平成 26 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
シンポジウム報告

平成 27 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）	マイクロソフト研究所アジア　首席研究員／ 東京大学大学院情報理工学系研究科　名誉教授
副委員長	横山　晶一（※2）	山形大学大学院 教授
	梶　　博行	静岡大学大学院 教授
委　　員	江原　暉将（※1）	山梨英和大学 教授
	宮澤　信一郎	秀明大学 教授
	黒橋　禎夫	京都大学大学院 教授
	宇津呂　武仁（※2）	筑波大学 教授
	二宮　崇	愛媛大学大学院 准教授
	越前谷　博（※2）	北海学園大学大学院 教授
	綱川　隆司	静岡大学大学院 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 プラットフォームソリューション事業部 ソフトウェア開発部 参事
	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
	潮田　明	元奈良先端科学技術大学院大学客員 准教授
	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
	隅田　英一郎	独立行政法人 情報通信研究機構 ユニバーサルコミュニケーション　研究所副所長 （多言語翻訳研究室室長兼務）（2014 年 12 月迄）
	今村　賢治	独立行政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 （2015 年 1 月から）
	中澤　敏明（※2）	科学技術振興機構 情報企画部 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
	呉　　先超	バイドゥ株式会社 プロダクト事業部 シニア RD
	磯崎　秀樹（※2）	岡山県立大学 教授
	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 主管研究員

	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　ソフトウェア事業部　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 特許情報研究所 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	野村　佳代子	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 26 年度 AAMT/Japio 特許翻訳研究会・活動履歴

平成 26(2014)年 5 月 9 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 26(2014)年 7 月 11 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 26(2014)年 9 月 5 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 26(2014)年 10 月 24 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 26(2014)年 11 月 28 日
第 3 回特許情報シンポジウム（於キャンパス・イノベーションセンター東京）

平成 26(2014)年 12 月 12 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 27(2015)年 1 月 23 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 27(2015)年 3 月 13 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 27(2015)年 3 月 31 日
『平成 26 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シンポジウム報告』完成

１．はじめに

マイクロソフト研究所アジア　首席研究員
東京大学大学院情報理工学系研究科　名誉教授
AAMT/Japio 特許翻訳研究会委員長

辻井　潤一

機械翻訳の構想は、20 世紀中葉にディジタル計算機が開発されたときとほぼ同時期に提案され、すでに 70 年近い研究開発の歴史を持っている。その間、この技術に対する過度の期待とその反動としての失望とが繰り返されてきたが、過去 10 年間は、この期待と失望の振れ幅が随分と小さくなってきた。Ｗｅｂで見かけた外国語の文章を機械翻訳にかけて意味を理解するという作業が、日常的に行われるようになってきている。英語は何とか原語で読める人も、それ以外のフランス語やドイツ語、中国語、韓国語などのテキストを読むのに機械翻訳を使う。完全に理解できなくとも、なんとかわかる翻訳が機械翻訳で提供される。このように、機械翻訳に日常的に接することで、過度な期待はなくなり、それだけ失望の度合いも少なくなったようだ。機械翻訳は、身の丈にあった形で、ひろく受け入れられてきている。この機械翻訳技術の日常化という傾向は、2020 年の東京オリンピックに向けて、多言語の音声機械翻訳サービスの提供を行うプロジェクトが開始され、さらに加速されていくと考えられる。

このような技術の進歩の影には、研究者、技術者の不断の努力があったことを忘れてはならないだろうし、この努力は、現在も続けられている。現在の機械翻訳技術は、1980 年代の終わりに現れた統計的機械翻訳を基盤にして、過去 20 年間、連続的にその性能を向上させてきた。オリンピックのための多言語音声翻訳も、この枠組みの中で取り組まれている。ただ、機械翻訳を使ったことがある人には、現在の機械翻訳が不完全な技術であることも、また自明であろう。意味不明の翻訳が出力されたり、読み手がなぞ解きのような過程を経て、やっと意味が取れる場合も多い。日本語と英語のように語族が離れた言語間の翻訳では、特許の審査をするのに十分な質の翻訳には程遠いというものであった。審査官が、関連する特許を同定するには機械翻訳が役に立つが、一旦、関連特許が同定されると、それを人間の翻訳家に翻訳してもらう必要がある。日本語と韓国語や、語族の近いヨーロッパ諸言語間の翻訳は別として、まだまだ混み入った情報内容を正確に伝えるような翻訳にはなっていない。

いまの機械翻訳の枠組みを大きく革新することで、語族が異なる言語間の翻訳の質を向上させることは、息の長い、次世代の機械翻訳技術の研究となる。また、それ自体では不完全な現在の機械翻訳をより使いやすいものにすることは、機械翻訳の研究者だけでは出来ない。AAMT/Japio 特許翻訳研究会は、このような両面からの技術開発を促進するために、（１）機械翻訳システムの開発に従事する技術者だけでなく、（２）機械翻訳の原理的な研究を行っている大学や研究機関の

研究者、また、（３）実際の特許の翻訳の工程を管理する機関の運営者、（４）翻訳業務にかかわる翻訳家など、背景の異なる人々に議論を深める場を提供している。また、公開の国際ワークショップやシンポジウムを企画することで、研究会の枠を超えて、特許翻訳の機械化に従事する人たちに連携の場を提供してきた。

本報告書は、本研究会の活動の成果を一般に公開するためのものである。本報告書が、知財の国際化に伴い、ますますその重要性を増している特許の多言語翻訳システムの開発、運用、利用に興味を持つ人たちの交流をさらに強めることに貢献できることを願っている。

2. 語彙・構文に関する技術

2．１　日中韓三言語の漢字情報を用いた訳語獲得

Kyoto University　Lu Yuanmei
Toshiaki Nakazawa
Sadao Kurohashi

2.1.1 Introduction

The quality of the statistical machine translation highly relies on the amount of parallel corpora available, and improving the lexical coverage of the parallel corpora seems to play an important role in reducing the number of out-of-vocabulary (OOV) words. However, the number of vocabularies of languages keeps growing, especially for technical terms. It is impossible to cover all the newly appeared words by augmenting the parallel corpora; therefore we need to prepare bilingual dictionaries for the new words, or translate them separately, for example, using the transliteration technique.

There are some parallel dictionaries available for limited language pairs and limited domains. In addition, we can extract parallel resources from Wikipedia. It offers hyper-linked pages of the same topic in different languages, and the title pairs of the linked pages can be used as a parallel dictionary. However, the coverage is not sufficient for both cases especially for technical terms. Although there may exist enough resources between English and the other language, there is less resources between two non-English languages, such as Korean, Chinese and Japanese.

As in the same linguistic area, Korean, Chinese and Japanese have much in common in their languages. One of the aspects is that these languages use Chinese characters. In Japan, they use Kanji, which is originated from Chinese, and Korean use Sino-Korean vocabularies, in which characters (Hangul) can be converted to corresponding Chinese characters (Hanzi). Even though the forms are different, most of the vocabularies in these three languages have one-to-one correspondence in character. In this paper, we propose a method of translating Korean words into Chinese using the Chinese character knowledge. We use the Hangul-to-Hanzi mapping table to generate translation candidates and rank the candidates considering the possibility of the character combination and contextual similarity.

2.1.2 Related Work

Since Korean characters are phonogram, we can find a corresponding Hangul for a given Hanzi. Actually, almost all of the Hanzi can be converted to one (or scarcely several) Korean character. (Huang,et,al. 2000) constructed a Chinese-Korean Character Transfer Table (CKCT Table) to reflect the correspondence between Hanzi and Hangul. The table contains 436 Hangul with corresponding 6763 Hanzi. The number of daily-used Hanzi in Korea is known as only 1800, and 3500 Hanzi are required to learn for practical Chinese character level test[1]. Obviously, many of the Hanzi in their table cannot be considered as practical ones.

2.1.3 Proposed Method

[1] Korea Foreign Language Evaluation Institute (http://www.pelt.or.kr/cs/10/main/main.aspx)

Figure 1 shows the overview of our Korean-to-Chinese word translation system. In this study, we only focus on the translation of Korean nouns because a large number of technical terms are nouns. Given a Korean sentence, we first apply morphological analyses to extract Korean nouns. Then we look up the Chinese translations of the Korean words in a Korean-Chinese parallel dictionary. The words not included in the parallel dictionary are passed to the next step: generating possible Chinese character combinations as the translation candidates using the Hangul-Hanzi mapping table. The candidates are ranked with the combination score and context similarity score. The combination score represents the possibility of the sequence of the Chinese characters calculated on the large Chinese web corpus. The context similarity score considers the context of the input sentence and that of the sentences in the large Chinese web corpus.

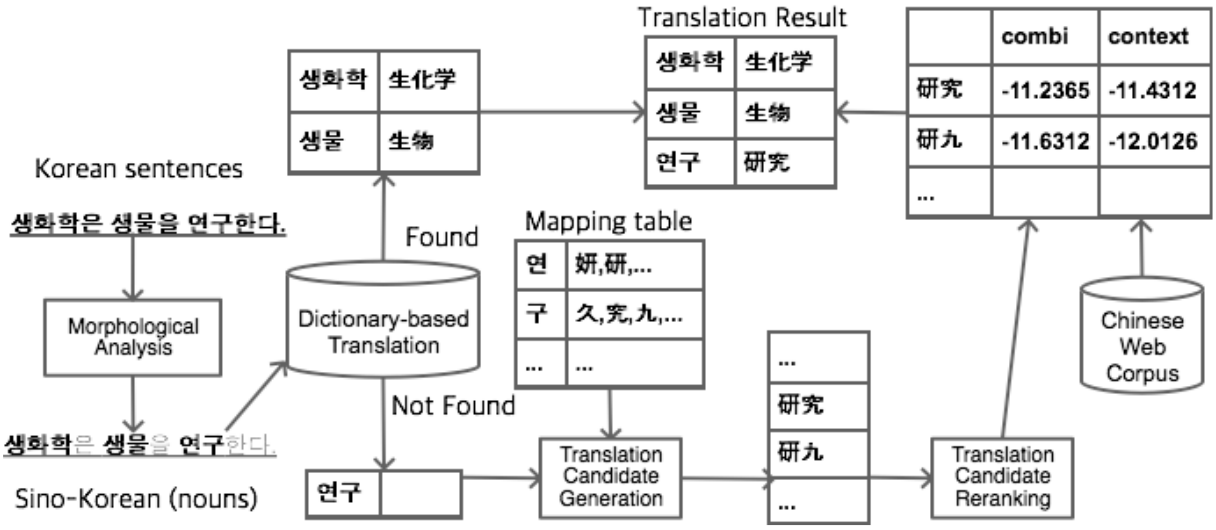

Figure 1: Lexicon Construction System

2.1.3.1 Extracting Nouns

Korean words are separated from each other with spaces, however, each words splitted by a space may contain one or more morphological elements. For example, Korean word "학교에서 is composed of "학교 (noun)" and "에서 (particle)". Since most of the Sino-Korean words are nouns, to obtain Sino-Korean words, we need a Korean morphological analyzer to extract nouns.

For the given Korean sentences that contain Sino-Korean vocabularies, apart from splitting words with spaces, we extracted nouns from these sentences with the help of a Java-library based morphological analyzer[2]. The precision of the analyzer is announced to be higher than 95%.

2.1.3.2 Translation by Dictionary Matching

Some of the Korean nouns are translated into Chinese with a parallel dictionary as the initial step. As is well known, Wikipedia offers a wide range of parallel data for many languages, among them is aligned

[2] KOMORAN ver 2.3 (Java Korean morphological analyzer)

Wikipedia titles. In our method, we use the aligned Wikipedia title pairs of Chinese and Korean as a parallel dictionary. In addition, we apply the following processes to improve the quality and coverage of the parallel dictionary:

- Make full use of redirect pages of each page, and validate the correctness using the first sentence of the definition part to augment the parallel dictionary.
- Convert Chinese characters of traditional Chinese into simplified Chinese.

The Korean nouns that cannot be translated with the parallel dictionary are passed to the next process. In addition, the Korean nouns that have multiple translation candidates (such as homonyms or ambiguous words) are also passed to the next process.

2.1.3.3 Generating Translation Candidates

The aim of this step is to generate possible translation candidates by combining Hanzi characters converted from the Hangul characters using the Hangul-Hanzi mapping table. For instance, using the mapping table in Table 1, we can generate the translation candidates for "한자 (Chinese character, 汉字)": 闲姐 韩姐 汉字 汉子.... Whether these combinations have the proper meaning or not is still unknown. Most of the combined words may have no practical meaning. So we need to select the most appropriate combination.

Table 1: A portion of the mapping table

한	闲	韩	恨	限	汉					
자	姐	字	磁	子	仔	姿	刺	自	资	瓷

2.1.3.4 Rank the Translation Candidates

Now we have a large amount of combinations of Hanzi characters. In order to select the most appropriate ones for each Korean word, we utilize combination score and context similarity score calculated using a large Chinese web corpus.

2.1.3.4.1 Combination Score

Combination score S_{combi} measures the strength of the link between the characters. For example, the combination score for "汉字" is calculated as

$$S_{combi}(\text{汉字}) = \log(P(\text{字}|\text{汉}) \times P(\text{汉}|\text{字})), \text{ where}$$

$$P(\text{字}|\text{汉}) = \frac{c(\text{汉字})}{c(\text{汉})}$$

2.1.3.4.2 Context Similarity Score

For each combination, character-based context vector is constructed using the web corpus. We use sentences, which contain the combination as the context window, and each element of the vector is the

co-occurrence count of Chinese characters. We ignore stop characters such as 的 and 了[3] and characters with less than 100 times of occurrence.

We also construct another context vector of the input Korean sentence using the formerly translated Korean words. The context similarity score $S_{context}$ are calculated as the cosine similarity of the two context vectors.

2.1.3.4.3 Interpolation

The combination score is useful to examine if the combination is appropriate or not, and the context similarity score is useful to select the appropriate one according to the context where two or more combinations have practical meanings. Therefore, we interpolate the two scores and calculate the score of the translation candidate $S(cand)$ as follows: of the translation candidate $S(cand)$ as follows:

$$S(cand) = \alpha S_{combi} + (1 - \alpha) S_{context}$$

The value of α ([0,…1]) is determined with 5-fold cross validation. We divide the into 5 parts and recursively select four of them to get best-performed α and use it and left ones to test the performance of translation. The character combination with the highest score is regarded as the final translation result.

2.1.4 Experiment

2.1.4.1 Settings

For the Hanja-Hanzi mapping table, [Chu, et, al. 2012] have produced a Chinese character mapping table for Japanese (Kanji), Traditional Chinese (TC) and Simplified Chinese (SC). We merged the table with the 3500 Chinese characters of practical use, and checked the compatibility with web-engined Hanja dictionary[4,5], and finally got the Hangul-Hanzi mapping table.

We used Wikipedia title dictionary (6.6M) and Web corpus (sentences, 45G) for querying the frequency of each combination and creating context vectors for them. For experiment data, we prepared 100 Korean sentences with 3281 words for test. After morphological analyzing, 1014 words among them returned analyzing result as nouns (38 of them are not Sino-Korean words). 466 words of them were found data from the Wikipedia. For the left words, we obtained the possible combination using the mapping table. For querying the web corpus, we used the KenLM model, which utilizes a character-based process.

2.1.4.2 Result

We conducted the experiment with 5-fold cross validation, and obtained the best-performed (highest precision) α for each test set, as shown in Table 2.

---

[3] https://code.google.com/p/verymatch/downloads/detail?name=stopwords.txt
[4] http://hanja.naver.com
[5] http://small.dic.daum.net/index.do?dic=hanja

Table 2: The α-Precision relation for each test set

testset	1	2	3	4	5
α	0.71	0.71	0.71	0.71	0.71
Precision(%)	63.16	72.16	69.36	68.82	69.47

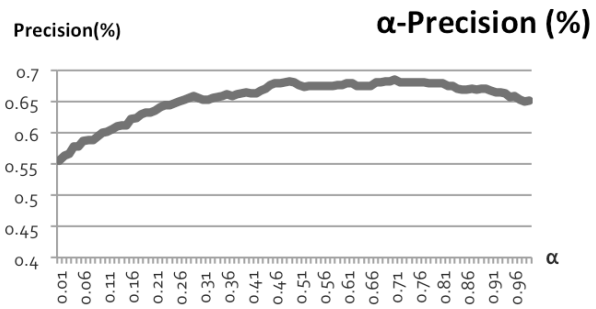

Figure 2: α-precision curve for testset 4

We selected a test set (testset 4) and obtained translation result for each α. Figure 2 shows the specific α-precision curve, and, the result of the translation with α = 0.71 is as shown in Table 3.

Table 3: Experiment result
(*Combination: considering combination score; Context: considering context similarity)

	Dictionary	Combination	Context	Combination + Context
Correct translation	418	208	237	344
Wrong translation	48	302	273	166
No translation	510	0	0	0
Precision (%)	89.70 (418/466)	40.78 (208/510)	46.47 (237/510)	67.45 (344/510)

2.1.4.3 Discussions

Table 4: Good example

Original Korean sentence

전자를 발견할 가능성이 있는 공간 영역을 궤도라고 부른다

Sino-Korean words

전자 电子 발견 发现 가능성 可能性 공간 空间 영역 领域 궤도 轨道

Extracted nouns

전자 발견 공간 영역 궤도

In dictionary

공간 空间 物理

Candidates to be generated combination

전자 발견 영역 궤도

Korean	Candidates	+Combi	+Context	+Combi+Context
영역	荣誉	-10.2171	-2.5632	-7.9975
	领域	-10.2388	-1.8687	-7.8115

Table 4, and 5 shows some good and bad examples of the translation result. In the experiment, if there are more than 2 words that cannot be translated by Wikipedia, we conducted the process of ranking translation candidates in sequence (from beginning to end of the sentence). Thus, for example, for word 영역 in the good example in Table 4, the final result of 전자, 발견, 공간 will be considered as context feature. The embedded tables shows ranking process in detail. Words in blue is the expected translation result, and orange ones indicate selected ones in each condition (+Combi: considering combination score only, +Context: considering context score only, +Combi+Context: considering both score)

Table 5: Bad example

Original Korean sentence

표면 조성은 여러 가지 이온화 방법을 이용해서 연구할 수 있다.

Sino-Korean words

표면(表面) 조성(组成) 이온화(电离) 방법(方法) 이용(利用) 연구(研究)

Extracted nouns

표면,조성,이온화,방법,이용,연구

In dictionary

이온화(电离),연구(研究)

Candidates to be generated combination

표면,조성,방법,이용

Korean	Candidates	+Combi	+Context	+Combi+Context
조성	组成	-11.0066	-2.3882	-8.5073
	造成	-10.4985	-2.7177	-8.2420

2.1.5 Conclusion

We conducted an automatic character-based Korean-to-Chinese translation. The ultimate aim of our system is to construct useful resources in MT among Japanese, Korean and Chinese. A Java library based morphology analyzer was induced to extract nouns as Sino-Korean words. In the translation step, we both considered context vectors and probabilities of each Hanzi combination. We used aligned Wikipedia title dictionary to get reference translations and used them to create context vectors for each Hanzi combination candidate. Some incorrect reference translation caused incorrect translation result. In the future, we will modify the Wikipedia title dictionary and use the model to conduct a Korean-Chinese-Japanese machine translation.

References

[1] Jin-Xia Huang and Key-Sun Choi. Chinese-Korean Word Alignment Based on Linguistic Comparison. In ACL, 2000.

[2] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Chinese characters mapping table of Japanese, Traditional Chinese and Simplied Chinese. In Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012), pages 2149-2152, Istanbul, Turkey, May 2012.

[3] Kenneth Heaeld. KenLM: Faster and Smaller Language Model Queries. In Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, pages 187-197, Edinburgh, Scotland, United Kingdom, July 2011.

２．２　パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定方式と統計的機械翻訳モデルとの間の比較・評価

筑波大学大学院システム情報工学研究科
龍　梓，董　麗娟，
宇津呂　武仁，山本　幹雄

2.2.1 はじめに

特許文書の翻訳は，他国への特許申請や特許文書の言語横断検索などといったサービスにおいて必要不可欠なものである．この特許文書翻訳の過程において，専門用語の対訳辞書は重要な情報源である．本論文では，日米パテントファミリーを情報源として，専門用語対訳辞書を生成する手法を提案する．従来より，日米パテントファミリーの対応特許文書中において，「背景」および「実施例」の部分の日英対訳文対の対応付けを行い，これを情報源として専門用語の対訳辞書を生成する手法が提案されている．例えば，NTCIR-7 特許翻訳タスク[1]において配布された日英 180 万件の対訳特許文を用いて，対訳特許文からの専門用語対訳対獲得を行った研究がある．この研究では，句に基づく統計的機械翻訳モデル[3]を用いることにより，対訳特許文から学習されたフレーズテーブル，要素合成法，Support Vector Machines(SVM)を用いることによって，専門用語対訳対獲得を行った．しかし，上述の方式では，対訳文対が抽出される部分は，「背景」及び「実施例」全体の約 30%であり，約 70%は利用されていなかった．そこで，本論文では，「背景」および「実施例」のうちの残りの 70%の部分を言語資源として，既存の対訳辞書を用いて専門用語の訳語推定を行う方式の有効性を実証する．特に，人手で作成された辞書である英辞郎及びその部分対応対訳辞書に加えて，全体の約 30%を訓練例として学習したフレーズテーブルを併用して要素合成法を適用し，専門用語の訳語推定を行う方式，および，フレーズテーブルのみを用いて要素合成法を適用し，専門用語の訳語推定を行う方式を提案する．また，要素合成法を適用する際には，訳語候補が相手言語側特許文書中に存在するか否かの検証を行うことにより，高精度な訳語推定を実現する．また，このとき，Moses と呼ばれる統計的機械翻訳モデルのツールキット 3) を用いて専門用語の訳語推定を行った場合と提案手法の要素合成法との比較実験を行う．その結果，比較実験において，提案手法の要素合成法により，Moses で訳語推定を行った場合よりも高い適合率・再現率を達成することができた．

2.2.2 日英対訳特許文

本論文では，フレーズテーブルの訓練用データとして，NTCIR-7 の特許翻訳タスク[1]で配布された約 180 万対の日英文対応データを使用した．なお，この文対応データは以下に示す手順で作成されたものである．

1. 1993-2000 年発行の日本公開特許広報全文と米国特許全文を得る．
2. 米国特許の中から日本に出願済みのものを優先権番号より得て，日英対訳特許文書を取得する．

3. 日英対訳特許において日英間で比較的直訳されている関係となっている度合が大きい「背景」及び「実施例」の部分を抽出する.
4. 抽出した部分に対して，文献[8]の手法によって日英間で文対応をつける.

2.2.3 句に基づく統計的機械翻訳モデルのフレーズテーブル

本論文で用いるフレーズテーブルでは，日英の句の組，及び，日英の句が対応する確率を推定し記述する．このとき，前節で述べた文対応データに対して，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]を適用する．Moses によってフレーズテーブルを作成する過程を以下に示す.

1. 単語の数値化，単語のクラスタリング，共起単語表の作成などの処理を文対応データに対する前処理として行う.
2. 文対応データを利用し，最尤な単語対応を英日・日英の両方向において得る.
3. 英日・日英両方向における単語対応を利用し，ヒューリスティックスを用いることにより，対称な単語対応を得る.
4. 対称な単語対応を用いて，可能な全ての日英の句の組を作成する．そして，各組に対して，「文単位の句対応制約」の条件に対する違反の有無をチェックする(違反しない句の組を有効な対応とみなす).
5. 文対応データにおける日英の句の対応の数を集計する．このとき，各句の対応に翻訳確率を付与する.

手順(4) について，以下に「文単位の句対応制約」の条件を示す.

日本語文の形態素列中の形態素を文頭から順に $V_1, V_2, ..., V_n$ 英文の単語列中の単語を文頭から順に $W_1, W_2, ..., W_n$ として，日本語句を $P_J (= V_p \cdots V_{p'})$ とし，英語句を $P_E (= W_q \cdots W_{q'})$ とする．ここで, 日英句の組 $\langle P_J, P_E \rangle$ が含まれるある一つの対訳文対 $\langle T_J, T_E \rangle$ 中において得られているあらゆる単語対応 $\langle V_i, W_j \rangle$ について，「$p \le i \le p' \Leftrightarrow q \le j \le q'$」が成り立つ場合に，$P_J$ と P_E は対訳文対 $\langle T_J, T_E \rangle$ において「文単位の句対応制約」に違反しない，と定義する.

2.2.4 要素合成法による訳語推定

2.2.4.1 既存の対訳辞書およびフレーズテーブル

本論文においては，既存の対訳辞書として，「英辞郎」[12] に加えて，英辞郎の訳語対から作成した部分対応対訳辞書[7]，及び，前節で述べたフレーズテーブルを用いる．両者における見出し語数および訳語対数を【表 1】に示す．部分対応対訳辞書生成の手順は以下のとおりである．まず，

1 http://www.eijiro.jp/

2 本論文では，英辞郎 Ver.79 及び Ver.131 を用いる.

