouns (except pronouns), dependent nouns (hijiritsu), adverbial nouns (fukushikanou), and adnomial adjective (rentaishi)*. We ran MeCab with `-F "%m:%h "` to get a Part-of-Speech tag ID for each word. We kept the sequences of words with the above POS tags as they are.

[1] `http://www.nactem.ac.uk/enju/`

[2] `http://code.google.com/p/mecab/`

[3] `http://code.google.com/p/cabocha/`

**Table 1: Comparison of preordering methods**

| System | average of $\tau$ |
|---|---|
| raw Japanese | 0.3960 |
| REV-like | 0.5373 |
| Submitted | 0.6283 |
| Bug-fixed | 0.6418 |

We used Moses[4] for training and decoding our translator. The training took 16 hours for Multi-threaded GIZA++[5] and five hours for MERT on a 12-core Xeon PC.

The distortion limit was 6. We did not tune the distortion limit, but we chose this value because our rough preordering rules will not yield perfect English word order but if our set of preordering rules is good enough, this value will suffice.

Table 1 compares the averages of Kendall's $\tau$ for each line of the `aligned.grow-diagonal-final-and` file. We also implemented CaboCha-based preordering methods, but we could not obtain a better $\tau$ value, and gave up this approach.

After the formal run submission, we found a bug in the above Part-of-Speech tag list. The POS tag ID for parallel case markers should be '`23`' but we mistakenly used '`14`'. By fixing this bug, the average of $\tau$ was slightly improved.

## 3. CONCLUDING REMARKS

We tried to improve MIT's REV method, and we obtained a better $\tau$ value. This research was supported by Okayama Prefectural University's Creative Research Supporting Fund.

## 4. REFERENCES

[FUYU08] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *Working Notes of the NTCIR Workshop Meeting (NTCIR)*, pages 389–400. National Institute of Informatics, 2008.

[GLC+11] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the NCTIR-9 workshop. In *Working Notes of the NTCIR Workshop Meeting (NTCIR)*, pages 559–578, 2011.

[GUS12] Isao Goto, Masao Utiyama, and Eiichiro Sumita. Post-ordering by parsing for Japanese-English statistical machine translation. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 311–316, 2012.

[ISTD12] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. HPSG-based Preprocessing for English-to-Japanese Translation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11 Issue 3, 2012.

[KBC08] Jason Katz-Brown and Michael Collins. Syntactic reordering in preprocessing for Japanese → English translation: MIT system description for NTCIR-7 patent translation task. In *Working Notes of the NTCIR Workshop Meeting (NTCIR)*, 2008.

[MT08] Yusuke Miyao and Jun'ichi Tsujii. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80, 2008.

[SDT+11] Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Masaaki Nagata, Xianchao Wu, Takuya Matsuzaki, and Jun'ichi Tsujii. NTT-UT statistical machine translation in NTCIR-9 PatentMT. In *Working Notes of the NTCIR Workshop Meeting (NTCIR)*, pages 585–592, 2011.

[SWD+11] Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Post-ordering in statistical machine translation. In *Proc. of the Workshop on Statistical Machine Translation*, 2011.

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

[5] `http://www.kyloo.net/software/doku.php/mgiza:overview`