平成 27 年度 AAMT/Japio 特許翻訳研究会
報 告 書

機械翻訳及び機械翻訳評価に関する研究
及び
海外調査

平成 28 年 3 月

一般財団法人　日本特許情報機構

目　　次

AAMT／Japio 特許翻訳研究会委員名簿

（敬称略・順不同）

委 員 長	辻井　潤一　（※2）	国立研究開発法人　産業技術総合研究所 人工知能研究センター　研究センター長/ 東京大学　名誉教授
副委員長	梶　　博行	静岡大学学術院 教授
	宇津呂　武仁（※2）	筑波大学大学院 教授
委　　員	横山　晶一	山形大学　名誉教授、客員教授
	江原　暉将（※2）	元・山梨英和大学 教授
	黒橋　禎夫	京都大学大学院 教授
	越前谷　博（※2）	北海学園大学大学院 教授
	磯崎　秀樹（※1）	岡山県立大学 教授
	二宮　崇	愛媛大学大学院 准教授
	綱川　隆司	静岡大学学術院 助教
	後藤　功雄（※2）	NHK 放送技術研究所 ヒューマンインターフェース研究部 専任研究員
	熊野　明	東芝ソリューション株式会社　プラットフォームセンター ソフトウェア開発部
	下畑　さより	沖電気工業株式会社 ソリューション&サービス事業本部 企画室
	須藤　克仁（※2）	NTT コミュニケーション科学基礎研究所 協創情報研究部 言語知能研究グループ 研究主任
	今村　賢治	国立研究開発法人　情報通信研究機構 先進的音声翻訳研究開発推進センター　専門研究員
	中澤　敏明（※2）	国立研究開発法人　科学技術振興機構 情報企画部 研究員/ 京都大学 大学院情報学研究科 知能情報学専攻 研究員
オブザーバー	中川　裕志	東京大学 情報基盤センター 教授
	範　　暁蓉	東京大学大学院　中川研究室
	潮田　　明	元・奈良先端科学技術大学院大学 客員准教授
	呉　　先超	バイドゥ株式会社 プロダクト事業部 シニア RD
	長瀬　友樹（※2）	株式会社富士通研究所 メディア処理システム研究所
	園尾　聡（※2）	株式会社東芝 研究開発センター 知識メディアラボラトリー
	高　京徹	株式会社高電社 経営企画部　部長

	守屋　敏道	一般財団法人日本特許情報機構　専務理事/ 特許情報研究所　所長
	河合　弘明	一般財団法人日本特許情報機構　特許情報研究所 調査研究部　部長
	大塩　只明	一般財団法人日本特許情報機構　特許情報研究所 調査研究部　総括研究主幹
	塙　　金治	一般財団法人日本特許情報機構　特許情報研究所 研究管理部　次長
	早川　貴之	一般財団法人日本特許情報機構　特許情報研究所 調査研究部　研究企画課　課長
	三橋　朋晴	一般財団法人日本特許情報機構　特許情報研究所 調査研究部　研究企画課　課長代理
	小川　直彦	一般財団法人日本特許情報機構　特許情報研究所 研究管理部　研究管理課係長
	土屋　雅史	一般財団法人日本特許情報機構　情報運用部 情報運用課　主任
	星山　直人	一般財団法人日本特許情報機構　情報運用部 情報整備課　主任
	王　　向莉	一般財団法人日本特許情報機構　調査研究部 研究企画課
		（※1：拡大評価部会部会長、※2：拡大評価部会メンバー）
事 務 局	小松　浩平	株式会社インターグループ
	大久保　あかね	株式会社インターグループ

平成 27 年度 AAMT/Japio 特許翻訳研究会・活動履歴

平成 27(2015)年 5 月 15 日
第 1 回 AAMT/Japio 特許翻訳研究会・拡大評価部会
（於キャンパス・イノベーションセンター東京）

平成 27(2015)年 6 月 26 日
第 2 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 27(2015)年 7 月 17 日
第 3 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 27(2015)年 9 月 25 日
第 4 回 AAMT/Japio 特許翻訳研究会・拡大評価部会
（於キャンパス・イノベーションセンター東京）

平成 27(2015)年 10 月 30 日
第 15 回機械翻訳サミット（於米国（マイアミ）Hyatt Regency Miami）

平成 27(2015)年 12 月 4 日
第 5 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 28(2016)年 1 月 29 日
第 6 回 AAMT/Japio 特許翻訳研究会・拡大評価部会
（於キャンパス・イノベーションセンター東京）

平成 28(2016)年 3 月 11 日
第 7 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 28(2016)年 3 月 31 日
『平成 27 年度 AAMT/Japio 特許翻訳研究会報告書　　機械翻訳及び機械翻訳評価に関する研究及び　海外調査』完成

１．はじめに

産業技術総合研究所人工知能研究センター　研究センター長
東京大学名誉教授
AAMT／Japio 特許翻訳研究会委員長

辻井　潤一

私事になって恐縮ですが、昨年 5 月に産業技術総合研究所の中に人工知能研究センターが設立され、マイクロソフト研究所からその研究センターに就任しました。4 年ぶりに日本に帰ってきたわけですが、大学の研究者から国立研究機関の研究センター長という立場の違いもあって、久しぶりの日本が随分と違って見えています。

人工知能がブームになっているということもあるのでしょうが、この分野を取り巻く状況が急激に活発化しています。機械翻訳をとってみても、東京オリンピックを目指した多言語の音声翻訳システムの研究が活況を呈していますし、マイクロソフトの音声翻訳も日本語を取り入れようと研究、開発に力を入れています。音声翻訳の音声処理部分は、人工知能からの深層学習技術の取り込みで性能が向上し、中核の翻訳部分にもこの技術の取り込みが始まっています。特許翻訳のような専門性の高い、複雑な構文構造をもった文の翻訳に神経回路的な技術がどこまでその能力を発揮できるのか注目すべきところでしょう。規則を中心として機械翻訳から統計モデルを使った機械翻訳への移行が 1980 年の末から始まり、20 年間で分野の主流となったのですが、同様な技術の変革が起こりつつあるように思えます。人工知能、言語理解などの研究が、これからさらに活発化していくことと思いますが、その中で機械翻訳の技術も大きく変化していくでしょう。

本委員会では、特許のような専門性の高い翻訳に不可欠な専門用語の取り扱いや、機械翻訳システムの評価の問題を中心的に取り扱ってきましたが、これらの課題の重要性は技術の変革にかかわらず重要な課題となるでしょう。ただ、技術の変遷は、専門用語の取り扱いや機械翻訳の質の評価の具体的な手法には影響を与えることになります。今回の報告書でのまとめが、この新たな変遷への準備となることは間違いありません。

この報告書は、我々の 1 年間の活動をまとめたものです。読者諸賢の参考になれば幸いです。

2. 語彙・構文に関する技術

2．1　Enhancing Function Word Translation with Syntax-Based Statistical Post-Editing

Kyoto University　John Richardson
Toshiaki Nakazawa
Sadao Kurohashi

2.1.1　Introduction

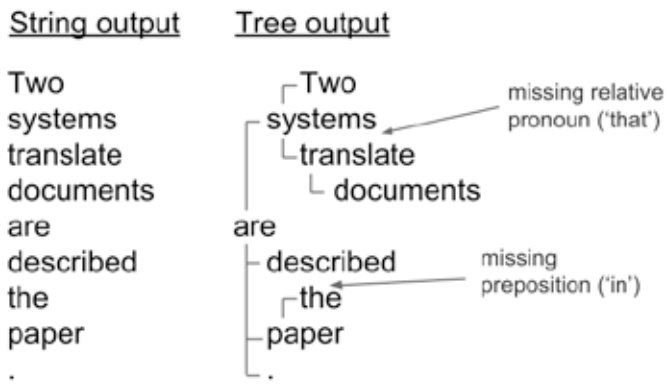
A high level of machine translation fluency is sought after in all subject domains. Translations with high adequacy however are especially important in patent and scientific translation, where it is particularly necessary to preserve the meaning of the input sentence in the generated translation.

Error analysis of state-of-the-art machine translation systems has shown that poorly translated function words are a major cause of loss in translation comprehensibility. For example, negation and passive structures can completely reverse their meaning when missing the correct function words, and it is important for understanding to select appropriate prepositions. We have also found that lack of (or incorrectly placed) relative pronouns has a large effect on preserving sentence meaning, and that badly formed punctuation impedes understanding.

Surprisingly few studies have been made specifically on improving function word translation for statistical machine translation systems, despite this having been looked at in rule-based systems (Arnold and Sadler, 1991). While we were unable to find any previous work on function word statistical post-editing, function words have been used to generate translation rules (Wu et al., 2011). The most similar approach to our method of editing function words used structural templates and was proposed for SMT (Menezes and Quirk, 2008). Statistical post-editing of MT output in a more general sense (Simard et al., 2007) and learning post-editing rules based on common errors (Elming, 2006; Huang et al., 2010) have shown promising results. The majority of statistical post-editing methods work directly with string output, however a syntactically motivated approach has been tried for post-editing verb-noun valency (Rosa et al., 2013).

We believe that the intended meaning of a sentence is often unclear from flat MT output. For example, in Figure 1, the intended meaning is much clearer from the dependency tree representation: we cannot tell easily that ‘translate documents’ is a relative clause (missing the relative pronoun ‘which’ or ‘that’) and that ‘the paper’ is a prepositional phrase (missing the preposition ‘in’) rather than the direct object of ‘described’. Based on this observation, we attempt to exploit the target structure of the output of a dependency tree-to-tree machine translation system in order to understand better the cause of the function word errors and therefore correct them more effectively.

Figure 1: String vs Tree Output



word language model.

We assume a set of function words F, a subset of the entire target-side vocabulary. We also define an empty token '<none>' which represents the lack of a function word. A root node and leaf nodes can be added to the tree to allow insertion of function words as the sentence root and leaves respectively.

A dependency tree can be decomposed into token–head pairs (t,t′). We derive a simple language model P(f | t,t′) approximating the probability of function word f ∈ F being inserted between t and t′. The model is estimated over the training data by counting the occurrence of (f,t,t′) tuples where f is a function word appearing between t and t′. Note that to make this definition well-defined, we strictly require that function words have only one child. The probability P(f | t,t′) is then calculated as:

$$P(f \mid t,t') = \frac{count(f,t,t')}{\sum_{g \in F \cup <none>} count(g,t,t')}$$

In our experiments we include part-of-speech tags inside tokens to reduce homonym ambiguity (e.g. use 'set-NN' instead of 'set'). We also split P(f | t,t′) into two cases, P_{left}(f | t,t′) and P_{right}(f | t,t′), to consider the difference between t being a left or right descendant of t′. We will write P_s to refer to whichever of P_{left} or P_{right} applies in each case.

2.1.2.1 Operations

For a token–head pair (t,t′), word insertion is performed when P_s(f | t,t′) > P_s(<none>| t,t′) for some function word f. We choose the function word with the highest probability if there are multiple candidates. Replacement of function word t is performed similarly if P_s(child(t) | f,t′) > P_s(child(t) | t,t′) for some other function word f. Similarly we choose the best f if there are multiple candidates. Deletion can be performed using the same method as for replacement by adding the function word '<none>' to F.

2.1.2.2 Filtering Replacements/Deletions with Word Alignments

In the majority of cases we found it counter-productive to replace or delete function words corresponding directly to non-trivial source words in the input sentence. For example, in a Chinese–English translation task, consider the two translations:

- 听/音乐 (listen/music) → listen to music
- 下面/100/米 (below/100/m) → 100m below

In the first sentence, the function word 'to' in the English translation has no corresponding word in the Chinese input and therefore its existence is based only on the target language model. In contrast, the preposition 'below' in the second sentence directly corresponds to '下面 (below)' in the input and care should be taken not to delete it (or change it to a completely different preposition such as 'above').

We therefore propose restricting replacement/deletion to function words that are aligned to trivial or ambiguous source-side words (function words without concrete meaning, whitespace, punctuation). This allows us to change for instance the unaligned 'to' in 'listen to' but not 'below' with an input alignment. The source–target word alignments are stored in the translation examples used by the baseline SMT system and kept track of during decoding.

2.1.3 Experiments

We performed translation experiments on the Asian Scientific Paper Excerpt Corpus (ASPEC) for Japanese–English translation. The data was split into 3 million training sentences, 1790 development sentences and 1812 test sentences.

We defined English function words as those tokens with POS tags of functional types such as determinants and prepositions, and treated Japanese particles as function words for the purposes of alignment-based filtering. The primary post-editing model was trained on the training fold of the ASPEC data. Since our model only requires monolingual data, for comparison we also trained a separate model on a larger (30M sentences) in-house monolingual corpus (Mono) of technical/scientific documents.

For the baseline SMT system we used KyotoEBMT (Richardson et al., 2014), a state-of-the-art dependency tree-to-tree translation system that can keep track of the input–output word alignments. Post-editing was performed on the top-1 translation produced by the tree-to-tree baseline system.

Japanese segmentation and parsing were performed with Juman and KNP (Kawahara and Kurohashi, 2006). For English we used NLParser (Charniak and Johnson, 2005), converted to dependency parses with an in-house tool. Alignment was performed with Nile (Riesa et al., 2011) and an in-house alignment tool. We used a 5-gram language model with modified Kneser-Ney smoothing built with KenLM (Heafield, 2011).

2.1.3.1 Evaluation

Human evaluation was conducted to evaluate directly the change in translation quality of function words. We found that automatic evaluation metrics such as BLEU (Papineni et al., 2002) were not sufficiently

sensitive to changes (the change rate is relatively low for post-editing tasks) and did not accurately measure the function word accuracy.

In human evaluation we asked two native speakers of the target language (English) with knowledge of the source language (Japanese) to decide if the system output was better, worse, or neutral compared to the baseline. A random sample of 20 edited sentences were selected for each experiment and the identity of the systems was hidden from the raters. The Fleiss' kappa inter-annotator agreement (Fleiss et al., 1981) for wins/losses was 0.663, and when including neutral results this was reduced to 0.285.

2.1.3.2 Tuning and Test Experiments

We first perfomed a preliminary tuning experiment on the development fold of ASPEC to investigate the effect of model parameters. The results in Table 1 show for each row the model settings, the number of wins (+), losses (–) and neutral (?) results compared to the baseline, and the change rate (CR) over the entire development set.

The first three settings ('OnlyIns', 'OnlyRep', 'OnlyDel') show the effects of allowing only insertions, replacements and deletions respectively without using source–target alignments. We can see that the quality for deletions is lower than insertions and replacements, and error analysis showed that the major cause was deletion of function words aligned to content words in the input.

We reran the experiments using the alignment-based filtering ('AlignA' and 'AlignB') and found the results improved. While possible to achieve a higher change rate by allowing all three operations, we could only achieve a slight increase in accuracy by disallowing replacements (the setting 'AlignB'). The difference was mainly due to alignment errors, which caused more serious problems for replacement as they were able to alter sentence meaning more severely.

The best settings in the tuning experiment ('AlignB') were used to conduct the final evaluation on the unseen test data from ASPEC. We also compared models trained on the ASPEC training fold and on our larger monolingual corpus. Table 2 shows the final evaluation results. The results on the test set show significant improvement on win/loss sentences at $p < 0.05$. There was no clear improvement gained by increasing the size of model training corpus, however the change rate could be improved by using more data.

2.1.4 Conclusion

The experimental results show that in general our proposed method is effective at improving the comprehensibility of translations by correctly editing function words. We found that using source–target alignments was effective in avoiding simple errors however there remained some trickier cases where the alignment information was not sufficient, for example when function words were null or incorrectly aligned. The remainder errors were primarily caused by incorrect parsing and sparsity issues.

In this study we have shown that target-side syntax can be used effectively to improve the quality of scientific domain machine translation through the automatic post-editing of function words. We have presented an algorithm for inserting/deleting/replacing function words based on a simple tree-based

Table 1: Results of tuning experiment on development set.

	Insert	Replace	Delete	Align	+	-	?	CR
OnlyIns	Yes	No	No	No	10	6	4	2.3
OnlyRep	No	Yes	No	No	11	7	2	5.5
OnlyDel	No	No	Yes	No	7	8	5	8.6
AlignA	Yes	Yes	Yes	Yes	11	7	2	10.5
AlignB	Yes	No	Yes	Yes	11	4	2	3.3

Table 2: Final evaluation results on unseen data.

	Insert	Replace	Delete	Align	+	-	?	CR
ASPEC	Yes	No	Yes	Yes	12	5	3	2.3
Mono	Yes	No	Yes	Yes	11	5	4	4.1
Both	Yes	No	Yes	Yes	23	10	7	3.9

language model and demonstrated the effectiveness of using source–target alignments to improve accuracy. In the future we plan to extend the model to provide more robustness against parsing/alignment errors and experiment with other language pairs.

2.1.5 Acknowledgements

We are grateful to Raj Dabre for his assistance in conducting the human evaluation.

References

[1] Arnold, D. and Sadler, L. (1991). EuroTra: An assessment of the current state of the ECs MT Programme. In Working Papers in Language Processing.

[2] Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 173–180. Association for Computational Linguistics.

[3] Elming, J. (2006). Transformation-based corrections of rule-based MT. In EAMT 11th Annual Conference.

[4] Fleiss, L., Levin, B., and Paik, M. C. (1981). The measurement of interrater agreement. In Statistical methods for rates and proportions (2nd ed), pages 212–236. Wiley.

[5] Heafield, K. (2011). KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation.
[6] Huang, A., Kuo, T., Lai, Y., and Lin, S. (2010). Discovering correction rules for auto editing. International Journal of Computational Linguistics and Chinese Language Processing, 15(3- 4).
[7] Kawahara, D. and Kurohashi, S. (2006). A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06, pages 176–183. Association for Computational Linguistics.
[8] Menezes, A. and Quirk, C. (2008). Syntactic models for structural word insertion and deletion during translation. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 735–744, Honolulu, Hawaii. Association for Computational Linguistics.
[9] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318. Association for Computational Linguistics.
[10] Richardson, J., Cromières, F., Nakazawa, T., and Kurohashi, S. (2014). KyotoEBMT: An example-based dependency-to-dependency translation framework. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 79–84, Baltimore, Maryland. Association for Computational Linguistics.
[11] Riesa, J., Irvine, A., and Marcu, D. (2011). Feature-rich language-independent syntax-based alignment for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 497–507. Association for Computational Linguistics.
[12] Rosa, R., Mareček, D., and Tamchyna, A. (2013). Deepfix: Statistical post-editing of statistical machine translation using deep syntactic analysis. In Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics.
[13] Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical phrase-based post-editing. In NAACL.
[14] Wu, X., Matsuzaki, T., and Tsujii, J. (2011). Effective use of function words for rule generalization in forest-based translation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 22–31, Portland, Oregon, USA. Association for Computational Linguistics.