既存の対訳辞書から，日本語及び英語の用語がそれぞれ2つの構成要素(具体的には，日本語の場合はJUMAN[3]による形態素解析によって得られる形態素列，英語の場合は単語列) からなる訳語対を抽出し，これを別の対訳辞書P_2 とする．次に，P_2 中の訳語対の第一構成要素から前方一致部分対応対訳辞書B_P を作成し，第二構成要素から後方一致部分対応対訳辞書B_S を作成する．本論文においては，英辞郎についてはVer.131を使用し，前方一致部分対応対訳辞書及び後方一致部分対応対訳辞書については，Ver.79 及びVer.131 を統合したものを用いた．

2.2.4.2 訳語候補のスコア

訳語候補のスコアを$Q(y_S, y_T)$とする．このとき，y_S は日本語専門用語を，y_Tは生成された訳語候補を表し，y_S は構成要素 $s_1,s_2,...,s_n$ に，y_Tは構成要素 $t_1,t_2,...,t_n$ に分解できると仮定する．すると，$Q(y_S,y_T)$は，対訳辞書スコア$\prod_{i=1}^{n} q(\langle s_i,t_i\rangle)$とコーパススコア$Q_{corpus}(y_T)$の積で定義される．実際には，ある訳語候補が２つ以上の系列の訳語対から生成される場合があるので，本論文では，以下に示すように，それぞれの系列のスコアの和によって$Q(y_S,y_T)$を定義する．

$$Q(y_S,y_T) = \sum_{y_S=s_1,s_2,...,s_n} \prod_{i=1}^{n} q(\langle s_i,t_i\rangle) \cdot Q_{corpus}(y_T)$$

このとき，対訳辞書スコアはこの構成要素同士のスコアの積によって求まり，コーパススコアは訳語候補が目的言語側のコーパスに出現するか否かによって求まる．

2.2.4.2.1 対訳辞書スコア

構成要素の訳語対$\langle s,t\rangle$の対訳辞書スコア$q(\langle s,t\rangle)$は，訳語対が英辞郎，前方一致部分対応対訳辞書 B_P，または，後方一致部分対応対訳辞書 B_Sに含まれる場合のスコアq_{man}及び訳語対がフレーズテーブルに含まれる場合のスコアq_{smt}の和によって求まる．

$$q(\langle s,t\rangle) = q_{man}(\langle s,t\rangle) + q_{smt}(\langle s,t\rangle)$$

$$q_{man}(\langle s,t\rangle) = \begin{cases} 1 & (\langle s,t\rangle\text{が英辞郎，}B_P\text{，または，}B_S\text{に含まれる場合}) \\ 0 & (\text{それ以外の場合}) \end{cases}$$

$$q_{smt}(\langle s,t\rangle) = \begin{cases} P(t|s) & (\langle s,t\rangle\text{がフレーズテーブルに含まれ、かつ、}P(t|s)\geq p_0\text{である場合}) \\ 0 & (\text{それ以外の場合}) \end{cases}$$

上記の定義においては，訳語対$\langle s,t\rangle$ が英辞郎，B_P，または，B_Sに含まれる場合，$q_{man}(\langle s,t\rangle)$は 1

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

となり，それ以外の場合は 0 となる．一方，$\langle s,t \rangle$ がフレーズテーブルに含まれる場合は，翻訳確率の下限 p_0 のパラメータに従い，スコアを決定する．このパラメータ p_0 は，6 節において，評価用データ以外の調整用データを用いて最適化される．

2.2.4.2.2 コーパススコア

コーパススコアは，訳語候補 y_T が目的言語側のコーパスに出現する場合にのみ 1 となり，出現しない場合には 0 となる．

$$Q_{corpus}(y_T) = \begin{cases} 1 & (y_T\text{が目的言語側コーパスに出現する場合}) \\ 0 & (y_T\text{が目的言語側コーパスに出現しない場合}) \end{cases}$$

2.2.4.2.3 例

例として，専門用語“並列態様”の対訳“parallel mode”を獲得する様子を【図 1】に示す．本論文では，まず，この日本語専門用語“並列態様”を構成要素 s_1 の“並列”と s_2 の“態様”に分解し，これらを既存の対訳辞書及びフレーズテーブルを利用して目的言語に翻訳する．そうすると，s_1 からは t_1 として“parallel”, “concurrent”, “multiple” が，s_2 からは t_2 として“aspect”, “mode”, “form” が生成され，さらに各々に訳語の参照元に応じたスコアが付与される．次に，前置詞句の構成を考慮した語順の規則にしたがって，それらの構成要素の訳語を結合し，訳語候補を生成する．このとき，各訳語候補の対訳辞書スコアは t_1 と t_2 の積となる．例えば，“parallel aspect”の対訳辞書スコアは(1.0+0.8)×1.0=1.8 となる．

最後に，これら訳語候補を対訳辞書スコア順に，目的言語側のコーパスに対して照合を行い，もし照合すればそのコーパススコアは 1，照合しなければ 0 となる．この場合，結果的に，訳語候補のスコアが最も高い“parallel mode”が獲得される．

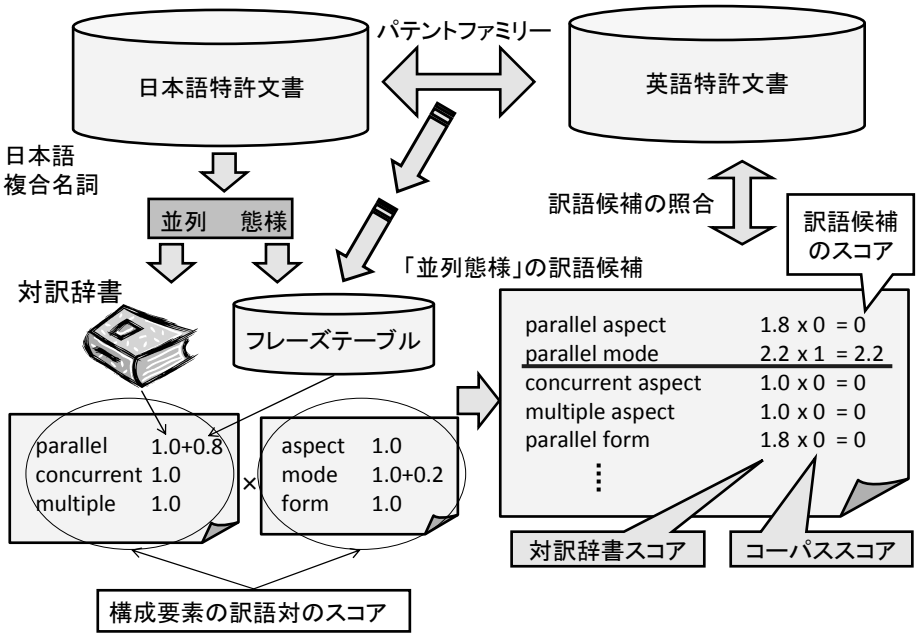

図 1:　要素合成法を用いた日本語専門用語「並列態様」の訳語推定

2.2.5 Moses による訳語推定

本論文では，Moses と呼ばれる統計的機械翻訳モデルのツールキットを用いて専門用語の訳語推定を行った場合と提案手法の要素合成法との比較実験を行う．Moses は翻訳モデル・言語モデル，そしてデコーダーによる翻訳確率の計算処理を組みあわせることで訳語推定を行う手法である．後述の，対訳文対抽出部分 PSD_E，及び，対訳文対非抽出部分 $NPSD_E$ から言語モデルを生成した．また，対訳文対抽出部分 PSD_J ，及び，PSD_E から翻訳モデルを生成した．訳語候補の対数尤度とコーパススコアの積を訳語候補のスコアとした．

2.2.6 対訳文対非抽出部分における訳語推定

本論文で用いる日米パテントファミリーの日本語側 D_J は，「背景」B_J ，「実施例」M_J，および，「背景・実施例以外の部分」 N_J から構成されている．そして，これらの部分のうち，「背景」B_J および「実施例」M_J は，対訳文対抽出部分 PSD_J，及び，対訳文対非抽出部分 $NPSD_J$ に分割される．また，英語側の特許文書の全体 D_E に対しても，同様に，「背景」B_E，「実施例」M_E，および，「背景・実施例以外の部分」N_E から構成され，「背景」B_E および「実施例」M_E は，対訳文対抽出部分 PSD_E，及び，対訳文対非抽出部分 $NPSD_E$ に分割される．この特許文書の構成の例を【図 2 】に示す.

$$D_J = \langle B_J, M_J, N_J \rangle$$
$$B_J \cup M_J = \langle PSD_J, NPSD_J \rangle$$
$$D_E = \langle B_E, M_E, N_E \rangle$$
$$B_E \cup M_E = \langle PSD_E, NPSD_E \rangle$$

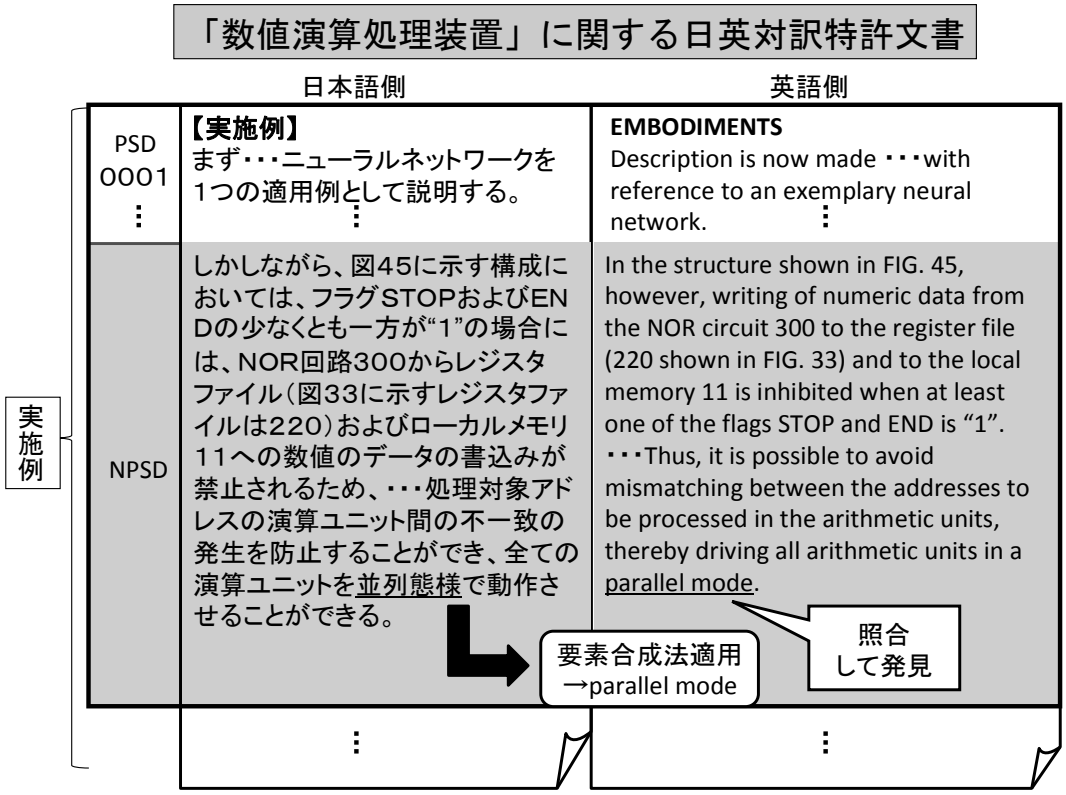

図２：日米パテントファミリー中の「実施例」における対訳文対非抽出部分の例

ここで，本論文では，英訳語推定対象となる日本語専門用語 t_J を抽出するにあたっては，対訳文対抽出部分 PSD_J 中の日本語専門用語の英訳語の多くは対訳文対から学習したフレーズテーブル中に含まれると予測し，「背景」B_J 及び「実施例」M_J における対訳文対非抽出部分 $NPSD_J$ を抽出元とした．

次に，その日本語専門用語 T_J に対して，英語側の「背景」B_J 及び「実施例」M_E を英語側コーパスとみなして要素合成法を適用し，英語訳語候補の集合 $TranCand(t_J, B_E \cup M_E)$ を作成した[4]．

[4] ここで，比較評価として，英語側の「背景」B_E 及び「実施例」M_E における対訳文非抽出部分 $NPSD_E$ のみを英語側コーパス

$$TranCand(t_J, B_E \cup M_E)$$
$$= \{t_E \in B_E \cup M_E \mid t_J\text{に対して要素合成法}\}$$
$$\text{により}t_E\text{を生成し}Q(t_J, t_E) > 0\}$$

そして，この$TranCand(t_J, B_E \cup M_E)$を用いて，以下の関数$CompoTrans_{\max}$によりスコア最大となる訳語候補を得る．

$$CompoTrans_{\max}(t_J, B_E \cup M_E) = \arg\max_{t_E \in TranCand(t_J, B_E \cup M_E)} Q(t_J, t_E)$$

以上の手順により，日英対訳特許文書の英語側の「背景」B_E及び「実施例」M_Eから英語専門用語t_Eを獲得する．

2.2.7 評価

2.2.7.1 評価対象日本語専門用語の選定

提案手法を評価するため，以下の 4 通りの比較を行った．

(i) 「英辞郎のみ」・・・要素合成法における対訳辞書として，英辞郎及びその構成要素から生成される辞書を用いる．

(ii) 「フレーズテーブルのみ」・・・要素合成法における対訳辞書として，フレーズテーブルを用いる．

(iii) 「英辞郎及びフレーズテーブル」・・・要素合成法における対訳辞書として，英辞郎及びフレーズテーブルを用いる．

(iv) 「Moses」・・・訳語推定の手法として Moses を用いる．

はじめに，パテントファミリー1,000 組を取り出し，そこから 61,133 例の日本語複合名詞を抽出した[6]．次に，これら 61,133 例の日本語複合名詞に対して 4 節で述べた要素合成法または Moses を訳語推定手法として適用し，表 2-(1) に示すように，以下の 5 つのカテゴリに分類した．また，このとき，(c) に関しては集合の大きさが最大になるように，要素合成法において用いるフレーズテーブルの翻訳確率の下限値 p_0(式(1))を 0 とした．同様に，Moses による訳語推定の場合は訳語候補のスコアに対して下限値を設けなかった．

(a) 日本語複合名詞が英辞郎に含まれ，その訳語が英語側特許文書中に含まれる．

(b) (a) 以外の場合で，日本語複合名詞がフレーズテーブルの日本語側と完全一致する[7]．

(c) (a)，(b) 以外の場合で，日本語複合名詞に対して訳語推定手法を適用した結果，その訳語が英語側特許文書中に含まれる．

(d) (a)，(b)，(c) 以外の場合で，日本語複合名詞に対する訳語が英辞郎及び，訳語推定手法によって生成されるが，英語側特許文書中に含まれない．

(e) (a)，(b)，(c)，(d) 以外の場合で，日本語複合名詞に対する訳語が英辞郎及び訳語推定手法に

とみなして要素合成法を適用する評価実験も行ったが，英語側コーパス中において適切な訳語候補を照合できる割合が下がったため，本論文においては，英語側の「背景」B_E 及び「実施例」M_E を英語側コーパスとみなして要素合成法を適用する方式を採用した．

より生成されない.

以下では，まず，表 2-(1) の(c) 欄に示すように，上記(i) ~(iv) の 4 通りの手法のうち(i)「英辞郎のみ」および(iv)「Moses」を適用した結果，訳語候補が英語側特許文書に含まれる日本語複合名詞の集合を求める.

• 対訳辞書として上記(i)「英辞郎のみ」を用いた場合について，求められた日本語複合名詞の集合を集合 E (4,003 例) とする.

• 上記(iv)「Moses」を用いた場合について，求められた日本語複合名詞の集合を集合 M とする. 次に，集合 E, $M-(E\cap M)=M-E$ より，E'(202 例)，M'(417 例) をそれぞれ作成する．ここで，集合 E 中の日本語複合名詞については，対訳辞書として上記(i)「英辞郎のみ」を用いることにより，大半の日本語複合名詞の英訳語を正しく推定できることが分かっているため，集合 M に対しては，集合 E に含まれる日本語複合名詞を除外し，集合 $M-(E\cap M)=M-E$ を作成した後，これらを母集団として評価用の日本語複合名詞の集合 M' を作成した．加えて，上記の集合に含まれる日本語複合名詞を以下のように分類した.

(c1) 専門用語であり，かつその適切な訳語が目的言語側に存在する.

(c2) 専門用語であるが，その適切な訳語が目的言語側に存在しない.

(c3) 専門用語以外の，形態素解析における区切り位置が誤っているもの，複合名詞抽出規則の誤り，などに該当する日本語複合名詞.

(c4) 専門用語以外の一般複合名詞．そして，評価用セット M'，E' を上述の分類に基づき，以下のように分割した．集合 M' のうち(c1) に属する日本語複合名詞を抽出した集合を M'_1($|M'_1|$ = 193) とする. また, 集合 M' のうち(c2) に属する日本語複合名詞を抽出した集合を M'_2($|M'_2|=31$) とする．集合 E' のうち(c1) に属する日本語複合名詞を抽出した集合を E'_1($|E'_1|$ = 181) とする．また，集合 E' のうち(c2) に属する日本語複合名詞を抽出した集合を E'_2($|E'_2|=4$) とする.

2.2.7.2 評価結果

要素合成法の場合はフレーズテーブルの翻訳確率に対して下限値を，Moses の場合は訳語候補のスコアに対して下限値を設けた．この節では，これらの下限値を変化させたときの再現率・適合率の推移をグラフにして報告する．まず，集合 E'_1，$E'_1\cup E'_2$，E'_2 を用いて，上記(i)「英辞郎のみ」の場合と上記(iv)「Moses」の場合との比較を行った. その評価結果を順に, 図 3(a), 図 3(b), 図 3(c) に示す. 集合 $E'1$ を用いた場合, 要素合成法は 90.1%の適合率・再現率を達成した. Moses は再現率約 14% ~86%の範囲で，適合率約 80% ~90%の間を推移し，推移している範囲の適合率は要素合成法に迫っているが，再現率では提案手法の要素合成法に及ばなかった．集合 $E'_1\cup E'_2$ を用いた場合，要素合成法は適合率 88.1%・再現率 90.1%を達成した．Moses は再現率約 14% ~86%の範囲で，適合率約 80% ~90%の間を推移し，推移している範囲の適合率は要素合成法に迫っているが，再現率では提案手法の要素合成法に及ばなかった．集合 E' を用いた場合，要素合成法は適合率 80.7%・再現率 90.1%を達成した．Moses は再現率約 10% ~86%の範囲で，適合率約 70% ~81%の間を推移し，推移している範囲の適合率は要素合成法に迫っているが，再現率では提案手法の要素合成法に及ばなかった．いずれの場合においても，要素合成法の再現率に Moses

は及ばなかった．つぎに，集合M'_1，$M'_1 \cup M'_2$，M'を用いて，上記(ii)「フレーズテーブルのみ」，および，上記(iii) 「英辞郎及びフレーズテーブル」の場合と(iv)「Moses」の場合との比較を行った．その評価結果を順に，図 4(a)，図 4(b)，図 4(c) に示す．集合M'_1を用いた場合，要素合成法(dic=P) は再現率 10% ～37%の範囲において，適合率 90%以上を達成した．要素合成法(dic=EP) は再現率 23% ～40%の範囲において, 適合率 90%以上を達成した. Moses は再現率 15% ～75%の範囲において，適合率 70% ～80%の間を推移した．再現率では Moses の方が高いが，適合率を比較すると，再現率 30% ～34%の範囲では要素合成法(dic=P) は適合率 95.2%，要素合成法(dic=EP)は適合率 92.9%，Moses は適合率 76.3%となり，要素合成法の方が高い適合率を示した．集合$M'_1 \cup M'_2$を用いた場合，要素合成法(dic=P) は再現率 9% ～44%の範囲において，適合率 80%以上を達成した．要素合成法(dic=EP)は再現率 23% ～43%の範囲において，適合率 80%以上を達成した．Moses は再現率 17% ～75%の範囲で，適合率 60% ～70%の間を推移した．再現率では Moses の方が高いが, 適合率を比較すると, 再現率 29% ～34%の範囲では要素合成法(dic=P) は適合率 89.4%，要素合成法(dic=EP) は適合率 86.7%，Moses は適合率 64.0%となり，要素合成法の方が高い適合率を示した．集合M'を用いた場合，要素合成法(dic=P) は再現率 27% ～37%の範囲において，適合率 50%以上を達成した．要素合成法(dic=EP) は再現率 23% ～48%の範囲において，適合率 50%以上を達成した．Moses は再現率 19% ～75%の範囲で，適合率 30% ～35%の間を推移した．再現率では Moses の方が高いが，適合率を比較すると，再現率 29% ～40%の範囲において要素合成法(dic=P)は適合率 52.7%，要素合成法(dic=EP) は適合率 55.8%，Moses は適合率 33.3%となり，要素合成法の方が高い適合率を示した．以上より，人手で作成された辞書である英辞郎を用いない場合においても，要素合成法の方が高精度で専門用語の訳語推定を行えることを示せた．

2.2.8 関連研究

文献[5][6][9]では，パテントファミリーから抽出された対訳特許文を用いて，訳語対の獲得を行っている．しかし，本論文では，対訳特許文が抽出されなかった残りの部分を利用して新たな専門用語訳語対の獲得を行っている点が異なる．また，文献[4] では，複数の対訳特許文において，ある日本語専門用語に対して複数の訳語が出現するという状況を考えて，同義対訳専門用語の同定と収集を行っている．上記の手法と本論文の手法を併用することは比較的容易であると考えられる．一方，提案手法と比較して，コンパラブルコーパスからの訳語対獲得手法(例えば，文献[2])においては，通常，文脈ベクトル等の類似性を言語間で測定した情報を手がかりとする点が特徴である．これに対して，本論文の手法において訳語推定の情報源として用いているパテントファミリーは，一般のコンパラブルコーパスと比較すると，対訳となっている部分の割合がかなり高い点にその特徴がある．本論文の手法においては，この利点を生かして要素合成法を適用することにより，比較的容易に訳語対の獲得を実現している．

2.2.9 おわりに

本論文では，パテントファミリーから専門用語の対訳辞書を生成する方法について述べた．

NTCIR-7 特許翻訳タスクにおいて配布された対訳特許文対を訓練例として学習したフレーズテーブル，および，既存の対訳辞書に訳語対が登録されていない日英専門用語を対象として，人手で作成された辞書である英辞郎及びその部分対応対訳辞書とフレーズテーブルを併用した要素合成法およびフレーズテーブルのみを用いた要素合成法を適用した．その結果，比較実験において，提案手法の要素合成法により，Moses で訳語推定を行った場合よりも高い適合率・再現率を達成することができた．今後の課題として，日本語専門用語獲得の際に，機械学習などを用いて，専門用語以外の語を効率よく除外していくことがあげられる．

表 1: 英辞郎における見出し語数及び訳語対数

辞書	見出し語数		訳語対数
	英語	日本語	
英辞郎	1,631,099	1,847,945	2,244,117
前方一致部分対応対訳辞書	47,554	41,810	129,420
後方一致部分対応対訳辞書	24,696	23,025	82,087
フレーズテーブル （日本語側エントリと完全一致する語の除外に使用）	33,845,218	33,130,728	76,118,632
フレーズテーブル(推定精度の評価に使用)	17,466,471	17,121,115	36,058,684

表 2: 対象日本語複合名詞の集合=全日本語複合名詞 61,133 個の場合

訳語推定手法		要素合成法			Moses
		英辞郎のみ	フレーズテーブルのみ	英辞郎及びフレーズテーブル	
(a)	英辞郎の英訳が英語側特許文書中に含まれる	5,450(8.9%)			
(b)	フレーズテーブルの日本語側と完全一致	32,516(53.2%)			
(c)	推定された訳語が英語側特許文書中に含まれる	4,003(6.5%)（集合 E）	14,500(23.7%)（集合 P）	14,846(24.3%)（集合 EP）	9,800(16.03%)（集合 M）
(d)	英辞郎または訳語推定手法により，英訳語候補生成可能であるが英語側特許文書中には含まれない	397(0.7%)	948(1.6%)	944(1.6%)	13,338(21.82%)
(e)	英辞郎または訳語推定手法により生成不能	18,767(30.7%)	7719(12.6%)	7,327(12.0%)	29(0.06%)
合計		61,133(100%)			

