様式第２号

平成２５年度　　独 創 的 研 究 助 成 費 実 績 報 告 書

平成２６年　３月３１日

<table>
<tr><td>申　請　者</td><td colspan="2">学科名</td><td>情報システム工学科</td><td>職 名</td><td>教授</td><td>氏 名</td><td colspan="2">磯崎　秀樹　　印</td></tr>
<tr><td>調査研究課題</td><td colspan="8">日英特許翻訳における翻訳品質の改善</td></tr>
<tr><td>交付決定額</td><td colspan="8">４５円</td></tr>
<tr><td rowspan="3">調査研究組織</td><td colspan="2">氏　　名</td><td colspan="2">所属・職</td><td colspan="2">専門分野</td><td colspan="2">役割分担</td></tr>
<tr><td>代表</td><td>磯崎　秀樹</td><td colspan="2">情報システム工学教授</td><td colspan="2">自然言語処理</td><td colspan="2">立案・実装・実験</td></tr>
<tr><td>分担者</td><td>天嵜　聡介<br>高地なつめ</td><td colspan="2">情報システム工学助教<br>情報システムＭ１</td><td colspan="2">ソフトウェア工学</td><td colspan="2">計算機環境支援<br>実験補助</td></tr>
<tr><td>調査研究実績の概要</td><td colspan="8">市販の日英翻訳ソフトは「ルールベース機械翻訳（ＲＢＭＴ）」という膨大な人手のかかる旧来のソフトウェア開発法で作成されている。これに対し欧米言語間の翻訳では、グーグル翻訳に代表される「統計的機械翻訳（ＳＭＴ）」という、人手のあまりかからない、統計に基づく新しいアプローチが考案されており、ＲＢＭＴの性能を抜き成功している。英日翻訳でも、磯崎の手法をベースにしたＳＭＴがＲＢＭＴを抜いた。<br><br>しかし、逆方向の日英翻訳は、省略などの日本語側の難しい問題と、英語側の構文に一貫性がない、という難しい問題の組み合わせのため、英日翻訳の場合のような簡単な解決方法が見つかっておらず、ＳＭＴはまだＲＢＭＴの翻訳品質を上回っていない。昨年度の独創的研究により、日英翻訳のＳＭＴ研究に本格的に着手し、岡山県立大学独自の日英特許翻訳のソフトの初版ができたが、その性能はまだまだＲＢＭＴに及ばない。<br>今年度は、昨年度作成したＳＭＴの翻訳結果を統計的に解析して、本質的な改良の手掛かりを見つけ、日英特許翻訳でＲＢＭＴの翻訳品質を上回ることのできるＳＭＴの仕組みを考案することを目指した。<br><br>まず、昨年度作成した事前並べ替え方式の翻訳機について、追加実験を行って英語論文にまとめ、国立情報学研究所が主催する国際会議ＮＴＣＩＲ－１０で発表した。この手法は、ＮＴＣＩＲ－７の日英翻訳で人手評価スコアが１位だった、ＭＩＴチームのＲＥＶ法を改良したものであり、実験により、並べ替えの基本性能はＲＥＶよりもよいことが判明</td></tr>
</table>

| | |
|---|---|
| 調査研究実績の概要 | した。しかし、ＳＭＴの並べ替えパラメタを抑制し過ぎたせいで、全体的性能はあまり高くなかった。<br><br>ＮＴＣＩＲ－１０の日英翻訳で上位を占めたのは、あいかわらずＲＢＭＴや、ＲＢＭＴとＳＭＴの組み合わせであり、ＳＭＴで１位だったのは、磯崎が考案した英日翻訳手法を逆方向に適用できるように工夫したＮＴＴチームのシステムである。<br><br>次に、日英翻訳がうまくいかない原因について、ＮＴＣＩＲ－７で参加者が提出した翻訳結果を人手で評価したデータを用いて探ることにした。卒論生の河田彰君に以下のような手法により、原因のマイニング（統計調査）をしてもらった。<br><br>・句読点の数や受け身など、悪影響を与えていそうな原因による成績への影響の可視化<br>・ＢＡＣＴというマイニングツールを用いた失敗原因の追究。<br><br>ＮＴＣＩＲ－７のＮＴＴチームの日英翻訳の人手評価の結果にＢＡＣＴを適用したところ、連体修飾節があると成績が下がることを示唆するパターンなどが見つかった。日本語の場合、連体修飾節が修飾される語句より前に来るのに対して、英語では後ろに来る。そのせいで、連体修飾では語順を大幅に入れ替えなければならないが、ＳＭＴは大幅な入れ替えが苦手なので、語順を入れ替えないで訳してしまうことがあり、そうした文で成績が下がっているらしい。<br>ＮＴＣＩＲ－１０の磯崎のシステムは事前並べ替え方式であり、連体修飾節はあらかじめ逆順になっているので、同じ問題はないと予想される。磯崎のシステムについては、人手評価の詳細な点数が公開されていないので、今年度は見送ったが、磯崎が以前考案したＲＩＢＥＳ（ライビーズ）という自動評価法は、人手評価との相関が高いので、ＲＩＢＥＳを人手評価の代用としてＢＡＣＴを適用することにより、まもなく失敗原因が判明するであろう。 |

| | |
|---|---|
| 成果資料目録 | Hideki Isozaki: OkaPU's Japanese-to-English Translator for NTCIR-10 PatentMT, In Proceedings of the NTCIR Conference, pp.348—349, 2013.<br><br>河田彰：翻訳ソフト改良のための可視化とマイニング、卒業研究論文要旨情報システム工学科C4、平成２５年度卒業論文および修士論文の要旨、岡山県立大学、2014. |

# 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`