２．２　機能語に着目した特許文の分割

山形大学名誉教授　横山晶一

2.2.1 はじめに

特許文において、課題や解決手段、請求項１の部分が、120 文字を超える長大な文になるとともに、複雑な係り受け構造を持つということは、すでに何度も言及してきた[1-4]。

これまでも、特許文解析に特徴的な複雑な係り受け構造を解明するため、並列接続詞[5]や並立助詞[6, 7]、入れ子構造[8]について調査し、誤り自動修正システムを構築してきた。また、並列に重要な役割を果たす名詞を、広く主辞（接尾辞）としてとらえることによって、特許文の係り受けを修正するシステムについても述べてきた[7, 9]。

本稿では、長い修飾句の中に含まれる機能語に着目し、その性質から修飾句を形式的に分割する可能性について調査した結果を述べる。

なお、本稿は、[10]に基づいて新たにデータを追加し、新しい知見を付け加えたものである。

2.2.2 目的と資料

土を供給した育苗容器を移送する移送台１の上方位置に、育苗容器２に供給した床土を育苗容器２の上面に摺接しながら均平するスクレーパー３と、育苗容器２に供給した床土を掃き出しながら均平する回転均平ブラシ１５とを有するもの**において**、前記スクレーパー３は前記移送台１に対して上下する上下移動体１１に高さ調節自在に取付け、前記回転均平ブラシ１５は前記上下移動体１１に回転のみ自在に設け、前記上下移動体１１に対して最も下方に位置させた前記スクレーパー３の下面は前記回転均平ブラシ１５の下縁と一致させた均平装置。

図１　長い修飾句を持つ特許文請求項（公開番号：特開 2003-180168）

図１に示すように、特許文の請求項１は、全体が一つの文で書かれることが多く、長大で複雑な文になりやすい。多くは、長い修飾句を有する並列構造になっている。これまで、この修飾句の構造を解明する手がかりとして、並列接続詞、並立助詞について調査し、係り受けの誤りを修正するシステムを構築してきた。本稿では長い修飾句を分割するための手段として、修飾句に含まれる機能語に着目した。

機能語とは、ここでは、日本語の複数の形態素から成る複合語の中で、いわゆる「つなぎ言葉」的な役割をになうものと定義し、「～において」、「～であって」、「～に関して」などを示す。図１では、「有するもの**において**」の部分がそれにあたる。

ここでは、2003 年の公開特許から約 100 を抽出し、その中で１文が 120 文字を超える請求項

１を持つもの（箇条書き等で複数の文章から成るものを除く）を 63 選んだ。内訳は、最も短いもので 121 文字、最も長いもので 422 文字、平均で 222 文字である。100 文字台が 29、200 文字台が 22、300 文字台が 10、400 文字台が 2 であった。

この中に含まれる機能語を調査した。調査の結果、図１にも示した「～において」が最も多く、63 例の中に 35 例含まれていた。また、「～であって」が 14 例含まれていた。63 例の中で、1 文の中にこれらが両方含まれている例は 2 例存在する（後述）。そこで、最も多く含まれていた「～において」について、詳細な分析を行った。

2.2.3 特許文に含まれる「～において」の性質

図１では、文頭から「～において」までの句を、名詞句全体を形作る文の最後の「均平装置」と切り離しても名詞句全体に対して影響を与えないと考えられる。図１は、「～において」が含まれる典型的な例ではないが、こうした例においてもこのようなことが言える。

「～において」が含まれる例として、特許文の請求項で最も典型的なのは、図２に示すようなものである。

> 鶏卵を等階級別に所定の集合場所に分配する分配手段と、分配手段から分配された鶏卵を受け取り所定の容器に充填する充填手段とを備えた鶏卵の選別充填装置**において**、少なくとも１つの等階級には集合場所が複数設けられており、分配手段の単位時間当たりの処理能力は少なくとも１つの充填手段の単位時間当たりの処理能力を上回り、分配手段は、集合場所が複数設けられている同一等階級の鶏卵を前記複数の集合場所のいずれの集合場所に分配すべきかを予め定められた各集合場所の優先順位に従って行う鶏卵の選別充填装置。

図２　「～において」を含む請求項の典型例（公開番号：特開 2003-18092）

図２も、図１の例と同様に、「～において」のところで、全体を分割しても名詞句の係り受けには影響を与えないことが分かる。図１と異なるのは、全体を受ける最後の名詞句である「**鶏卵の選別充填装置**」が、前の「～において」の前にも現れていることである。すわなち、この名詞句は、全体として「～名詞句 A において、…した（である）名詞句 A」という構造を持っている。この調査で得られた 35 例の「～において」を含む請求項のうち、23 例がこの構造を持っていた。したがって、この構造を持つ句（文）を解析するためには、「～において」の前後で分離した、より短い句を解析すればよいことが分かる。

図１のような例は、7 例が確認された。残りの 5 例のうち 3 例は、やや微妙で、分離することができる可能性もあるが、修飾句の中に含まれていて、係り受けの意味を考慮しないと扱えない可能性も排除できない。その例を図３に示す。

図３では、「～において」までの句が、修飾句として最後の名詞句にかかっているようにも見えるし、副詞句的に、切り離しができるようにも見える。こちらはさらに分析が必要である。また、「～において」の後に読点を含まない 2 例（そのうち 1 例は「～であって」の後ろに読点を含む

（次節で示す)）は、いずれもその部分では分割できず、比較的短い修飾句の一部をなしているものであった。

結論として、図２のような典型例を機械的に分割すれば約 63%、図１，２を分割すれば 83%近くが、より短い句に分割して解析を行うことができることが分かった。

摺動自在の刈刃を装着した切断部を機体の進行方向に対して横設し､駆動装置を前記切断部の後方に設けて切断した穀稈を揚上搬送装置で脱穀部に供給するコンバインの刈取部**において**、刈刃（１）と一体化して往復運動をするナイフヘッド（６）の駆動点（X）を前記刈刃（１）の摺動方向と同方向の平行移動をする受動構造にしたことを特徴とするコンバインの刈刃駆動装置。

図３　「～において」を含む請求項の非典型例（公開番号：特開 2003-180109）

2.2.4 「～であって」の分析

ロール状に巻かれた農用マルチシートを畦に沿って敷設する農用治具**であって**、ロール状の農用マルチシートの巻き芯の孔に挿通する棒状体と、棒状体の両端部に着脱自在に取付けてロール状の農用マルチシートを畦の上に転動させながら引き出す手引き用の紐状体を備えて成ることを特徴とする農用治具。

図４　「～であって」を含む請求項の例（公開番号：特開 2003-180174）

合成樹脂フィルムを含むシートにより作られ、内部にきのこ種菌を収納するための袋**であって**、前記シートを筒状にして、筒状の周方向の一部**において**シートの両端部同志を互いにオーバーラップさせあるいはシートの両端部の端面を突き合わせ状態とし、このシートの両端部のオーバーラップ部分あるいは突き合わせ部分の内側に通気性を備えた不織布製のシートを位置せしめ、この不織布製のシートの幅方向両端部を前記袋のシートの周方向の両端部近傍の内面側を構成するフィルム層に重ねて、ヒートシールにより融着してなり、前記筒状の軸芯方向の両端部にあってはヒートシールにより閉じられるように構成したことを特徴とするきのこ種菌収納用袋。

図５　「～であって」と「～において」をともに含む請求項の例（公開番号：特開 2003-180157）

「～であって」を含む句は、調査した例の中に 14 例と、「～において」よりかなり少ないので、あまり断定的なことは言えないが、図４のように、「～において」と同様に、「～名詞句 A であって、…した（である）名詞句 A」という構造を持つものが 8 例見出された。また、図１のタイプ、すなわち同じ名詞句が 2 回現れないものが 2 例（1 例はやや微妙）あった。残りの 4 例は、機能

語が比較的文頭に近いところに現れる例で、前節で述べたものとやや似た構造を持ち、後ろの名詞に係るとも、分離できるとも取れるものであった。

また、前節で述べた、「～において」と「～であって」を両方含む例は図５のようなものであるが、この例においては、「～であって」で分割し（余り意味はないが)、「～において」では分割できないと思われる。もう 1 例では逆に、「～において」の後の読点で分割し、「～であって」の後には読点がないので分割しない方が望ましい。

このように、読点の有無や修飾句内の位置によっても分割しうるかどうかが左右される。

2.2.5 問題点と今後の検討

調査した例が 63 と少ないうえに、並列助詞の「と」との関係や、長い修飾句内の他の形態素との位置関係や意味的な関係については、まだ詳細な分析を行っていない。これらの機能語を含まずに、並列構造のみで長い文を構成する例も多くみられる。また、調査した機能語も、ここに述べた 2 つのみで、他の機能語についての調査は今後の検討課題である。

しかしながら、この調査によって、長い修飾句が機能語を境としてもう少し短い修飾句や名詞句に分割できる可能性が示唆された。今後は、さらに多くのデータに当たって本稿で得られた考察を確認するとともに、修飾句内の細かい構造に踏み込んで調査、解析することによって、長い修飾句の文法的、意味的な構造や係り受け構造をさらに明らかにしていきたい。

参考文献

[1] 横山晶一、高野雄一：語のグループ化を用いた特許文動詞の自動訳し分けに関する調査、Japio YEAR BOOK (2011) pp.234-237

[2] 横山晶一、高野雄一：特許文の英語への訳し分けと述語の関係、Japio YEAR BOOK (2010) pp.274-279

[3] 横山晶一：特許文の英語への訳し分けと格フレームとの関係、Japio YEAR BOOK (2009) pp.262-265

[4] 横山晶一：動的シソーラスを用いた特許文の解析システム、科学技術研究費成果報告書(2007～2009)

[5] 横山晶一：特許文における接続詞と係り受けの構造、Japio YEAR BOOK (2008) pp.68-73

[6] 横山晶一：特許文解析誤り自動修正システムと正確な翻訳のための特許文の分割、Japio YEAR BOOK (2007) pp.228-233

[7] 髙橋尚矢、横山晶一：接続詞と主辞に着目した特許文の並列構造解析、Japio YEAR BOOK (2014) pp.242-245

[8] 髙橋尚矢、横山晶一：特許文における入れ子構造の調査、Japio YEAR BOOK (2013) pp.266-270

[9] 横山晶一：接尾辞に着目した特許文の並列構造解析、Japio YEAR BOOK (2012) pp.250-253

[10] 横山晶一：機能語に着目した特許文の調査、Japio YEAR BOOK (2015) pp.314-316

２．３　対訳語抽出における Decipherment 法と文脈に基づく手法の比較

東京大学　範 暁蓉
愛媛大学　二宮 崇

2.3.1　はじめに

対訳辞書は機械翻訳において非常に重要な言語資源であり、二言語のコーパスは対訳辞書の自動抽出のための重要なリソースである。パラレルコーパスから自動的に質の高い対訳辞書を抽出できることは知られているが、大規模なパラレルコーパスが利用できる分野は非常に限られており、そのため、得られる対訳辞書の分野も限られてしまう問題がある。近年、この問題を解消するため、コンパラブルコーパスから自動的に対訳辞書を抽出する研究が盛んに行われている。

コンパラブルコーパスから対訳辞書を自動的に抽出する手法は、様々な手法が提案されている。基本対訳辞書がある場合には、文脈に基づく手法が主な手法として用いられているが、基本対訳辞書がない場合は、Decipherment 法が有効な対訳辞書抽出手法として考えられる。文脈に基づく手法は一定量以上のコーパスを用いる必要があるが、一般的に、対訳語の抽出精度は、コーパスサイズが一定量になるまで、コーパスサイズの拡大と共に増加することが知られている (Darja Fišer ら、2011)。一方、Decipherment 法は、大量のコーパスを必ずしも必要とはしないが、コーパスサイズの拡大と共に精度が増加するかどうかまだ確認されていない。

本稿では、まず、コンパラブルコーパスのサイズに対し、この二つ手法の性能がどのように変化するか、その影響について実験で詳しく調べる。次に、この二つの手法を組み合わせた方法について実験を行い、その実験結果を報告する。

本稿の構成は以下のようになっている。2.3.2 節では、文脈に基づく手法と Decipherment 法によるコンパラブルコーパスからの対訳語抽出手法を説明する。2.3.3 節は、コーパスのサイズを変化させ、二つの手法による対訳語抽出の性能を評価する実験を行い、その結果について報告する。2.3.4 節は、二つ手法を組み合わせる手法と実験について報告する。2.3.5 節で本稿の主旨をまとめ、今後の課題について述べる。

2.3.2　コンパラブルコーパスからの対訳語抽出

本節では，今回の実験で使用する文脈に基づく手法と Decipherment 法による対訳語の抽出手順をそれぞれ詳しく説明する。

2.3.2.1　文脈に基づく手法による対訳語抽出

文脈に基づく手法による対訳語抽出は、一般に「ある言語で共起する語があれば，翻訳後の言語でもそれらの翻訳語は共起する」という仮説に基づき、単語の文脈情報を用いて訳語を推定する方法である。この手法は次の三つのステップにより実現される。

ステップ１　文脈情報（文脈ベクトル）の収集と正規化

原言語コーパス F、目標言語コーパス E から、それぞれ対象単語の文脈情報を収集し、文脈ベクトルを生成する。文脈情報にはさまざまな情報が用いられており、対象単語の前後にでてくる単語や構文解析、係り受けの結果などが文脈情報としてよく用いられている。特に、構文解析や係り受け解析の結果を文脈の情報として用いることにより、精度の高い対訳語が抽出されることが知られているが、構文解析と係り受け解析は、文の長さと文の読みづらさが精度に大きく影響を与える問題がある。今回の実験は対象単語の前後にでてくる N 個の単語を文脈情報として利用する（今回の実験の場合、$N = 5$）。対象単語の出現頻度と周りの単語の共起頻度をそれぞれカウントする。また、相関指標である discounted 対数オッズ（LO）の値を用いて文脈ベクトルを正規化する。LOの値は以下の式で表される。

$$LO(i,j) = \log \frac{\left(cooc(i,j) + \frac{1}{2}\right) \times \left(cooc(\neg i, \neg j) + \frac{1}{2}\right)}{\left(cooc(i, \neg j) + \frac{1}{2}\right) \times \left(cooc(\neg i, j) + \frac{1}{2}\right)} \qquad \text{式 (1)}$$

ただし、i は対象単語であり、$cooc(i,j)$ は対象単語 i と単語 j の共起頻度である。

ステップ 2　文脈ベクトルの同一言語化

基本対訳辞書を用いて原言語の文脈ベクトルを目標言語に翻訳する。原言語の単語に対して複数の訳語が存在する場合、すべての訳語に対し重みを付けて使用する。

ステップ 3　分布類似度で訳語候補の選択

分布類似度の計算に用いられる距離尺度はいくつか存在する。典型的な尺度として Jaccard 係数や Cosine 距離、Dice 係数などが存在する。Weeds は博士論文（Weeds, 2003）の中で、これらの類似度アルゴリズムに関する詳細な説明を与えている。今回の実験では Cosine 距離を用いて分布類似度を計算する。

2.3.2.2　Decipherment 法による対訳語抽出

Decipherment 法（Ravi ら 2011）は非パラレルコーパスを利用して、片方の言語の暗号解読(deciphering) とみなして機械翻訳を行う統計的機械翻訳手法である。この手法では原言語文 f が目標言語文 e の暗号化されたものとして設定する。式(2) に示す生成モデルで e が周辺化されることにより f の生成確率が求まる。EM アルゴリズムを用いて式(2) を最大化することで翻訳モデルが得られる。

$$\underset{\theta}{\operatorname{argmax}} \prod_{f} \sum_{e} p(e) \cdot p_\theta(f|e) = \underset{\theta}{\operatorname{argmax}} \prod_{f} \sum_{e} p(e) \cdot \prod_{i=1}^{n} p_\theta(f_i|e_i) \qquad \text{式(2)}$$

ただし、$p(e)$ は言語モデル、$p_\theta(f_i \mid e_i)$ は翻訳確率、$f = f_1 \cdots f_n$、$e = e_1 \cdots e_n$ である。

Ravi らの Decipherment 法は基本対訳辞書を必要とせず、翻訳モデルと並び替えモデルを学習する手法であるが、コーパス中の長い文に対して、高い計算量を要する。Dou ら（2012）により、bigram を用いて文を近似することで十分精度の高い decipherment 法が実現されることが示されているため、今回の実験は次の式(3)に示す Bigram Decipherment モデルを利用する。

$$\underset{\theta}{\operatorname{argmax}} \prod_{f} \sum_{e} p(e_1 e_2) \cdot \prod_{i=1}^{2} p_\theta(f_i|e_i) \qquad \text{式(3)}$$

本研究では、Ravi ら（2011）の研究と同じ生成ストーリーで翻訳モデルを生成する。Ravi ら（2011）の EM アルゴリズムでの生成過程は次のとおりである。

（1）目標言語の文 e が確率 $p(e)$で生成する。

（2）隣接する二つの単語間に NULL 文字を挿入する。

（3）文 e の中の各単語 e_i（NULL を含む)と原言語の文 f の各単語 f_i に翻訳確率 $p_\theta(f_i|e_i)$を付ける。

（4）原言語の**文 f の中に、**隣接する二つの単語 f_{i-1} と f_i の順を変更する。

（5）NULL **を削除して、**原言語の**文 f を出力する。**

2.3.3 コーパスサイズの影響

Emmanuel Morin ら（2014）は、コンパラブルコーパスが変化するとき、文脈に基づく手法による対訳語の抽出精度への影響を実験で詳しく調べた。今回、コンパラブルコーパスのサイズが変化するとき、Decipherment 法へどのように影響を与えるか、実験で評価する。

2.3.3.1 実験用データ

実験用データは、以下のように用意した。

（1）原言語を日本語、目標言語を英語とする。実験データは 2 種類、バランスコーパスと非バランスコーパスを設定する。バランスコーパスは、原言語のコーパスサイズと目標言語のコーパスサイズがほぼ同じである。非バランスコーパスは、目標言語のコーパスサイズが原言語のコーパスサイズと大きく異なる。本稿では、言語資源として日英新聞記事対応付けデータ(JENAAD)[1]（374,085 対訳文対）の内、日本語 1,000 文を抽出し、原言語コーパスとする。英語コーパスから、日本語 1,000 文の訳語文以外の 5,000 文を目標言語のコーパスとする。目標言語のコーパスを 5 つに分け、各部分コーパスは英語 1,000 文から成る。この 5 つの英語コーパスをそれぞれ日本語コーパスに対するバランスコーパスとする。非バランスコーパスは英語の 5,000 文から、1,000 文、2,000 文、…、5,000 文を抽出して生成する。各コーパスに含まれるタイプとトークンの数を表 1 にまとめる。

（2）基本辞書は JMDict[2]の英日辞書（580,077 対訳）を使用した。

（3）コーパスの前処理

日本語文に対して、単語分割、全角符号は半角に変更する、POS タグ付けとストップワードの除去など前処理とを行った。英単語に対して、トークナイザ（tokenizer)、小文字化

[1] http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/index-ja.html
独立行政法人情報通信研究機構作成

[2] http://www.edrdg.org/jmdict/edict_doc.html

表 1: コーパスの諸元

コーパス	バランスコーパス		非バランスコーパス	
	types	tokens	types	Tokens
Japanese	11,061	267,602	11,061	267,602
English corpus1	8,794	221,472	8,794	221,472
corpus2	9,648	237,188	12,288	458,660
corpus3	9,975	246,254	14,921	704,914
corpus4	10,239	249,149	17,172	954,063
corpus5	10,170	235,261	19,033	1,189,324

(lowercase)、見出し語化 (lemmatization)、POS タグ付けとストップワードの除去など前処理は行わなかった。以下に使用したツールの一覧を示す。

- 言語モデル：Srilm[3]
- 日本語の単語分割と POS タグ付け：Mecab[4]
- 英語の前処理：Stanford CoreNLP 3.6.0[5]

2.3.3.2 実験

実験の手順を以下に示す。

(1) 各バランスコーパスから 2.3.2 節に説明した二つの手法で対訳語を抽出する。

(2) 各非バランスコーパスから 2.3.2 節に説明した二つの手法で対訳語を抽出する。

2.3.3.3 実験の結果

評価尺度として、上位 1 位における精度(Top1 精度)を用いて実験結果を評価した。評価用の正解となる辞書（115 英日単語対）は手作業で作成した。今回の実験の結果は図 1 に示す。

実験の結果から、どちらの手法でも、コーパスサイズの拡大と共に対訳語の抽出精度は高くなる傾向がわかる。Darja Fišer ら（2011）の研究によると、コーパスサイズはある量 (1,800 万単語) になると、文脈に基づく手法の抽出精度は増加しなかったことが報告されている。今回の実験では、実験データが小さかったため、精度が増加しなくなるコーパスの量については判断できなかった。

2.3.4 組み合わせ実験

文脈に基づく手法と Decipherment 法を組み合わると、対訳語の抽出精度にどのように影響を与

[3] http://www.speech.sri.com/projects/srilm/
[4] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[5] http://stanfordnlp.github.io/CoreNLP/

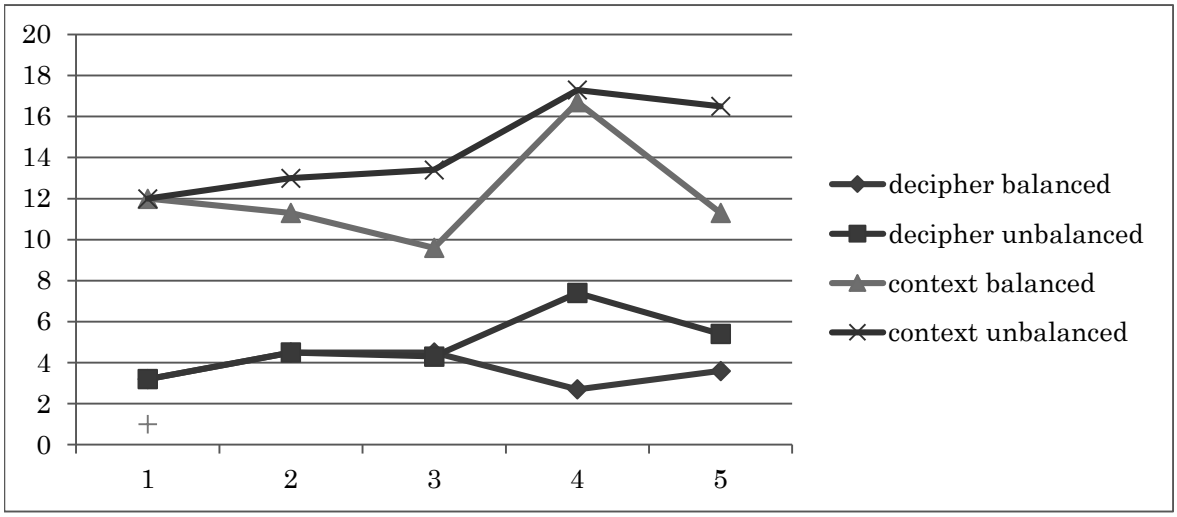


図 1: 対訳語の抽出精度

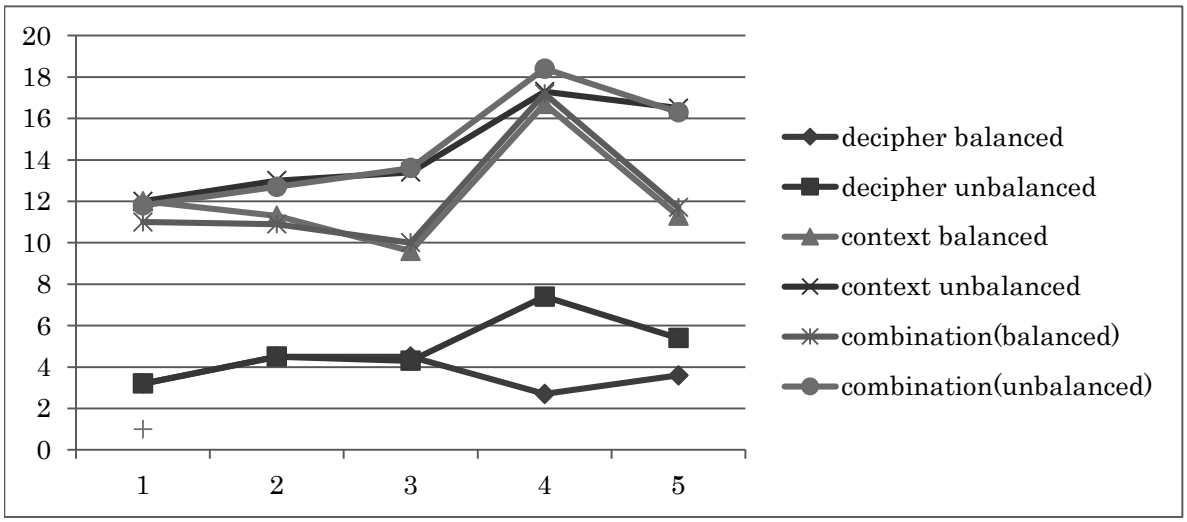


図 2: 組み合わせ手法の精度

えるか、実験で評価する。組み合わせ方法は、二つ手法の重み付け線形和で与えられ、次の式(4)で表される。

$$Sim_{comb}(e_i, f_j) = \gamma Sim_{context}(e_i, f_j) + (1-\gamma) Sim_{decipherment}(e_i, f_j) \quad \text{式(4)}$$

ただし、γ は 0.8 と設定した。

実験の結果は図 2 に示す。実験の結果から、組み合わせ手法は、それぞれ単一の手法よりも高い精度を実現することがわかった。しかし、Decipherment 法は時間がかかるので、改善する必要がある。表 2 は日本語単語「歴史」に対する各手法の訳語候補スコアを示す。

表 2: 日本語「歴史」の訳語候補

Candidate	$Sim_{context}$	$Sim_{decipherment}$	Sim_{comb}
history	0.186	0.3075	0.2
friendship	0.183	0	0.1464
rich	0.1353	0.2478	0.1578
certification	0.147	0.	0.147

2.3.5 まとめ今後の課題

本稿では、対訳語抽出のための文脈に基づく手法と Decipherment 法に対し、コーパスサイズが変化した場合の対訳語抽出への影響について、実験により調査した。実験結果より、対訳語の抽出精度はコーパスサイズの増加と共に高くなることがわかった。また、二つ手法の重みつけ線形結合も有効であることがわかった。今後は、精度向上のため、基本辞書を使った Decipherment 法による対訳語抽出を行う。

参考文献

Dou, Qing and Kevin Knight. "Large scale decipherment for out-of-domain machine translation." Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.