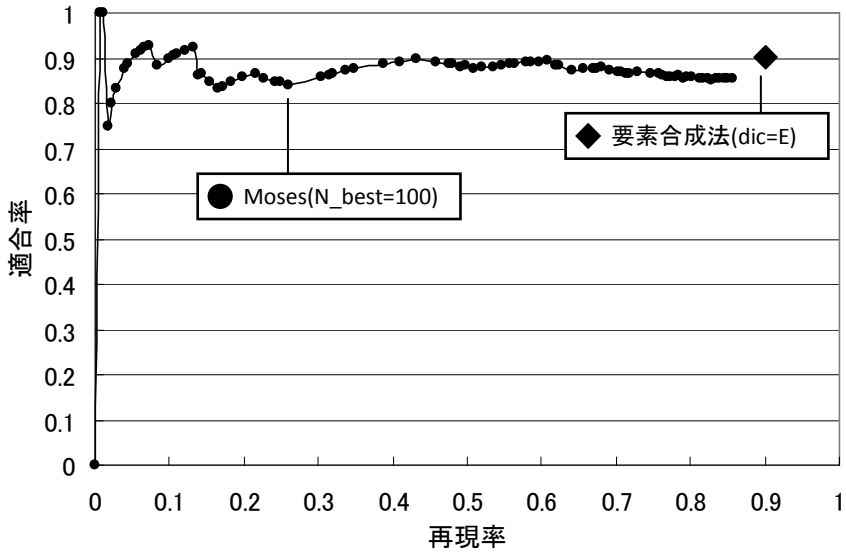

図 3: 集合 E_1' における評価結果

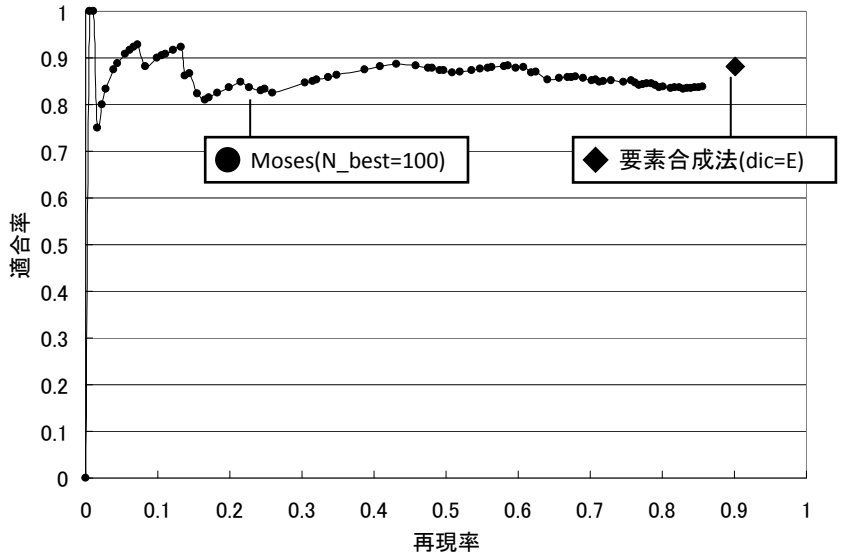

図 4: 集合 $E_1' \cup E_2'$ における評価結果

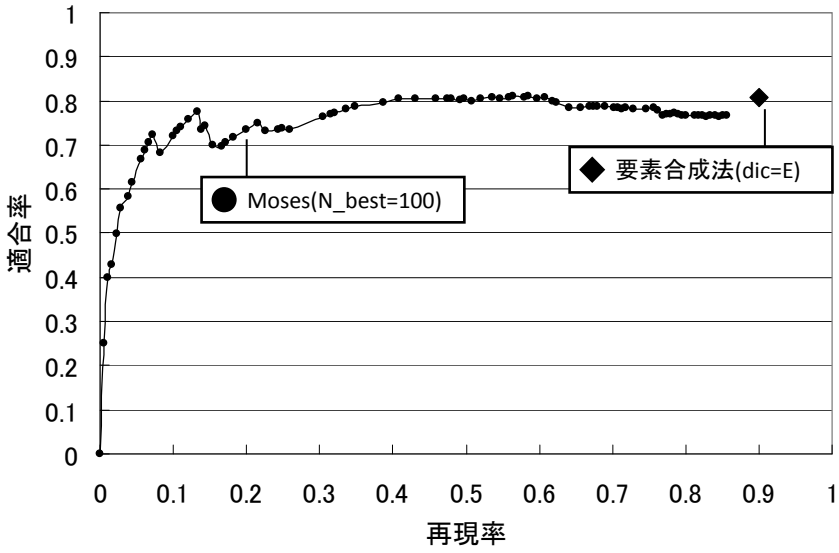

図 5: 集合 E' における評価結果

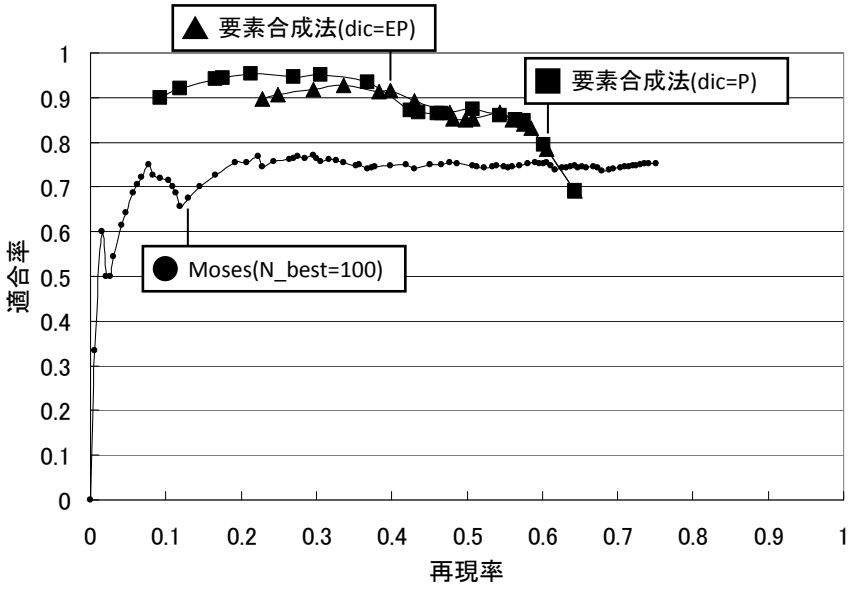

図 6: 集合 M_1'における評価結果

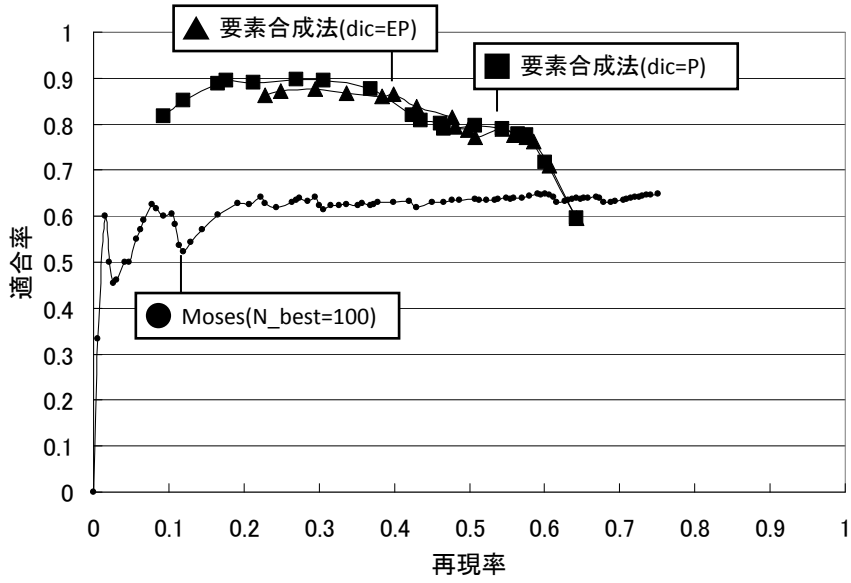

図 7: 集合 $M_1' \cup M_2'$における評価結果

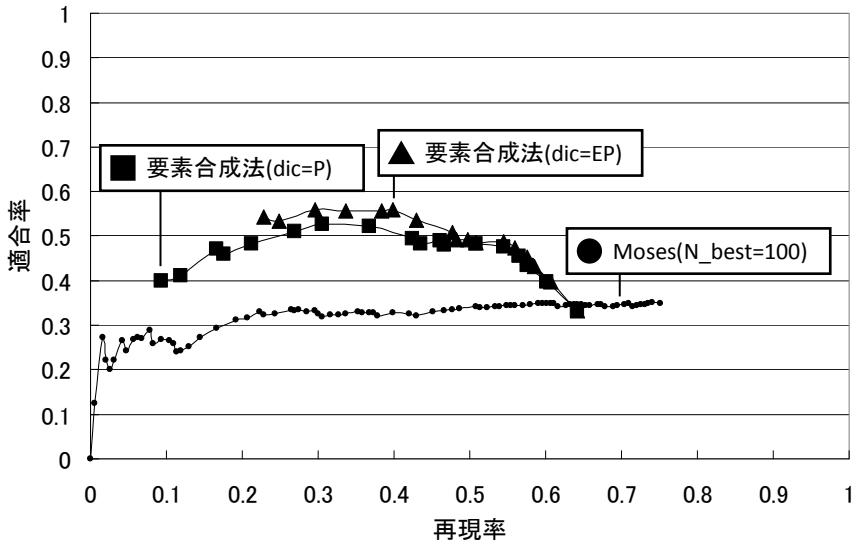

図 8: 集合 M'における評価結果

参考文献

[1] A. Fujii, M. Utiyama, M. Yamamoto, andt. Utsuro. Overview of the Patent translation task attheNTCIR-7 Workshop. In *Proc. 7th NTCIR WorkshopMeeting*, pp. 389-400, 2008.

[2] P. Fung and L. Y. Yee. An IR approachfor translating new words from nonparallel comparable texts. In *Proc. 17th COLING and 36th ACL*, pp. 414–420, 1998.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open sourcetoolkit for statisticalmachinetranslation. In *Proc. 45th ACL, CompanionVolume*, pp. 177-180, 2007.

[4] 梁冰, 宇津呂武仁, 山本幹雄. 対訳特許文を用いた同義対訳専門用語の同定と収集. 言語処理学会第 17 回年次大会論文集, pp. 963-966, 2011.

[5] B. Lu and B. K. Tsou. Towards bilingual term extraction in comparable patents. In Proc. 23rd PACLIC, pp. 755–762, 2009.

[6] 森下洋平, 梁冰, 宇津呂武仁, 山本幹雄. フレーズテーブルおよび既存対訳辞書を用いた専門用語の訳語推定. 電子情報通信学会論文誌, Vol. J93-D, No. 11, pp. 2525-2537, 2010.

[7] 外池昌嗣, 宇津呂武仁, 佐藤理史. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. 自然言語処理,Vol. 14, No. 2, pp. 33–68, 2007.

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

[9] K. Yasuda and E. Sumita. Building a bilingual dictionary from a Japanese-Chinese patent corpus.

In *Computational Linguistics and Intelligent Text Processing*, Vol.7817 of *LNCS*, pp. 276–284. Springer, 2013.

２．３　特許文書から Wikipedia 記事へのリンク自動付与

静岡大学　綱川 隆司
梶 博行

2.3.1 はじめに

テキストに出現する語句の中から重要な語句を選び出し、その語句について説明した記事へのハイパーリンクを付与することを “wikification” と呼び、研究課題として注目を集めている (Roth et al., 2014)。Wikification を自動化することにより、任意のテキストから容易に Wikipedia 記事にアクセスすることが可能になる。Wikipedia は科学技術分野に関する記事も充実してきており、特許文書の wikification により、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。

Wikification 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のステップからなる。いずれも、Wikipedia 記事に存在する他の記事へのハイパーリンク（内部リンク）を訓練データとして利用することで、コストをかけずに教師付き学習によってアンカーテキストの分類器およびリンク先記事決定器を構成できる。これは Wikipedia 記事やそれに近いテキストに対しては有効であるが、特許文書へ適用すると、Wikipedia 記事と異なる性質のテキストであることから精度の低下が予想される。

本稿では、Wikipedia 記事の訓練データを用いて学習した wikification のリンク先記事決定器を特許文書の wikification に適用したときに生じる具体的な問題を示し、特許文書の wikification の実現のための可能なアプローチを探る。

2.3.2 決定リストによる Wikification

本稿では、特許文書に適用する wikification モデルとして、Wikipedia 記事の内部リンクを訓練データとする決定リストによる方法 (袁ら, 2015) を用いる。決定リストは、リンク元となるアンカーテキスト毎に生成し、共起アンカーテキストをリンク先決定のための手掛かりに用いる。以下、決定リストの学習と、決定リストによるリンク先記事の決定について述べる。

(1)　決定リストの学習

アンカーテキスト a の共起アンカーテキスト $a_i\ (i = 1, 2, \dots)$ は、Wikipedia の同一記事中で a と共起した頻度が閾値 θ 以上のアンカーテキストとする。また、アンカーテキスト a のリンク先記事候補 $D_j\ (j = 1, 2, \dots)$ は、Wikipedia 記事中で a のリンク先となったことのある記事とする。共起アンカーテキスト a_i と、アンカーテキスト a から記事 D_j への内部リンク $l(a, D_j)$ の関連度として、下記の t スコア $\mathrm{T}\left(a_i, l(a, D_j)\right)$ を採用する[1]。

[1] 袁ら (2015) による結果から、Wikipedia 記事に対しては、最も良い結果が得られた対数尤度比を関連度に採用したが、対数尤度比の計算は煩雑であるため、特許文書に対しては、容易に計算でき、かつ対数尤度比に近い精度が得られた t スコアを採用する。

表 1　アンカーテキスト “Jaguar” の決定リスト

共起アンカーテキスト	リンク先記事	関連度（t スコア）	順位
Porsche	Jaguar Cars	14.97	1
Ford	Jaguar Cars	14.21	2
Ferrari	Jaguar Cars	13.43	3
BMW	Jaguar Cars	12.10	4
Chevrolet	Jaguar Cars	10.50	5
Mercedes-Benz	Jaguar Cars	10.44	6
Formula One	Jaguar Racing	9.97	7
Chrysler	Jaguar Cars	9.78	8
Toyota	Jaguar Cars	9.66	9
Aston Martin	Jaguar Cars	9.49	10

$$\mathrm{T}\left(a_i, l(a, D_j)\right) = \frac{m - n_1 n_2 / N}{\sqrt{m}},$$

ここに、m はアンカーテキスト a_i と内部リンク $l(a, D_j)$ の共起頻度、n_1 は a_i の出現頻度、n_2 は $l(a, D_j)$ の出現頻度、N はすべてのアンカーテキストと内部リンクの組合せの共起頻度の総和とする。

各共起アンカーテキスト a_i について、関連度が最大のリンク先を選択するルール

IF （a_iがaと共起）THEN （aを$\hat{D}(a, a_i)$にリンクする）

を生成する。ここに、$\hat{D}(a, a_i) = \mathrm{argmax}_{D_j} \mathrm{T}\left(a_i, l(a, D_j)\right)$ とする。各ルールを関連度 $\mathrm{T}\left(a_i, l\left(a, \hat{D}(a, a_i)\right)\right)$ の降順に並べ、その順位を各ルールに付与する。

決定リストの末尾に、すべての共起アンカーテキストが現れなかった場合に適用するデフォルトルールを追加する。デフォルトルールのリンク先記事は、アンカーテキスト a が指すリンク先記事候補のうち、Wikipedia において最も多くリンクされたもの（最頻リンク先）とする。

表 1 に、アンカーテキスト “Jaguar” に対する決定リストの例を示す。

(2)　決定リストによるリンク先記事の決定

入力テキストの中で、Wikipedia において少なくとも 1 回アンカーテキストとして用いられた語句が出現する部分をすべて特定し、これらを入力テキスト中に現れるアンカーテキストとして、それぞれのリンク先記事を決定する。

各アンカーテキストについて、それ以外のアンカーテキストを共起アンカーテキストと考え、決定リストの上位から順に、共起アンカーテキストが入力テキストに現れるルール（以下、ヒットしたルールと呼ぶ）を k 個とり、 それらの中のリンク先記事から多数決で出力リンク先記事を決定する。Wikipedia 記事を用いた交差検定による評価では、$k = 3$ のときにリンク先記事決定の正解率が最も高かった(袁ら, 2015)。図 1 に、リンク先記事決定の例を示す。$k = 3$ の場合は、最も多いリンク先記事 “Jaguar Racing” を出力する。

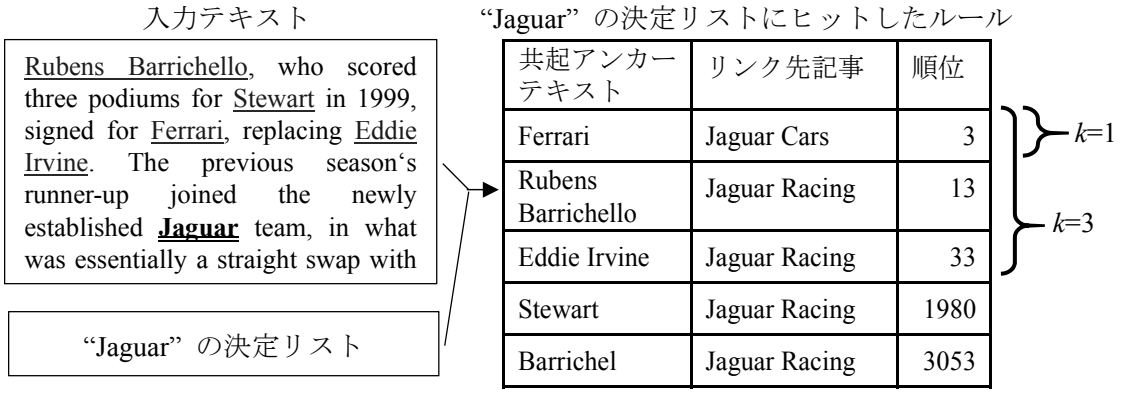

図 1　決定リストによるリンク先記事決定の例

2.3.3　実験

2.3.3.1　実験設定

Wikipedia 記事から生成した決定リストを用いて、特許文書および Wikipedia 記事のそれぞれを入力テキストとする wikification を実施し、その差異を確かめる実験を行った。決定リストを生成するため、英語版 Wikipedia の 2014 年 2 月 3 日付けのダンプデータを用いた。特許文書を入力テキストとする場合、訓練データとしてダンプデータ全体を用いた。また、Wikipedia 記事を入力とする場合、ダンプデータの 5 分の 4 を訓練データとし、残りを入力テキストとした。それぞれの訓練データについて、出現頻度 100 以上、かつ、曖昧性のある（リンク先記事候補が 2 つ以上ある）アンカーテキストを 10000 語選定し、決定リストを生成した。決定リスト生成のためのパラメータとして、$\theta = 2$、$k = 3$ を採用した。

Wikification の入力とする特許文書は、米国特許庁の 2014 年 9 月 4 日付けの 1 週間分の特許出願書類全文データ 5039 件[2]から、各出願書類の abstract, description, claims の範囲のテキスト部分を利用した。

Wikipedia 記事および特許文書のテキストは、句読点等の削除のみを行い、大文字・小文字の正規化や形態素解析器は適用しなかった。

2.3.3.2　特許文書におけるアンカーテキストの出現数

リンク先記事を決定する手掛かりは、共起するアンカーテキスト数が多いほど増える。決定リストにおいては上位 k 個のルールをリンク先記事決定に用いるが、共起アンカーテキスト数が少なければ、上位に入る有力な手掛かりも少なくなると思われる。そこで、Wikipedia 記事と特許文書のそれぞれについて利用できる共起アンカーテキスト数を調べた。

表 2 に、テストデータにおける Wikipedia 記事 10000 件および特許出願書類 5039 件について、1 件当たりのアンカーテキスト出現数と単語数（延べ数(Token)と種類数(Type)）、およびそれらの比率を示した。特許文書は Wikipedia 記事に比べて 1 件の長さが 18 倍程度あるため、1 件当たりのアンカーテキスト出現数も多くなる。決定リストによる wikification では、同一記事・文書内のアンカーテキストはすべて共起するものとして扱うため、特許文書では Wikipedia 記事に比べ、

[2] http://patents.reedtech.com/downloads/ApplicationFullText/2014/ipa140904.zip

表 2　Wikipedia 記事・特許文書中のアンカーテキスト出現数および単語数

単語カウント方法	入力テキスト	1 件当たりのアンカーテキスト出現数	1 件当たりの単語数	アンカーテキスト出現数／単語数
Token-based	Wikipedia 記事	361.7	681.5	0.5307
	特許文書	6582.2	12330.9	0.5338
Type-based	Wikipedia 記事	170.9	233.5	0.7319
	特許文書	1014.5	1195.8	0.8484

表 3　ヒットしたルールの順位・相対順位の平均

	入力テキスト	最初にヒット	2 番目にヒット	3 番目にヒット
順位の平均	Wikipedia 記事	262.9	607.2	1058.8
	特許文書	87.0	204.3	336.7
相対順位の平均	Wikipedia 記事	6.10%	13.44%	22.58%
	特許文書	7.81%	17.33%	27.75%

共起アンカーテキストを 6 倍程度多く利用できることになる。

長い特許文書では、wikification 対象のアンカーテキストとその共起アンカーテキストが離れて出現することがあり得る。しかし、特許文書は一つの発明に関する説明であり、Wikipedia 記事と同様、内容に一貫性があると考えられる。また、アンカーテキスト出現数の延べ数と種類数から、特許文書では同一のアンカーテキストを何度も用いる傾向があることが明らかであり、このことも内容の一貫性を示唆している。これらのことから、特許文書においても共起関係の距離の影響は低いと仮定し、Wikipedia 記事による共起関係の距離を考えない決定リストを特許文書に適用している。

2.3.3.3　ヒットしたルールの順位

アンカーテキストに対するリンク先決定のための決定リストが有効に働くためには、入力テキスト中の共起アンカーテキストが決定リストに現れる必要がある。このとき、より上位のルールの方がアンカーテキストとの関連度が高く、手掛かりとしての確信度が高いと考えられる。そこで、決定リストにおけるリンク先決定過程において、最初にヒットしたルール、2 番目にヒットしたルールおよび 3 番目にヒットしたルールのそれぞれの順位の平均について、入力テキストが Wikipedia 記事と特許文書の場合で比較した。また、ヒットしたルールの決定リスト内での相対的な位置を調べるため、順位を決定リスト内のルール数で割った相対順位の平均も求めた。表 3 はその結果を示す。

絶対的な順位では特許文書の方が Wikipedia 記事より順位が高いが、これは特許文書で wikification の対象となるアンカーテキストの決定リストのルール数が少ないことから生じる。相対順位を比較すると、特許文書の方がやや下がっているが、大きな差はみられない。これらのことから、特許文書における共起アンカーテキストを用いた場合でも、リンク先決定に用いる上位 3

個のルールは Wikipedia 記事の場合と同程度に有効であることが示唆された。

2.3.3.4 Wikificationの結果

表 4 と表 5 に、Wikipedia 記事および特許出願書類に対する wikification の結果の抜粋をそれぞれ示した。各アンカーテキストに対し出力されたリンク先記事について、人手による正誤の判定結果と、誤っている場合に適切なリンク先記事を記した。Wikipedia 記事に対しては、記事中で内部リンクになっている各アンカーテキストのみに対してリンク先記事を出力したのに対し、特許文書に対しては、テキストに出現する全てのアンカーテキストをリンク先記事出力の対象とした。このため、後者には一般の名詞・形容詞に対する実行結果が含まれる。

特許文書に対する wikification においてまず課題とみられるのは、“default” や “walk” といった一般語の扱いである。このような語は、Wikipedia 記事においてもしばしばアンカーテキストとして出現し、Wikipedia において説明が必要な事柄を表す語として出現する。例えばアンカーテキスト “default” に対しては、 “Default (finance)” （債務不履行）などのリンク先記事候補がある。しかし、一般語としての“初期設定”といった辞書的意味に対応する記事は必ずしも存在せず、その場合にはアンカーテキストとして抽出したことが不適切となる。そのため、特許文書においてはアンカーテキストの選択のためのモデルがより重要になると考えられる。

Wikipedia 記事に頻出する固有表現は特許文書にはあまり現れず、抽出されたアンカーテキストは専門用語と一般的な内容語がほとんどである。専門用語については、分野によって意味が異なるアンカーテキストに対してはリンク先の決定が比較的うまく行われている。例えば、 アンカーテキスト“condenser” は “Condenser (heat transfer)”（復水器）の他に “Condenser (optics)”（集光レンズ）等の他のリンク先候補記事を持つが、正しくリンク先が決定されている。一方で、 “acrylic” のように同じ分野で異なるリンク先記事候補があるときのリンク先決定は難しい。

特許出願書類中には、“PCT”（特許協力条約）のような特許に関する語が直接現れることはあまりなく、このために、特許に関係するリンク先記事が選択されないという結果がみられた。特許文書に頻出する表現については、Wikipedia 記事から得られた決定リストを用いずにリンク先を予め人手で定めておく方法も考えられる。

2.3.4 関連研究

Wikipedia 記事を用いて学習した wikification モデルを他の文書に適応する研究としては、マイクロブログ (Cassidy et al., 2012) や 文化遺産データ (Fernando and Stevenson, 2012) への適応が挙げられる。He et al. (2011) は、放射線医学の報告書に対して wikification を適応させている。放射線医学の報告書に現れる医学用語の中には、複数の修飾語と並列関係から、一つの語句に複数の概念が表されていることがあり、アンカーテキスト抽出を難しくしている。例えば、ある語句 “acute cerebtal and cerebellar infarction” には “acute cerebral infarction” （急性脳梗塞）と “cerebellar infarction” （小脳梗塞）の二つの意味を表すのに用いられている。He らはアンカーテキスト抽出のための系列ラベル付けモデルを提案し、その性能を改善した。

2.3.5 おわりに

本稿では、Wikipedia 記事から学習した wikification のための決定リストの特許文書への適用可能性とその問題点について検討した。決定リストの適用時に手掛かりとなる共起アンカーテキストは特許文書においても十分な数があり、Wikipedia 記事から学習した決定リストは特許文書にも適用可能である。特許文書の wikification の精度を向上させるためには、特許文書中の語句を専門用語とそれ以外に分類し、wikification の対象とするアンカーテキストを専門用語に限るといった、アンカーテキストの抽出を適切に行うことが特に必要と考えられる。

今後の課題として、特許文書の訓練データとしての利用が挙げられる。特許文書にはアンカーテキストもそれらのリンク先記事も付いていないが、リンク先記事の初期値を適当に設定し、決定リストの特許文書への適合とリンク先記事の決定を反復させる方法が考えられる。また、リンク先記事決定のための大域的手掛かり (Milne and Witten, 2008) を反復学習に加える方法も有望と考えられる。

参考文献

Cassidy, T, Ji, H., Ratinov, L., Zubiaga, A., and Huang, H. 2012. “Analysis and enhancement of wikification for microblogs with context expansion,” In *Proc. of COLING 2012*, pages 441-456.