Emmanuel, Morin and Amir Hazem. "Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). 2014.

Fišer, Darja et al. "Building and using comparable corpora for domain-specific bilingual lexicon extraction." Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. Association for Computational Linguistics, 2011.

Ravi, Sujith and Kevin Knight. "Deciphering foreign language." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

Rapp, Reinhard. "Automatic identification of word translations from unrelated English and German corpora." Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999.

Weeds, Julie Elizabeth. "Measures and applications of lexical distributional similarity." University of Sussex, 2003.

２．４ パテントファミリーを用いた日中対訳専門用語の同定

筑波大学大学院システム情報工学研究科
龍　梓, 宇津呂　武仁, 山本　幹雄

2.4.1 はじめに

近年，中国国内における特許出願数は大幅に増加している．ここで，特許文書の言語横断検索等のサービスを実現するためには，中国特許文書の翻訳が不可欠である．しかし，特許翻訳において，機械翻訳や人手による翻訳を行う場合，高品質な翻訳を行うためには，大規模かつ高精度な対訳辞書が必須である．しかし，各国では，年々新しい技術開発が行われ，新しい専門用語が作られ，特許が申請されている．そのため，人手を介して高精度な対訳辞書を作成するためには，膨大な時間と労力を要する．よって，自動もしくは半自動的に日中専門用語対訳辞書を構築する手法が必要となる．

この問題に対して，文献［1］では，日中パテントファミリーから抽出された 360 万件の日中対訳特許文を対象として，統計的機械翻訳モデルより学習されるフレーズテーブルを利用し，さらに，機械学習手法として Support Vector Machine (SVM)［6］を用いて対訳専門用語を獲得する手法を提案した．そこで，本論文では特に，文献［1］の手法において用いられた素性の組み合わせに対して，評価実験によって最適な性能を達成する素性の組み合わせを同定する．さらに，日中対訳専門用語の同定において最も有効な単一の素性を同定する．評価結果においては，再現率 60%以上の条件のもとで，95%以上の適合率，および，87%以上の再現率，または，85%以上の F 値を達成した．そして，単一の素性としては，「要素合成法の確率」の素性が最も有効であることを示した．

2.4.2 日中対訳特許文

本論文では，約 360 万対の日中対訳特許文をフレーズテーブルの訓練用データとして使用した．この日中対訳特許文は，2004-2012 年発行の日本公開特許広報全文と 2005-2010 年中国特許全文を対象として，文献［2］の手法によって日中間で文を対応付け，スコア降順で上位の 360 万文対を抽出したものである．

2.4.3 句に基づく統計的機械翻訳モデルのフレーズテーブルを用いた訳語推定

本論文では，文献［1］と同様に，句に基づく統計的機械翻訳モデルのフレーズテーブルを用いて，日中専門用語対訳対の候補集合を選定する．

2.4.3.1 フレーズテーブルの作成

句に基づく統計的機械翻訳モデルのツールキットである Moses［3］を用いて，2 節で述べたデータからフレーズテーブルを作成した．フレーズテーブルを作成する際の準備として，Mecab［4］によって，日本語文の形態素解析を行い，一形態素を単語の単位とする．一方，中国語文に対しては，Chinese Penn Treebank を用いた Stanford Word Segment［5］を適用して，

形態素解析を行った．以上の準備を行った日中対訳文に対して，Moses を適用し，日中の句対および対応する確率を示したフレーズテーブルを作成した．

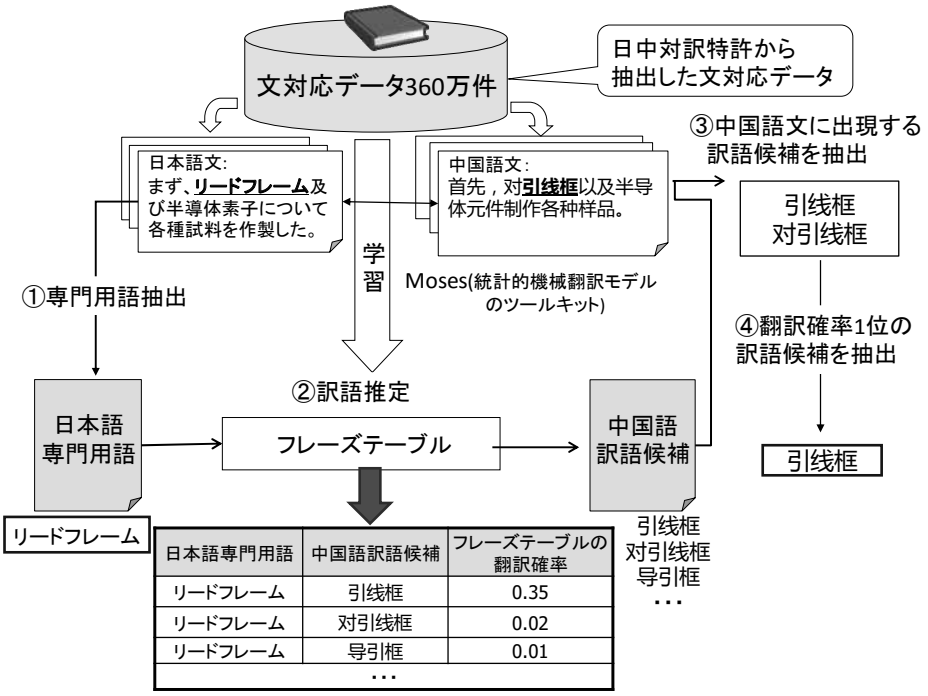


図 1: フレーズテーブルを用いた日中対訳専門用語の生成の流れ

2.4.3.2 一組の日中対訳文およびフレーズテーブルを用いた訳語推定

日中対訳文から対訳専門用語を推定する手順を図 1 に示す．訳語推定手法において，日本語専門用語 t_Jに対して，t_Jが出現する一つの日中対訳文<S_J, S_C>から，その日中対訳文に出現する t_Jの日中対訳対<t_J, t_C>を推定する．ここで，日本語専門用語 t_Jの中国訳語候補 t_Cは t_Jの訳語としてフレーズテーブルに存在し，かつ，日中対訳文<S_J, S_C>の中国語文 S_Cに出現するもので，フレーズテーブルにおける翻訳確率 $P(t_C|\ t_J)$が最大の訳語である．

2.4.4 SVM を用いた日中対訳専門用語の同定

2.4.4.1 参照用対訳対集合の作成

本論文では，文献［1］と同様に，人手で選定した578例の日本語専門用語を評価対象として用いた．次に，日本語専門用語が出現する全日中対訳文を収集し，前節の手順によって訳語推定を行う．その結果，2,533例の日中対訳専門用語を獲得した．最後に，人手で専門用語の対訳対としての適切さを判定し，正例を1,531例，負例を1,002例とした．

2.4.4.2 SVM の適用

前節で生成した 2,533 例の日中対訳専門用語を事例集合として，互いに素な 5 つの部分集合に分割した．ただし，各日本語専門用語に対しする複数の日中対訳専門用語は，同一の部分集合に分割された．また，本論文では，LIBSVM［7］を利用して，評価実験を行った．カーネル関数としては，動径基底関数、シグモイド関数、一次多項式カーネルおよび二次多項式カーネルを評価し，相対的によい動径基底関数カーネルを用いた．また，LIBSVM の出力した評価事例が各クラスに属する確率に下限を設定した．具体的には，5 個の部分集合のうち，4 個を訓練用事例集合として SVM の訓練を行った．そして，残りの 1 個を再び 2 分割して，1 個を調整用事例集合とし，もう 1 個を評価用事例集合とした．調整用事例集合を用いたパラメータの調整においては，評価用事例が正例クラスに属する確率の下限のパラメータの調整を行った．本論文では，日中対訳専門用語の適合率、再現率および F 値を最大化する調整を行った．ただし，適合率を最大化する場合は，再現率が 60%以上となるという条件のもとで調整を行った．再現率を最大化する場合は，適合率が 80%以上となるという条件のもとで調整を行った．

2.4.4.3 素性

本論文の手法には，表 1 に示すように，文献［1］と同じ素性を用いた．

表 1 日中対訳専門用語同定のための素性（文献［1］）

分類	素性名	定義
単言語素性	f_1：日本語専門用語の頻度	日本語専門用語が属する頻度レンジの番号(1~13)
	f_2：中国語専門用語の頻度	中国語専門用語が属する頻度レンジの番号(1~13)
二言語素性	f_3：翻訳確率	フレーズテーブルにおける翻訳確率
	f_4：訳語候補の順位(翻訳確率の降順)	同一日本語専門用語に対する訳語候補の順位(翻訳確率の降順)
	f_5：日中対訳専門用語の頻度	日中対訳専門用語が属する頻度レンジの番号(1~13)
	f_6：日本語専門用語と対訳共起頻度の頻度差	日本語専門用語の頻度−日中対訳共起頻度が上限値(本論文では105)以下の場合1，上限値を超える場合0
	f_7：訳語数	同一の日本語専門用語に対する中国語訳語候補数
	f_8：文単位の句対応制約の違反のない対訳文の割合	文単位の句対応制約の違反のない対訳文対の数/当該日中対訳専門用語の共起頻度
	f_9：要素合成法の確率	要素合成法により出力された訳語候補の確率

対訳文に対する日本語専門用語の頻度(f_1)と中国語専門用語の頻度(f_2)は単言語素性である．

二言語素性としては，フレーズテーブルによって各訳語候補の翻訳確率 (f_3)，同一日本語専門用語に対する訳語候補を翻訳確率の降順の順位 (f_4)，および日中対訳専門用語の共起頻度 (f_5)を用いた．また，日本語専門用語の頻度と日中対訳専門用語の共起頻度の差 (f_6)，同一日本語専門

用語に対する中国語訳語候補の数 (f_7)，文単位の句対応制約の違反のない対訳文の割合の素性(f_8)を用いた．さらに，要素合成法に基づき，フレーズテーブルを用いて，日本語専門用語と中国語専門用語それぞれに対して，フレーズテーブル中の要素によって専門用語を分割し，各々の専門用語を翻訳した場合の各要素の翻訳確率の積を要素合成法確率(f_9)としての素性を用いた．ただし，本論文では，フレーズテーブルを用いて構成要素の訳語推定を行う際，訳語の翻訳確率に下限値(0.005)を設定した．そして，同一の日本語専門用語を同じ分割の仕方によって同一の中国語専門用語に翻訳した場合はそれらの要素合成法の確率の和を用い，異なる分割の仕方によって同一の中国語専門用語に翻訳した場合はそれらの要素合成法確率の相加平均を用いた．次節の評価結果において示すように，素性f_9は性能に大きな影響を持つ重要な素性である．

2.4.5 評価結果

同定の性能評価の結果を表 2 に示す．

表 2　対訳専門用語同定の評価結果（%）

手法		素性	適合率	再現率	F値
ベースライン			60.4	100	75.3
SVM	適合率最大	全素性	**93.8**	62.8	75.2
		最適な素性の組み合わせ：$f_{2\sim3}$+f_9	**95.2**	63.3	76.1
	再現率最大	全素性	78.2	**86.9**	82.3
		最適な素性の組み合わせ：f_1+f_4+$f_{6\sim7}$+f_9	80.3	**87.2**	83.6
	F値最大	全素性	84.6	81.3	**82.9**
		最適な素性の組み合わせ：f_1+$f_{5\sim6}$+f_9	85.9	85.7	**85.8**

ベースラインとして，2.4.4.1 節で生成した全事例が正しいと判定した場合，適合率は 60.4%，再現率は 100%，F 値は 75.3%となった．全素性を用いた場合，正例クラスに属する確率の下限のパラメータの調整を行った．適合率を最大化する調整を行った場合の適合率は 93.8%，再現率を最大化する調整を行った場合の再現率は 86.4%，F 値を最大化する調整を行った場合の F 値は 82.9%となった．さらに，適合率を最大化する最適な素性の組み合わせ($f_{2\sim3}$+f_9)を用いた場合には，適合率は 95.2%である．再現率を最大化する最適な素性の組み合わせ(f_1+f_4+$f_{6\sim7}$+f_9)を用いた場合には，再現率は 87.2%である．F 値を最大化する最適な素性の組み合わせ(f_1+$f_{5\sim6}$+f_9)を用いた場合には，　85.8%の F 値を達成した．

表 3: 適合率が最大となる場合で,「全素性の場合」との間で有意差（有意水準 5%）のない適合率となる最少数(2 個)の素性組とその評価結果 (%)

素性	適合率	再現率	F 値
f_1+f_9	92.5	60.5	73.1
f_2+f_9	92.3	62.8	74.7
f_6+f_9	90.9	71.0	79.7
f_8+f_9	92.8	60.2	73.0

表 4: 全素性から一素性を取り除いた場合の評価結果（%）

素性	適合率	再現率	F 値
f_1 以外の全素性	93.6	64.6	76.5
f_2 以外の全素性	91.2	65.6	76.3
f_3 以外の全素性	91.4	58.3	71.2
f_4 以外の全素性	91.5	56.1	70.0
f_5 以外の全素性	93.2	63.5	75.5
f_6 以外の全素性	92.6	66.1	77.1
f_7 以外の全素性	92.8	65.1	76.5
f_8 以外の全素性	92.8	60.2	73.0
f_9 以外の全素性	**87.9**	63.5	73.7

性能に大きな影響を持つ素性を同定するために，適合率が最大となる場合で,「全素性の場合」との間で有意差（有意水準 5%）のない適合率となる素性の組み合わせのうち，最少数 (2 個)の素性を用いる場合の性能を表 3 に示す．素性f_9「要素合成法の確率」は日中対訳専門用語の同定における最も有効な素性であることが分かる．さらに，表 4 に示すように，全素性からただ一つだけの素性を取り除いた場合，f_9を取り除いた場合のみ，全素性を用いた場合と比較して、有意差のある(有意水準 5%)適合率を達成した．

一例として，素性f_6とf_9のみを用いた場合の訳語候補の正解例を表 5 に示す．正解の日中専門用語対訳対「ポリエチレン/樹脂」および「聚乙烯/树脂」に対しては，日本語専門用語の頻度(jf)は 156，対訳対の共起頻度(*jcf*)は 151 であり，その差は 5 で上限値 105 以下を満たすため，素性 f 6 を 1 と設定した．また，要素合成法によって生成された訳語候補の確率は 0.80 であった．この二つの素性によって，正しい訳語候補であると判定された．その一方で，誤りの日中専門用語対訳対「ドーピング/濃度」および「掺杂质/浓度」においては，日本語専門用語の頻度と対訳対の共起頻度の差(f_6)は 0 であり，要素合成法の確率(f_9)は最小値の 0 であるので，誤り対訳対であると判定できた．このように，日本語専門用語の頻度と対訳対の共起頻度の差(f_6)が小さく，日中対訳専門用語の要素合成法の確率(f_9)が高くなるほど，入力日中対訳専門用語が適切な対訳対である

可能性が高くなると言える.

表 5: SVM による正解例（*jf* は日本語専門用語の頻度で，*jcf* は日中対訳専門用語の共起頻度である）

（表 3 において素性f_6とf_9のみを用いるモデル）

日本語専門用語	中国語専門用語	素性f_6	素性f_9	人手による判断	SVMによる判断
ポリエチレン/樹脂	聚乙烯/树脂	1 (*jf*=156, *jcf*=151)	0.86	正解	正解
ドーピング/濃度	掺杂质/浓度	0 (*jf*=107, *jcf*=1)	0	誤り	誤り

2.4.6 おわりに

本論文においては，文献［1］において提案された日中対訳特許文からの対訳専門用語獲得の枠組において，SVM における素性の組み合わせを列挙し，網羅的な評価および各素性の有効性に関する詳細な評価を行った．評価結果から，素性f_9「要素合成法の確率」が性能に大きな影響を持つ重要な素性であることを示した．今後は，日中二言語の間の音素［8］および文字［9］の対応を情報として要素合成法(f_9)に導入することによって，日中対訳専門用語同定の性能を改善する方式に取り組む.

参考文献

［1］ Dong, L., Long, Z., Utsuro, T., Mitsuhashi, T., and Yamamoto, M. Collecting bilingual technical terms from Japanese-Chinese patent families by SVM. In *Proc. PACLING*. 2015.

［2］ M. Utiyama and H. Isahara, A Japanese-English patent parallel corpus, In Proc. MT Summit XI, pp. 475–482, 2007.

［3］ P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In Proc. 45th ACL, Companion Volume, pp. 177–180, 2007.

［4］ http://mecab.sourceforge.net

［5］ H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter for Sighan bakeoff 2005. In Proc. 4th SIGHAN Workshop on Chinese Language Processing, pp. 168-171, 2005.

［6］ V. N. Vapnik. Statistical Learning Theory. Wiley-Interscience, 1998.

［7］ https://www.csie.ntu.edu.tw/~cjlin/libsvm/

［8］ L. Xu, A. Fujii, and T. Ishikawa. Modeling impression in probabilistic transliteration into Chinese. In Proc. 2006 EMNLP, pages 242–249, 2006.

［9］ C. Chu, T. Nakazawa, D. Kawahara, and S. Kurohashi. Chinese-Japanese machine translation

exploiting Chinese characters. ACM Transactions on Asian Language Information Processing, 12(4):16:1–16:25, 2013.

２．５　国際特許分類を用いた特許文書のクロスリンガル wikification

静岡大学　綱川 隆司
梶 博行

2.5.1　はじめに

特許文書には幅広い分野の専門用語が多く含まれており、特許文書の読者はしばしばそれらについて調べる必要が生じる。Web 上で特許文書を閲覧するときには、特許文書の専門用語からその用語を説明する Web ページへのリンクがあると便利であり、特許審査官や特許を利用する技術者が特許の内容を効率よく理解する助けになると期待される。

一般に、テキスト中の語句から Wikipedia 記事へのリンクを張ることを wikification と呼び、近年さかんに研究されている (Roth et al., 2014) 。Wikipedia には一般的な概念や人名・地名などの固有名詞だけでなく、特許文書に現れる幅広い分野の専門用語に関する記事も充実してきており、リンク先として Wikipedia 記事は有用であると考えられる。

Wikification を実現するにあたり課題となるのは、リンク元となる重要な語句（アンカーテキスト）の抽出、および、各アンカーテキストのリンク先記事の決定の二つである。特許文書においてはリンク元となる専門用語を特定し、その用語に複数の意味があるときは適切なリンク先記事を決定する必要がある。そこで本研究では、特許明細書に付与された国際特許分類 (IPC) を手掛かりに用いる。Wikipedia 記事を整理するために各記事に対応付けられるカテゴリと、各 IPC タグを関連付け、特許明細書の IPC タグと関連の強いカテゴリに属する記事を特定することで、専門用語の抽出とリンク先記事決定を行う方法を提案する。

また、Wikipedia は多言語百科事典であり、各言語版は独立して編集されるために規模が異なっている。例えば英語版の記事数は日本語版の 5 倍以上であり、専門用語に関する記事もより充実している。リンク先とする Wikipedia 記事を、特許文書の言語版のものに限定せず他の言語版に広げることで、リンクできる専門用語を増やすことができる。このようにテキストと異なる言語へのリンク付けはクロスリンガル wikification (McNamee et al., 2011) と呼ばれる。本研究では、特許の日英パラレルコーパスから得た用語の対訳フレーズテーブルを用い、抽出した専門用語を日本語から英語に訳して英語記事に対応付ける方法を検討する。

2.5.2　関連研究

Wikification におけるアンカーテキストの特定には、アンカーテキストとなる語句が入力テキストの中で重要かどうか判断することが必要である。重要性の主な指標として、Wikipedia 全体において語句がリンクのアンカーテキストになっている確率（キーフレーズネス (Mihalcea and Csomai, 2007)）が挙げられる。リンク先記事の決定には、アンカーテキストの語義曖昧性解消 (Navigli, 2009) が必要となる。アンカーテキストから各リンク先記事候補にリンクされる確率 (Milne and Witten, 2008)、周辺文脈の類似度 (Mihalcea and Csomai, 2007)、および、周辺リンクのリンク先記事との関連性 (Ratinov et al., 2011) が曖昧性解消に有効な特徴である。Miao et al. (2013) は、

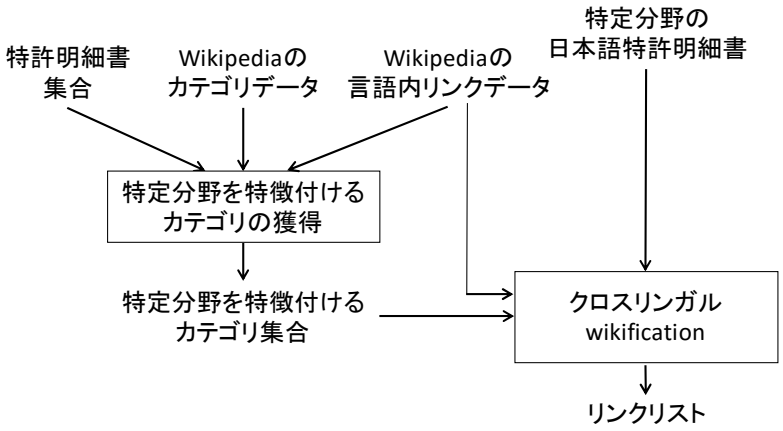


図 1　提案方法の概要

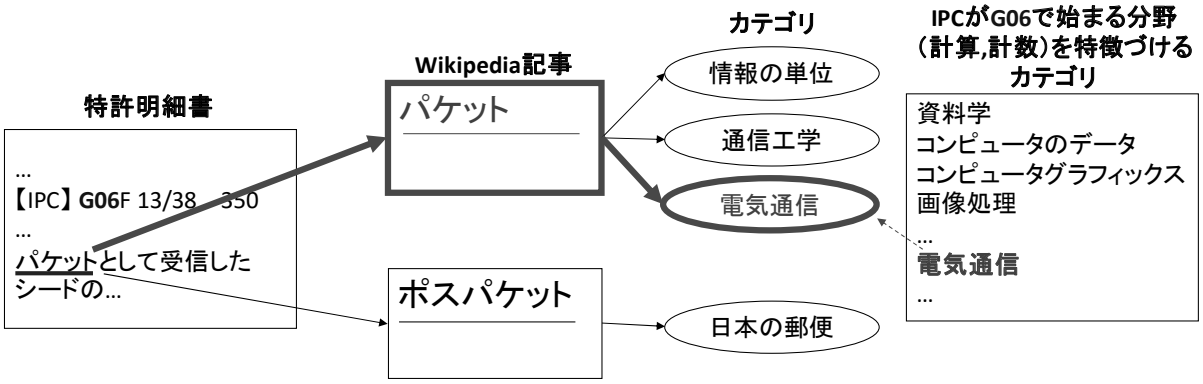


図 2　カテゴリを用いた wikification

Wikipedia の言語間リンク情報および語彙統語パターンから得られる対訳辞書を用いて用語を翻訳することでクロスリンガル wikification を行った。

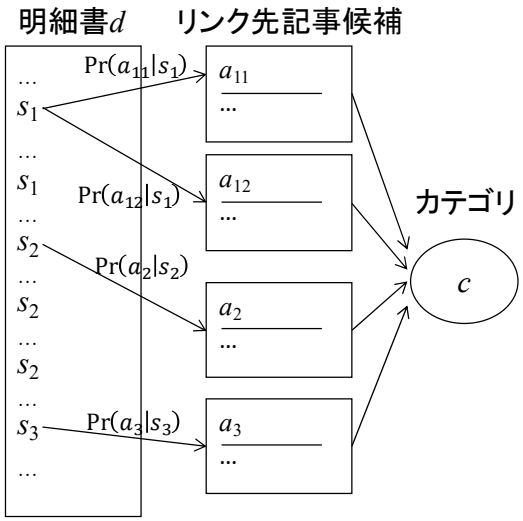
2.5.3　提案方法

2.5.3.1　基本アイデア

2.5.1 節で述べたように、本研究では、特許文書に出現する専門用語のうち、特許文書の持つ国際特許分類 (IPC) に関連するものを抽出する。図 1 に提案方法の基本アイデアの概要を示す。まず特定の IPC タグが示す分野を特徴付ける Wikipedia カテゴリを特定する。その後、専門用語の中でそれらの Wikipedia カテゴリに対応付くものを抽出し､そのカテゴリに属する記事をリンク先として決定する。図 2 にカテゴリを用いた wikification の方法を示す。IPC タグが G06 で始まる分野（計算, 計数）を特徴付けるカテゴリを求めておき、同じ分野の特許明細書に出現する各語句について、対応付く可能性のある Wikipedia 記事を列挙し、それぞれの記事が属するカテゴリが G06 を特徴付けるカテゴリに含まれれば、この語句をアンカーテキストとし、リンク先記事もそのカテゴリに属する記事に決定する。これにより、特許明細書の分野に関係する専門用語を特定でき、かつ、関連する記事をリンク先として決定できる。

2.5.3.2 特許の特定分野を特徴付ける Wikipedia カテゴリの獲得

特許の特定分野を特徴付ける Wikipedia カテゴリを求めるために、IPC タグと各 Wikipedia カテゴリの関連度を、文書を特徴付ける語句の重み付け方法の一つである tf-idf に基づいて計算する。語句の tf-idf 値は語句の文書内の出現頻度 (tf) と逆文書頻度 (idf) の積で求められる。これに倣い、ある IPC タグを持つ特許明細書に出現する各カテゴリの出現頻度 (cf) と、全分野における各カテゴリの逆文書頻度 (idf) を求め、それらの積をその IPC タグに対するカテゴリの重みとする。このとき、カテゴリは特許明細書に直接出現するわけではないので、次のように各頻度を推定する。



$$\mathrm{cf}_d(c) = 2 \times (\Pr(a_{11}|s_1) + \Pr(a_{12}|s_1)) + 3 \times \Pr(a_2|s_2) + 1 \times \Pr(a_3|s_3)$$

$$\mathrm{df}_d(c) = \max(\Pr(a_{11}|s_1) + \Pr(a_{12}|s_1), \Pr(a_2|s_2), \Pr(a_3|s_3))$$

図 3　カテゴリの出現頻度・文書頻度の計算例

特許明細書の集合を D とし、ある IPC タグ t を持つ特許明細書集合 $D(t)\,(\subset D)$ を抽出し、各明細書 $d \in D(t)$ に出現する全ての名詞列の集合を $S(d)$ とする。各名詞列 $s \in S(d)$ のうち、Wikipedia においてアンカーテキストとして用いられたことのあるものをアンカーテキスト候補として出現頻度 $\mathrm{freq}_d(s)$ とともに列挙する。ここで、アンカーテキスト s が記事 a をリンクする確率 $\Pr(a|s)$ をすべての組合せについて予め Wikipedia から求めておく。

ある特許明細書 d におけるカテゴリ c の出現頻度 $\mathrm{cf}_d(c)$ を、以下の式で推定する。

$$\mathrm{cf}_d(c) = \sum_{s \in S(d)} \left(\mathrm{freq}_d(s) \times \sum_{a \in A(c)} \Pr(a|s) \right).$$

ここに、$A(c)$ はカテゴリ c に属する全ての記事の集合とする。次に、特許明細書 d におけるカテゴリ c の文書頻度 $\mathrm{df}_d(c)$ は以下の式で求める。

$$\mathrm{df}_d(c) = \max_{s \in S(d)} \sum_{a \in A(c)} \Pr(a|s).$$

これらを用いて、カテゴリ c が IPC タグ t の分野を特徴付ける重み $\text{cf-idf}_t(c)$ を以下の式で求める。

$$\text{cf-idf}_t(c) = \sum_{d \in D(t)} \mathrm{cf}_d(c) \times \log \frac{|D|}{\sum_{d \in D} \mathrm{df}_d(c)},$$

図 3 は、ある特許明細書 d にアンカーテキスト s_1, s_2, s_3 が現れ、各アンカーテキストがカテゴリ c と図のように対応しているときのカテゴリ c の出現頻度 $\mathrm{cf}_d(c)$ および文書頻度 $\mathrm{df}_d(c)$ の計算例を示している。

全ての Wikipedia カテゴリについて cf-idf 値を計算し、cf-idf 値が上位 θ %のカテゴリを IPC タグ t の分野を特徴付けるカテゴリ集合として得る。

2.5.3.3 特定分野の特許明細書に対するクロスリンガル Wikification

IPC タグ t の分野の特許明細書から、アンカーテキストとなる専門用語を抽出し、それぞれリンク先記事を決める。まず、特許明細書中の名詞列をアンカーテキスト候補として抽出する。各アンカーテキスト候補 s がリンクする可能性のある記事のうち、記事が属するカテゴリが前節で求めた IPC タグ t の分野を特徴付けるカテゴリ集合に含まれるものがあるかどうかを調べる。そのような記事がない場合、もとのアンカーテキスト候補は IPC タグ t の分野と関連がないとみなし、リンクを付与しない。IPC タグ t の分野を特徴付けるカテゴリに属する記事がある場合は、それらの記事 a の中で確率 $\Pr(a|s)$ が最も高い記事を s のリンク先記事としてリンクを付与する。

さらに、英語記事へのリンクを付与するため、アンカーテキスト候補のうち、対応する日本語記事がないものについて、日英特許パラレルコーパス (Utiyama and Isahara, 2007) からフレーズベース統計的機械翻訳に基づく方法 (Koehn et al., 2007) により得た日英対訳フレーズテーブルを用いて英語に翻訳する。このとき、適切な訳を得るために以下の手順で翻訳する。まず、フレーズテーブルから下記の条件を満たす対訳対 (j, e) （j は日本語フレーズ、e は英語フレーズ）を削除する。

> e が e' の部分単語列で、フレーズの翻訳確率が $\Pr(e'|j) > \Pr(e|j)$ であるような対訳対 (j, e') がフレーズテーブルに存在する。

例えば、$\Pr(\text{mobile phone}|\text{携帯電話}) > \Pr(\text{phone}|\text{携帯電話})$ のときは、対訳対 (携帯電話, phone) をフレーズテーブルから削除する。これにより、日本語のアンカーテキスト候補に対してより広い概念を表す英語フレーズに訳すことを防ぐ。

フレーズテーブルに日本語のアンカーテキスト候補を含む対訳対が存在するときに限り、翻訳確率が上位 10 件の英語フレーズについて、それぞれ英語版 Wikipedia において英訳したアンカーテキスト候補からリンクされたことのある英語版記事を列挙する。以下、上位の英語フレーズから順に以下のいずれかの条件を満たすものを探し、最初に条件を満たした日本語版記事または英語版記事をリンク先とする。

- 英語版記事に日本語版が存在し、その日本語版記事が属するカテゴリの少なくとも一つが IPC タグ t の分野を特徴付けるカテゴリに含まれる
- 英語版記事に日本語版が存在せず、英語版記事が属するカテゴリに日本語版があるものがあり、かつその日本語版カテゴリの少なくとも一つが IPC タグ t の分野を特徴付けるカテゴリに含まれる

2.5.4 評価実験

2.5.4.1 実験設定

本提案方法により、日本語の特許明細書に対してクロスリンガル wikification を行う実験を行った。特許明細書として NTCIR-7 PATMT テストコレクションに含まれる日本語・英語特許明細書

表 1　IPC タグ G06（計算, 計数）の分野を特徴付けるカテゴリ

カテゴリ	cf-idf ($\times 10^3$)	カテゴリ	cf-idf ($\times 10^3$)	カテゴリ	cf-idf ($\times 10^3$)
資料学	1131	ソフトウェア	512	コンピュータの利用	387
コンピュータのデータ	1116	情報・ワイドショー番組	511	知識	387
コンピュータグラフィックス	980	記憶	468	CPU	385
画像処理	972	コンピュータの形態	454	生態域	373
コンピュータの仕組み	966	検索	452	草原	372
情報学	774	出力機器	452	検索アルゴリズム	371
コンピュータネットワーク	663	サーバ	434	ロシア語由来の外来語	364
コンピュータのユーザインタフェース	616	情報技術史	422	情報処理	361
記憶装置	576	文字	401	パソコンの周辺機器	354
ラジオの情報・ワイドショー番組	516	OS のファイルシステム	388	入力機器	337

を用い、各明細書の【要約】以降の範囲のテキスト部分を利用した。本実験では、特許の特定分野として IPC タグが G06 から始まる“計算, 計数”クラスを採用し、2000 年に出願された日本の特許明細書のうち、IPC に G06 から始まるものを含むもの 10000 件、および、全分野から 10000 件をそれぞれ任意に抽出した。Wikification の対象として、2001 年に出願された G06 で始まる IPC タグを持つ特許明細書 15 件を選択し、人手で正解となるリンクを付与したものをテストセットとして、提案方法により得られるリンクと比較した。

特許明細書に含まれる名詞列を抽出するため、MeCab による形態素解析を行い、名詞（非自立等を除く）と接頭詞からなる単語列をすべて名詞列として扱った。アンカーテキスト候補を日英翻訳するためのフレーズテーブルは、上記の NTCIR-7 PATMT コレクションに含まれる日英特許パラレルコーパス (Utiyama and Isahara, 2007) から構築されたものを用いた。

Wikipedia の記事・カテゴリデータは 2013 年 3 月時点のダンプデータを用い、記事についてはすべてを、カテゴリについては隠しカテゴリ以外のすべてのカテゴリを用いた。特定分野を特徴付けるカテゴリの閾値 θ については、$\theta = 10$ (%) を用いた。

2.5.4.2　特許の特定分野を特徴付ける Wikipedia カテゴリの獲得

表 1 に、提案方法により IPC タグ G06 の分野に対して cf-idf 値が上位となった日本語 Wikipedia カテゴリを示した。1 位のカテゴリは“資料学”であり、これは当該分野で頻出する語“データ”のリンク先記事“データ”が属しているカテゴリである。以下、上位にはコンピュータ関連のカテゴリが多く並んでおり、当該分野にコンピュータを用いたシステムの特許が多いことを反映していると考えられる。

一部、“ラジオの情報・ワイドショー番組”、“情報・ワイドショー番組”や、“生態域”、“草原”といった、一見して当該分野と関連のないカテゴリが見られる。前者は、“スタンバイ”や“アクセス”といった当該分野に関連する意味を持つ語がラジオの番組名に用いられているために、カテゴリの出現頻度および文書頻度の推定において番組名としての頻度がカウントされた

表 2　提案方法で得られたリンクとテストセットのリンクの比較

アンカーテキスト	リンク先記事	リンク数
テストセットと提案方法の両方で抽出	テストセットと一致	302
	テストセットと異なる	144
テストセットのみで抽出		204
提案方法のみで抽出		4055

表 3　アンカーテキスト“インタフェース”のリンク先記事選択

リンク先記事候補	記事が属するカテゴリ	cf-idf ($\times 10^3$)
インタフェース (情報技術)	ソフトウェア	497
	電子工学	207
	インタフェース規格	121
	インタフェース	104
グラフィカルユーザインタフェース	コンピュータグラフィックス	**970**
	コンピュータのユーザインタフェース	589
	グラフィカルユーザインタフェース	161
	ソフトウェアアーキテクチャ	39

結果、cf-idf 値が上昇した例である。このため、“情報・ワイドショー番組”に結び付く可能性のある他のアンカーテキスト（例えば、“ローカル”など）が現れたときに本来の意味と異なる番組に関する記事にリンクされるおそれが生じる。後者も同様に、当該分野の頻出語“ステップ”から得られたと考えられる。段階・手順を示す一般語である“ステップ”に関する記事はなく、リンク先記事候補として“ステップ (植生) ”があるため、その記事が属するカテゴリの cf-idf 値が上昇した。この結果、特許明細書中に現れる“ステップ”はすべて記事“ステップ (植生) ”にリンクされてしまい、不適切な結果となる。

これらの問題に対処するためには、カテゴリの出現頻度を求める際に単独のアンカーテキストから得られたカテゴリを無視する、文脈に応じたリンク先記事の選択を行う、といった方法が考えられる。

2.5.4.3 Wikification 結果

表 2 に、提案方法によって得られたリンクに対して、テストセットのリンクと比較した結果を示した。得られたリンクがテストセットに含まれないケースが多いが、これはテストセットに比べ、提案方法は一般語に近い“情報”のような語にもリンクしているためである。テストセットで抽出したアンカーテキストのうち、提案方法によって抽出できたものは 68.6%あった。テストセットと提案方法の両方で抽出されたリンクについて、提案方法によりリンク先記事も一致した割合は 67.7%であった。テストセットで抽出したアンカーテキスト 650 件のうち、日本語 Wikipedia においてアンカーテキストとして現れたものは 582 件あり、それぞれに対して最もリンクされやすい記事を常に選択した場合は一致率 80.9%であった。本提案方法ではアンカーテキストからリンク先記事への対応確率は主として用いていないため一致率は単純に比較できないが、対応確率を考慮した方法の開発が今後の課題である。

表 4　アンカーテキストを英訳して得られたリンクの例

例	特許明細書（下線はアンカーテキスト）	提案方法で得られた英語リンク先記事	対応する日本語版記事
1	…この発明は、例えばセルラ無線通信システムの加入者が…	Mobile phone	携帯電話
2	…シードのアシンクロナスパケットを受信して…	Data transmission	データ転送
3	…音声データが記録されるミニディスク（商標）１１は、スピンドルモータ１２により回転駆動される。…	Hard disk drive	ハードディスクドライブ
4	…当該シードに基づいて復号用の ODD 鍵または EVEN 鍵が生成されて…	Parsing	構文解析
5	出力処理部２６でレベル調整やインピーダンス調整等が行われて、	Guru Gobind Singh	グル・ゴービンド・シング

表 3 に、タイトルが“データ通信装置及び方法、並びに媒体”である特許明細書に含まれるアンカーテキスト“インタフェース”について、リンク先記事候補の一部と記事が属するカテゴリおよびその cf-idf 値を示した。テストセットにおいて適切とされたリンク先記事は“インタフェース (情報技術)”であるが、提案方法では最も大きい cf-idf 値をもつカテゴリ“コンピュータグラフィックス”に属する記事“グラフィカルユーザインタフェース”が選択された。

本提案方法は分野のみに依存してリンク先記事を決定するため、この例のようにより細かい分類が必要な曖昧性解消を行うには、国際特許分類の細分類（サブクラスなど）から得られた cf-idf 値を組み合わせる、あるいは文脈情報など他の wikification 手法と組み合わせるといった改善法が考えられる。

提案方法において、アンカーテキストを英語に翻訳することによって得られたリンクの例を表 4 に示す。例 1 は、“セルラ”という語からは適切な日本語記事が見つからないため、英訳して“cellular”にすることで携帯電話を表す記事と対応付いたもので、概ね適切である。例 2 は、データ転送の記事が選択されており、記事中で非同期転送についての記述があることから、関連した記事に対応付けることができたものである。一方で、アンカーテキストを英訳した時点で異なる意味の記事に対応しやすくなるために不適切な記事が選ばれる例も散見された。例 3 は“スピンドルモータ”の英訳“Spindle motor”が、英語 Wikipedia において記事“Brushless DC electric motor”（無整流子電動機）および記事“Hard disk drive”のアンカーテキストになっており、後者がコンピュータ関連の記事であることから選択されたものである。例 4 も同様に“復号用”の英訳“decoding”に対して記事“Parsing”が対応付いた例である。例 5 は、“インピーダンス調整”からフレーズテーブルで得られる英訳の一つに“after”があり、英語 Wikipedia で“after”をアンカーテキストとするリンクが存在するために、その中で特徴付けるカテゴリに関係する記事が選択される例である。得られたフレーズテーブルからより適切な訳を選択するとともに、日本語のア

ンカーテキストと関連性の低い英語のリンク先記事を排除する方法を開発する必要がある。

2.5.5 おわりに

本研究では、特許明細書中の専門用語の理解を容易にするため、明細書の内容とリンク先記事の関連性を国際特許分類と Wikipedia カテゴリの対応付けから得ることによるクロスリンガル wikification 方法を提案した。

今後の課題として、一般語の頻出から得られる不適切なカテゴリの除去が挙げられる。このようなカテゴリは同一明細書中の他のカテゴリとの関連性が低いため、カテゴリ間の関連性を求めることで除去できる可能性がある。また、リンク生成のときにも他のリンクのカテゴリとの関連性が高いものを優先的に選ぶことで入力特許明細書ごとに適した記事を選ぶ方法も有効と考えられる。

謝辞

本研究は JSPS 科研費 15K16096 の助成を受けたものです。本研究を進めるにあたり、NTCIR データセットに含まれる日英特許パラレルコーパスから構築した日英対訳フレーズテーブルをご提供頂いた筑波大学宇津呂武仁教授および山本幹雄教授に深く感謝致します。

参考文献

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). “Moses: open source toolkit for statistical machine translation,” In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL) on Interactive Poster and Demonstration Sessions*, pages 177-180.