Fernando, S. and Stevenson, M. 2012. “Adapting wikification to cultural heritage,” In *Proc. of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 101-106.

He J., de Rijke, M., Sevenster, M., van Ommering, R., and Qian, Y. 2011. “Generating links to background knowledge: a case study using narrative radiology reports,” In *Proc. of CIKM 2011*, pages 1867-1876.

Milne, D. and Witten, I.H. 2008. “Learning to link with Wikipedia,” In *Proc. of CIKM 2008*, pages 509-518.

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.

袁楊，綱川隆司，梶博行．2015．“決定リストの機械学習による wikification,” 言語処理学会第 21 回年次大会発表論文集.

表 4　特許文書に対する wikification の結果

入力特許出願書類タイトル	アンカーテキスト	出力リンク先記事タイトル	判定	適切なリンク先記事	判定に用いた共起アンカーテキスト
Exotect（スポーツ用品）	disabilities	Disability	○		association, trauma, assembly
	options	Option (aircraft purchasing)	×	なし（一般名詞）	n/a
	elasticity	Elasticity (physics)	○		motion, elastic, strength
	span	Span (architecture)	○		flap, tension, tracks
	default	Default (finance)	×	なし（一般名詞）	risk, Time, options
Novel Oxidation Dye Precursors（酸化染料）	May 29	Historical anniversaries/May 29	×	なし（単なる日付）	German, oil, Aqua
	PCT	Won/lost percentage	×	Patent Cooperation Treaty	n/a
	acrylic	Acrylic paint	×	Acrylic acid	German, organic, synthetic, phase, oil
	substance	Substance theory	×	Chemical substance	German, substances, object
	MHz	Hertz	○		commercial, alternative, dark
	mol	Mole (unit)	○		IV, water, moles
	violet	Violet (color)	○		German, Violet, orange
Aerosol Separator Assembly; Components; and, Methods（フィルター）	aperture	Aperture (mollusc)	×	なし（一般名詞）	n/a
	upstream	Upstream (petroleum industry)	×	なし（一般形容詞・副詞）	n/a
	PCT	Won/lost percentage	×	Patent Cooperation Treaty	n/a
	20 mm	Oerlikon 20 mm cannon	×	なし（単なる長さ）	General, 5 in, 30 mm
	acrylic	Poly(methyl methacrylate)	×	Acrylic resin / Acrylic acid / Acrylic fiber	phase, oil, porous, Glass
Steam Power Cycle System（蒸気機関）	condenser	Condenser (heat transfer)	○		Ranking cycle, condenses, Ammonia
	upstream	Upstream (petroleum industry)	×	なし（一般形容詞・副詞）	power, sea, production
	substance	Substance theory	×	Chemical substance	necessary, properties, Mass
	walk	Base on balls ※野球の四球	×	なし（句動詞の一部）	loss, single, second
	GWS	Overtime (ice hockey)	×	なし（変数名）	ヒットルールなし
	Pump	Pump	○		production, Pressure, Turbine
In-situ Detection and Analysis of Methane …（メタン検出）	variables	Variable (mathematics)	×	なし（一般名詞）	set, variable, function
	substance	Substance theory	×	Chemical substance	substances, object, quality
	Raman spectroscopy	Raman spectroscopy	○		emission, Raman scattering, UV
	streamlined	Streamliner	×	なし（一般動詞）	n/a
	upstream	Upstream (petroleum industry)	×	なし（一般形容詞・副詞）	production, bandwidth, Gas
	span	Span (architecture)	×	なし（一般動詞）	feet, beam, tension
Router and Rapid Response Network（ルータ）	VPN	Virtual private network	○		Internet, IPsec, Ethernet
	MHz	Hertz	○		Wi-Fi, GHz, FCC
	IP address	IP address	○		Internet, IPv4, IPv6
	IPv6	IPv6	○		Internet, IP address, IPv4
	sight	Sight (device)	×	Line-of-sight propagation / Non-line-of-sight propagation	safety, receiver, scope
	variables	Variable (mathematics)	○		function, set, functions

表 5　Wikipedia 記事に対する wikification の結果

入力記事タイトル	アンカーテキスト	出力リンク先記事タイトル	判定	適切なリンク先記事	判定に用いた共起アンカーテキスト
Paint	acrylic	Acrylic paint	×	Acrylic resin	primer, USA, ISO, ASTM, Lacquer
Alpena Light	acrylic	Poly(methyl methacrylate)	○		Light, Lake Huron, Michigan
	U.S. Coast Guard	United States Coast Guard	○		Lake Huron, Michigan, built
Brunei	Seria	Seria	○		Jawi, Borneo, Sarawak
	Unitary	Unitary state	○		Borneo, Southeast Asia, Asia
Culture of Yorkshire	batter	Batter (cooking)	○		Yorkshire, England, dish
	luxury	Luxury goods	○		n/a
	White Rose	White Rose	×	White Rose of York	Sheffield, London, Barnsley
XBMC4Xbox	DLL	Dynamic-link library	○		fork, 10-foot user interface, APIs
	plugins	Plug-in (computing)	○		free and open source, audio, applications
9th Youth in Film Awards	Babes in Toyland	Babes in Toyland (band)	×	Babes in Toyland (1986 film)	The Believers, Stars, King
	Over the Top	Over the Top (film)	○		River Phoenix, The Mosquito Coast, The Lost Boys
John Mozeliak	Sports Illustrated	Sports Illustrated	○		January 18, General Manager, GM
	batting	Batting average	○		free agency, runs batted in, MVP
Lex Hilliard	Touchdown	Touchdown	○		fullback, National Football League, drafted
	fullback	Fullback (American and Canadian football)	○		National Football League, Dolphines, Draft
Alberta, Canada	Wild West	American frontier	○		Alberta, British Columbia, Calgary
	Fort Saskatchewan	Fort Saskatchewan	○		Alberta, Canada, province
	mass graves	2011 San Fernando massacre	×	Bone bed	Ukraine, Mexico, United States
2009 Challenge Cup	Craven Park	Craven Park, Hull	×	Craven Park, Barrow-in-Furness	Cup, League, BARLA
	Catalans Dragons	Catalans Dragons	○		Cup, rugby league, England
	Warrington Wolves	Warrington Wolves	○		Cup, rugby league, England
Wu Shaocheng	Chang'an	Chang'an	○		Wu Shaocheng, Zhangyi, Zhumadian
	Yangzhou	Yangzhou	○		Wu Shaocheng, Puyang, Zhumadian
	Emperor Xuanzong	Emperor Xuanzong of Tang	○		Wu Shaocheng, Puyang, Zhumadian

２．４　接続詞と主辞に着目した特許文の並列構造解析

山形大学　髙橋 尚矢
横山 晶一

2.4.1　はじめに

近年、国際的な特許の共有化に伴い国際特許の申請数も増加し続けている。特許文の検索や翻訳などの作業には多くの人手が必要であり、そのため作業を自動化または半自動化することが求められている。これらを解決するためには特許文中に含まれる情報を的確に抽出することが要求される。したがって、特許文に対する正確な係り受け解析が不可欠である。

特許文の課題や解決手段の部分は、200 文字を超える長大な一文になることが多い。しかも単語同士の係り受けが複雑であり、意味が明確でないことがある。これは特許文独特の記述や専門用語の多さなどが原因である。そのため通常文に比べると係り受けが曖昧になりやすく、解析の誤りが発生しやすいという特徴がある。

本稿では、特許文中にあらわれる助詞の「に」に着目する。助詞の「に」には格助詞的用法と並列助詞（並立助詞という表現の方が正しいが、今回は並列構造を捉えるので並列助詞と記す[1]）的用法があるが、特許文において並列的な用いられ方をすることは少ないと思われる。しかし KNP[2]では「に」を並列的用法だと誤解析する例が見受けられたため、原因の解明と修正する方法を検討した。

本稿は、主として[3]に基づき、その内容に加筆したものである。

2.4.2　関連研究

松山ら[4]は特許文ではなく法令文書を対象にした並列構造解析をおこなっている。あらかじめ法令文書の重要語を決め、それらをもとに並列構造を検出した。この研究は並列構造を pfi(前方句、i は前方句の数だけ増える) ,key(並列となるキーワード) ,pb(後方句)と定義し、この型に当てはまる部分を検出するというものである。具体例を挙げると、

“四百八十から保険料納付済期間の月数(pf3) 、保険料四分の一免除期間の月数(pf2) 、保険料半額免除期間の月数(pf1) 及び(key) 保険料四分の三免除期間の月数(pb) を合算した月数を控除して得た月数を限度とする”

のように法令文では 3 つ以上の句が並列関係にある場合は 2 つ以上の前方句を読点で並べた後，並列キー，後方句が続くという特徴がある。

しかし、階層的並列構造は検出が難しい、前方句と後方句の長さが違いすぎると並列構

造の検出に失敗しがちである、など改良の余地がある。また法令文書では並列を意味する語が細かい規則で設定されている（岩本ら[5]）のに対し、特許文ではそのような規則が無いため並列階層の特定が困難である。

我々は、これまでに特許文の並列構造解析について様々な手法で解析を試みてきた[6][7]。ただし主辞に限定した解析のため、特許文全体を捉えた場合の精度がさほど上がらなかった。今回は主辞をベースにしつつ、助詞（格助詞・並列助詞）の観点から精度向上を目指した。

2.4.3　助詞「に」の解析および考察

助詞・助動詞の辞典[8]によれば、並列助詞の「に」とは“「AにB」の形で両者を対比したり両者の取り合わせを問題にするようなセットとなることを表す”とある。これはペアであることが重要視される文節間にあらわれるが、これらの表現は文語的であり特許文（説明文）中に登場するとは考えにくい。

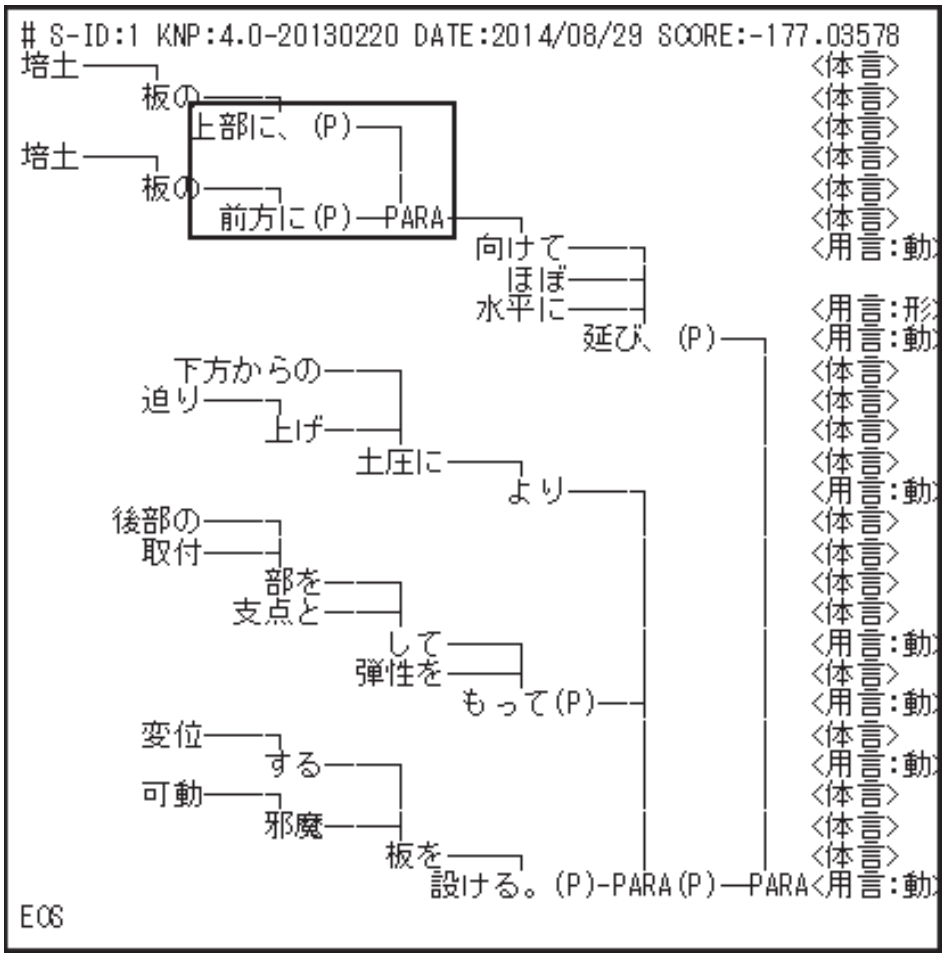

図 1. 助詞「に」を誤解析した例

図 1 に示したのは助詞の「に」が KNP で誤って出力した例である。元の文は、

培土板の上部に、培土板の前方に向けてほぼ水平に延び、下方からの迫り上げ土圧により後部の取付部を支点として弾性をもって変位する可動邪魔板を設ける。”

とあり、『培土板の上部に』が副詞句に、『培土板の前方に向けてほぼ水平に延び、下方からの迫り上げ土圧により後部の取付部を支点として弾性をもって変位する可動邪魔板を設ける』が主文になるはずである。しかし図 1 の太線で囲った部分を見ると、『培土板の上部』と『培土板の前方』が並列構造を持つ、と誤解析していることがわかる。この誤解析の原因は『上部に』の「に」を並列助詞として解釈しているからだと考えられる。

今回はひとつの特許文中に 2 つ以上の助動詞「に」が出現する場合に着目した。特許文は長文になる場合が多いため 1 文中に複数の助動詞「に」があらわれることは珍しくないが、それが入れ子構造内にあらわれた場合、並列助詞と誤判定することがある。特許文では部品の名称など同じ単語が重複して使用されるケースが多いという特徴があり、今回の例では助詞「に」の直前に“培土板”が共通してあらわれているため並列助詞と判定したと予測される。“培土板”を適当な単語に置き換えた場合は図 2 のように正しい判定をしている。

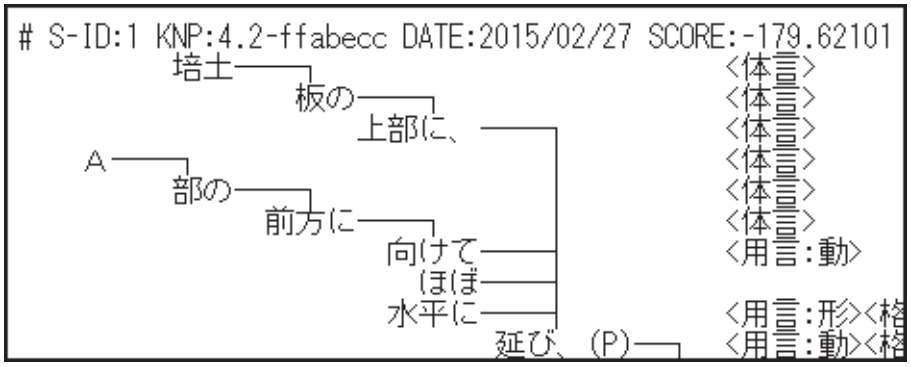

図 2. 助詞「に」が格助詞として判定されている例

2.4.4 問題点と今後の課題

構文解析誤りに関して助詞「に」のみ対象に実験をおこなったが、原因がそれ以外のものが見受けられた。図 3 に示したのは並列構造を誤判定した例である。元の文は、

容器本体の取付機構に対する搬送部品の取付作業の作業性を向上させ、取付機構から搬送部品が外れ易くなるおそれを排除し、取付機構や搬送部品に洗浄水等の液体が残存するのを抑制できる基板収納容器を提供する。”

とあり、『容器本体の取付機構に対する搬送部品の取付作業の作業性を向上させ』と『取付機構から搬送部品が外れ易くなるおそれを排除し、取付機構や搬送部品に洗浄水等の液体が残存するのを抑制できる』が並列関係になる。KNP の解析結果では助詞「に」はいずれも正しく解析できたが、一方で向上・排除・残存の 3 つの動詞句が並列扱いと誤解析している。上記のように助詞「に」以外による解析誤りが多数あった。

今回の調査で、構文解析システムは長文の解析が不完全であることが改めて判明した。今後は KNP で特許文を解析する際に助詞「に」を正しく解析できるような手法を引き続き検討する。それに加え並列助詞「と」と主辞の関係性を探る。KNP で並列修飾句を的確に扱えるような手法を検討し、特許文における修飾構造を正しく解析するシステムを作成する予定である。また文法や文の構造といったアプローチを中心に調査してきたが、汎用性に欠けるため意味的アプローチが必要である。さきほどの図 3 の例においても、意味解析を加えれば正しい解析を可能にすることが期待できる。

これまで KNP を使用し実験をおこなってきたが、特許文解析において KNP では限界がある。今回は助詞にのみ着目したが、長大な文のために構文解析が困難な特許文が非常に多く見受けられたので他の手法も検討していく予定である。長文に対してより頑健な解析器として南瓜[9]や茶筌[10]などを利用することも考えている。

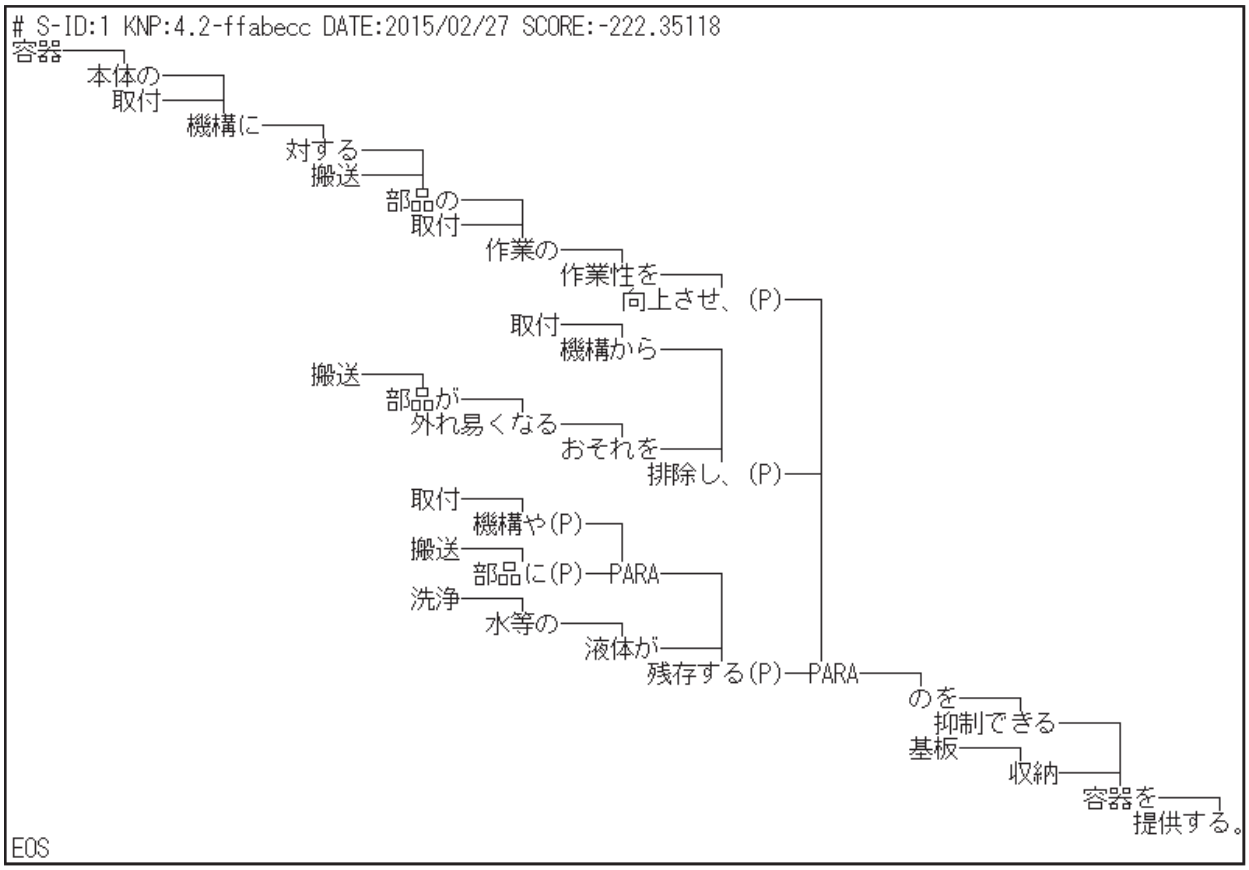

図 3. 並列構造を誤判定した例

参考文献

[1]日本語文法学会　編：日本語文法辞典、大修館書店（2014）
[2]日本語構文・格解析システム KNP：（http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP）
[3]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 Yearbook(2014) pp.242-245
[4]松山宏樹、白井清昭、島津明：法令文書を対象にした並列構造解析、
言語処理学会 第 18 回年次大会(2012)
[5]岩本秀明、長野馨、永井秀利、中村貞吾、野村浩郷：法律文における並列構造の特徴とそれに基づく制限言語モデルについて、情報処理学会自然言語処理研究会(1993)
[6]横山晶一：接尾辞に着目した特許文の並列構造解析、Japio Yearbook(2012) pp.250-253
[7]髙橋尚矢、横山晶一：特許文における入れ子構造の調査、Japio Yearbook(2013)
pp.266-270
[8]森田良行　著：助詞・助動詞の辞典、東京堂出版(2007)
[9]南瓜：（https://code.google.com/p/cabocha/）
[10]形態素解析システム茶筌：（http://chasen.naist.jp/hiki/ChaSen/）

3. 機械翻訳および翻訳支援技術

3.1　統計機械翻訳のためのリサンプリングを用いたドメイン適応の調査

愛媛大学　田中 飛馬
愛媛大学　二宮 崇

3.1.1　はじめに

近年、機械翻訳の分野では統計的機械翻訳（SMT）の研究が盛んに行われている。SMT は、ある分野に存する大量の対訳データから翻訳モデルを学習する手法である。しかし、ある分野（例えば新聞記事分野）で学習された SMT モデルは異なる分野（例えば化学、電子などの分野）で利用する時に、性能が大きく低下するという問題がある。元々翻訳の対象としていた分野（新聞記事分野）のデータ分布と新しい翻訳対象の分野（化学分野や電子分野）のデータ分布が違うことが原因である。分野によって異なるデータ分布の問題を解決する技術は一般にドメイン適応（domain adaptation）と呼ばれ、機械学習から機械翻訳を含む自然言語処理の諸タスクにおいて広く盛んに研究されている(Daumé & Marcu 2006; Daumé 2007; Koehn & Schroeder 2007)。本研究では機械翻訳のためのドメイン適応について研究を行う。ドメイン適応において、元々解析の対象としていた分野は「ソースドメイン（source domain)」と呼ばれ、新しい解析対象の分野は「ターゲットドメイン（target domain)」と呼ばれる。

本稿は、特許の対訳文書と新聞の対訳文書に対しドメイン適応を用いることにより、特許ドメインでの翻訳精度と新聞ドメインでの翻訳精度の向上について調査を行う。一般に、ドメイン適応の問題は、ソースドメインに大量の学習データが存在し、ターゲットドメインには少量の学習データしか存在しないことを仮定し、ソースドメインのモデルやデータをターゲットドメインに適応することが目標となっている。本研究では、ターゲットドメインが少量であることを仮定せず、ソースドメインのデータをターゲットドメインに付加することでターゲットドメインでの性能向上を図ることを目的とする。従って、新聞ドメインの対訳データを特許ドメインに適応した場合の翻訳性能と、特許ドメインの対訳データを新聞ドメインに適応した場合の翻訳性能の双方向のドメイン適応を実験において評価する。本研究は、ソースドメイン対訳データに対してリサンプリングによる文量調節を行い、ターゲットドメインの機械翻訳の翻訳精度を評価する。リサンプリングの量を適切に与えることで、学習データ増加による精度向上とソースドメインのデータ追加による副作用を適切に制御する。

ドメイン適応の多くの手法は、学習後に得られるモデルを適応するか、モデルの適応を学習と同時に行う手法となっているが（Daumé & Marcu 2006)、本研究はデータを適応することによりドメイン適応を実現する手法（Daumé 2007; Koehn & Schroeder 2007）を採用する。モデルを適応する従来手法では、モデルを変更する必要があるため、個々のタスクに応じて学習すべきモデルを実装レベルで変更する必要があり、実現には多大な開発コストを要する。提案手法は、データの分布を変更するだけでドメイン適応を実現できるため、非常に少ない開発コストで実現でき、特に、ドメイン適応の対象となるタスクのためのツールをブラックボックスとして扱うことができるため、機械翻訳にも容易にドメイン適応を実現することができる。

Procedure $Resampling\left(\left(f^{(i)}, e^{(i)}\right)_{i=1}^{n}, m\right)$:

$D' := \phi$（ただし、D'は重複を許す多重集合）

for $i \in \{1, \cdots, m\}$:

$r :=$ 0〜$n-1$までの乱数

$D' := D' \cup \left(f^{(r)}, e^{(r)}\right)$

return D'

図 3.1 リサンプリングのアルゴリズム

本稿の構成は以下のようになっている。3.1.2 節では、本研究の提案手法であるリサンプリングによるドメイン適応手法について説明する。3.1.3 節では実験の結果について述べる。3.1.4 節で本稿の主旨をまとめ、今後の課題について述べる。

3.1.2 リサンプリングを用いたドメイン適応

本研究では、二つのドメインデータを使って機械翻訳学習を行う際、リサンプリング法を用いてドメイン適応を行い精度を向上させる手法を提案する。

3.1.2.1 リサンプリング

統計学において、標本データを反復し、統計量の母集団への近似を図る手法をリサンプリングと呼ぶ。標本データを増加させるときはオーバーサンプリングと呼び、減少させるときはアンダーサンプリングと呼ぶ (He & Garcia 2009)。統計的機械翻訳では、対訳集による学習において特定部分のデータの学習を反復させることで、その特定部分の翻訳により特化した統計モデルを生成することが出来る。標本のリサンプリングを用いた不均衡データからの学習が研究されており (He & Garcia 2009; Weiss et al. 2007)、リサンプリングを用いることによる機械学習の高精度化が実現されている。

手法としてはまず、学習用データに対し文章全体から無作為抽出を行うためにランダムサンプリングを行う。図 3.1 は、文書ファイルからm文のサンプリングデータをリサンプリングにより取得するアルゴリズムを示している。ただし、$\left(f^{(i)}, e^{(i)}\right)_{i=1}^{n}$は対訳データ、$n$はデータサイズ、$m$はサンプリング回数を表す。100 万文対の学習用データを作る場合、ランダムサンプリングデータの先頭から 100 万行目までを抽出することで、学習用リサンプリングデータを生成する。

3.1.2.2 データ結合によるドメイン適応

解析対象としない分野のデータや学習モデルを解析対象の分野に適応させる技術は一般にドメイン適応（domain adaptation）と呼ばれ、機械学習から自然言語処理の諸タスクにおいて広く盛んに研究されている。ドメイン適応において、解析対象の分野は「ターゲットドメイン（target domain)」と呼ばれ、解析対象としない適応元の分野は「ソースドメイン(source domain)」と呼

Procedure $ResamplingDomainAdaptation(D_S, D_T, m)$:
$D'_S := Resampling(D_S, m)$ (ただし、D'_Sは重複を許す多重集合)
$D'_T := D_T \cup D'_S$ (ただし、D'_Tは重複を許す多重集合)
return D'_T

図 3.2 リサンプリングを用いたデータ結合によるドメイン適応アルゴリズム

ばれる。ドメイン適応は、ソースドメインのデータや学習モデルをターゲットドメインに適応させることで、ターゲットドメインでの性能を向上させることを目的とする。本研究においては、ソースドメインの対訳データとターゲットドメインの対訳データを取得し、ソースドメインの対訳データをリサンプリングにより新しく取得し、ターゲットドメインの対訳データと結合することにより、ドメイン適応を実現する。リサンプリングの量を調節することにより、ソースドメインとターゲットドメインとの間の文量の比を調節することができる。図 3.2 は対訳データのためのリサンプリングによるドメイン適応のアルゴリズムを示している。ただし、D_Sはソースドメインの対訳データ、D_Tはターゲットドメインの対訳データ、mはリサンプリングの回数を表す。mはリサンプリングにより得られるソースドメイン対訳データの文数と等しくなる。得られた結合対訳データを用いて、統計機械翻訳の学習を行うことでドメイン適応された統計機械翻訳のモデルが得られる。

3.1.3 実験

特許ドメインの全ての対訳データを用いた学習を特許ドメインのベースラインとし、新聞ドメインの全ての対訳データを用いた学習を新聞ドメインのベースラインとする。これに対し、提案手法では、ターゲットドメインの対訳データとリサンプリングにより文量を調節したソースドメインの対訳データを結合することで新しい学習用対訳データを生成し、統計機械翻訳で学習する。

3.1.3.1 実験設定

特許ドメインには NTCIR10 PatentMT タスクの日英対訳データ、新聞ドメインには読売新聞電子版の日英対訳データ JENAAD (Utiyama and Isahara 2003) を用いた。学習用データは、特許ドメインが 3,187,626 文対、新聞ドメインが 178,500 文対から成り、チューニングデータは各ドメインそれぞれ 500 文対から成り、テストデータは特許ドメイン 899 文対と新聞ドメイン 922 文対から成る。統計機械翻訳ツールは Moses を用いた (Koehn et al. 2007)。翻訳精度は BLEU で評価した。

リサンプリングの量は様々な量を実験で試しており、ターゲットドメインが特許ドメイン、ソースドメインが新聞ドメインの場合には、特許ドメイン 300 万文対に対し、新聞ドメインはリサンプリングにより 1 万文、 2 万文、4 万文、8 万文、16 万文に減らした場合(アンダーサンプリング)、18 万文全てを用いた場合(フル)、リサンプリングにより 50 万文、100 万文、200 万文、

表 3.1　日英方向の翻訳精度

実験 A: 日英特許翻訳

	特許 (万文)	新聞 (万文)	BLEU (%)
ベースライン	300	0	29.82
アンダーサンプリング		1	29.69
		2	29.61
		4	29.60
		8	29.68
		16	30.07
フル		18	**30.19**
オーバーサンプリング		50	29.73
		100	29.25
		200	29.19
		300	29.57
		400	29.57

実験 B: 日英新聞翻訳

	特許 (万文)	新聞 (万文)	BLEU (%)
ベースライン	0	18	9.07
アンダーサンプリング	1		8.92
	2		**9.18**
	4		9.04
	8		8.91
	16		8.76
	18		8.67
	50		8.55
	100		8.54
	200		8.51
フル	300		7.96

300 万文、400 万文に増やした場合(オーバーサンプリング)を試した。ターゲットドメインが新聞ドメイン、ソースドメインが特許ドメインの場合には、新聞ドメイン 18 万文に対し、特許ドメインの対訳データをリサンプリングにより、1 万文、2 万文、4 万文、8 万文、16 万文、18 万文、50 万文、100 万文、200 万文に減らした場合(アンダーサンプリング)、300 万文全てを用いた場合(フル)を試した。

3.1.3.2 実験結果

日英方向と英日方向をそれぞれ試し、ターゲットドメインが特許ドメインとなる場合と新聞ドメインとなる場合を試した。従って、次の 4 つの実験を行ったことになる。

実験 A 日英方向の特許ドメインの翻訳
実験 B 日英方向の新聞ドメインの翻訳
実験 C 英日方向の特許ドメインの翻訳
実験 D 英日方向の新聞ドメインの翻訳

表 3.1 は日英方向の機械翻訳(実験 A,B)の実験結果を表わしている。「フル」と書いてある箇所はソースドメインのリサンプリングを行わず、ソースドメインの対訳データを全て追加した場合

表 3.2　英日方向の翻訳精度

実験 C: 英日特許翻訳

	特許 (万文)	新聞 (万文)	BLEU (%)
ベースライン	300	0	32.31
アンダーサンプリング		1	**33.00**
		2	32.47
		4	32.45
		8	32.91
		16	32.15
フル		18	32.71
オーバーサンプリング		50	32.68
		100	32.65
		200	32.72
		300	31.85
		400	32.26

実験 D: 英日新聞翻訳

	特許 (万文)	新聞 (万文)	BLEU (%)
ベースライン	0	18	8.91
アンダーサンプリング	1		**9.13**
	2		8.97
	4		9.02
	8		8.66
	16		8.41
	18		8.33
	50		8.48
	100		7.97
	200		7.77
フル	300		7.77

を表わしている。「アンダーサンプリング」はリサンプリングによりソースドメインのデータ量をフルのデータ量よりも減らした場合を表しており、「オーバーサンプリング」はソースドメインのデータ量をフルのデータ量よりも増やした場合を表わしている。表よりわかるように特許翻訳の場合(実験 A)、新聞ドメインの対訳データを全て 18 万文追加した場合(フル)が最も良い精度を与え、ベースラインに比べ BLEU 値が 0.37%ポイント上昇した。新聞翻訳の場合(実験 B)、特許ドメインの対訳データを 2 万文追加した場合に最も良い精度を与え、ベースラインに比べ、BLEU 値が 0.11%ポイント上昇した。

表 3.2 は英日方向の実験結果(実験 C,D)を表わしている。表よりわかるように、特許翻訳の場合(実験 C)、新聞ドメインの対訳データを 1 万文追加した場合が最も精度が良く、ベースラインに比べ、BLEU 値が 0.69%ポイント上昇した。新聞翻訳の場合(実験 D)、特許ドメインの対訳データを 1 万文追加した場合が最も精度が高く、ベースラインに比べ BLEU 値が 0.22%ポイント上昇した。

3.1.4 まとめ

本研究は、統計的機械翻訳におけるドメイン適応の手法として、ソースドメインのデータにリサンプリング処理を施し、ターゲットドメインのデータと統合してモデル学習を行う手法を提案し、その性能を調査した。提案手法は、データの分布を変更するだけでドメイン適応を実現できるた

め、非常に少ない開発コストで実現でき、特に、ドメイン適応の対象となるタスクのためのツールをブラックボックスとして扱うことができるため、複雑な処理を要する機械翻訳にも容易にドメイン適応を実現することができる。実験には学習用データに特許ドメイン約 300 万文、新聞ドメイン約 18 万文を用意し、提案手法での英日／日英方向、及び特許ドメイン／新聞ドメイン方向の翻訳精度を調査した。いずれの場合にも BLEU 値の向上が確認できたが、ベースラインに比べ大きな性能向上はみられなかった。将来の課題として、より高精度の機械翻訳を実現するリサンプリング手法として、ランダムサンプリング方法の変更、両ドメインデータの同時リサンプリング、密度比推定による重み付けなどの手法が考えられる。

参考文献

Hal Daumé III and Daniel Marcu (2006) Domain adaptation for Statistical Classifiers. Journal of Artificial Intelligence Research, 26:101–126.

Hal Daumé III (2007) Frustratingly Easy Domain Adaptation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), pp. 256-263.

Philipp Koehn and Josh Schroeder (2007) Experiments in Domain Adaptation for Statistical Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation (StatMT'07), pp. 224–227.

Haibo He and Edwardo A. Garcia (2009) Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 9, pp. 1263-1284.

Gary M. Weiss, Kate McCarthy and Bibi Zabar (2007) Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? In Proceedings of the 2007 International Conference on Data Mining (ICDM2007), pp. 35-41.

Masao Utiyama and Hitoshi Isahara (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), pp. 72-79.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007) Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL) on Interactive Poster and Demonstration Sessions, pp. 177–180.

３．２　新方式による英文作成支援システム
―システム構成を中心として―

秀明大学　宮澤信一郎　　山梨英和大学　江原　暉将
秀明大学　松山　宏樹　　創価大学　　岡田　　勇
秀明大学　宮﨑　瑞之　　早稲田大学　Dawn L. Miyazaki

3.2.1　はじめに

英文作成支援システムとは、コンピュータによって人間の英文作成を支援するシステムである。この点、自動的に翻訳を行う機械翻訳とは異なる。英文作成支援システムは、機械翻訳がまだ完全でない現況下において、その重要性は変わらない。特に大量の対訳コーパスを利用できる現代においては英文作成支援システムは新しい段階に入っているといえる。我々は、将来的に英文作成支援システムを機械翻訳システムと相互に補完するものと考えている。

本論では我々が提案している新方式の英文作成支援システム（日英翻訳支援システム）で使用する文末類似文辞書構築のシステム構成を中心に述べる。

3.2.2　既存システム

実用的な英文作成支援システムは大きく二つのタイプに分類できる。一つはテンプレート型であり、英文メール作成支援システムに多い方式である。このようなタイプは、ビジネスレターといった定型化されている分野に適しているが、そのような分野は限られている。

もう一つはコーパス利用型であり、その方式は、日本語のキーワードを入力して、日英対訳コーパスから引いた英語の例文を参考にして、人力で英文を作成するものである。このタイプではコーパスの品質がシステムの性能を左右する。実用的なコーパスサイズは数十万文でも不十分といわれている。このタイプではコーパスのサイズを拡大すると、作成コストよりも品質の向上が見込めなくなるので、おのずとサイズは限界を迎えることになる。コーパスを利用する第二の方式は、キーセンテンスを入力してコーパス中に存在する類似文を検索するものであり、翻訳メモリーとして知られている(1)。 コーパスには単言語コーパスと対訳コーパスの２種類がある。単言語コーパスを用いる方法は、大量のコーパスデータを入手できるため、表現の一般性を調べるのに適している。対訳コーパスを用いる方法は、文の対応付けがされた日本文と英文を比較しながら用法を参照できるので、ユーザが意味を把握しやすいという利点を持つ。

3.2.3　新たなシステム構想の提案

コーパスを利用するタイプで生じる最大の問題は、結局は人手で翻訳するために翻訳者の能力にかなり依存するというところにある。機械翻訳がまだ実用段階にない現在においては、より優

れた英文を作成するためには、どのような例文を参照すればよいのか。このような観点は言語の特性を考慮した専門的な議論が求められる。しかし、一般にコーパス利用型はキーワード一致による参照文がほとんどである。しかし、単語そのものは辞書を用いて適切な単語に翻訳できる場合が多い。我々の関心である日英翻訳に限ると、文型や語順に対する適切な提案がコーパスなどのシステムからなされることが重要なのではないだろうか。

例をあげて説明する。「私はあなたがどこから来たのか知らなかった」という日本文を、単語の辞書のみを用いて語順をそのままに「翻訳」すると" I you where from came know didn't." となってしまう。この場合は、SV(=not+verb.)O(=wh-phrase) という文型の例示が重要で、これを利用することで正しい英文に近づき、構造が正確になり、ネイティブにとっても理解しやすくなる。

我々は本稿において、コーパス利用型のうち、現在主流である「キーワード抽出型」の日英翻訳支援システムに代わり、「文型提示型」の日英翻訳支援システムを提案する。文型提示型のシステムには、文型類似文辞書が必要となる。本論文で提案する翻訳支援システムも一種の翻訳メモリーであるが、キーセンテンスとコーパス中の文との間の類似性の計算方法に独自性がある。そのために、従来の翻訳メモリーでは利用していない係り受け解析を用いている。

3.2.3.1 方法の概要

本システムは以下の３段階で構成される。

(1)日英対訳コーパスを用意しておく。

(2)日本文を入力させる。

(3)日英対訳コーパスを日本文の類似度順に表示する。

この方式であれば、テンプレート型の様に場面を考えたり、キーワード検索型コーパス利用方式のようにキーワードを考えたりする必要がない。また、コーパスを類似度順に得られるので、ユーザの負担が少ない。

1) 日英対訳コーパス

日英対訳コーパスには特許コーパスを使う。特許コーパスには二つの特徴があるからである。第一の特徴は日本文と英文の作成者がそれぞれ科学技術の専門家と翻訳の専門家ということである。これによって良質な文を得られる。第二の特徴はデータが多いということである。これによって多様な文を得られる。

2) 日本文の類似度

日本文の類似度は文末の類似度によって評価する。日本語の言い回しが文末で決まるからである。文末の類似度は形態素解析、文節解析、ならびに係り受け解析と経験式によって評価する。

具体的には以下の通りである。翻訳対象の和文を a、データベースから取り出した文を b と定義する。先ず、a と b を形態素解析、文節解析、ならびに係り受け解析する。これによって、a と b が文節係り受け関係に分解される。次に、文節に対して、受け属性と係り属性（表 1）を自動付与する。属性は、日本語の係り受けの性質を考慮して筆者らが独自に設定したものである。

最後に、語形と属性を使いながら文間の距離（相違度）を経験式によって計算する。経験式では、文間の距離を「文節間の非類似性に基づく距離」と「文節が深くなるほど小さくなる値」の積の総和の最小値で表す。最小値は動的計画法で求める。なお、文節の深さとは、根文節（日本語では文末の文節）からの係り受けの深さである。これによって、深い文節の距離を軽視するような評価が可能となる。すなわち、浅い文節（文末に近い文節）の距離を重視するような評価が可能となる。以上の方法によって、文型の類似性を大枠で捕らえることができる。

表 1　文節属性

受け種別

意味	記号
名詞	N
述語	V
サ変動詞	NV
述語名詞	NV
副詞・連体詞	E
接続詞	NV

係り種別

意味	記号
連用修飾	y
連体修飾	t
連用または連体修飾	ty
係助詞「は」	h
終止	s

3.2.3.2　高速化

入力文を全てのコーパスと比較していては処理が膨大になり実用的な速度を出しにくい。そこで、次の様にする。先ず、コーパス同士を文末の類似度によってクラスタリングしておく。次に、入力文を各クラスタでの代表的な文と比較する。これによって、入力文とクラスタの類似度が分かる。次に、類似度の高いクラスタを選ぶ。次に、入力文を該当クラスタでの全文と比較する。これによって、入力文と全文の類似度が分かる。最後に、該当クラスタの全文を類似度順に並べ替えて表示する。

3.2.3.3　クラスタリング

文間の距離を用いて文型類似文辞書を作成する。作成手順の概要を図 1 に示す。図 1 でクラスタとは、クラスタリングの手法(最遠隣法)で距離の近い文を集めた集合で、これらのクラスタを構築することで文型ごとの代表的な文を抽出し辞書を構築することができる。クラスタは全ての文と文との間の距離から作られる距離行列に基づいて行われるべきであるが、計算コストが大きいと考えられるので、我々は文末表現の一致する文を集めて（文末でクラスタリング)、仮の文集合を作り、この仮の文集合の中で多くの文を含むものだけに対してクラスタリングを行うことで計算時間を現実的なものに抑える方針である。すなわち、ここで作成されたクラスタが文型となる。クラスタごとに距離行列を用いて最も中心となる文を選び、これを当該クラスタの代表的な文として選択する。これを集めることで文型類似文辞書の索引を作ることができる。それぞれの索引にはクラスタ内の文が表示される。

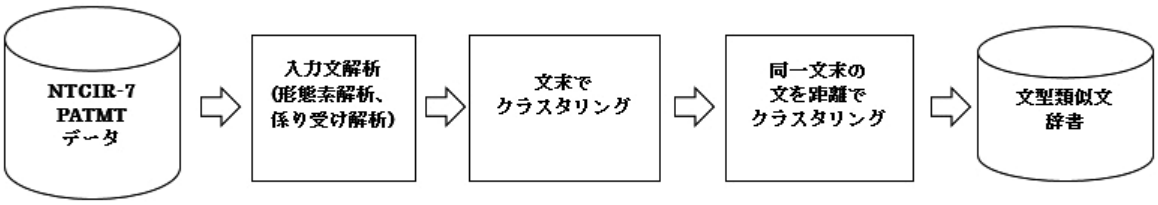

図 1 文型類似文辞書作成過程

3.2.3.4 クラスタを用いたシステム

前節で構成した文型類似文辞書を用いて日英翻訳支援システムを構築する構想を述べる。概念図を図 2 に示す。まず、翻訳したい日本語文を入力すると形態素解析、係り受け解析が行われる。次に、文型類似文辞書の文型とパターンマッチングが行われ、同一文型の中の日本語文と上記方法で距離計算がなされる。ここで全ての索引のうち、距離が近い順番に日英文のペアが表示される。それぞれのペアから当該クラスタ内の全ての文を閲覧することもでき、それを用いて翻訳者は英文作成を行う。

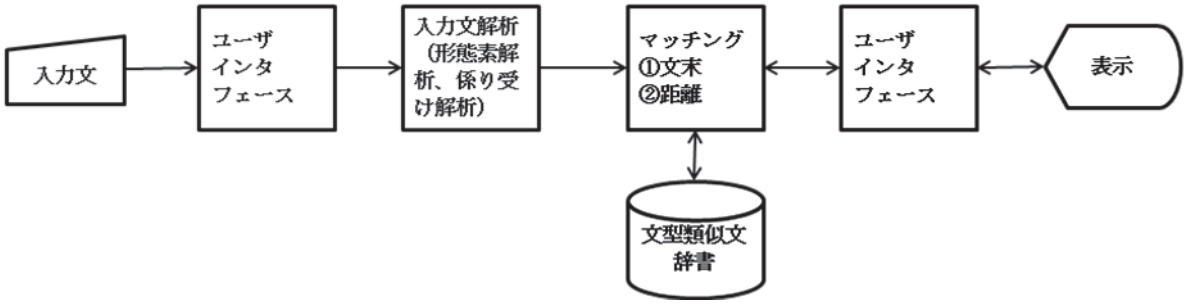

図 2　日英翻訳支援システムの構成

3.2.3.5 システムの特徴

本節で、本システムの特性について考察する。はじめに文型類似文辞書を用いる利点について述べる。

(1)　文型の網羅性

本提案の文型類似文辞書は上述の膨大な特許コーパス（180 万文例）から自動抽出されるので、文型の網羅性、完全性が高い。この網羅性はもちろんコーパスの質に負うところが大きいが、翻訳における有益性を示している。

(2)　文型の自動分析

文型によって英文作成支援をする本があるが(2)(3)、この場合は、翻訳者自身で該当文型を探さなければならず、時間のロスなどコストが大きい。これに対し、提案したシステムでは入力文に対応する文型を自動的に計算するため、複数の候補が表示されユーザの負担が少ない。

(3)　文型類似文の多さ

クラスタによって類似文を集めているため、当該文に対する類似文を複数表示することができる。

(4)　文型類似文の近さによる並び替え

文型類似文が入力文に近い順に自動的に距離計算されて表示されるので、ユーザは参照する類似文を見つけやすい。

(5)　文型とクラスタとの関係

クラスタの生成は文間距離を用いて自動的になされるため、これがいわゆる言語学的な文型と一致するかどうかは別問題となる。しかし、我々は、コーパスが網羅的であるなど適切であれば、我々の定義によって形成したクラスタは対象となる言語分野において使用される文型（特に文末表現）を網羅的に表現する手段として適切であると考える。

(6)　特許文との関係

本提案システムが対象としている科学技術文（特に特許文）の特性との関連について議論する。科学技術文は複文や長文が多く、現段階では機械翻訳で高品質の英文を作成するのが困難であるにも関わらず、年々増加する特許の国際的な戦略に対応するため、その翻訳コストは重要な課題となっている。人間翻訳の場合はその科学技術分野の専門外の人が英訳する場合も多いが、分野特有の表現があり難しい。こういった点は、既存のテンプレートによる英文作成支援システムや、キーワードを入力してヒットする文を表示する英文作成支援システムでは困難であることを示している。それに比べ、提案した日英翻訳支援システムは、プロの人間翻訳の実際の類似文を参考に出来るため、既存の専門用語辞書も援用しながら品質の良い英文を作成しやすい。

3.2.4　システム構成の概要

ここでは文型類似文末辞書作成（図 1）のシステム構成の概要を述べる。

3.2.4.1　NTCIR－7　PATMT データ

今回のシステムで採用した NTCIR－7 PATMT コーパスは本来、特許情報を対象として機械翻訳を評価するためのテストコレクションである。

以下のデータから構成されている。

① 日英パテントファミリーから自動抽出された約 180 万のモデル訓練用日英対訳データ
② 自動抽出された 5200 対訳ペアを人手でクリーニングしたテストセット
③ マルチリファレンスによる自動評価を行うために人手翻訳を行った追加の正解文
④ 言語横断検索による Extrinsic な評価を行うための 124 の検索課題
⑤ NTCIR-7 参加チームの翻訳結果に対する人手による評価結果

この内使用しているのは①の約 180 万のモデル訓練用日英対訳データである。

3.2.4.2　入力文解析

NTCIR－7 PATMT コーパス約 180 万文のモデル訓練用日英対訳データの形態素解析、係り受け解析を行い、形態素単位の文節係り受け形式に変換する。形態素解析には ChaSen、係り受け解析には CaboCha を使用している。次に形態素単位の文節係り受け形式ファイルを、文節単位で処理できるように文節単位の文節係り受け形式ファイルに変換する（図 3)。