McNamee, P., Mayfield, J., Lawrie, D., Oard, D.W., and Doermann, D. (2011). “Cross-language entity linking,” in *Proc. of the 5th International Joint Conference on Natural Lan-guage Processing (IJCNLP)*, pages 255-263.

Miao, Q., Lu, H., Zhang, S., and Meng, Y. (2013). “Cross-lingual link discovery between Chinese and English wiki knowledge bases,” In *Proc. of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 374-381.

Mihalcea, R. and Csomai, A. (2007). “Wikify!: linking documents to encyclopedic knowledge,” In *Proc. of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pages 233–242.

Milne, D. and Witten, I.H. (2008). “Learning to link with Wikipedia,” In *Proc. of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 509-518.

Navigli, R. (2009), “Word sense disambiguation: a survey,” *ACM Comput. Surv.*, 41(2):10:1-10:69.

Ratinov, L., Roth, D., Downey, D. and Anderson, M. (2011). “Local and global algorithms for disambiguation to Wikipedia,” in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 1375–1384.

Roth, D, Ji, H., Chang, M.-W., and Cassidy, T. (2014). “Wikification and Beyond: The Challenges of Entity and Concept Grounding,” Tutorial at ACL 2014.

Utiyama, M. and Isahara, H. (2007). “A Japanese-English patent parallel corpus,” In *Proc. of Machine Translation Summit XI*, pages 475-482.

3. 機械翻訳評価手法

３．１　拡大評価部会の活動概要

岡山県立大学　磯崎　秀樹

2012 年度から、本研究部会の下部組織として「拡大評価部会」を設置し、機械翻訳の評価に関する議論を深めてきた[1][2][3]。本部会での議論の焦点は以下の 5 点である。

1. 「技術調査目的」のために特許文書を機械翻訳する場合の評価
2. 人手評価、自動評価、半自動評価
3. 評価用テストセット
4. 対象とする言語の範囲：日本語、英語、中国語
5. 評価手法の理想形、理想を実現するための課題、課題克服への道程

昨年度に引き続き今年度も 3 回の部会を開催した。

・2015 年 5 月 15 日　今年度の活動計画の策定
・2015 年 9 月 25 日　中間報告と今後の活動内容についての議論
・2016 年 1 月 29 日　最終報告と年度報告書の執筆について

活動は、人手評価、自動評価、テストセットの 3 つのグループに分かれて行った。概要を以下に示すが、詳細については本章の各記事をご覧いただきたい。

人手評価に関しては、WAT2015 の中日・韓日特許翻訳の人手評価結果の分析を行った。この評価ではクラウドソーシングを利用した評価の実験も行った。

自動評価に関しては、語順を評価する RIBES が、語順が比較的自由な日本語では、参照訳と異なるがよい語順の文が不当に低く評価されるという問題を解決するために、参照訳を係り受け解析して、意味が変わらない別の語順の文だけを自動生成する手法を提案した[4]。また、IMPACT に大局的な評価を導入する試みを行った。後者については３・５で詳述する。テストセットとは、ある特定のパターンが正しく翻訳されるかどうかを確認するための文の集合である。本年度は中日翻訳のためのテストセットを作成した。

[1] 拡大評価部会員：機械翻訳評価、平成 24 年度 AAMT/Japio 特許翻訳研究会報告書、6 章、pp.37—104、2013 年 3 月.
[2] 拡大評価部会員：機械翻訳評価、平成 25 年度 AAMT/Japio 特許翻訳研究会報告書、6 章、pp.61—82、2014 年 3 月
[3] 拡大評価部会員：拡大評価部会活動報告、平成 26 年度 AAMT/Japio 特許翻訳研究会報告書、5 章、pp.79—110, 2015 年 3 月.
[4] Hideki Isozaki and Natsume Kouchi: Dependency Analysis of Scrambled References for Better Evaluation of Japanese Translation, Proc. of WMT-2015, pp.450—456, 2015.

３．２　翻訳自動評価法の改良に関する２つの提案

岡山県立大学　磯崎　秀樹
北海学園大学　越前谷　博
ＮＴＴコミュニケーション科学基礎研究所　須藤　克仁

提案１：日本語訳の RIBES による採点について、昨年度考案した語順の入れ替え（スクランブリング）への対応について、ルールを用いない手法を検討する。

日本語訳を RIBES で採点する場合に、日本語には語順の自由度があるために、よい訳なのに、不当に低い点数がついてしまうことがある。たとえば以下のような場合である。

参照訳　：提案手法を図３に示す。
機械訳１：提案手法を図３に示す。RIBES = 1.000
機械訳２：図３に提案手法を示す。RIBES = 0.679

機械訳２は参照訳と同じ意味であり、特に問題はないので、もっと点がよくてもよいはずであるが、RIBES は語順を評価するので、点が悪くなる。このような文は**スクランブリング**と呼ばれる。

そこで、WMT-2014 では、このような語順の入れ替えに対応するため、参照訳を係り受け解析し、得られた係り受け木をポストオーダーで出力することによって、日本語らしい主辞後置の語順の別の文を自動生成する方法を提案した。ただし、誤解を招く文が生成されることがあるため、ルールによって誤解を招く語順の文を排除しようとした。(Isozaki et al. 2014)

しかし、非常に厳しい制約を用いたせいで、ほとんどの文でスクランブリングが生成できない、という問題があった。そこで、WMT-2015 ではルールを用いないスクランブリングへの対応法を考案し、リスボンで開催された WMT-2015 で口頭発表した。(Isozaki and Kouchi 2015)

新しい手法は、「**係り受け比較法**」といい、参照訳を係り受け解析して得られた係り受け木をポストオーダーで出力して得られる文を係り受け解析し、元の文と同じ係り受け木が得られるかどうかで、新しい参照訳として採用するかどうか判定する手法である。NTCIR-9 の英日翻訳のデータを用いた実験によると、去年 WMT-2014 のルールによる手法よりカバーできる文が多くなり、文レベル相関が向上した。17 システム中全システムで文レベル相関が向上し、符号検定で

p=0.0000153 となり、有意差があった。

江原ら 2009 は、BLEU より Word Error Rate (WER) の方が人間の評価に近いことを示している。そこで同じ「係り受け比較法」を WER に適用したところ、17 システム中 12 システムで文レベル相関が向上したが、これは符号検定で p=0.1435 であり、有意差はなかった。(門田 2016)

RIBES と WER のこの差がなぜ生じたかを考えてみると、RIBES は直接語順を測定しているが、WER は語順ではなく、参照訳と一致させるための操作の数を計算していることが原因であろう。

たとえば、以下の参照訳と２つの機械訳を考えると、どちらの機械訳も WER では、１単語を削除して１単語を追加すればよいので、２回の操作が必要である。

参照訳　：The quick brown fox jumps over the lazy dog .
機械訳１：The brown fox jumps over the quick lazy dog .
機械訳２：The brown quick fox jumps over the lazy dog .

つまり、この２つの機械訳は、どちらも参照訳との編集距離が２なので、WER では同じスコア 2/10 = 0.2 になる。
しかし、RIBES では、quick が大きく動いた機械訳１の方が、あまり動かなかった機械訳２よりも語順の変化が大きく、成績が悪くなる。実際に計算すると、機械訳１が 0.911、機械訳２が 0.978 で、機械訳２の方がよいと判断される。

RIBES では有意差が出たのに、WER では有意差が出なかったのは、このように、語順を考慮せず、編集の手間だけを問題にする WER の性質によるものと考えられる。

提案２：多言語のための大局的評価を用いた自動評価法

多言語に容易に適用可能な自動評価法は基本的に単語を最小単位としているため、長文においては局所的な評価に陥りやすいという問題を抱えている。このような問題を解決するために、より大きな単位に基づく大局的な評価を導入することは有効と考えられる。そこで、様々な言語に適用可能であり、かつ、大局的な観点での評価を考慮した、新たな自動評価を提案する。提案手法では、翻訳文と参照訳をそれぞれいくつかの部分に分割し、その部分を最小単位とした大局的な

評価を行う。そして、その大局的な評価結果を、従来の単語を最小単位とした局所的な観点からの自動評価法 IMPACT の評価結果に対する重みとして用いる。いくつかのデータを用いた評価実験の結果、提案手法の有効性を確認した。本提案手法の詳細については「３．５　多言語のための大局的評価を用いた自動評価法」で述べているため、ここでは割愛する。

参考文献

Hideki Isozaki, Natsume Kouchi, and Tsutomu Hirao: Dependency-based Automatic Enumeration of Semantically Equivalent Word Orders for Evaluating Japanese Translations, Proc. of WMT-2014, pp.287—292, 2014.

Hideki Isozaki and Natsume Kouchi: Dependency Analysis of Scrambled References for Better Evaluation of Japanese Translation, Proc. of WMT-2015, pp.450—456, 2015.

江原 暉将、越前谷 博, 下畑 さより, 藤井 敦, 内山 将夫, 山本 幹雄, 宇津呂 武仁, 神門 典子：機械翻訳精度の各種自動評価の比較、Japio 2009 Year Book, pp.272—275, 2009.

門田 悠一郎：日本語の語順の自由度を考慮した編集距離による翻訳自動評価、岡山県立大学卒業論文, 2016.

３．３　中国語特許文献の中日翻訳評価のためのテストセットの拡充

元・山梨英和大学　江原　暉将
(株)富士通研究所　長瀬　友樹
筑波大学　宇津呂武仁
筑波大学　龍　　梓
(財)日本特許情報機構　王　　向莉

3.3.1　はじめに

機械翻訳評価の一手法として、表現パターン別に評価用例文を用意しておき、翻訳結果に対して対応する表現パターンがうまく訳されていることをピンポイントでチェックする「テストセット評価」が提案されている[1)2)3)]。

筆者らは、中国語特許文献の中日機械翻訳評価のためにテストセットの検討を行い、昨年までに以下のことを実施した[4)5)]。

・中日特許文平行コーパスの作成
・テストセットの作成
・評価用サイトの整備[1]

昨年度までのテストセットの作成において、515 個の中国語表現パターンとそれを含む中国語特許文の収集および中国語表現パターンに対する日本語翻訳パターン設問の作成を行った。これらの作業は、主として既発表資料からのデータ抽出および、翻訳先言語である日本語の文末表現に着目して、それに対応する中国語表現パターンを探しテスト文とするという手法を用いて行った。今年度は、逆に中国語側の表現パターンを直接収集し、対応する日本語の翻訳パターンを作成することでデータの拡充を図った。

3.3.2　中国語表現パターンの収集

別途収集した 360 万文対からなる中日特許文平行コーパスから Moses[6)]を用いてフレーズテーブルを作成し、複合名詞などのフレーズを除くフィルタリングを実施した。フィルタリングは、以下の条件をすべて満たすフレーズのみを抽出するものである。

・頻度が 500 以上
・形態素数は 2 以上
・文字数は 3 以上
・フレーズの先頭と末尾は「的」でない
・2 形態素に対して複合名詞とみなせない(品詞列が NN NN, VV NN, NN VV 以外)

上記のフィルタリングの結果得られた 5458 個のフレーズを 5 個の頻度レンジに分け、各頻度レンジから約 80 個ずつ、合計 365 個のフレーズを中国語パターンとして抽出した。この最後のステップでは、特許文特有のパターンである傾向が強く、一般の文にはあまり出現しないパターン

[1] 本部分は、AAMT 課題調査委員会で整備したサイトを利用させてもらっている。

であるものを中心に集めた。

3.3.3　中国語表現パターンを含む中国語文の収集

3.3.2 で収集した各中国語パターンに対して、それを含む中国語文を昨年度までに作成した中日特許文平行コーパスから収集した。その結果、全部で 6515 文が収集できた。これらの文の中から各中国語パターンが主要な役割を持っている文をテストセット用の中国語文として選択した。

3.3.4　中国語表現パターンに対応する日本語翻訳パターン設問の設定

3.3.2 で収集した中国語パターンに対応する日本語翻訳パターン設問を設定した。設定にあたっては 3.3.3 で用いた中日特許文平行コーパスでの翻訳を参考にした。また翻訳のバラエティを吸収できるように Perl の正規表現パターンとして設定した。中国語文、日本語参照訳文、中国語パターン、日本語翻訳パターン設問の例を表 1 に示す。

パート	中国語文	日本語文	CN_パターン	JP_パターン
DES	高压处理设备中容器内用的是40下水。	高圧処理装置の容器中で利用した水は40°Fであった。	容器内	容器中(の|で|に)
ABS	该遮蔽物包括可以在打开位置和折叠位置之间折叠的伞。	シェルターは開いた状態と閉じた状態との間で折り畳み可能な傘を備える。	打开位置	開(いた)?状態
DES	3.上述的各种碳掺杂源一般置于石英安瓿130的下端125的各个位置上，如图1-5所示。	3．図1～図5に示されるように、一般的に、上述された様々な炭素をドープするドープ源が、石英アンプル（130）の下端（125）の様々な場所に配置される。	上述的各个	(前|上)(記|述)された(様々な|各種の)

表 1 抽出された中国語文、日本語参照訳文、中国語パターン、日本語翻訳パターン設問の例

3.3.5　AAMT 自動評価サイトでの試験

3.3.4 までで作成したテストセットを AAMT 自動評価サイトにアップし、動作確認を行った。

3.3.6　まとめと今後の課題

昨年度までのデータ作成と今年度の作成を合わせて 635 の設問設定ができた。中国語パターンとしてはかなりの程度が収集できたのではないかと考える。今後の課題としては以下のことがあげられる。

・日本語翻訳パターンのバラエティが不足している部分があり、より適切な設問とすることが必要である。
・これまでは中国語パターンとして主として連続パターンを集めてきたが不連続パターンについても収集する必要がある。
・数式や化学式、数量表現など特許に特有な表現パターンが不足している。
・自動評価や人手評価とテストセット評価との比較を行い、双方のメリット・デメリットを明らかにする。

今後、これらの課題を解決して、より良い中日特許文テストセットとしていきたい。

参考文献

1) Isahara, H. 1995. JEIDA's Test-Sets for Quality Evaluation of MT Systems –Technical Evaluation from the Developer's Point of View–. *Proc. of MT Summit V*.

2) Uchimoto, K., K. Kotani, Y. Zhang and H. Isahara. 2007. Automatic Evaluation of Machine Translation Based on Rate of Accomplishment of Sub-goals. *Proc. of NAACL HLT*, pp.33-40.

3) Nagase, T., H. Tsukada, K. Kotani, N. Hatanaka and Y. Sakamoto. 2011. Automatic Error Analysis Based on Grammatical Questions . *Proc. of PACLIC*.

4) 長瀬友樹, 江原暉将, 王向莉. 2014. 中日特許文評価用テストセットの作成, *平成 25 年度 AAMT/Japio 特許翻訳研究会報告書*, pp.78-82.

5) 長瀬友樹, 江原暉将, 王向莉. 2015. 中国語特許文献の中日翻訳評価のためのテストセットの改良と評価サイトの作成, *平成 26 年度 AAMT/Japio 特許翻訳研究会報告書,* pp.104-109.

6) Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, 2007. Moses: Open Source Toolkit for Statistical Machine Translation, *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

３．４　特許文の中日・韓日機械翻訳の人手評価結果の分析

科学技術振興機構　中澤　敏明
(株)東芝　園尾　　聡
NHK 放送技術研究所　後藤　功雄

3.4.1　はじめに

今年度、拡大評価部会人手評価グループでは、WAT2015[1]で行われた中日及び韓日特許翻訳の人手評価結果の分析を行った。WAT2015 ではクラウドソーシングを利用した一対比較による評価と、特許庁が提案している「特許文献機械翻訳の品質評価手順」のうち「内容の伝達レベルの評価」に従った翻訳の専門家による評価の 2 種類の人手評価を実施した。

クラウドソーシング評価では 400 文に対して、各システムとベースラインとなるシステムとの間で、1 文ずつ、どちらの翻訳の方がより良いか（もしくは同程度か）を判定し、その勝敗数をスコア化して各システムをランキングする。システムの出力がベースラインより良い場合は+1、悪い場合は-1、同程度の場合は 0 とし、5 人の異なる評価者の判断を足し合わせる。足しあわせた結果が+2 以上ならばその文ペアについては Win、-2 以下ならば Lose、それ以外ならば Tie と判定する。400 文に対してそれぞれ判定を行い、最終的にクラウドソーシングスコア（Crowd）は以下の式で計算される。

$$Crowd = 100 \times \frac{Win - Lose}{Win + Lose + Tie}$$

内容の伝達レベルの評価は、各文について以下の基準での絶対評価を行う。なお内容の伝達レベルの評価は、クラウドソーシング評価の対象である 400 文のうち、ランダムに選択された 200 文に対して行った。また各言語対ごとに、クラウドソーシング評価の上位 3 チームに対してのみ内容の伝達レベルの評価を行った。

評価	基準
5	すべての重要情報が正確に伝達されている。(100%)
4	ほとんどの重要情報は正確に伝達されている。(80%～)
3	半分以上の重要情報は正確に伝達されている。(50%～)
2	いくつかの重要情報は正確に伝達されている。(20%～)
1	文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。(～20%)

本稿ではこれらの評価結果に対して分析を行ったので、報告する。

3.4.2　クラウドソーシングによる人手評価の分析

クラウドソーシングによる人手評価では、評価タスクを遂行するクラウドソーシングワーカー（以下、ワーカー）の作業品質を一定に保つことが課題となる。本節では、評価タスク全体のマクロな視点と、特定の言語現象に関するミクロな視点から個々のワーカーについての分析を行い、その分析結果について述べる。

3.4.2.1 評価言語に対するワーカーの作業分布（マクロ分析）

人手評価タスク全体におけるワーカーの作業分布について調査した。今回の人評価タスクでは、日本語⇔英語(E)、日本語⇔中国語(C)、韓国語(K)⇒日本語の機械翻訳文(トータル 148,000 文)を計 192 名のワーカーが評価タスクを行った。各ワーカーは、複数の評価タスクを担当することが可能である。評価タスクの言語（原言語または目的言語）と担当したワーカー数およびタスク数（評価文数）を表 1 に示す。

各ワーカーが評価した言語に着目すると、日本語を含む 2 言語間のみの評価タスクを担当したワーカーが全体の半数以上であり、特に E(日英翻訳文および英日翻訳文)のみを評価したワーカーが最も多かった。一方で、日本語を含む 3 言語間(E&C, C&K, E&K)の評価タスクを担当したワーカーに比べて、全翻訳方向である日本語を含む 4 言語間(E&C&K)の評価タスクを担当したワーカーの方が多かった。さらに、評価文数に着目すると、4 言語間を評価した約 20%のワーカーによって、タスク全体の約 80%が評価された。4 言語間を評価したワーカーの実際の言語スキルは未知であるが、特定のワーカーによって大多数の作業が行われたという傾向が明らかとなった。

今回のクラウドソーシングによる人手評価の枠組みでは、2 つの翻訳結果についてより適切な訳文を選ぶタスクであるので、必ずしも原言語と目的言語に熟知している必要がなく、また、言語スキルに応じたワーカーのフィルタリングを行っていないため、ワーカーは言語に関係なくより多くの評価タスクを担当する傾向にあったと思われる。

表 1 評価タスクの言語に対するワーカー数およびタスク数

言語	ワーカー数		タスク数	
	(人)	(%)	(文)	(%)
E	90	46.9	14,945	10.1
C	20	10.4	1,794	1.2
K	10	5.2	684	0.5
E & C	25	13.0	13,402	9.1
C & K	3	1.6	1,033	0.7
E & K	1	0.5	2	0.0
E & C & K	43	22.4	116,140	78.5
	192		148,000	

3.4.2.2 表記揺れに対する評価のロバスト性検証（ミクロ分析）

続いて、クラウドソーシングによる人手評価における評価結果のロバスト性について調査した。韓日特許翻訳タスク(JPCko-ja)に含まれる一部の翻訳結果について、英数字の全角/半角が混在した翻訳文(unnormalized)および英数字が全て全角に正規化された翻訳文(normalized)を用意し、クラウドソーシングによる人手評価結果を比較した。

図 1 に同一ワーカーによる評価結果の変化（unnormalized に対する評価結果から normalized に対する評価結果への変化）を示す。同一ワーカーによって評価された評価文の内、正規化によ

って訳が変化した 220 文については評価結果に変化は見られなかった。一方で、48 文が改善方向(Lose->Tie, Lose->Win, Tie->Win)、95 文が悪化方向 (Win->Lose, Win->Tie, Tie->Lose)、へと評価結果が変化した。

5 人のワーカーによる最終評価結果の変化を図 2 に示す。英数字の正規化によって評価が変化しなかった評価文が大半であったが、124 文(全体の約 40%)については評価が変化した。特に、正規化前に比べて評価が改善した文(32 文)よりも、悪化した文(92 文)の方が圧倒的に多く、最終的な HUMAN スコアは、unnormalized が 29.75、normalized が 3.00 と、評価文の表記揺れに対するロバスト性が課題であることが判明した。評価が悪化方向に変化した原因としては、韓国語原文が半角英数字を使用しているため、より原文に忠実な翻訳結果が選ばれたためだと推測される。

クラウドソーシングによる人手評価において、ワーカーの作業品質を一定に保つことが課題であり、基準となる評価指針（未知語、固有表現、表記揺れをどう扱うか等）をワーカーに提示するなど、評価タスクの設計が重要となる。