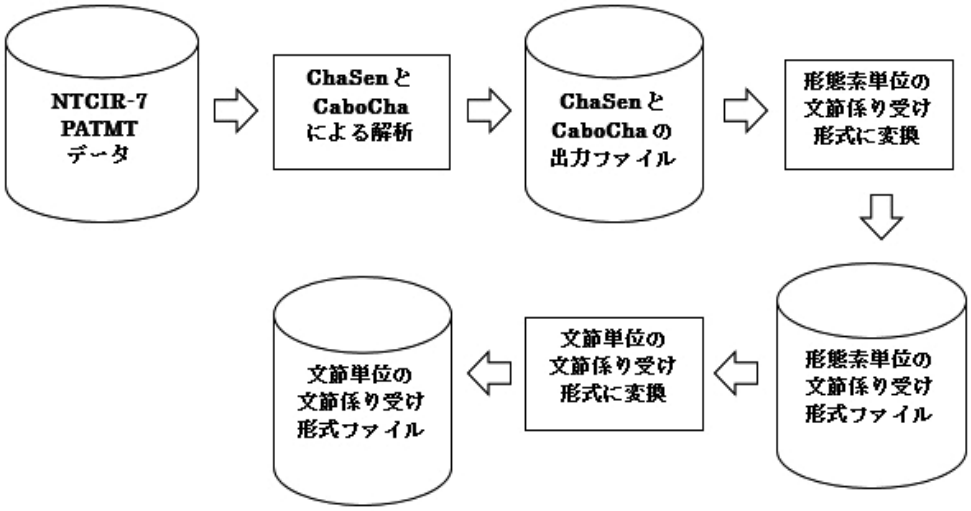

図 3　入力文解析

また後の処理で文節属性を付加するために、NTCIR－7 PATMT コーパスから文節属性を抽出して、文節に対して、受け属性と係り属性（表 1）を付与し文節属性抽出ファイルを作成する（図 4)。

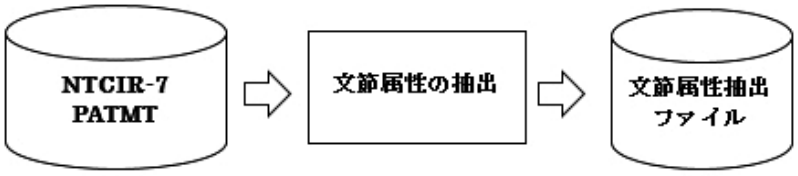

図 4　文節属性抽出ファイルを作成

前の処理で出来上がった文節単位の文節係り受け形式ファイルと文節属性抽出ファイルをマージし、係り受けと文節属性のマージファイルを作成する。このファイルから文末表現を抽出し文末表現ファイルを作成する。また係り受け深さデータを付加して係り受け深さのファイルを作成する（図 5)。

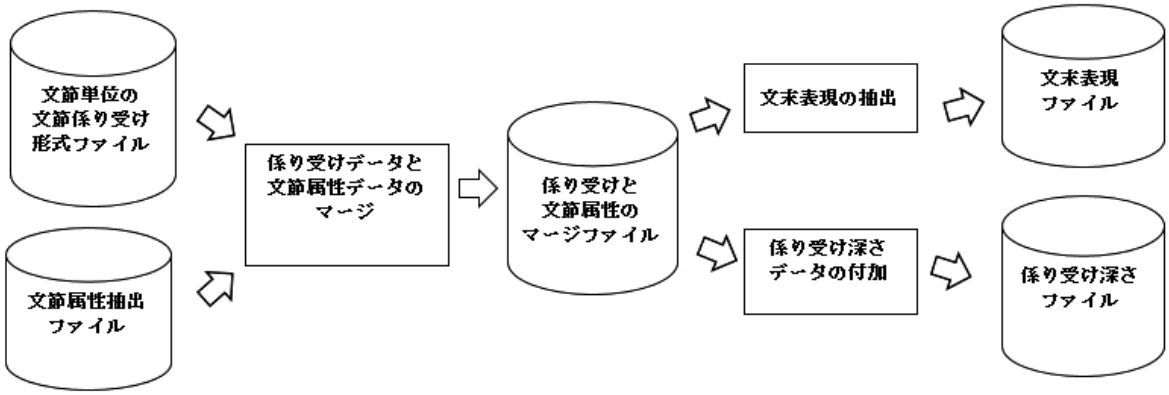

図５　文末表現ファイルと係り受け深さのファイルの作成

3.2.4.3　文末クラスタリング

文末表現を集めたクラスタを作成する（図 6)。まず先に作成した文末表現ファイルの文末表現を文節数単位で分割する。文節数は１文節、２文節、３文節単位に分割してファイルする。４文節以上は計算量の関係で無視する。その後、文節数ごとに文末表現を収集して文末表現のクラスタを作成する（図 6)。

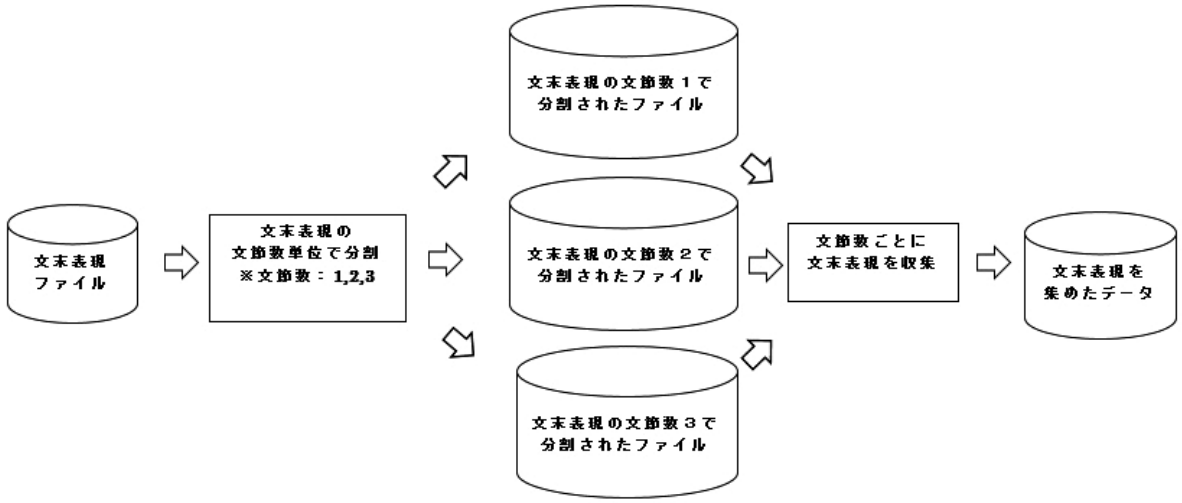

図６　文末クラスタリング

3.2.4.4　同一文末の文を距離でクラスタリング

同一文末の文を距離計算してクラスタリングする。

3)　距離でクラスタリングしたファイルの作成

係り受け深さファイルと文末表現を集めたデータから文末表現単位のファイルを作成する。距離計算を現実的なものにするために、このうち深さ３以下の文節のみを残し、ファイル内の各文

間の距離を計算し文間距離計算結果ファイルを作成する。さらに距離計算したデータをクラスタリングし、距離でクラスタリングしたファイルを作成する（図 7）。

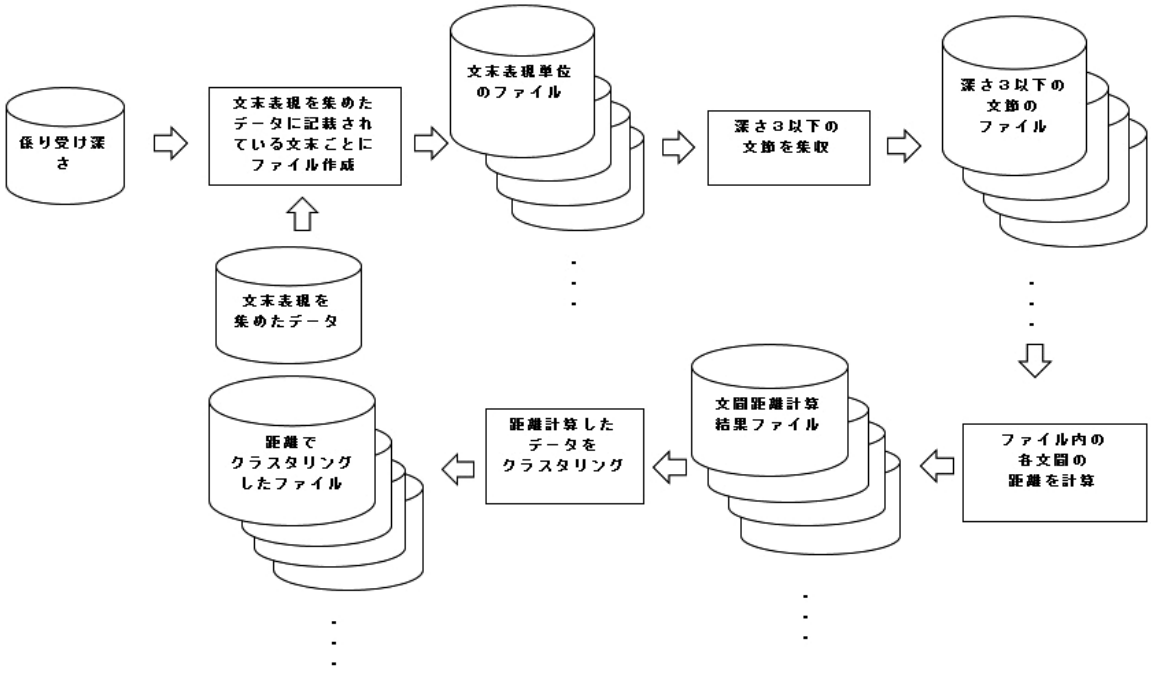

図 7　距離でクラスタリングしたファイルの作成

4)　文末表現を含む文の ID を抽出したファイルの作成

作成された係り受け深さファイルと文末表現を集めたデータから、文末表現を集めたデータに記載されている文末を含む文の ID を文末ごとに抽出し、文末表現を含む文の ID を抽出したファイルを作成する（図 8）。

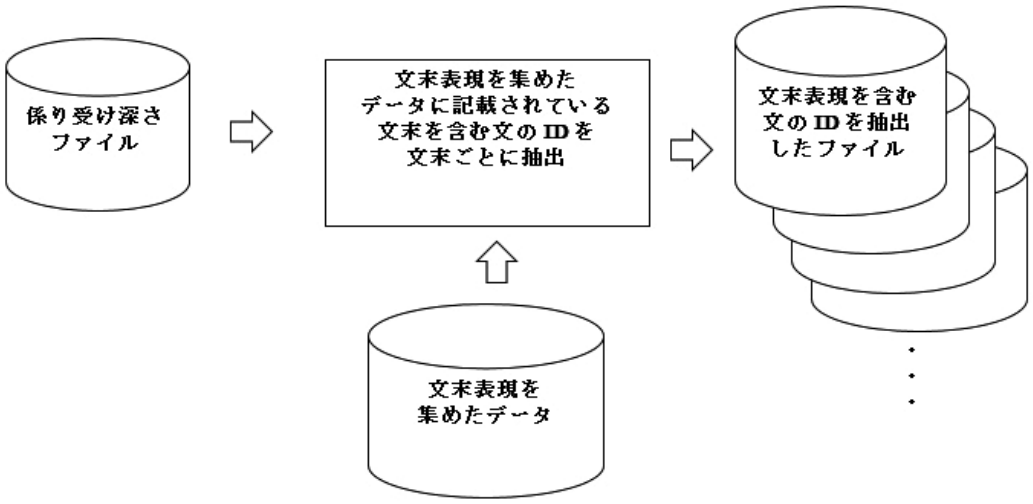

図 8　末表現を含む文の ID を抽出したファイルの作成

5)　同一文末の文を距離でクラスタリング

解析対象文のファイル（NTCIR-7 PATMT から解析対象を抽出した文集合）と距離でクラスタリングしたファイル、文末表現を含む文の ID を抽出したファイルの３つのファイルをマージして、文型類似文辞書（同一文末を距離でクラスタリングしたファイル）を作成する（図 9）。

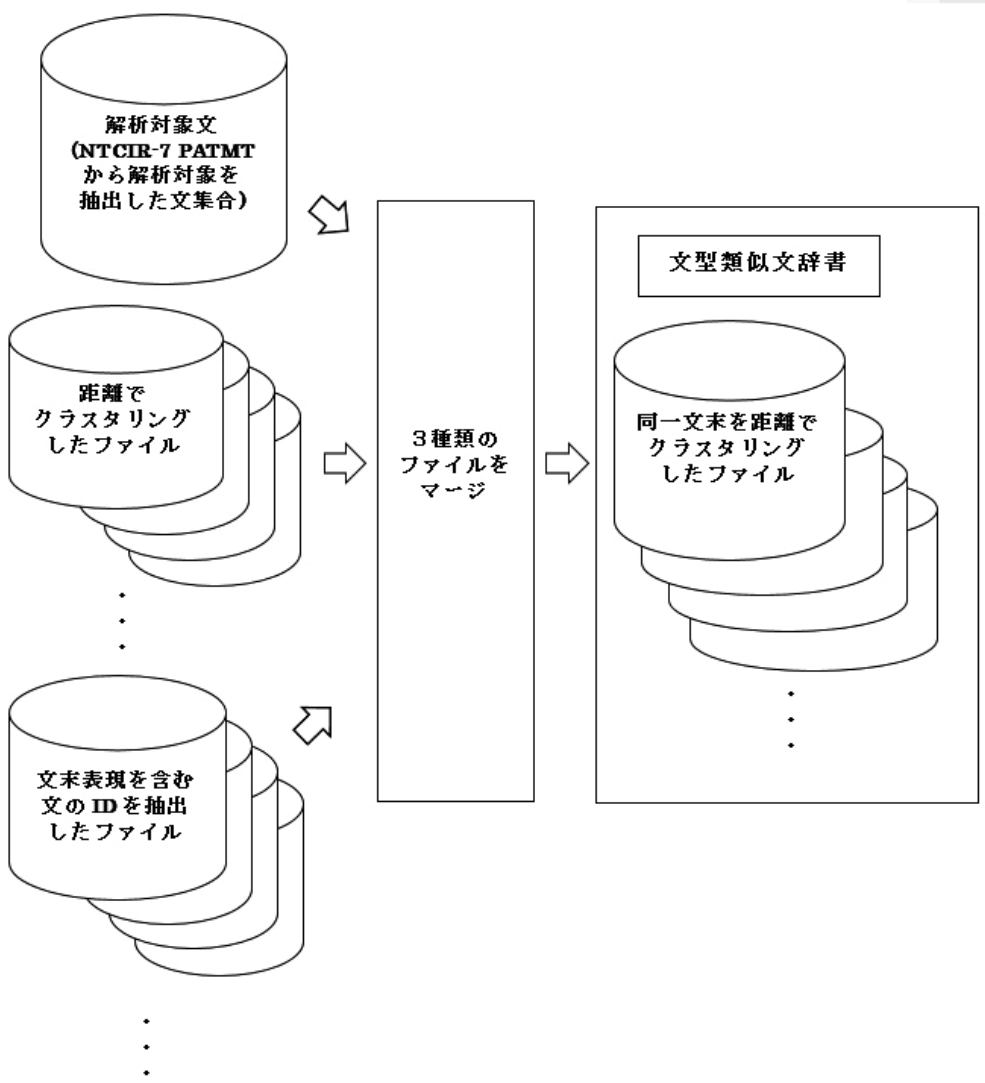

図 9　同一文末の文を距離でクラスタリング

3.2.5　まとめと今後の展望

我々は、日英翻訳を行う際に重要となる文型の抽出を容易に行えるようにするため、コーパスを用いる方法のうち、キーワードの一致のみで候補文を抽出する従来型に代わって、文末に注目する方法を提案している。この方法では、文末構造が似ている文をクラスタリング手法でグルーピングし、代表的な文を索引とする文型類似文辞書を構築する。今回はこの文型類似文辞書構築

のシステム構成を中心に述べた。

この辞書を用いると、翻訳したい文と同じ文末構造を持つ文の集合を優先的に表示するために、同じコーパスであっても異なる文を候補文とすることができる。このような方法では、いわゆる機械翻訳のように直訳的に翻訳するのではつくれないような、こなれた英文を作成できると予想される。それを確かめるために、作成した文型類似文辞書を用いた日英翻訳支援システムのプロトタイプを構築し、被験者を用いた評価実験を行う必要がある。

この新方式は機械翻訳にとっても有用と考えている。例えば機械翻訳の前処理に利用できる可能性をもっている。将来的には機械翻訳や特許文の翻訳への応用を図る予定である。

謝辞

支援をいただいた(財)日本特許情報機構に感謝します。

参 考 文 献

(1) Somers, H.: “Translation memory systems,” in Computers and Translation: a Translator’s Guide, John Benjamins Publishing Co.: pp.31-48 (2003)

(2) 佐藤元志著、田中宏明監修、古米弘明監修、鈴木穣監修:“英語論文表現例集 with CD-ROM すぐに使える 5,800 の例文”、技報堂出版 (2009)

(3) 佐藤洋一編著:“科学技術英語論文英借文用例辞典 英作文から英借文へ簡単！英語論文作成法”、オーム社 (2010)

4. 機械翻訳評価手法

４．１　機械翻訳の評価について

山梨英和大学　江原暉将

4.1.1　はじめに

いかなる技術においてもその進歩のためには客観的な評価が必要であり、機械翻訳も例外ではない。しかし「翻訳」とはきわめて主観的な作業であり、その評価も主観的になることが多い。本研究会の拡大評価部会では、自動評価、人手評価、テストセットの三種類の評価手法について研究を進めている。これらの評価手法に対して客観性の有無を筆者の考えで付与したのが表 4.1.1 である。客観的な評価手法とは、万人あるいは少なくとも多くの人が認める評価手法であろう。

	試験文選択	評価基準 （設問設定）	評価値決定
自動評価	○	×	○
人手評価	○	○	×
テストセット	×	○	○

表 4.1.1　3 種類の評価手法に対する客観性の有無
（○：客観性あり、×：客観性なし）

評価の手順を試験文選択、評価基準の設定（設問設定)、実際の評価値の決定、の 3 段階に分けて考える。試験文選択の段階では、自動評価と人手評価が翻訳対象から無作為に試験文を選択することが多いのに対して、テストセットでは後の設問設定を考えて作為的に試験文を選択する。そのため、自動評価と人手評価は、試験文選択に客観性があるといえる。一方、テストセットは客観性が少ない。評価基準の段階では、人手評価が Adequacy や Fluency など多くの人が認める基準であり客観性がある。テストセットでも翻訳精度に影響する事項を網羅的に基準(設問)として設定するため客観性は大きい。一方、自動評価は、その性質上、自動化が可能な範囲での評価基準であり、翻訳の観点からは「単純」すぎる基準である。そのため特に翻訳者が認める基準となっておらず、客観性が低い。最後の評価値決定の段階では、自動評価はもちろん、テストセットもパターンマッチを用いて自動的に評価値を決めており、客観性がある。一方、人手評価は、評価者の能力や当該分野への知識の深さ、翻訳品質に対する考え方に差があり、主観的となる。このような状況の中で各評価手法について客観性をできるだけ確保する努力がなされている。本文では、機械翻訳の利用現場で多く使われている人手評価に着目し、評価値決定段階での客観性について考察する。

4.1.2　人手評価における客観性の度合い

人手評価において異なる評価者間の評価の一致率P(A)は評価の客観性を測る指標として利用できる。また偶然に一致する可能性を考慮した指標である Kappa 係数(K)も良く用いられる。K は

$$K=\frac{P(A)-P(E)}{1-P(E)} \qquad (1)$$

で与えられる。ここでP(E)は、異なる評価者が偶然に同じ評価値を与える確率である。Adequacy や Fluency は 5 段階評価であるからP(E) = 1/5になる。WMT2007 で求められている Adequacy と Fluency に対するP(A)と K の値を表 4.1.2 に示す(Callison-Burch et al., 2007)。

	P(A)	K
Adequacy	0,380	0.226
Fluency	0.400	0.250

表 4.1.2 WMT2007 における評価者間の一致率(P(A))と Kappa 係数(K)

K の値に対して Landis and Koch (1977)は表 4.1.3 のような解釈を与えている。

Kの範囲	解釈
0 - 0.2	slight
0.21 - 0.4	fair
0.41 - 0.6	moderate
0.61 - 0.8	substantial
0.81 - 1	almost perfect

表 4.1.3 Kappa 係数の値に対する解釈

これに従うと WMT2007 における Adequacy と Fluency の Kappa 係数は fair となる。

越前谷ほか(2014)が自動評価のメタ評価を行ったときに用いた NTCIR-7 の JE サブタスク人手評価データ(Fujii et al., 2008)について P(A)や K を求めたところ表 4.1.4 のようになった。NTCIR-7 では 100 文の試験文に対して 15 システムの機械翻訳結果が評価されており、それら 1500 文の機械翻訳結果を 3 人の評価者で評価した。評価者の組み合わせは 3 通りあるため、ここで用いたデータ数は Adequacy と Fluency に対してそれぞれ 4500 評価対となる。

	P(A)	K
Adequacy	0.441	0.301
Fluency	0.339	0.174

表 4.1.4 NTCIR-7 の JE サブタスクにおける評価者間の一致率(P(A))と Kappa 係数(K)

WMT2007 と NTCIR-7 を比較すると、Adequacy では NTCIR-7 の方が P(A)や K が大きく、Fluency は WMT2007 のほうが大きい。NTCIR-7 の値は Adequacy で fair だが、Fluency では slight となっている。いずれにしても Kappa 係数はあまり大きくなく評価値決定段階での客観性が確保できているとはいえない。

4.1.3 人手評価値の確率化

人手評価における評価値決定に客観性が期待できないので、評価値を確率的に扱うことを考える。本節以降では、一致率が比較的に高かった NTCIR-7 での Adequacy のデータについてのみ考察する。表 4.1.5 には、NTCIR-7 での評価者間の評価値の各組み合わせに関する度数を示す。表 4.1.5 から、例えば第 1 の評価者(E1 とする)が 3 と評価したとき、同一出力に対して第 2 の評

価者(E2 とする)が 2 と評価した度数は 529 である。表 4.1.5 では、対称性を確保するために、評価者の組み合わせを入れ替えた事例も用いている。そこで評価対の総数は 9000 となる。

E2＼E1	1	2	3	4	5
1	1324	736	319	59	0
2	736	736	529	108	15
3	319	529	734	366	70
4	59	108	366	528	313
5	0	15	70	313	648

表 4.1.5 NTCIR-7 の JE サブタスクにおける
異なる評価者間(E1 と E2)の Adequacy 評価値の頻度

E1 が例えば 3 と評価したときでも E2 の評価値はばらつくことになる。そこで、E1 の評価値を確率変数とする。E1 が i と評価するとき E2 の評価値を値にとる確率変数を $X^{(i)}$とする。$X^{(i)}$が{1, … ,5}の各値をとる確率分布は表 4.1.6 のように計算できる。表 4.1.6 には平均値(μ_X)と標準偏差(σ_X)も示してある。

x＼X	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$
1	0.5431	0.3465	0.1581	0.0429	0.0000
2	0.3019	0.3465	0.2621	0.0786	0.0143
3	0.1308	0.2491	0.3637	0.2664	0.0669
4	0.0242	0.0508	0.1814	0.3843	0.2992
5	0.0000	0.0071	0.0347	0.2278	0.6195
μ_X	1.64	2.03	2.67	3.68	4.52
σ_X	0.80	0.93	1.05	1.04	0.69

表 4.1.6 E1 の各評価値に対する E2 の評価値の確率変数化と
その確率分布、平均値(μ_X)、標準偏差(σ_X)

例えば E1 が 3 と評価したときには E2 が平均で 2.67 と評価し、標準偏差は 1.05 である。$X^{(1)}$から $X^{(5)}$の確率分布を図 4.1.1 の左側の棒グラフ(青色)に示す。

次に、この経験分布を二項分布でモデル化する。二項分布 $B_N(n, p)$は n, p の二つのパラメータを持ち、$i \in \{0, \dots, n\}$ に対する確率 P(i) は

$$p(i)=\binom{n}{i} p^i(1-p)^{(n-i)} \qquad (2)$$

で与えられる。Adequacy の場合、値は{1, … ,5}であるからこの範囲を{0, … ,4}に平行移動することによって n=4 とすることができる。二項分布の平均値 μ_N と標準偏差 σ_N は

$$\mu_N=n\,p \qquad (3)$$

$$\sigma_N=\sqrt{n\,p\,(1-p)} \qquad (4)$$

で与えられる。$\mu_N=\mu_X$とおくことによって、表 4.1.6 の μ_Xから式(3)に従ってパラメータ p を求めることができる。この二項分布モデルによる確率変数を $Y^{(1)}$から $Y^{(5)}$とすると、パラメータ n, p は表 4.1.7 のように計算できる。

y＼Y	$Y^{(1)}$	$Y^{(2)}$	$Y^{(3)}$	$Y^{(4)}$	$Y^{(5)}$
n	4	4	4	4	4
p	0.1590	0.2564	0.4181	0.6689	0.8810

表 4.1.7 得られた二項分布モデルのパラメータ

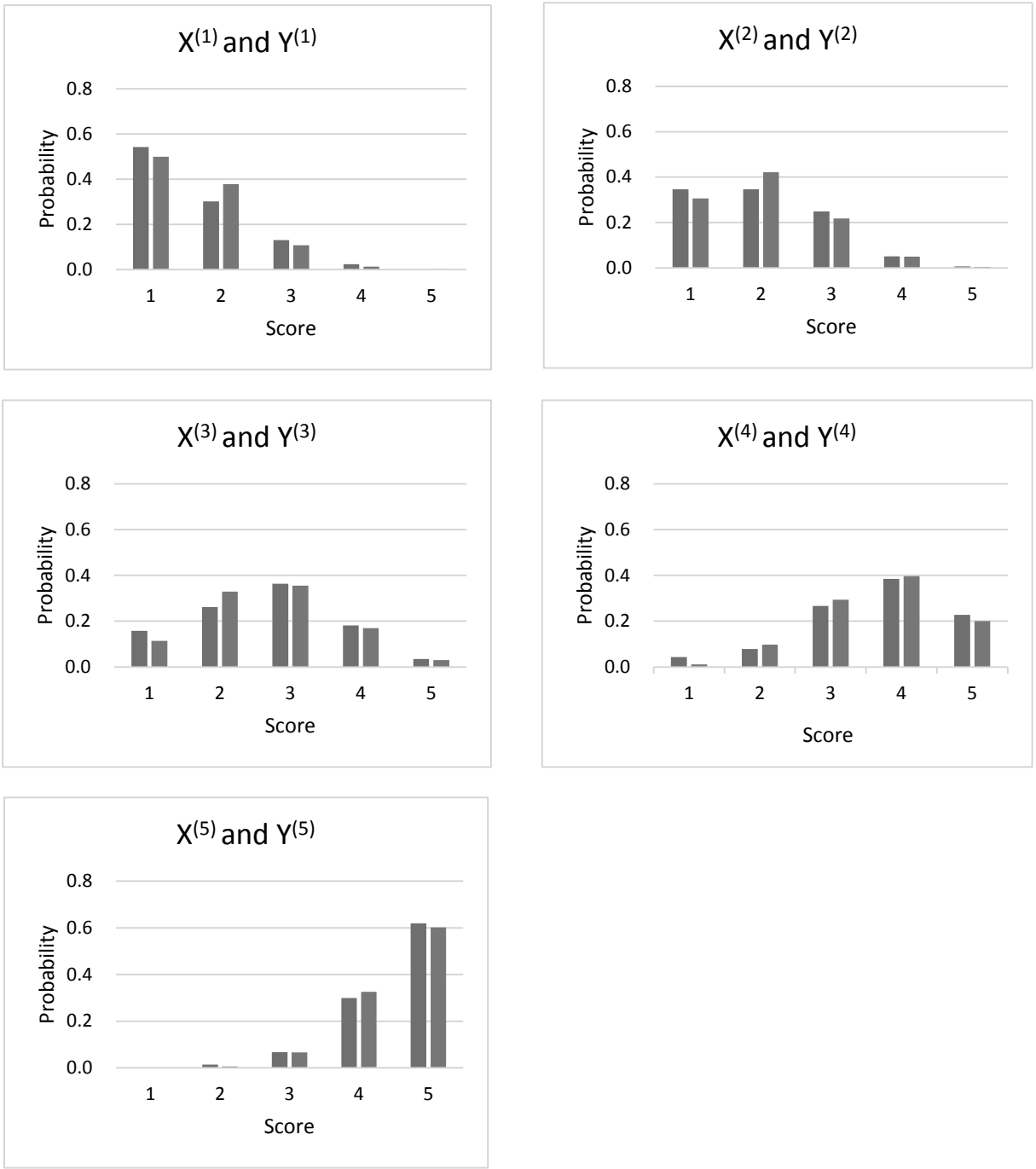

図 4.1.1 $X^{(1)}$から $X^{(5)}$の確率分布(左側、青色)と
二項分布モデル($Y^{(1)}$から $Y^{(5)}$)による確率分布(右側、茶色)

この二項分布モデルを用いた確率分布と平均値(μ_N)と標準偏差(σ_N)を表 4.1.8 に示す。表 4.1.6

と表 4.1.8 を比較すると、二項分布モデルの方が、標準偏差が小さくなっている。さらに、図 4.1.1 には $Y^{(1)}$から $Y^{(5)}$の確率分布も右側の棒グラフ(茶色)として示してある。X と Y の分布は比較的に類似している。

y＼Y	$Y^{(1)}$	$Y^{(2)}$	$Y^{(3)}$	$Y^{(4)}$	$Y^{(5)}$
1	0.5001	0.3058	0.1146	0.0120	0.0002
2	0.3784	0.4217	0.3295	0.0972	0.0059
3	0.1073	0.2181	0.3552	0.2943	0.0660
4	0.0135	0.0501	0.1701	0.3963	0.3255
5	0.0006	0.0043	0.0306	0.2001	0.6024
μ_N	1.64	2.03	2.67	3.68	4.52
σ_N	0.73	0.87	0.99	0.94	0.65

表 4.1.8 二項分布モデルによる確率分布、平均値(μ_N)、標準偏差(σ_N)

4.1.4 人手評価値の客観性の向上方法

人手評価値の客観性を向上させる一つの方法として、複数の評価者により評価を行い、その評価値を組み合わせるという方法がある。多くの評価者によって得られた評価値は、多くの人が認める評価値であり、客観的である。評価値の組み合わせ方には種々あるが、ここでは、複数の評価者の評価値の平均値を用いる方法を考察する。k+1 人の評価者が評価を行うとき、ある評価者が評価値 m を与えたとき他の k 人のうち i 番目の評価者が与える評価値を確率変数として $Y^{(m)}i$ (m=1, … ,n+1; i=1, ... ,k)とする。 すべての i について $Y^{(m)}i$ が表 4.1.7 に示す $Y^{(m)}$と同一の確率分布 $B_N(n, pm)$に従うと仮定し[1]、かつ $i \neq j$ のとき $Y^{(m)}i$ と $Y^{(m)}j$ は独立であると仮定する。この時、

$$Z^{(m)} = \sum_{i=1}^{k} Y^{(m)}i \qquad (m=1, \ldots, n+1) \qquad (5)$$

とおくと、$Z^{(m)}$は{0, … ,kn}に値をとり、再び二項分布 $B_N(kn, pm)$に従う。$Z^{(m)}$の平均と標準偏差は式(3)と(4)において n を kn で p を pm で置き換えたものとなる。そこで $Z^{(m)}$から[1,5]に値をとる確率変数

$$U^{(m)} = \frac{1}{k} Z^{(m)} + 1 \qquad (6)$$

に変数変換すると、$U^{(m)}$は平均値が n pm+1 であり標準偏差は$\sqrt{n\,pm\,(1\text{-}pm)}/\sqrt{k}$となる。

NTCIR-7 では、3 人の評価者によって評価が行われたので、k=2 とした場合の実測値が表 4.1.9 のように得られる。

E1 の評価値に対する(E2 の評価値+E3 の評価値)/2 の値を確率変数化したものを W とするとその確率分布が表 4.1.10 のようになる。この表には平均値と標準偏差も示してある。一方、式(6)に示す二項分布モデルによる確率分布、平均値、標準偏差は表 4.1.11 のようになる。これらの表より二項分布モデルによる標準偏差が経験分布による標準偏差より小さいことが分かる。

[1]ここで pm は $Y^{(m)}$に対する p の値。

(E2+E3)/2＼E1	1	2	3	4	5
1	420	170	65	7	0
1.5	340	274	108	14	0
2	267	243	153	22	2
2.5	122	184	211	64	7
3	48	117	190	104	16
3.5	13	50	163	128	27
4	9	20	86	141	78
4.5	0	4	27	132	150
5	0	0	6	75	243

表 4.1.9 NTCIR-7 の JE サブタスクにおける
異なる評価者間の Adequacy 評価値(E1 の評価値と(E2 の評価値+E3 の評価値)/2)の頻度

w＼W	$W^{(1)}$	$W^{(2)}$	$W^{(3)}$	$W^{(4)}$	$W^{(5)}$
1	0.3445	0.1601	0.0644	0.0102	0.0000
1.5	0.2789	0.2580	0.1070	0.0204	0.0000
2	0.2190	0.2288	0.1516	0.0320	0.0038
2.5	0.1001	0.1733	0.2091	0.0932	0.0134
3	0.0394	0.1102	0.1883	0.1514	0.0306
3.5	0.0107	0.0471	0.1615	0.1863	0.0516
4	0.0074	0.0188	0.0852	0.2052	0.1491
4.5	0.0000	0.0038	0.0268	0.1921	0.2868
5	0.0000	0.0000	0.0059	0.1092	0.4646
μ_W	1.64	2.03	2.67	3.68	4.52
σ_W	0.63	0.76	0.90	0.90	0.59

表 4.1.10 E1 の各評価値に対する(E2 の評価値+E3 の評価値)/2 の確率変数化と
その確率分布、平均値(μ_W)、標準偏差(σ_W)

u＼U	$U^{(1)}$	$U^{(2)}$	$U^{(3)}$	$U^{(4)}$	$U^{(5)}$
1	0.2501	0.0935	0.0131	0.0001	0.0000
1.5	0.3785	0.2579	0.0756	0.0023	0.0000
2	0.2505	0.3112	0.1900	0.0165	0.0001
2.5	0.0948	0.2146	0.2731	0.0667	0.0009
3	0.0224	0.0925	0.2453	0.1685	0.0085
3.5	0.0034	0.0255	0.1410	0.2722	0.0501
4	0.0003	0.0044	0.0507	0.2749	0.1854
4.5	0.0000	0.0004	0.0104	0.1586	0.3922
5	0.0000	0.0000	0.0009	0.0401	0.3628
μ_U	1.64	2.03	2.67	3.68	4.52
σ_U	0.52	0.62	0.70	0.67	0.46

表 4.1.11 二項分布モデルの確率値と平均値(μ_U)、標準偏差(σ_U)

図 4.1.2 に W の分布(左側の棒グラフ、青色)と U の分布(右側の棒グラフ、茶色)を示す。$W^{(1)}$と$U^{(1)}$の乖離が 1 と 1.5 のところで大きい。これは、図 4.1.1 からわかるとおり、$P(X^{(1)}=1)>P(Y^{(1)}=1)$である一方、$P(X^{(1)}=2)<P(Y^{(1)}=2)$であることが原因であろう。二項分布モデルの不備と考えられる。

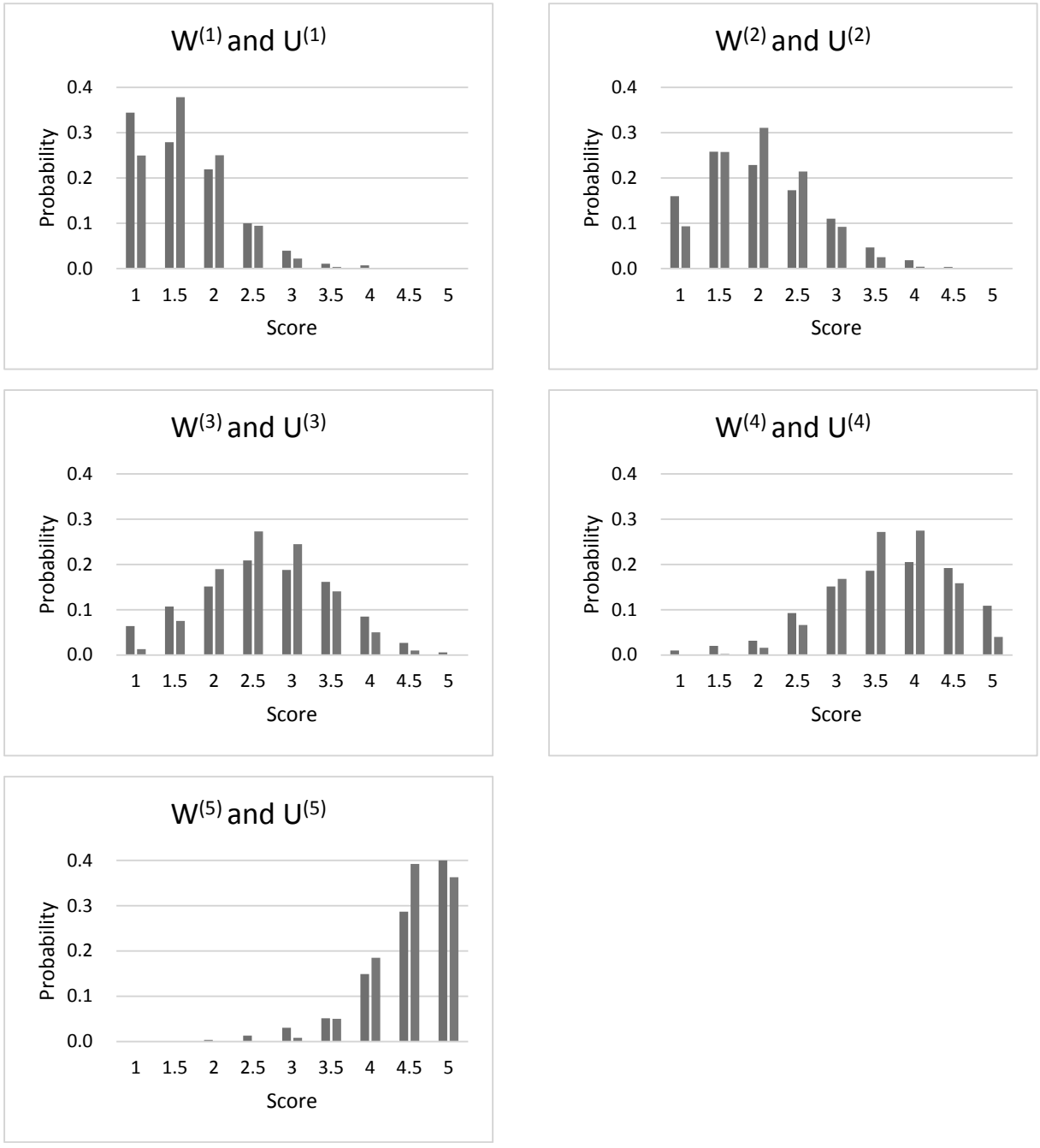

図 4.1.2 二人の評価者の平均に対する $W^{(1)}$から $W^{(5)}$の確率分布(左側、青色)と
二項分布モデル($U^{(1)}$から $U^{(5)}$)による確率分布(右側、茶色)

NTCIR-7 では 3 人で評価したので、k=3 の場合を二項分布モデルから予測する。k=3 の場合のモデル確率変数を V とすると、図 4.1.3 が得られる。$V^{(m)}$は(仮想的な)第 4 の評価者が評価値 m を

与えた時、他の 3 人の評価者が与える評価値の平均値を表す確率変数である。

評価者数 k を増やした時の標準偏差の変化を図 4.1.4 に示す。k を増やすことで標準偏差が減少することが分かる。モデル分布では k=1 での標準偏差を σ_N とするとき k>1 に対する標準偏差は $\sigma_N/\sqrt{k}$に減少するが、経験分布では減少の度合いが少ない。

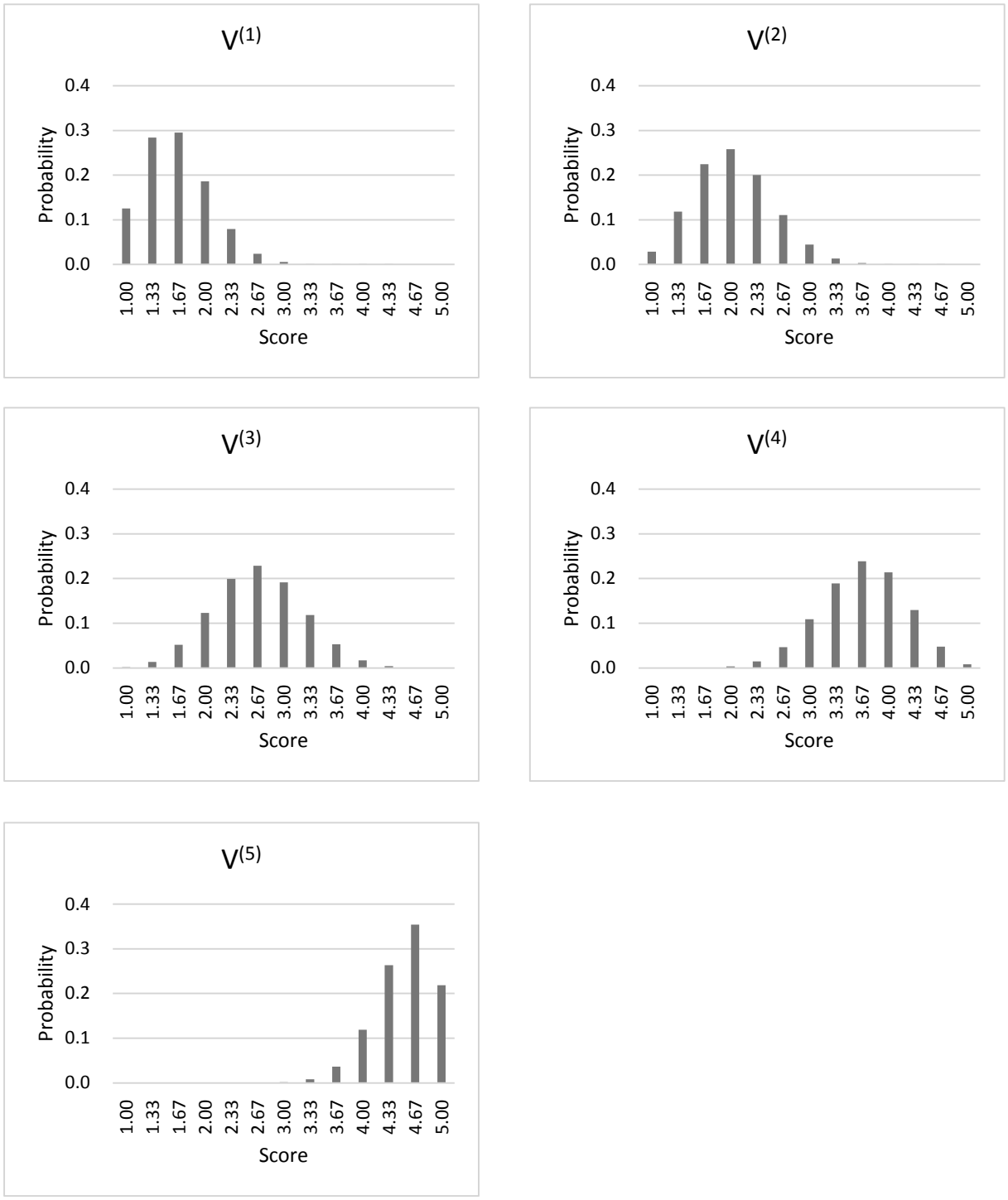

図 4.1.3 k=3 の場合の二項分布モデルによる確率分布

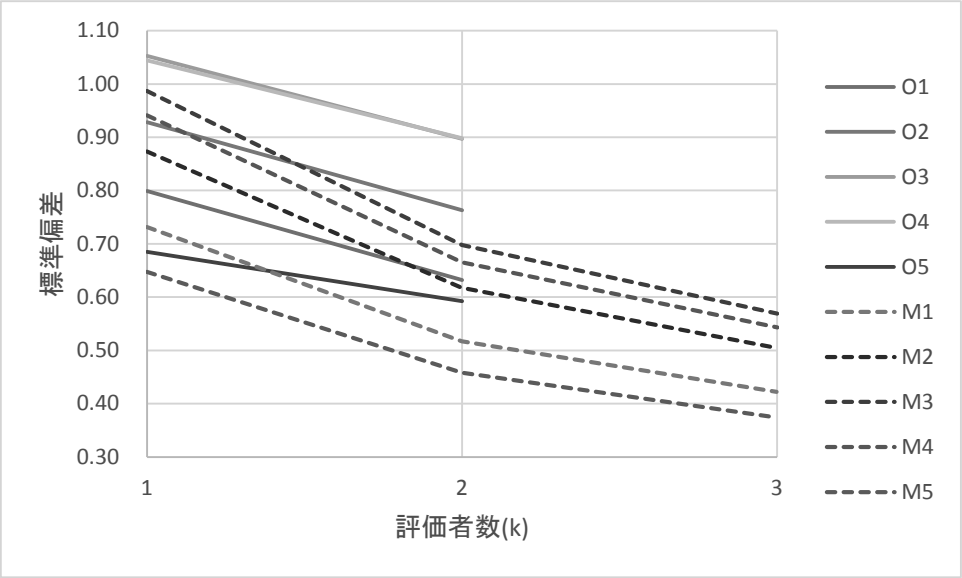

図 4.1.4 評価者数の増加による標準偏差の減少
(Om:評価値 m の経験分布、Mm:評価値 m のモデル分布)

4.1.5 おわりに

人手評価の客観性を確保することを目的として、人手評価の結果を確率変数としてとらえ、二項分布モデルを用いて解析した。評価者数を増やすことで評価値の標準偏差が減少し、評価の客観性が向上することが確認できた。しかしながらモデル分布と経験分布との間に乖離が見られる。モデルの改良が今後の課題である。

人手評価における客観性を向上させる工夫として Acceptability (Goto et al., 2011)や Ranking (Callison-Burch et al., 2007; Nakazawa et al., 2014)などの評価基準が提案されている。評価基準設定段階の客観性を向上させるのに加えて、評価値決定段階での客観性を向上させる必要があろう。

参考文献

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder. 2007. Meta-Evaluation of Machine Translation, *Proceedings of the Second Workshop on Statistical Machine Translation*, pp.136—158.

越前谷博, 須藤克仁, 磯崎秀樹, 江原暉将. 2014. NTCIR PATMT データと WMT METRICS TASK データにおける英文データを対象とした自動評価法のメタ評価, *平成 25 年度 AAMT/Japio 特許翻訳研究会報告書*, pp.62-70.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proceedings of NTCIR-7 Workshop Meeting,* pp.389-400.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou. 2011. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, *Proceedings of NTCIR-9 Workshop Meeting,* pp. 559-578.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics,* Vol. 33, pp.159–174.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi and Eiichiro Sumita. 2014. Overview of the 1st Workshop on Asian Translation, *Proceedings of the 1st Workshop on Asian Translation*.