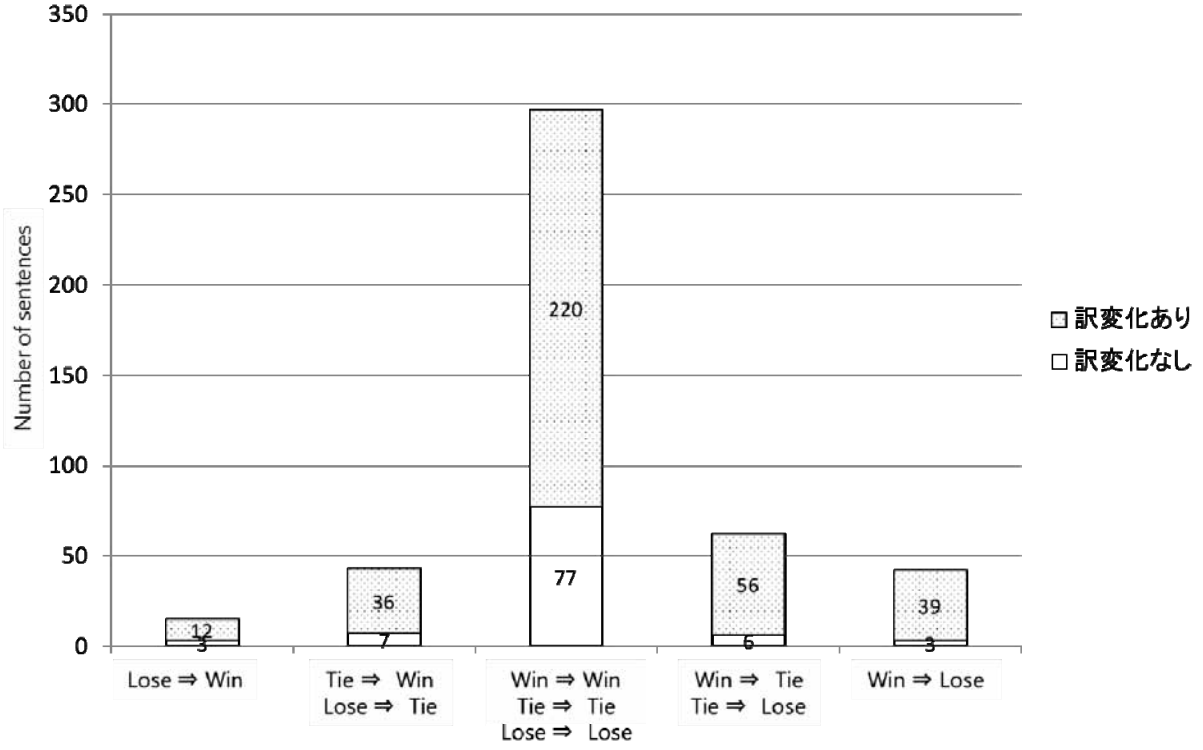


図 1　同一ワーカーによる評価結果の変化

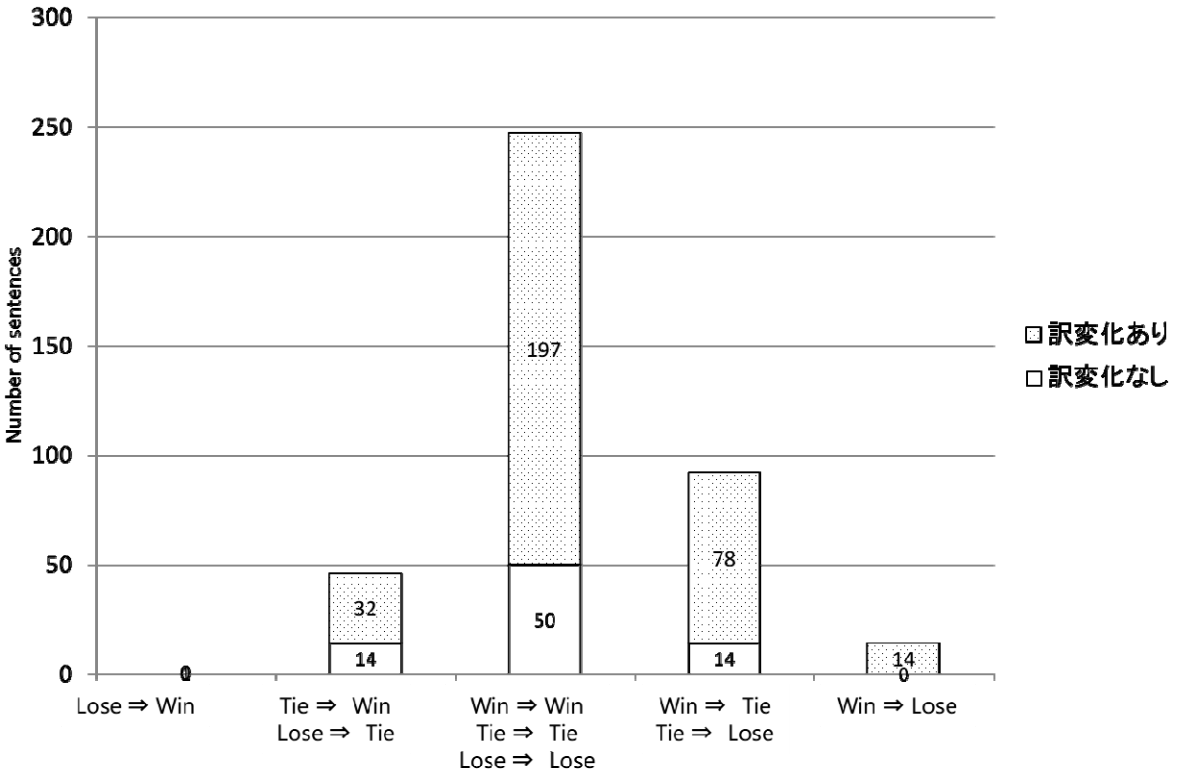


図 2 複数ワーカーによる最終評価結果の変化

3.4.3 自動評価が高く人手評価が低い韓日機械翻訳結果の分析

韓日特許翻訳タスク(JPCko-ja)での評価結果では、BLEU スコアが他のシステムより 13 ポイント以上高かったが、クラウドソーシング評価（Crowd）のスコアが低かったシステムがあった。そのため、韓日特許翻訳タスクでは特別に、Crowd が上位３チームのシステムに加えて、自動スコアが最も高かった１システムを加えた４システムが内容の伝達レベルの評価（Adequacy）の対象となった。

韓日特許翻訳タスクで内容の伝達レベルの評価を行ったシステムの評価結果を図 3 に示す。図 1 において、MT4 の BLEU スコアは、MT1～MT3 の BLEU スコアより 13 ポイント以上高い。しかし、Adequacy の値はこれら４つの中で最も低い。また、フレーズベース SMT システムのベースラインシステムとの比較に基づく Crowd のスコアはマイナス、すなわち、ベースラインシステムより翻訳品質が低いという評価になっている。

MT4 はシステム説明論文[2]によるとフレーズベース SMT システムで、語順並べ替えの最大値の設定である distortion limit は 20 に設定されている。それに対してベースラインシステムの distortion limit は 0 である。ベースラインシステムでこの値が 0 に設定された理由は日本語と韓国語の語順がほぼ同じであるためである。

この設定の違いから、MT4 の人手評価が低くなった理由は語順に問題があったためと推測される。この推測が正しいかどうかを検証するために、Adequacy 評価が低い翻訳結果を抽出して、翻訳誤りの原因を調べた。Adequacy 評価では 1 文の翻訳結果に対して、2 人の評価者が 1～5 の 5 段

階の評価値を独立に付与している。Adequacy 評価が低い翻訳結果として、2 つの評価値がどちらも 3 以下の翻訳結果の文を抽出した。MT4 で Adequacy 評価を実施した 200 文のうち、抽出された文数は 28 文であった。

この抽出した 28 文について、訳質低下の原因を調べたところ、いずれの訳文も語順に問題があることを確認した。また、語順の他に訳語選択に問題があるものも一部見受けられた。語順に問題がある訳文の例を以下に示す。

参照訳 1：２）０．５％D-アロース処理時の平均寿命は１５．８日（コントロール１２．５日）となり，平均寿命は２６％延長した。
MT 出力 1：処理時の平均寿命は１５．８日（コントロール１２．５０．５％D-２のソースとして、平均寿命は２６％延長したことである。

参照訳 2：上記内部電極用導電性ペースト組成物は、本発明の一実施形態によるものを使用することができ、具体的な成分及び含量は、上述した通りである。
MT 出力 2：上記内部電極用導電性ペースト組成物は、本発明の一実施形態による具体的な成分及び含量は、上述した ものを使用することができる。

上記の MT 出力 1 では、参照訳 1 の文頭部分に対応する部分が文中に位置している。それ以外の部分はほぼ参照訳と一致している。また、MT 出力 2 では、前半部分は完全に一致しているが、後半部分の語順が参照訳 2 と比べて局所的に一致していないために同じ意味になっていない。なお、MT 出力 1 の Sentence BLEU スコアは 0.532、MT 出力 2 の Sentence BLEU スコアは 0.7877 である。抽出した 28 文の Sentence BLEU スコアを図 4 に示す。スコアが 0.1 以下の低いものもいくつかあるが、多くのものは 0.5 以上あり、半数以上は 0.6 以上ある。これらのことから、語順の誤りは訳質に大きな影響がある場合があるが、BLEU ではこの影響の大きさは十分にスコアに反映されていないといえる。

3.4.4 まとめ

本稿では WAT2015 の中日及び韓日特許翻訳の人手評価結果に対して、2 種類の分析を行った。クラウドソーシング評価の信頼性についての分析の結果から、クラウドワーカーはこちらの意図しない点（評価対象とは考えていない点）を翻訳評価のポイントとしてしまい、望んだ結果が得られないことがあることがわかった。より適切な評価結果を得るためには、ワーカーへの作業説明をより詳細化する必要がある。

また自動評価と人手評価との相関が低くなる例の分析の結果からは、特に平均的な翻訳精度が高い状況においては、人手評価に大きく影響するような語順の誤りが、自動評価手法では適切に捉えられず、正しく評価が行えないことがわかった。これまで自動評価スコアが高ければ高いほど、人手評価も高いという認識が一般的であったが、今回得られた結果はこれを覆すものであり、面白い知見が得られたと思う。

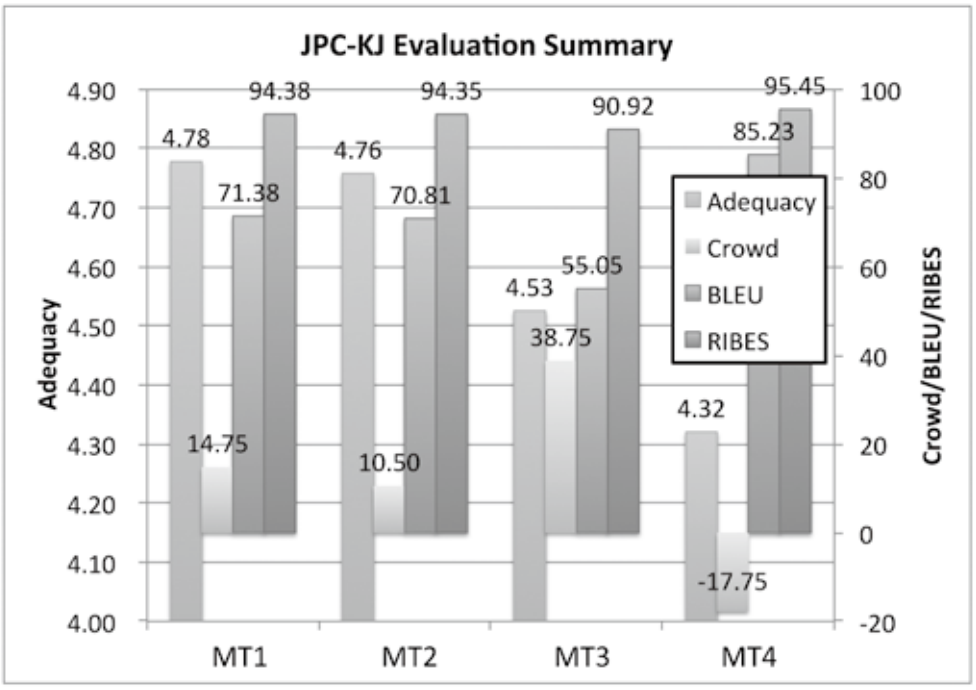


図 3 韓日特許翻訳タスク評価結果

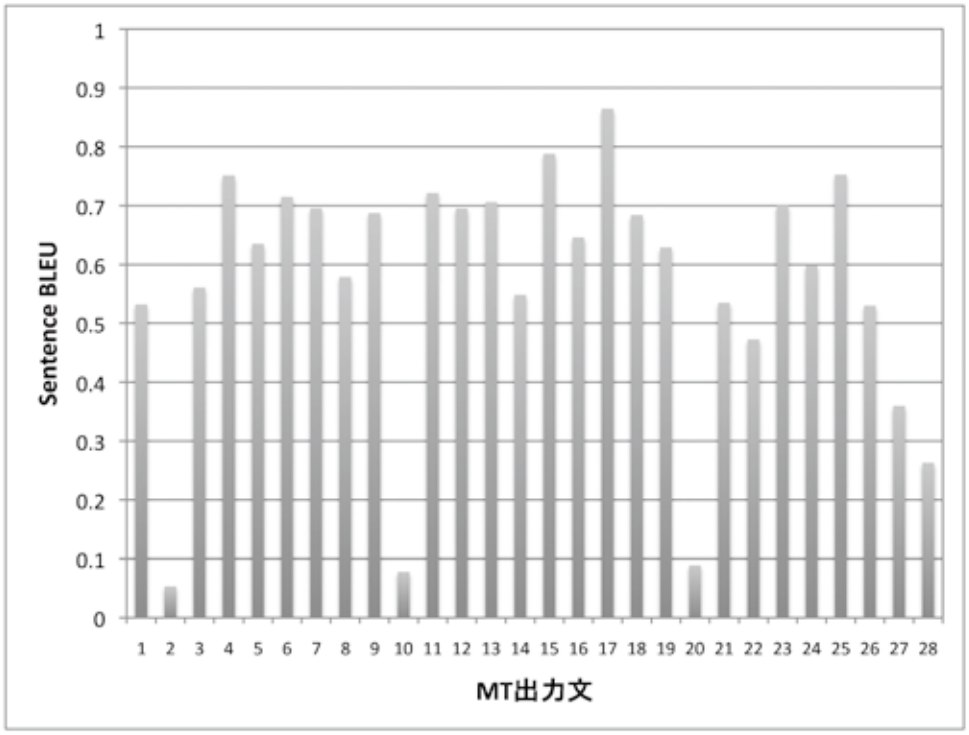


図 4 Sentence BLEU スコア

References

[1] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi and Eiichiro Sumita, Overview of the 2nd Workshop on Asian Translation, WAT2015, pages 1-28.

[2] Liling Tan, Jon Dehdari, Josef van Genabith, An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation, WAT2015, pages 74-81.

３．５　多言語のための大局的評価を用いた自動評価法

北海学園大学　越前谷　博

3.5.1　はじめに

統計翻訳やニューラルネットに基づく翻訳の発展に伴い、自動評価の重要性が一層高まっている。そうした背景よりこれまでに様々な自動評価法が提案されてきた。これまでの自動評価法は大きく 2 つのタイプに分別できる。一つは BLEU[1]や NIST[2]に代表されるような言語非依存の自動評価法である。これらの手法は単語単位でのマッチングに基づいているため、翻訳文と参照訳が共に単語分割されていれば特定の言語に依存することなく容易に評価することが可能である。このことが現在 BLEU がスタンダードな自動評価法として利用されている理由の一つになっている。しかし、これらの手法においては単語を最小単位とした単語マッチングのみに基づいているため、長文においては局所的な評価のみとなる。即ち、大局的な観点からの評価という点においては不十分と考えられる。

もう一つの自動評価法のタイプとしては様々な言語リソースを用いることを前提とした言語依存の自動評価法である。例えば、METEOR[3]は評価の際に言語リソースとして stemming、WordNet、そして、paraphrase table などを利用する。また、構文解析に基づく手法[4][5]、意味的知識を用いた自動評価法[6]も提案されている。更には、談話構造解析に基づく手法[7][8]が提案されている。これらの自動評価法では、様々な言語リソースを利用することにより、局所的な観点だけでなく大局的な観点からの評価も可能である。しかし、言語リソースが十分ではない言語の翻訳文に対しては評価対象とすることは難しいという問題点がある。

このような先行研究に対して、本報告では言語リソースに依存することなく大局的な評価も考慮した、新たな自動評価法を提案する。提案手法では、機能語に相当するストップワードを自動的に全参照訳から抽出する。その際の全参照訳とは、全翻訳文に対応する全ての参照訳を意味する。そして、そのストップワードは文を部分に分割するための区切り単語として利用される。分割された部分はフレーズに相当するため、大局的な評価を行うために有効である。このように提案手法では特定の言語に依存した情報を用いることなく、分割部分に基づく大局的な情報を翻訳文の評価に反映させる。性能評価実験の結果、提案する自動評価法の有効性を確認することができた。

3.5.2 関連研究

提案手法ではフレーズに相当する部分を自動的に決定することで大局的な情報を得る。フレーズチャンクを用いた自動評価法はこれまでにいくつか提案されている。Giménez and Màrquez[9] は shallow syntactic similarity に基づく手法を提案している。その際、フレーズチャンキングはパーセプトロン学習に基づく shallow parsing[10]により得る。更に、チャンクの順列を用いたスコア計算には NIST が用いられている。性能評価実験では、WMT2006[11]と NIST2005[12]のデータを用いているが、対象言語は英語のみである。

Echizen-ya and Araki[13]は、名詞句のチャンクに着目した大局的な評価を導入している。その際には、名詞句のチャンクのみを抽出し、チャンク単位で類似度計算を求めた結果を評価スコアに反映させている。名詞句のチャンクは条件付き確率場（CRF）による shallow parser を用いて決定している。性能評価実験では、Japanese-to-English 翻訳により得られた英語のみを評価対象としている。LiangYou ら[14]はフレーズ間の類似度、フレーズの重み付け、最大類似度マップの探索の 3 つの処理に基づく自動評価法を提案している。その際には名詞句のチャンクだけではなく、動詞句のチャンクも利用している。更に、全てのフレーズは CRF を用いたチャンカーを用いて得ている。性能評価実験においては、Chinese-to-English 翻訳により得られた英語のみを評価対象としている。

このようにフレーズチャンクに基づく自動評価法は基本的には言語依存の手法である。チャンカーはコーパスに基づく統計的手法を用いているが、容易に多言語に適用することは困難である。そのため関連研究では対象言語が英語のみになっていると考えられる。このような問題を踏まえ、提案手法では、フレーズに相当する部分を全参照訳のみから自動的に決定する。そのため、提案手法は言語非依存の手法として、大局的な情報を利用した評価が可能である。

3.5.3 大局的評価を用いた自動評価法

提案手法は主に次の 5 つの処理（①ストップワードの抽出、②文の分割、③大局的評価、④局所的評価、⑤大局的評価と局所的評価の組み合わせによる最終的なスコア計算）より構成されている。以下に、それぞれの詳細について述べる。

3.5.3.1 ストップワードの抽出

本報告では、文を分割するために、機能語に相当する単語としてストップワードを抽出する。ストップワードは評価対象の全翻訳文に対応する全ての参照訳より、単語出現頻度に基づき決定される。したがって、複数の参照訳の使用を前提としている。始めに以下の式(1)を用いて、全参照訳中の全ての単語に対して *tf・idf* を付与する。

$$tf \cdot idf = \log\left(tf(w,|R|)\right) \times \frac{|R|}{df(w)} \qquad (1)$$

式(1)の $tf(w,|R|)$は任意の単語 w の全参照訳数$|R|$に対する出現頻度を示している。また、$|R|/df(w)$は全参照訳数に対する、任意の単語 w が出現する参照訳の数の逆数である。$tf(w,/R/)$は出現頻度が高い単語ほど大きな値となり、$|R|/df(w)$は多くの参照訳に出現する単語ほど小さな値となる。更に、$tf(w,|R|)$に対しては log を付与しているため、出現頻度の高い単語についてはその値は抑えられる。そして、$|R|/df(w)$に対しては log を用いていないため、出現頻度の低い単語についてはその値は高くなる。したがって、式(1)により、機能語のような多くの参照訳に出現する単語の *tf・idf* は小さくなり、内容語のような限られた参照訳に出現する単語の *tf・idf* は大きくなる。

次いで、提案手法では式(1)より得られた *tf・idf* に対して閾値を設け、その閾値より小さな *tf・idf* を持つ単語をストップワードとして抽出する。閾値は以下の式(2)より求める。

$$閾値 = \log\left(\frac{|R|}{\mu}\right) \times \mu \qquad (2)$$

式(2)は閾値が参照訳数|R|に応じて動的な値となることを示している。μは1以上のパラメータである。例えば、μの値が10である場合、全参照訳数|R|中の10分の1に出現する単語がストップワードになることを意味する。全単語が個々の参照訳に一度のみ出現することを前提とした場合、μが10であれば式(1)の$tf(w,|R|)$は|R|/10となり、$|R|/df(w)$は10（=|R|/(|R|/10)）となる。したがって、閾値は式(2)においてμに10を用いた場合のlog(|R|/10)×10より得られることになる。このように閾値を決定するためにはパラメータμの値を与えなければならない。しかし、式(2)による閾値は参照訳数に応じて動的に変化することで適切な値を導き出せることから、固定的な閾値を設けるよりも有効と考えられる。

3.5.3.2 文の分割

前節で述べたストップワードを用いて翻訳文と参照訳を分割することでフレーズに相当する部分を得る。始めに、翻訳文と参照訳文間で最長共通部分列（LCS）に基づき共通部分を決定する。共通部分とは共通単語が翻訳文と参照訳共に連続している部分である。図１に決定された共通部分の具体例を示す。

翻訳文：

I was waiting at the gate to see the gnashing of teeth and argument .

参照訳：

I was expecting to see gnashing of teeth and a fight breaking out at the gate .

図１　LCSに基づく共通部分の例

図１の翻訳文と参照訳間においては下線部の“I was”、“to see the gnashing of teeth and”、そして、“.”が共通部分となる。LCSでは語順の異なる共通単語は選択されないため、“at the gate”には下線部が付与されていない。

次いで、3.5.3.1で述べた処理より抽出されたストップワードを用いて翻訳文と参照訳を複数の部分に分割する。共通部分にストップワードが含まれている場合に、そのストップワードを境界として文を分割する。図２にストップワードを用いた文分割の具体例を示す。図２では、共通部分中のストップワードは“to”、“of ”、“and”、そして、“.”である。したがって、これらのストップワードを境界として翻訳文と参照訳を分割する。その結果、翻訳文は“I was waiting at the gate”、“see the gnashing”、“teeth”、“argument”の４つに分割され、参照訳は“I was expecting”、“see gnashing”、“teeth”、“a fight breaking out at the gate”の４つに分割される。

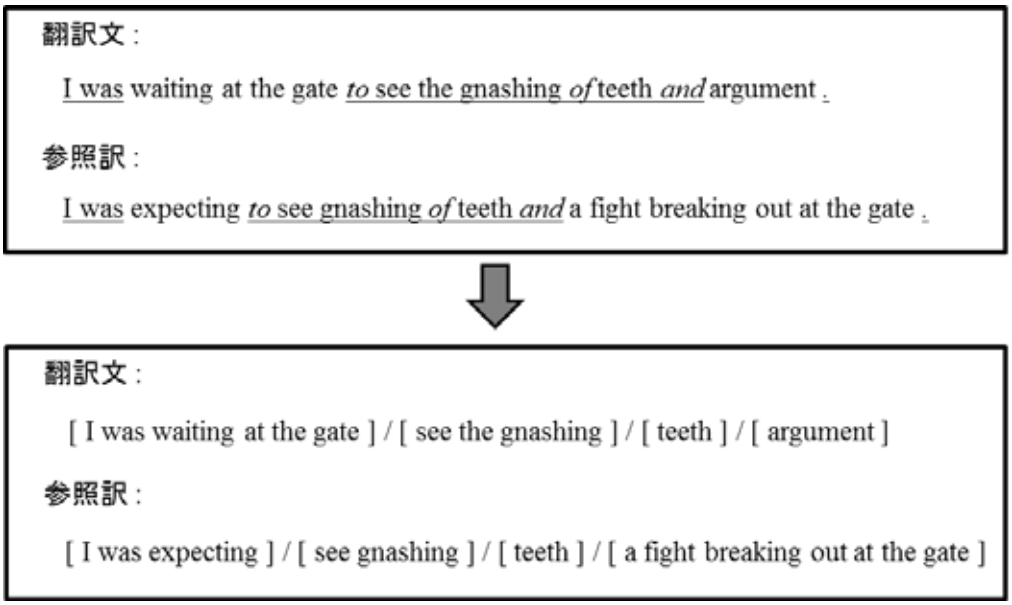


図２　ストップワードを用いた文分割の例

3.5.3.3 大局的評価

提案手法では前節で得られた文分割の結果を用いて大局的な評価を行う。即ち、大局的な観点からのスコア（global_score）を算出する。図３に global_score の算出の具体例を示す。

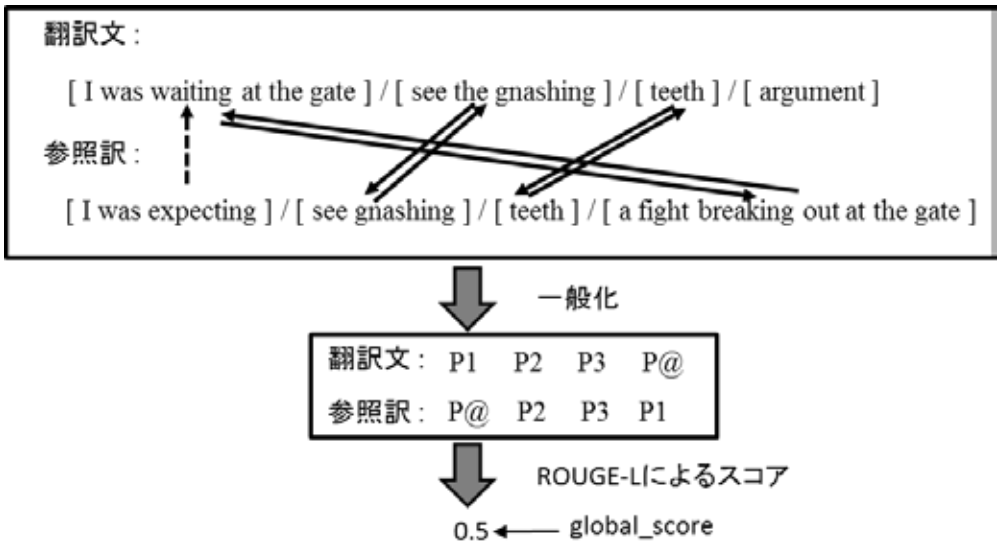


図３　global_score の算出例

大局的な評価である global_score を求めるために始めに、ストップワードを用いて分割された翻訳文と参照訳の部分間の対応関係を求める。部分間の対応関係は類似度を用いて行う。具体的には以下の式(3)を用いて部分間の類似度を求める。

$$sim = \frac{LCS}{length(P)} \qquad (3)$$

式(3)の *length(P)*は分割された部分 *P*の構成単語数であり、*LCS*は部分間における *LCS*の値である。例えば、翻訳文中の部分“see the gnashing”と参照訳文中の全部分との類似度を式(3)より求めると、“I was excepting”との類似度は 0.0、“see gnashing”との類似度は二つの単語“see”と“gnashing”が一致するため 0.66(=2/3)となり、“teeth”との類似度は 0.0、そして、“a fight breaking out at the gate”との類似度は“the”のみが一致するため 0.33(=1/3)となる。したがって、最も類似度が高い部分として“see gnashing”が選択される。同様に参照訳中の“see gnashing”と翻訳文中との全部分との類似度を求めると、最も類似度が高い部分は“see the gnashing”となる。したがって、翻訳文中の部分“see the gnashing”と参照訳文中の部分“see gnashing”はお互いに最も類似度の高い部分として選択されることになるため、対応関係が成り立つと位置付けられる。このように部分間の対応関係を求めると図３のように、“I was waiting at the gate”と“a fight breaking out at the gate”、“see the gnashing”と“see gnashing”、そして、“teeth”と“teeth”の間で対応関係が成立する。

参照訳中の部分“I was expecting”については、翻訳文中の部分“I was waiting at the gate”が最も類似度の高い部分として選択される。しかし、翻訳文中の部分“I was waiting at the gate”は参照訳中の“I was expecting”との類似度は 0.33(=2/6)であり、“a fight breaking out at the gate”との類似度は 0.50(=3/6)となるため、“I was expecting”ではなく“a fight breaking out at the gate”が選択される。したがって、参照訳中の部分“I was expecting”は対応する部分が存在しないことになる。翻訳文中においても部分“argument”は対応関係が存在しない部分となる。

このように翻訳文の部分と参照訳の部分の対応付けを行った後、部分を一般化することにより、単語単位ではなく、フレーズに相当するより大きな単位での評価が可能となる。具体的には対応関係が成立する部分間においては翻訳文と参照訳で同じ番号を付与する。対応する部分が存在しない場合には、便宜上番号ではなく“@”を付与する。例えば、図３では翻訳文は“P1 P2 P3 P@”として一般化される。参照訳は“P@ P2 P3 P1”として一般化される。

最後に一般化された翻訳文と参照訳を用いて global_score を求める。その際には、自動評価法の一つである ROUGE-L[15]を用いる。ROUGE-L は LCS に基づいているため出現順に厳しい評価基準であり、文の構造を大局的に捉えることに適していると考えられる。また、スコアは 0.0～0.1 に正規化されているため算出されたスコアを直感的に捉えやすい。図３においては一般化された翻訳文と参照訳との間で ROUGE-L を用いた場合、global_score として 0.5 が得られる。global_score が低下した要因は“at the gate”の位置が翻訳文と参照訳で大きく異なっているためである。単語を最小単位とした局所的評価だけでは、このような場合にそれほどスコアに反映されないが、より大きな単位に基づく大局的評価を用いることで文の構造の違いをスコアに反映することが可能となる。

3.5.3.4 局所的評価

局所的な評価は単語を最小単位として行う。その結果得られたスコアを local_score と呼ぶこととする。局所的な評価スコア local_score は著者が従来より提案している自動評価法の IMPACT[16]を用いる。IMPACT は翻訳文と参照訳間の共通部分を LCS に基づき決定するが、語順の異なる共通部分についてもスコアに反映させるために共通部分の決定処理を再帰的に行っている。そのためには共通部分列を一意に決定する必要があるが、IMPACT では個々の共通部分の相対的な位置と共通部分の長さに基づき一意に共通部分列を決定している。また、語順の異なる共通部分をスコアにどの程度反映させるかはパラメータを用いて制御可能となっている。このように IMPACT は全ての共通単語をスコアに反映させながらも、語順を考慮した柔軟な自動評価法である。

3.5.3.5 大局的評価と局所的評価の組み合わせによる最終的なスコア計算

提案手法では、大局的な評価スコアである global_score を局所的な評価スコアである local_score の重み付けとして用いる。具体的には以下の式(4)を用いて最終的なスコアを求める。

$$score = (global\_score)^{\delta} \times local\_score \qquad (4)$$

式(4)の δ はパラメータである。パラメータ δ の値としては global_score が local_score に過度に影響を及ぼすことを避けるために本報告では 0.1 を用いる。

3.5.4 性能評価実験

3.5.4.1 実験データ

実験データには、NTCIR-7 データ[17]及び WMT14 Metrics Task データ[18]、更には WMT15 Metrics Task データ[19]を用いた。NTICR-7 データは英日、日英両方向の翻訳文、参照訳が提供されている。翻訳文は英日においては、5 つの機械翻訳システムがそれぞれ 100 文の英文を日本文に翻訳した結果が用いられており、計 500 の翻訳文が提供されている。日英においては、15 の機械翻訳システムがそれぞれ 100 文の日本文を英文に翻訳した結果が用いられており、計 1500 文の翻訳文が提供されている。参照訳には正解訳として日本文、英文それぞれ 100 文ずつが提供されている。人手評価は 3 名の評価者が adequacy と fluency の観点より 1 から 5 までの 5 段階での絶対評価を実施した結果が提供されている。なお、5 段階評価においては、評価値が高いほど高い評価となる。今回は 3 名の評価値の平均値を用いている。

WMT14 Metrics Task データはチェコ語（cs）―英語（en）、ドイツ語（de）―英語、フランス語（fr）―英語、ヒンディー語（hi）―英語、そして、ロシア語（ru）―英語間の双方向でのシステム訳が提供されている。機械翻訳システムの数は cs-en が 5、de-en が 13、fr-en が 8、hi-en が 9、ru-en が 13、en-cs が 10、en-de が 18、en-fr が 13、en-hi が 12、そして、en-ru が 9 の計 110 である。WMT15 Metrics Task データについてはチェコ語（cs）―英語（en）、ドイツ語（de）―英語、フランス語（fr）―英語、フィンランド語（fi）―英語、そして、ロシア語（ru）―英語間の双方向でのシステム訳が提供されている。機械翻訳システムの数は cs-en が 16、de-en が 13、fr-en が 7、fi-en が 14、ru-en が 13、en-cs が 15、en-de が 16、en-fr が 7、en-fi が 10、そして、

en-ru が 10 の計 121 である。

3.5.4.2 評価方法

評価は、自動評価法のスコアと人手評価のスコアと間の相関係数を求めることで行った。その際には、system-level と segment-level の両方について相関係数を求めた。NTCIR-7 データについては、system-level と segment-level に対して Pearson の相関係数、Spearman の順位相関係数、そして、Kendall の順位相関係数を求めた。また、WMT14 Metrics Task データと WMT15 Metrics Task データにおいては、system-level は Pearson の相関係数、segment-level は 2 つの自動評価法のスコアと人手評価のスコアの大小比較に基づく Kendall の τ を求めることで評価を行った。system-level の人手評価は TrueSkill[20]を用いて求めている。このような WMT14 Metrics Task と WMT15 Metrics Task の評価方法は文献[18]と文献[19]に準拠している。日本文に対しては MeCab[21]を用いて、分かち書きを行った。

また、今回は自動評価法として IMPACT、BLEU、SENTBLEU[18][19]、CDER[22]、そして、提案手法を使用した。

3.5.4.3 実験結果

表 1 から表 4 に NTCIR-7 データを用いた実験結果、表 5 から表 8 に WMT14 Metrics Task データを用いた実験結果、そして、表 9 から表 12 に WMT15 Metrics Task データを用いた実験結果を示す。表 5、表 7、表 9、表 11 の“()”内の数値は、機械翻訳システムの数を示している。表 6、表 8、表 10、表 12 の“()”内の数値は、スコアの大小比較を行った際のペアの数を示している。また、表 5 から表 12 の太字の数値は自動評価法の中で最も相関係数が高かったことを示している。

表 1　NTCIR-7 データにおける英日翻訳での system-level の相関係数

	Pearson		Spearman		Kendall	
	adequacy	fluency	adequacy	fluency	adequacy	fluency
提案手法	**0.284**	**0.514**	**0.300**	**0.500**	**0.200**	**0.400**
IMPACT	0.254	0.489	**0.300**	**0.500**	**0.200**	**0.400**
BLEU	−0.199	0.184	−0.100	0.200	0.000	0.200

表 2　NTCIR-7 データにおける英日翻訳での segment-level の相関係数

	Pearson		Spearman		Kendall	
	adequacy	fluency	adequacy	fluency	adequacy	fluency
提案手法	**0.665**	**0.586**	**0.637**	**0.578**	**0.471**	**0.424**
IMPACT	0.657	0.583	0.624	0.572	0.461	0.419

表 3　NTCIR-7 データにおける日英翻訳での system-level の相関係数

	Pearson		Spearman		Kendall	
	adequacy	fluency	adequacy	fluency	adequacy	fluency
提案手法	**0.836**	**0.943**	**0.873**	**0.766**	**0.760**	**0.625**
IMPACT	0.814	0.936	0.810	0.695	0.689	0.555
BLEU	0.730	0.881	0.567	0.552	0.498	0.440

表 4　NTCIR-7 データにおける日英翻訳での segment-level の相関係数

	Pearson		Spearman		Kendall	
	adequacy	fluency	adequacy	fluency	adequacy	fluency
提案手法	**0.639**	0.640	**0.636**	**0.627**	**0.474**	**0.469**
IMPACT	0.631	**0.646**	0.625	0.624	0.466	0.467

表 5　WMT14 Metrics Task データにおける英語から多言語翻訳の system-level の相関係数

	en-fr(13)	en-hi(12)	en-cs(10)	en-ru(9)	Avg.	en-de(18)
提案手法	0.943	0.961	**0.984**	**0.933**	0.955	0.252
IMPACT	0.943	**0.965**	**0.984**	**0.933**	**0.956**	0.253
CDER	**0.948**	0.953	0.973	0.932	0.952	**0.428**

表 6　WMT14 Metrics Task データにおける英語から多言語翻訳の segment-level の相関係数

	en-fr (33,350)	en-de (54,660)	en-hi (28,120)	en-cs (55,900)	en-ru (28,960)	Avg.
提案手法	**0.272**	**0.218**	0.233	0.305	0.404	0.286
IMPACT	**0.272**	0.215	**0.234**	**0.306**	**0.406**	**0.287**
SENTBLEU	0.239	0.193	0.197	0.272	0.368	0.254

表 7　WMT14 Metrics Task データにおける多言語から英語翻訳の system-level の相関係数

	fr-en(8)	de-en(13)	hi-en(9)	cs-en(5)	ru-en(13)	Avg.
提案手法	**0.953**	0.823	0.920	**0.977**	**0.806**	**0.896**
IMPACT	0.952	0.818	0.914	0.976	0.805	0.893
BLEU	0.952	**0.832**	**0.956**	0.909	0.789	0.888

表 8　WMT14 Metrics Task データにおける多言語から英語翻訳の segment-level の相関係数

	fr-en (26,090)	de-en (25,260)	hi-en (20,900)	cs-en (21,130)	ru-en (34,460)	Avg.
提案手法	**0.397**	**0.296**	0.361	0.235	0.295	**0.317**
IMPACT	0.393	0.292	**0.362**	**0.237**	**0.296**	0.316
SENTBLEU	0.352	0.261	0.257	0.189	0.249	0.262

表 9　WMT15 Metrics Task データにおける英語から多言語翻訳の system-level の相関係数

	en-fr(7)	en-fi(10)	en-de(16)	en-cs(15)	en-ru(10)	Avg.
提案手法	0.952	**0.722**	0.531	**0.956**	**0.879**	0.808
IMPACT	**0.954**	0.720	0.529	0.954	0.872	0.806
CDER	0.953	0.640	**0.660**	0.929	0.863	**0.809**

表 10　WMT15 Metrics Task データにおける英語から多言語翻訳の segment-level の相関係数

	en-fr (34,512)	en-fi (32,694)	en-de (54,447)	en-cs (136,890)	en-ru (49,302)	Avg.
提案手法	**0.327**	**0.259**	**0.300**	0.381	0.364	0.3262
IMPACT	**0.327**	**0.259**	0.299	**0.382**	**0.365**	**0.3264**
SENTBLEU	0.318	0.227	0.294	0.360	0.347	0.309

表 11　WMT15 Metrics Task データにおける多言語から英語翻訳の system-level の相関係数

	fr-en(7)	fi-en(14)	de-en(13)	cs-en(16)	ru-en(13)	Avg.
提案手法	0.973	0.949	**0.902**	**0.983**	**0.931**	**0.948**
IMPACT	0.973	0.951	0.897	0.982	0.928	0.946
CDER	**0.983**	**0.966**	0.890	0.962	0.920	0.944

表 12　WMT15 Metrics Task データにおける多言語から英語翻訳の segment-level の相関係数

	fr-en (29,770)	fi-en (31,577)	de-en (40,535)	cs-en (85,877)	ru-en (44,539)	Avg.
提案手法	0.369	**0.349**	**0.378**	0.407	0.350	0.3706
IMPACT	**0.371**	0.346	0.377	**0.408**	**0.353**	**0.3710**
SENTBLEU	0.358	0.308	0.360	0.391	0.329	0.349

3.5.4.4 考察

表 1 から表 4 の NTCIR-7 データにおいては、提案手法は大局的評価を適用していない IMPACT に比べ高い相関係数を示している。IMPACT よりも相関係数が低かったのは表 4 の日英翻訳での segment-level における Pearson の fluency のみであった。したがって、提案手法は日英間の翻訳文及び特許翻訳文においては IMPACT よりも高い評価精度を有すると考えられる。また、表 1 において、いずれの自動評価法も相関が非常に低くなっている。NTCIR-7 データの英日翻訳では機械翻訳システムが 5 つと非常に少ないため、1 つでも人手評価と異なると著しく相関係数が低下してしまう。したがって、他のデータと比べて極端に評価精度が低かったということにはならないと考えられる。

表 5 から表 8 の WMT14 Metrics Task データにおいては、“Avg.”を比較すると、多言語から英語への翻訳においては表 7 と表 8 より提案手法は IMPACT に対して高い相関係数を示している。それに対して、英語から多言語への翻訳においては表 5 と表 6 より IMPACT の方が高い相関係数を示している。しかし、表 5、表 6 共に差はわずかである。また、表 7 の多言語から英語への翻訳の system-level においては、提案手法は IMPACT に対して全ての言語で高い相関係数を示している。したがって、英語を評価対象とした system-level において提案手法は IMPACT に比べ有効と考えられる。

表 9 から表 12 の WMT15 Metrics Task データにおいては、“Avg.”を見ると表 11 の多言語から英語への翻訳の system-level のみ提案手法の相関係数は最も高く、他の“Avg.”は他手法が高い相関係数を示した。この傾向は、WMT14 Metrics Task データと同様であり、やはり提案手法は英語を評価対象とした system-level において有効であると考えられる。表 10 と表 12 の segment-level においては提案手法は IMPACT の相関係数よりも低い値となっているが、その差は非常に小さく、評価精度が著しく低いという訳ではない。一方、表 9 の英語から多言語への翻訳の system-level では提案手法は CDER の“Avg.”よりは低いが IMPACT の“Avg.”よりも高く、大局的評価の効果が見られる。

このように WMT Metrics Task データでは常に提案手法の相関係数が他の手法の相関係数と比べて高いわけではないが、多言語から英語への翻訳においては system-level で高い相関係数を示すなど大局的評価の効果を確認することができた。

3.5.5 まとめ

本報告では大局的評価も考慮した、多言語に適用可能な自動評価法を提案した。提案手法では、文を分割する際に使用するストップワードを参照訳のみから自動抽出することで、対象言語に依存することなく、様々な言語の翻訳文に対して文分割が可能である。そして、分割により得られたフレーズに相当する部分を用いて大局的な評価を行う。更に、分割部分を最小単位として得られるスコア global_score は、単語を最小単位とした評価手法より得られる局所的スコア local _score の重み付けに用いられる。このように提案手法は大局的評価と局所的評価の両方の観点に着目した自動評価法となっている。性能評価実験の結果、提案手法の有効性が確認された。

今後は、自動抽出されたストップワードを局所的評価にも利用するなど、より良い自動評価法

を実現するための改良を行う予定である。

謝辞

この研究は国立情報学研究所との共同研究に関連して行われた。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu(2002) “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 311-318.

[2] NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, 2002, http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

[3] Satanjeev Banerjee, Alon Lavie(2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pp.65-72.

[4] Ding Liu, Daniel Gildea(2005) “Syntactic Features for Evaluation of Machine Translation,” Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp.25-32.

[5] Hui Yu, Qingsong Ma, Xiaofeng Wu, Quu Liu(2015) “CASICT-DCU Participation in WMT2015 Metrics Task,” Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 417-421.

[6] Lo, Chi-kiu, Anand Karthik Tumuluru, Dekai Wu(2012) “Fully Automatic Semantic MT Evaluation,” Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 243-252.

[7] Jesús Giménez, Lluís Màrquez, Elisabet Comelles, Irene Castellòn, Victoria Arranz(2010) “Document-level Automatic MT Evaluation based on Discourse Representations,” Proceedings of the Joint fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 333-338.

[8] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Preslav Nakov(2014) “Using Discourse Structure Improves Machine Translation Evaluation,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 687-698.

[9] Jesús Giménez, Lluís Màrquez(2007) “Linguistic Features for Automatic Evaluation of Heterogenous MT Systems,” Proceedings of the Second Workshop on Statistical Machine Translation, pp. 256-264.

[10] Xavier Carreras, Lluís Màrquez, Jorge Castro(2005), “Filtering-Ranking Perceptron Learning for Partial Parsing,” Machine Learning, 60(1), pp. 41-71.

[11] Philipp Koehn, Christof Monz(2006) “Manual and Automatic Evaluation of Machine Translation between European Languages,” Proceedings of the Workshop on Statistical

Machine Translation, pp.102–121.
[12] Audrey Le, Mark Przybocki(2005) “NIST 2005 Machine Translation Evaluation Official Results,” Technical Report, NIST.
[13] Hiroshi Echizen-ya, Kenji Araki(2010) “Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking,” Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp.108-117.
[14] Li LiangYou, Gong ZhengXian, Zhou GuoDong(2012) “Phrase-Based Evaluation for Machine Translation,” Proceedings of the 24th International Conference on Computational Linguistics, pp. 663-672.
[15] Chin-Yew Lin, Franz Josef Och(2004) “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics,” In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pp.605-612.
[16] Hiroshi Echizen-ya, Kenji Araki(2007) “Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum,” Proceedings of the Eleventh Machine Translation Summit, pp.151-158.
[17] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro(2008) “Overview of the Patent Translation Task at the NTCIR-7 Workshop,” Proceedings of NTCIR-7 Workshop Meeting, pp.389-400.
[18] Matouš Macháček, Ondřej Bojar(2014) “Results of the WMT14 Merics Shared Task,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.293-301.
[19] Matouš Macháček, Amir Kamran, Philipp Koehn, Ondřej Bojar(2015) “Results of the WMT15 Merics Shared Task,” Proceedings of the Tenth Workshop on Statistical Machine Translation, pp.256-273.
[20] “TrueSkill,” http://en.wikipedia.org/wiki/TrueSkill
[21] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,”
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[22] Gregor Leusch, Nicola Ueffing, Hermann Ney(2006) “CDER: Efficient MT Evaluation Using Block Movements,” Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics, pp.241-248.

4. 第15回機械翻訳サミット参加報告

4　第 15 回機械翻訳サミット参加報告

ＮＴＴコミュニケーション科学基礎研究所　須藤　克仁

本研究会が中心となって開催する特許・技術文書翻訳ワークショップのオーガナイザとしてワークショップに出席するとともに、機械翻訳・特許翻訳に関する調査を目的として第15回機械翻訳サミットに参加した。

４．１　本会議等参加報告

第 15 回機械翻訳サミット (Machine Translation Summit XV)は、2015 年 10 月 30 日～11 月 3 日に米国フロリダ州マイアミの Hyatt Regency Miami で開催された。初日と最終日はチュートリアル及びワークショップ、本会議は 10 月 31 日～11 月 2 日の 3 日間であった。

本会議は参加者数約 200 名（事前登録数は 169)、最終日の翻訳後編集ワークショップは参加者約 60 名(事前登録数は 52)であった。本会議の全体構成としては、3 つのパラレルセッション（MT researchers' track, commercial MT users & translators' track, government MT users' track）の他に各開催日とも招待講演が 1-2 件行われ、本会議 2 日目の午後に企業製品等の展示が行われる技術展示(Technology Showcase)が、また本会議最終日の午後にパネルディスカッションが開催された。
本会議の研究トラックの採択率は約 45%（投稿数 60、口頭発表 17、ポスター発表 10）であった。商用ユーザ・翻訳者トラック、政府ユーザトラックの採択率は公開されなかったようだが、それぞれ 20 件、11 件の発表があった。IAMT によって機械翻訳の発展への貢献を表彰する IAMT Award of Honor は、句に基づく統計翻訳の考案とオープンソースの統計翻訳ツールキット Moses の公開・発展に大きく貢献した、ジョンズ・ホプキンス大学の Philipp Koehn 教授に授与された。
なお、最終日に次回の機械翻訳サミットは 2017 年 9 月 18 日～22 日に名古屋大学で開催されることが AAMT 会長の中岩浩巳先生(名古屋大学)から発表された。また、AMTA の会議は 2016 年 10 月 29 日～11 月 2 日に米国テキサス州オースティンで、自然言語処理の国際会議 EMNLP との連続開催となることが発表された。

招待講演は以下の 5 件であった。
・ニューヨーク大学の Kyunghyun Cho 氏によるニューラルネットワーク機械翻訳に関する講演
・ジョンズ・ホプキンス大学（元 NAIST 助教）の Kevin Duh 氏によるニューラ

ルネット機械翻訳のための自然言語のモデル化に関する講演
・欧州委員会(EC)の Spyridon Pilos 氏による EC での機械翻訳プロジェクトに関する講演
・Google の Macduff Hughes 氏による今後 10 年の機械翻訳の展望に関する講演
・ジョンズ・ホプキンス大学の Matt Post 氏による話し言葉翻訳についてのサマーワークショップに関する講演
全体的に研究者寄りの講演が多かったが，現時点での最先端技術であるニューラルネットワーク機械翻訳に関する話題は研究者以外の参加者からも注目を集めており、活発に議論されていた。

パネルディスカッションは David Rumsey (ATA)、Jost Zetzsche (翻訳者)、Jose Palomares (Venga) の 3 氏により、翻訳者と機械翻訳の関係、主に機械翻訳が翻訳者にとって有用であるかどうかについての議論が行われた。従来の翻訳メモリの利用に加え、後編集やコンピュータ補助翻訳(computer aided translation)の導入によって今後翻訳者の業務の習慣 (work habit) には変化が訪れるだろうという認識が示された。筆者にとって興味深かった議論としては、機械翻訳を利用した人手の翻訳の生産性の評価のあり方として単純に時間だけで測ることは望ましくないのではないか、人手翻訳と機械翻訳が協調していけるはずだが秘匿すべき情報の管理については課題がある、といったものがあった。

一般講演は 3 セッションあったが、筆者は大部分研究トラックを聴講していた。ACL や EMNLP といった自然言語処理の難関国際会議においてはニューラルネットワーク機械翻訳の改善に関する論文がかなり多数を占めるようになってきつつあるが、機械翻訳サミットでは少しタイプの違う問題を扱う研究など内容が多岐に渡る印象を受けた（ただし、技術的な動向としてニューラルネットワークがある種当たり前のツールとして利用されつつあることは間違いない)。また、翻訳後編集に関する研究発表が聴衆からの質疑の活発さという面で目立っていたことが印象に残っている。以下、筆者が注目した論文をいくつか簡単に紹介する。

まず、機械翻訳のフレーズの分散表現に関する発表が 2 件（中国蘇州大学の Wang 氏ら、NICT の美野氏ら）あり、ニューラルネットワークを利用して句の類似度を適切に評価することで、従来型の句に基づく統計翻訳を改善している。対応する両言語の句の分散表現を一致させるような学習を行う、という意味では両者は非常に似通っているが、Wang 氏らが再帰型(recursive)ニューラルネットワークで句の分散表現を得る手法であるのに対し、美野氏らは回帰結合型(recurrent)ニューラルネットワークを利用している。単に表層のみで句を区別するとデータの不足やノイズの混入に弱くなってしまうため、似た意味の句を分散表現で近い

領域に射影し機械翻訳の頑健性を向上させることは重要である。現在のところ、精度向上の度合いは決して大きいとは言えないが、その評価方法も含めて検討する価値がありそうである。

また、複雑な形態素変化の起こる言語に対する機械翻訳に関する発表が 2 件（コロンビア大学の Kholy 氏ら、アムステルダム大学の Tran 氏ら）あった。Kholy 氏らは形態素変化の多い言語間の翻訳において、中間言語として形態素変化の少ない言語を挟むピボット翻訳を利用すると情報の不足によって正しい対応関係が取れなくなる問題に対し、ピボット翻訳時の句の対応における形態素の対応制約をルールで定義してフレーズテーブルのクリーニングを行うことでヘブライ語からアラビア語への翻訳が改善できることを示した。Tran 氏らは形態素変化を決定するための性・数・時制などの情報を予測するためのニューラルネットワークモデルを提案し、英語からロシア語への翻訳を改善している。性・数・時制による形態素変化は日英の間でも正しく対応付けることが容易でなく、単に内容語が一致していればよいという段階からさらに進んでいくためにこうした技術が必要であることは間違いない。

今回の研究論文の中で異色と感じたのはニューヨーク市立大学の Zhai 氏らによる、従来のパイプライン型処理ではなく end-to-end の処理によって統計翻訳のモデル化を行う研究である。句に基づく統計翻訳では通常単語の対応付け、句の対応付け、句の翻訳モデルの推定、モデル重みの最適化、と段階を踏んで最終的なモデルを学習するが、Zhai 氏らの研究では、簡単な初期モデルから始めて、対訳データに対する強制(forced)デコーディング結果からの対訳句抽出と構造化パーセプトロンによるモデル更新の繰り返しによって一挙にモデルを学習する方法を提案している。現在は従来型の手法と同程度の結果が得られたに過ぎないが、従来の統計翻訳の複雑さを解決する一つの試みとして非常に興味深かった。

技術展示では、機械翻訳システムのデモ（Systran や Microsoft 等）、機械翻訳を自社サービスに組み込んでいる企業の技術紹介（IBM、SDL、eBay 等）、など計 20 ブースなどが技術や製品の紹介を行っていた。多くの参加者を集めていたのは Microsoft の Skype Translator であり、接話マイクを利用してはいたものの、騒がしい会場で英語・ドイツ語の音声から音声への翻訳がかなりの精度で動作していたことが印象的であった。

会議最終日の午前中に開催された翻訳後編集ワークショップでは、2 件の招待講演と 5 件の技術論文講演があり、およそ 60 名の翻訳者と研究者が集まり、活発な議論が行われた。特に注目されたトピックはいかにして翻訳者に使いやすい・役立つ後編集の仕組みを作るか、という点であった。単に機械翻訳の精度が向上す

ればよいということではなく、適切なインタフェースの設計や情報提示のやり方はどうあるべきか、どういった機械翻訳誤りが後編集しづらいのかという分析や後編集しやすさの評価はどうするか、といった、機械翻訳の精度向上の研究とは異なる方向性の存在を強く感じた。

会議を通じて筆者が最も強く感じたのは、機械翻訳は産業上の応用が明確に存在する技術であって、応用によって重視されるポイントが大きく異なる、ということである。特に機械翻訳サミットや AMTA、EAMT 等の会議では機械翻訳の研究者・開発者だけでなく翻訳者や翻訳業者からの参加者も多く、応用の視点から機械翻訳の現在を見つめ直せることを再確認した。研究トラックだけに注目すると必ずしも完成度の高い研究が並んでいるというわけではないのだが、応用、特に翻訳後編集やコンピュータ補助翻訳の利便性向上のための研究やユーザからの問題提起といった内容は機械翻訳専門の会議であるからこそのものであると感じた。

４．２　特許・技術文書翻訳ワークショップ開催報告

本研究会の活動の一環として、2005 年の第 1 回から数えて 6 回目のワークショップを、機械翻訳サミットに併設する形で会議初日の 10 月 30 日に開催した。今回は特許翻訳に限定せず幅広く技術文書に関する機械翻訳の課題を扱うという観点から、特許・技術文書翻訳ワークショップ（Workshop on Patent and Scientific Literature Translation）と題した。ワークショップ co-chair は梶博行先生（静岡大学学術院教授）と須藤の 2 名が担当し、予稿集の編集を綱川隆司先生（静岡大学学術院助教)、プログラム委員を研究会委員全員と海外の関連研究者 9 名にご担当いただいた。ワークショップは、テーマごとの 4 つのセッションに分けて、それぞれの内容に関連する招待講演（計 5 件）と技術論文の講演（計 4 件）を行う形式であった（なお、技術論文は計 6 件を採択したが、ビザ等の問題で中国からの発表 2 件が取消となった)。以下、各セッションごとの内容について報告する。

セッション 1: MT in Patent Organizations

本セッションでは、公的な知財担当機関における機械翻訳の活用についての 2 件の招待講演をお願いした。公的機関において機械翻訳が有効に活用され、必要な知的財産の情報に簡単にアクセスできるようになることは技術や産業の発展に重要であり、特にこの数年での大きな進展を知ることができ非常に有益であった。

1 件目の招待講演は世界知的財産機構（World Intellectual Property Organization: WIPO)の Bruno Pouliquen 氏による、WIPO の機械翻訳についての講演であった。WIPO では多言語の特許文書から自動的に対訳コーパスを構築

し、統計的機械翻訳によって発明の名称と概要の機械翻訳を行い、他言語での特許検索を可能にするサービスを提供している。多言語化によって扱うデータの規模が膨大であることから学習はすべて自動化している、また高速・省メモリな翻訳を実現するためのデータ選択や、ドイツ語・日本語等一部言語での事前並べ替え等も行っている、といった、実際にシステムを運用する上での様々な工夫について紹介された。

2 件目の招待講演は特許庁の加藤啓氏による、特許庁での機械翻訳に関する取り組みについての講演であった。特許庁では機械翻訳の評価基準の策定、また要望が年々高まっている中国語特許への対応を見据えた日中対訳用語辞書の整備などを進めている。また今年(2015 年)からは中国・韓国の特許に対応した特許検索システムを公開していること、米国・欧州・中国・韓国の各特許庁との情報共有を行う OPD (Open Portal Dossier)において日本語から英語への機械翻訳が活用されていることなどが紹介された。

セッション 2: Effective Use of Patent MT

本セッションでは機械翻訳の活用による翻訳者支援をテーマとし、Iconic Translation Machines Ltd. の John Tinsley 氏に招待講演をお願いした。実際のビジネスとして考えた場合には生産性やユーザにとっての使いやすさが重要であり、後編集に渡す機械翻訳がいかにあるべきか、ということについて考えさせられた。

講演では、Google のようにターゲットを絞らない機械翻訳とは対極的に、分野適応をすることで特許のような特殊な分野の機械翻訳の精度を大きく向上させることができ、それによって後編集による翻訳の生産性が大きく向上することについて、実例を示しながら説明があった。彼らのシステムでは、用語辞書やキーワード抽出、ルールベース翻訳、統計翻訳などの様々なモジュールを組み合わせ、後編集を含めたワークフローの生産性が高まるように調整が行われている。生産性の評価には TAUS の Dynamic Quality Framework (DQF)を利用し、後編集が直接翻訳よりも生産性が高まることを確認している。また、参照訳に対する翻訳編集率(Translation Edit Rate: TER)が 40%を下回るくらいになると後編集の効率がよくなるなどの分析結果が示された。

セッション 3: Challenges for Advanced Patent MT

本セッションでは、特許機械翻訳のさらなる改善について、1 件の招待講演と 2 件の技術論文の講演が行われた。これまでのセッションでは機械翻訳の使われ方が議論の対象であったが、本セッションは機械翻訳の改善のための技術が焦点となった。

最初に、独ハイデルベルク大学の Stefan Riezler 教授による、ユーザフィードバックに基づく統計翻訳の改善についての講演をお願いした。翻訳後編集結果を利用して統計翻訳を改善する研究はこれまでにも様々行われているが、彼らの研究では、従来研究が前提としていた専門翻訳者の後編集結果ではなく、ユーザの局所的なフィードバック（例えば翻訳結果に対する品質推定値，実験では参照訳に対する文単位 BLEU 値）のみを利用する方法を提案している。基本的な考え方は構造化パーセプトロンと同様で、フィードバックで与えられる損失が小さくなるようにモデルパラメータを更新するというものである。この手法により、Europarl（議会議事録）から News Commentary（ニュース）への分野適応を行った場合の翻訳精度が改善することが示されている。

技術論文の 1 件目は、京都大学の John Richardson 氏らによる、機能語の翻訳誤り訂正の研究であった（発表は共著者の中澤氏）。機械翻訳において前置詞や関係詞といった機能語はしばしば正しく翻訳できないことがあるため、構文木から構文木への翻訳における出力構文木に対する編集操作によって機能語の誤り訂正を行う手法が提案された。

技術論文の 2 件目は、筑波大学の龍梓(Long, Zi)氏らによる、日中パテントファミリからの専門用語対訳知識獲得の研究であった。日本語と中国語の間で対訳用語獲得を行う場合、それぞれを単語分割して対応する用語を探すという手法があるが、本研究では中国語側を文字単位に分割することで、単語分割の誤りや曖昧性によって対訳用語の検出漏れを減らす手法を提案した。

セッション 4: Beyond Patent Translation

最終セッションは本ワークショップの名称変更とも関連し、他の技術文書等の機械翻訳に向けての取り組みをテーマとして、1 件の招待講演と 2 件の技術論文の講演が行われた。

招待講演は、科学技術振興機構(JST)の中澤敏明氏による、中国と日本の間での機械翻訳を通じた科学技術情報交換の実現に向けた取り組みについての紹介であった。JST と京都大学，中国科学技術信息研究所(ISTIC)が共同で進めている日中間の機械翻訳のための言語資源や言語解析エンジン、機械翻訳エンジンの整備・開発を行うプロジェクトの計画と現状についての説明と、2015 年 10 月に開催された第 2 回アジア言語ワークショップ(WAT)の結果紹介が行われた。

技術論文の 1 件目は、愛媛大学の野口敬輔氏による、大量の特許分野対訳データを利用した他分野への分野適応についての研究であった。特許分野では比較的容易に非常に大量の対訳データが得られるため、共変量シフトを用いたデータ重み付けを利用して適用先分野に近い特許対訳データ中の文を選択し、適用先分野の翻訳を改善する手法が提案され、新聞の英日翻訳精度が改善することが示され

た。

技術論文の 2 件目は、綱川隆司先生による、特許文書中の専門用語の Wikification（Wikipedia エントリへの対応付け）についての研究であった。本研究では、日本語の特許文書中の専門用語に対して適切な日本語版 Wikipedia エントリが存在しない場合でも対応する英語版 Wikipedia エントリへの対応付けができるように、日本語の専門用語の英訳候補を統計翻訳のフレーズテーブルから生成して対応する英語版 Wikipedia エントリを探す、という方法が提案され、日本語版のみの場合よりも多くの Wikipedia エントリへの対応が取れることが示された。

本ワークショップでは様々な立場の招待講演者からの講演があり、特許・技術文書翻訳の現状や課題について議論された。Pouliquen 氏の講演で指摘されたように、現状の特許機械翻訳は assimilation（同化: 他国語を母語に翻訳し自らが理解できるようにすること）の目的であれば、いくつかの言語対で十分実用に耐える水準の翻訳が可能になりつつある一方、dissemination（異化: 母語を他国語に翻訳し、他国語話者に理解してもらえるようにすること）の目的であれば読者が特定されないゆえに比較的高水準の翻訳が求められることから、まだ少し先の未来の話であると筆者も認識している。国際出願や詳細な技術調査の観点では高水準の翻訳が求められることは間違いなく、人手での翻訳は欠かせない。昨今の機械翻訳、特に統計翻訳の急速な進化によって、特許のような対訳資源が豊富な分野においては翻訳後編集やコンピュータ補助翻訳等が人手での翻訳の生産性を大きく向上できる可能性が高いことが明らかであり、また科学技術論文も特許との内容の類似性の観点から同じように機械翻訳の有用性が期待できる。そういった意味では本ワークショップの領域は非常に有用なものであり、今後もユーザの視点、翻訳者の視点、システム開発者や研究者の視点などを多角的に捉えるという意味で、本ワークショップは今後も継続して開催すべきものと考える。

平成 27 年度 AAMT/Japio 特許翻訳研究会報告書

発 行 日　平成 28 年 3 月

発　　行　一般財団法人　日本特許情報機構（Japio）
〒135−0016　東京都江東区東陽町 4 丁目1番 7 号
佐藤ダイヤビルディング
TEL：(03) 3615−5511　FAX：(03) 3615−5521

編　　集　アジア太平洋機械翻訳協会（AAMT）

印　　刷　株式会社インターグループ