４．２　文の長さを考慮したチャンクに基づく自動評価法

北海学園大学　越前谷　博

4.2.1　はじめに

統計翻訳の発展に伴い自動評価に対するニーズが増加している。そのような状況において、現在、様々な自動評価法が提案されている。自動評価法は、BLEU[1]や NIST[2]のように n-gram 一致率に基づく手法、WER[3]、ROUGE シリーズ[4]、METEOR[5]、そして、IMPACT[6]のようにチャンクに基づく手法、そして、RIBES[7]のように語順に着目した手法に大きく分類される。本報告では、チャンクに基づく自動評価法において、文の長さを考慮した評価が可能な自動評価法を提案する。更に、提案手法を自動評価法 IMPACT に適用し、提案手法の有効性を確認するための性能評価実験を行った。

4.2.2 チャンクに基づく自動評価法における問題点

METEOR や IMPACT などのチャンクに基づく自動評価法では、機械翻訳システムが出力するシステム訳と人手で作成される正解訳、即ち、参照訳との間で一致単語を決定する。そして、一致単語が連続する場合、それを一つの塊、即ち、チャンクと見なすことで評価結果であるスコアを算出する。更に、スコアを算出する際、チャンクに対して数値化された値をシステム訳及び参照訳の構成単語数で割ることで、スコアを 0.0 から 1.0 に正規化している。このようなスコアの算出においては、文の長さが短いほど、過度に低いスコアが付与される傾向がある。

例えば、参照訳が“doctor cured a patient”、システム訳が“doctor treated a patient”の場合、チャンクは“doctor”と“a patient”の 2 つである。一方、不一致単語は参照訳中の“cured”とシステム訳中の“treated”である。参照訳とシステム訳の構成単語数は 4 であるため、単純化すると個々の不一致単語の重みは 0.25（=1/4）となる。長文になると、例えば、構成単語数が 10 である場合、不一致単語の重みは分散されて 0.1 となる。即ち、短文と長文の間の不一致単語の重みが大きく異なっていることを意味する。その結果、短文の方が不一致単語 1 つにおける重みが大きいため、スコアの大幅な低下を招く。上述の例では、不一致単語が 1 つ存在するだけで、精度は 0.75（=3/4）となる。短文の場合、一致単語も不一致単語も構成単語数が少ないため、単語の重みは長文に比べて大きくなることは当然である。しかし、自動評価のタスクではこの点について配慮が必要と考えられる。例えば、不一致単語“cured”と“treated”は、表層的には不一致であっても意味的には全く異なっているという訳ではない。そのため自動評価タスクにおいては、不一致単語の重みが過度に大きくなることは問題と考えられ、短文になるほどこの問題は深刻になる。したがって、短文の場合には、不一致単語の重みを小さくし、スコアが過度に低くなることを避ける方法はチャンクに基づく自動評価法において有効と考えられる。このような観点より、本報告では、文の長さを考慮したチャンクに基づく自動評価法を提案する。

文の長さを考慮したチャンクに基づく自動評価法

本節ではチャンクに基づく自動評価法 IMPACT に対して、文の長さを考慮した重みづけを付与する。IMPACT はチャンクを決定する際に、不一致単語の多義性を解消するために、個々のチャンクの長さと文中でのチャンクの位置情報を用いている。そして、一意に決定されたチャンクに基づいて、スコアを算出する。一意に決定されたチャンクは式(3)に基づいて数値化する。*ch* は個々のチャンクを表している。β はパラメータでチャンクの長さに対する重みづけであり、1.0 以上の値である。デフォルト値には 1.2 を用いている。この *Ch_score* を用いて、式(1)より参照訳に対する再現率 *R*、式(2)よりシステム訳に対する適合率 *P* を求める。式(1)と(2)のパラメータ α はチャンクの並びが参照訳とシステム訳において異なる場合、そのチャンクをどのようにスコアに反映させるかを制御するために用いる。1.0 以下として与えられ、1.0 であれば全てのチャンクを同様の重みでスコアに反映させる。即ち、語順の違いに寛容な評価であることを意味する。逆に、0.0 に近いほど語順の異なるチャンクはスコアにはほとんど反映されない。即ち、語順に厳しい評価であることを意味する。式(1)、(2)より得られる *R* と *P* を用いて、式(4)よりスコアを求める。γ は *P/R* より得られる値である。

$$R = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right)}{m^\beta} \right)^{\frac{1}{\beta}} \quad (1)$$

$$P = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right)}{n^\beta} \right)^{\frac{1}{\beta}} \quad (2)$$

$$Ch\_score = \sum_{ch \in ch\_num} length(ch)^\beta \quad (3)$$

$$score = \frac{(1+\gamma^2)RP}{R+\gamma^2 P} \quad (4)$$

このような IMPACT に対して、文の長さに基づく重みづけを付与する。そして、重みづけを付与された自動評価法を APAC と呼ぶこととする。その際、基本的に重みづけは式(1)と(2)に反映する。また、重みの付与の方法としては複数考えられるが、今回は予備実験に基づき重みの計算式の異なる 2 つの自動評価法を用いる。本報告ではそれぞれ APAC_I[7]と APAC_II[8]と呼ぶこととする。

APAC_I は以下の式よりスコアを求める。式(7)が文の長さに基づく重みの計算式となっている。*m* と *n* は参照訳とシステム訳それぞれの構成単語数である。*Ch_score* は式(3)より得られる値で

ある。この式(7)より得られる重みを R と P の分母と分子に加えることで文の長さに応じた重みづけが行われる。そして、式(5)と(6)の結果を式(4)の R と P としてスコアを求める。

$$R = \left(\frac{\left(\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right) \right) + weight}{m^\beta + weight} \right)^{\frac{1}{\beta}} \quad (5)$$

$$P = \left(\frac{\left(\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right) \right) + weight}{n^\beta + weight} \right)^{\frac{1}{\beta}} \quad (6)$$

$$weight = \begin{cases} \left(\frac{1.0}{\log(m+n)} \right)^\beta, & Ch\_score > 0.0 \\ 0.0, & Ch\_score = 0.0 \end{cases} \quad (7)$$

次いで、APAC_II について述べる。APAC_II では、参照訳とシステム訳の構成単語数 m と n それぞれに基づく重みを計算し、R と P それぞれにその重みを付与する。以下にその計算式を示す。式(10)は参照訳に基づく重みであり、式(8)に用いる。そして、式(11)はシステム訳に対する重みであり、式(9)に用いる。このようにして得られる R と P の値を式(4)に用いることにより、最終的なスコアを求める。

$$R = \frac{\left\{ \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right)}{m^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times weight\_m \right\}}{2.0} \quad (8)$$

$$P = \frac{\left\{ \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right)}{n^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times weight\_n \right\}}{2.0} \quad (9)$$

$$weight\_m = \frac{1.0}{\log(m) + 1} \quad (10)$$

$$weight\_n = \frac{1.0}{\log(n) + 1} \quad (11)$$

4.2.4 性能評価実験

4.2.4.1 実験データ

実験データには、NTCIR-7 データ[10]及び WMT14 Metrics Task データ[11]を用いた。NTICR-7 データは英日、日英両方向のシステム訳、参照訳が提供されている。システム訳は英日においては、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 のシステム訳が提供されている。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 文のシステム訳が提供されている。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。また、これらの機械翻訳システムが生成したシステム訳の数は cs-en が 15,015、de-en が 39,039、fr-en が 24,024、hi-en が 22,563、ru-en が 39,039、en-cs が 30,030、en-de が 49,266、en-fr が 39,039、en-hi が 30,084、そして、en-ru が 27,027 の計 315,126 である。

参照訳はこれらのシステム訳に対する正解訳として、cs-en が 3,003、de-en が 3,003、fr-en が 3,003、hi-en が 2,507、ru-en が 3,003、en-cs が 3,003、en-de が 2,737、en-fr が 3,003、en-hi が 2,507、そして、en-ru が 3,003 提供されている。人手評価は評価者が 5 つのシステム訳を比べ、相対評価として 1 から 5 までの 5 段階評価を行った結果が提供されている。その際、高い評価の場合には小さな数値を付与される。

IMPACT、APAC 共にパラメータの値にはデフォルト値を用いている。また、日本文に対しては MeCab[12]を用いて、分かち書きを行っている。英文に対しては、提供されたテキストをそのまま使用している。BLEU には“mteval-13a.pl”を用い、オプションには“- -international -tokenization”を使用している。

4.2.4.2 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[13]を用いて求めている。このような WMT14 Metrics Task の評価方法は文献[11]に準拠している。

4.2.4.3 実験結果

表 1 から表 4 に NTCIR-7 データを用いた実験結果、そして、表 5 から表 8 に WMT14 Metrics Task データを用いた実験結果を示す。表 5 と表 7 の“()”内の数値は、機械翻訳システムの数を示している。表 6 と表 8 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 8 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
IMPACT	0.178	0.449	0.300	0.500	0.200	0.400
APAC_Ⅰ	0.169	0.442	0.300	0.500	0.200	0.400
APAC_Ⅱ	0.155	0.434	0.300	0.500	0.200	0.400
BLEU	−0.199	0.184	−0.100	0.200	0.000	0.200
NIST	−0.471	−0.013	−0.300	0.100	−0.200	0.000

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
IMPACT	0.661	0.587	0.632	0.583	0.468	0.429
APAC_Ⅰ	0.669	0.593	0.641	0.591	0.476	0.436
APAC_Ⅱ	0.684	0.604	0.657	0.603	0.491	0.447
BLEU	0.399	0.415	0.329	0.389	0.231	0.276
NIST	0.301	0.342	0.257	0.335	0.180	0.235

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
IMPACT	0.849	0.919	0.838	0.824	0.670	0.651
APAC_Ⅰ	0.849	0.918	0.838	0.824	0.670	0.651
APAC_Ⅱ	0.841	0.921	0.838	0.824	0.670	0.651
BLEU	0.730	0.881	0.567	0.552	0.498	0.440
NIST	0.688	0.874	0.577	0.549	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
IMPACT	0.593	0.598	0.577	0.582	0.429	0.433
APAC_I	0.608	0.609	0.594	0.597	0.442	0.445
APAC_II	0.617	0.619	0.612	0.613	0.457	0.459
BLEU	0.446	0.470	0.435	0.463	0.315	0.338
NIST	0.412	0.507	0.387	0.431	0.279	0.314

表 5　WMT14 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.
IMPACT	**0.963**	0.831	0.921	0.962	**0.817**	**0.899**
APAC_I	0.962	0.826	0.915	0.968	**0.817**	0.898
BLEU	0.952	**0.832**	**0.956**	0.909	0.789	0.888
APAC_II	**0.963**	0.817	0.790	**0.982**	0.816	0.874

表 6　WMT14 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.
IMPACT	0.363	**0.275**	0.287	**0.205**	**0.278**	**0.282**
APAC_I	**0.364**	**0.275**	0.287	0.200	0.277	0.281
APAC_II	**0.364**	0.271	**0.288**	0.198	0.276	0.279
sentBLEU	0.352	0.261	0.257	0.189	0.249	0.262

表 7　WMT14 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-de(18)	en-hi(12)	en-cs(10)	en-ru(9)	Avg.
IMPACT	0.948	**0.373**	0.958	**0.976**	**0.929**	**0.837**
APAC_I	0.948	0.372	0.954	**0.976**	**0.929**	0.836
APAC_II	**0.950**	0.346	0.940	0.973	**0.929**	0.828
BLEU	0.937	0.216	**0.973**	**0.976**	0.915	0.803

表 8　WMT14 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.
IMPACT	**0.254**	0.211	**0.213**	0.291	**0.390**	**0.272**
APAC_I	0.253	**0.212**	0.209	**0.293**	0.388	0.271
APAC_II	0.253	0.210	0.203	0.292	0.388	0.269
sentBLEU	0.239	0.193	0.197	0.272	0.368	0.254

4.2.4.4　考察

表 1 から表 4 より、NTCIR-7 データにおいては system-level の相関係数は IMPACT、APAC_I、そして、APAC_II の間で大きな差はない。Pearson の相関係数に多少の差があり、IMPACT が若干高い相関係数を示しているが、その差は小さい。それに対して、segment-level では、全てにおいて、APAC_II が IMPACT、APAC_I よりも高い相関係数を示した。また、APAC_I においても全ての相関係数が IMPACT よりも高い値を示しており、文の長さに基づく重みづけが有効に働いたと考えられる。

表 5 から表 8 より、WMT14 Metrics Task においては、IMPACT が最も高い相関係数を示し、APAC_II が最も低い相関係数を示した。しかし、表 5 の多言語から英語の system-level においては、IMPACT と APAC_II との間の“Avg.”の差は約 0.025 であるが、その他の“Avg.”の差は 0.01 以下となっている。したがって、基本的には文の長さに基づく重みづけによる変化は見られなかったと考えられる。

表 1 から表 8 より、文の長さに基づく重みづけの導入は、segment-level の評価精度の向上に有効であったといえる。WMT14 Metrics Task においては提案手法による大きな変化は見られなかったが、NTCIR-7 データにおいては、提案手法の有効性を確認できた。

4.2.5　まとめ

本報告では、文の長さを考慮したチャンクに基づく自動評価法 APAC を提案した。提案手法ではチャンクに基づく自動評価法が短い文に対して過度に低いスコアを付与するという問題点を解決するために、短い文に対しては不一致単語の重みを軽減することを目的として重みづけを行っている。本報告では自動評価法 IMPACT に対して重みづけを導入し、有効性を検証するための性能評価実験を行った。その結果、NTCIR-7 データの segment-level において評価精度の向上が見られた。

今後は、更なる評価精度、特に system-level の評価精度の向上のための自動評価法についての研究に取り組む予定である。

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] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul(2006) "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, pp.223-231.

[4] Chin-Yew Lin and 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 2014), pp.605-612.

[5] Satanjeev Banerjee and 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.

[6] Hiroshi Echizen-ya and 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.

[7] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada(2010) “Automatic Evaluation of Translation Quality for Distant Language Pairs,” Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 944–952.

[8] Hiroshi Echizen’ya, Kenji Araki, Eduard Hovy(2013) “Automatic Evaluation Metric for Machine Translation that is Independent of Sentence Length,” Proceedings of Recent Advances in Natural Language Processing, pp.230-236.

[9] Hiroshi Echizen’ya, Kenji Araki, Eduard Hovy(2014) “Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.381-386.

[10] 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.

[11] Matouš Macháček, Ondřej Bojar(2014) “Results of the WMT14 Merics Shared Task,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.293-301.

[12] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,” http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[13] “TrueSkill,” http://en.wikipedia.org/wiki/TrueSkill

5. 拡大評価部会活動報告

５．１　拡大評価部会の活動概要

山梨英和大学　江原暉将

一昨年度から本研究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進めてきた[1][2]。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。

・2014 年 5 月 9 日　今年度の活動計画の策定
・2014 年 10 月 24 日　中間報告と今後の活動内容についての議論
・2015 年 1 月 23 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、NTCIR-9[3]の英日、日英のデータを対象にして、クラウドソーシングによる評価を実施した。これは WAT2014[4]での評価手法を特許文書に適用したものであり、同手法の有効性を考察した。自動評価に関しては、昨年と同様に WMT[5]および NTCIR における機械翻訳結果に対して各種の自動評価を適用し人手評価との相関を分析した。今年度は部会員が考案した新しい自動評価手法を組み入れメタ評価した。テストセットに関しては、中国語から日本語への機械翻訳評価を対象に、テスト文と設問パターンを拡充した。また、WEB ベースで自動的にテストができるシステムにテストセットのデータを組み込んだ。

参考文献

[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*、chap. 6、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*、chap. 6、pp.61-82、2014 年 3 月。

[3] Isao Goto et al. : Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, *Proceedings of NTCIR-9 Workshop Meeting*, pp.559-578, Dec, 2011.

[4] Toshiaki Nakazawa et al. : Overview of the 1st Workshop on Asian Translation, Proceedings of the 1st Workshop on Asian Translation, pp.1-19, Oct, 2014.

[5] Ondrej Bojar et al. : Findings of the 2014 Workshop on Statistical Machine Translation, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, June, 2014.

５．２　NTCIR-7 PATMT データの日本文データ及び WMT14 METRICS TASK データを用いた自動評価法のメタ評価

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
岡山県立大学　磯崎　秀樹
山梨英和大学　江原　暉将

5.2.1　はじめに

本研究会では、これまで翻訳対象文を英文に限定した自動評価法の性能評価について述べてきた[1]。英語は世界中で最も広く利用されている言語であり、かつ、他の言語に比べ検証が比較的容易であるとの考えからである。しかし、自動評価研究の発展に伴い、英語以外の言語についても自動評価法に必要なシステム訳、参照訳、そして、人手評価などのデータが充実してきており、メタ評価を容易に実施できる状況にある。個々の自動評価法の詳細な検証は英語に比べ困難であると考えられるが、メタ評価の結果のみからであっても新たな知見を得ることは可能である。そのような観点より、英語から多言語についてもメタ評価を実施することとした。本報告では、NTCIR-7 PATMT データ[2]を用いた英語から日本語への翻訳に対する自動評価及び WMT14 Metrics Task データ[3]を用いた英語からヨーロッパ言語への翻訳とヨーロッパ言語から英語への翻訳に対する自動評価について、複数の自動評価法を用いて行ったメタ評価の結果について述べる。

5.2.2 NTCIR-7 PATMT データを用いたメタ評価

本報告では、NTCIR-7 PATMT データの英語から日本語へのシステム訳を対象に、自動評価法に対するメタ評価を行った。自動評価法には IMPACT[4]と RIBES[5][6][7]、そして、改良版 RIBES[8][9]を用いた。NTCIR-7 PATMT データでは 5 つの機械翻訳システムにより得られた英日翻訳のシステム訳がそれぞれ 100 文存在する。参照訳にはシステム訳 100 文に対応した人手による正解訳 100 文が提供されており、それらを使用した。なお、自動評価法は単語単位で処理を行っているため、日本文であるシステム訳と参照訳を分かち書き文に変換する必要がある。そこで、本報告では、日本文に対する分かち書きは日本語形態素解析器 MeCab（MeCab 0.996）[10]を用いて行った。人手評価は adequacy と fluency の 2 つの観点より 3 名の評価者が 5 段階による絶対評価を行った結果が提供されている。本報告では、メタ評価を実施するにあたり、3 名の評価値の平均値をシステム訳に対する最終的な人手評価として使用した。

次いで、メタ評価に用いた IMPACT と RIBES、そして、改良版 RIBES について簡単に紹介する。IMPACT は大きく 2 つの特徴を有する自動評価法である。IMPACT は一致単語が連続する単語列、即ち、チャンクに基づきスコアを算出するが、文が長くなるほどチャンクを決定する際に多義性が生じ、一意に決定することが困難となる。そこで、IMPACT の 1 つ目の特徴として、

チャンクの長さと文全体におけるチャンクの位置を考慮し、文に対する大局的観点より一意にチャンクを決定することが挙げられる。その際には以下の式(1)より算出される *score* の値が大きいチャンク列が選択される。*length(c)*は 1 つのチャンクの長さ、即ち、単語数を意味する。*c_num* は文中におけるチャンクの数である。また、β はチャンクの長さに対する重みづけパラメータであり、デフォルト値は 1.2 である。式(2)は個々のチャンクのシステム訳と参照訳における相対位置のずれを表している。ずれが大きいほど *pos* の値は小さくなり、ずれが小さいほど *pos* の値は大きくなる。式(2)の c_i と c_j はそれぞれ参照訳とシステム訳における個々のチャンクの先頭単語の位置である。また、*m* と *n* は、参照訳とシステム訳それぞれの構成単語数である。このようなチャンク列の決定処理は全ての一致単語について再帰的に行われる。

$$score = \left(\sum_{c \in c\_num} \left(length(c)^{\beta} \times pos \right) \right)^{\frac{1}{\beta}} \quad (1)$$

$$pos = \left(1.0 - \left| \frac{c_i}{m} - \frac{c_j}{n} \right| \right) \quad (2)$$

2 つ目の特徴は、システム訳と参照訳間におけるチャンクの出現順の違いをどのようにスコアに反映させるかをパラメータで柔軟に制御できる点にある。チャンクが一意に決定された後には、以下の式(3)から(5)に基づきスコアを算出する。式(3)は参照訳に対するスコア、式(4)はシステム訳に対するスコアである。そして、式(5)がスコアの計算式であり、*R* と *P* の F 値を求めることで得られる。式(3)と(4)の α は語順をスコアに反映させるためのパラメータである。α の値が大きければ、異なる語順のチャンク列であっても区別されることなくスコアに反映される。即ち、語順の違いを考慮しない評価となる。それに対して α の値が小さい場合には、異なる語順のチャンク列はスコアに反映されなくなる。即ち、語順の違いに厳しい評価となる。パラメータ α のデフォルト値は 0.1 である。また、*RN* はチャンク列の決定の再帰処理の回数を示している。更に式(5)の γ は *P/R* より得られる値である。IMPACT が出力するスコアは 0.0～1.0 であり、大きな値ほど評価が高い。

$$R = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \sum_{c \in c\_num} length(c)^{\beta} \right)}{m^{\beta}} \right)^{\frac{1}{\beta}} \quad (3)$$

$$P = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \sum_{c \in c\_num} length(c)^{\beta} \right)}{n^{\beta}} \right)^{\frac{1}{\beta}} \quad (4)$$

$$IMPACT\ score = \frac{(1+\gamma^2)RP}{R+\gamma^2 P} \quad (5)$$

RIBES はシステム訳と参照訳間の語順の近さを測定することでスコアを出力する。英日間のよ

うに語順の大きく異なる言語間の自動評価において、人手評価との強い相関が得られる。語順の近さの測定には Kendall の τ を用いている。更に、一致単語に対する適合率を Kendall の τ に対する重みづけとして用いている。RIBES が出力するスコアは 0.0～1.0 であり、IMPACT と同様に大きな値ほど評価が高い。しかし、このようなスコアの算出方法では、システム訳が明らかに誤っていても文が非常に短い場合、高い適合率が得られスコアが高くなることがある。そこで、スコアに対してペナルティを付与し、システム訳が短く、参照訳との間で文の長さが大きく異なる場合にスコアを抑える。RIBES のスコアは以下の式(6)と(7)より得られる。式(7)の NKT は Kendall の τ を 0.0～1.0 に正規化したものである。式(6)の P は適合率、BP は文の長さに対するペナルティである。これらの重みづけに用いられるパラメータ α と β のデフォルト値はそれぞれ 0.25 と 0.10 である。

$$\text{RIBES} \overset{\text{def}}{=} \text{NKT} \quad P^{\alpha} \quad \text{BP}^{\beta} \quad (6)$$

$$NKT \overset{\text{def}}{=} \frac{\tau + 1}{2} \quad (7)$$

更に、改良版 RIBES では、日本語の語順が比較的自由で、「スクランブリング」と呼ばれる、同じ意味の別の語順に対して、人間は高得点を与えるのに、語順を評価する RIBES が低いスコアを与える問題に対処している。

この手法では、日本語参照訳を係り受け解析し、その係り受け木に基づいて、意味が変わらない語順の文だけを自動生成している。こうして生成される文を新たな参照訳として、それぞれの参照訳に対する RIBES スコアの最大値を、その訳文のスコアとするものであり、特に segment-level での評価精度の向上を図っている。

5.2.3 WMT14 Metrics Task データを用いたメタ評価

本報告では更に、WMT14 Metrics Task データの多言語から英語へのシステム訳及び英語から多言語へのシステム訳を用いて自動評価法に対するメタ評価を行った。使用した自動評価法は WMT14 Metrics Task においてベースラインとして使用された BLEU[11]、NIST[12]、TER[13]、WER[14]、PER[15]、CDER[16]、そして、sentBLEU をベースラインとして用いた。BLEU と NIST については　mteval-13a.pl”を使用し、TER、WER、PER、CDER、そして、sentBLEU については統計翻訳の翻訳エンジンとして広く利用されている mosesdecoder[17]に含まれているシステムを用いた。ただし、sentBLEU は segment-level のメタ評価のみに使用される。これらのベースラインに加え、上述した IMPACT、RIBES、そして、APAC[18]を使用した。

APAC は IMPACT の改良版である。IMPACT と同様に、チャンクに基づきスコアを計算するが、IMPACT では文が短い場合に不一致単語の比重が大きくなるという問題点を解決するために、文の長さに応じて不一致単語の重みを自動的に変更することができる。即ち、文が短い場合には、不一致単語の比重を軽くすることで過度にスコアが小さくなることを避けている。

メタ評価実験を行うにあたり使用したシステム訳、参照訳、人手評価について述べる。システ

ム訳はWMT14のTranslation Taskの参加システムが出力したシステム訳が提供されているため、それらを用いた。Translation Taskに提出された機械翻訳システムの数は110である。これら110の機械翻訳システムの内訳はcs(Czech)-en(English)への翻訳を行ったものが5システム、de(German)-enへの翻訳を行ったものが13システム、fr(French)-enへの翻訳を行ったものが8システム、hi(Hindi)-enへの翻訳を行ったものが9システム、ru(Russian)-enへの翻訳を行ったものが13システム、en-csへの翻訳を行ったものが10システム、en-deへの翻訳を行ったものが18システム、en-frへの翻訳を行ったものが13システム、en-hiへの翻訳を行ったものが12システム、そして、en-ruへの翻訳を行ったものが9システムの計110システムである。これらの機械翻訳システムのシステム訳をメタ評価に使用した。使用したシステム訳の数はcs-enが15,015文、de-enが39,039文、fr-enが24,024文、hi-enが22,563文、ru-enが39,039文、en-csが30,030文、en-deが49,266文、en-frが39,039文、en-hiが30,084文、そして、en-ruが27,027文の計315,126文である。

参照訳はMetrics Taskより提供されたものを用いた。内訳はcs-enが3,003文、de-enが3,003文、fr-enが3,003文、hi-enが2,507文、ru-enが3,003文、en-csが3,003文、en-deが2,737文、en-frが3,003文、en-hiが2,507文、en-ruが3,003文である。

次いで、人手評価について述べる。システム訳に対する人手評価には、評価者が原文と参照訳を用いて5つのシステム訳に対して1から5までの5段階の相対評価を実施しており、その結果を用いた。その際には、良いシステム訳ほど小さな値が付与される。

5.2.4 メタ評価実験

5.2.4.1 実験方法

NTCIR-7 PATMTデータを用いたメタ評価実験は自動評価法より得られたスコアと人手評価の評価値（1～5）の間のスコアとの相関係数を求めることで行った。その場合、system-levelとsegment-levelの両方について相関係数を求めた。人手評価のsystem-levelの評価値には全segment-levelの評価値の平均値を用いた。自動評価法の評価は、system-levelは5つの自動評価法と人手評価のスコア、segment-levelは500（=5×100）の自動評価法と人手評価のスコアに基づき相関係数を算出することで得られる。相関係数については、Pearsonの相関係数、Spearmanの順位相関係数、Kendallの順位相関係数の3つの相関係数を求めた。

WMT14 Metrics Taskデータを用いたメタ評価実験でもsystem-levelとsegment-levelのそれぞれについて相関係数を求めた。system-levelの相関係数は自動評価法によるスコアと人手評価による評価値との間でPearsonの相関係数を求めた。その際、人手評価の評価値は前述した相対評価のスコアを用いてTrueSkill[19]より算出したものを用いた。TureSkillはMicrosoft Researchが開発した、ベイズ理論に基づくランキングアルゴリズムである。このTrueSkillに基づき、system-levelの人手評価の評価値が全機械翻訳システム110に対して得られる。したがって、system-levelでは110の自動評価法と人手評価のスコアに基づきPearsonの相関係数を算出することになる。次いで、segment-levelの相関係数は自動評価法によるスコアと人手評価による評価

値との間で Kendall の順位相関係数を求めた。以下に Kendall の τ の計算式を示す。

$$\tau = \frac{Concordant - Discordant}{Concordant + Discordant} \quad (8)$$

segment-level では、自動評価法のスコアと人手評価のスコアそれぞれにおいて常に 2 つのスコアを比較し、スコアの大小関係が一致した場合には、*Concordant* がカウントアップされ、一致しない場合には *Discordant* がカウントアップされる。例えば、システム訳のスコア A と B があり、人手評価はスコア B が高く、自動評価法もまたスコア B が高い場合、一致となり *Concordant* がカウントアップされる。このような大小比較の結果、同じスコアであった場合は、WMT14 においては、分母のみにその数を反映させている。したがって、等しいスコアが多数出現する場合には、相関係数は低くなる。

また、自動評価法においては、BLEU と NIST ではオプションとして“- -international -tokenization”を付与してスコアを出力している。IMPACT（ver4.0.2）と RIBES（ver1.03.1）では生データをそのまま使用している。パラメータ値はデフォルトを用いている。

5.2.4.2 実験結果

メタ評価における実験結果を表 1 から表 7 に示す。表 1 と表 2 は NTCIR-7 データにおける英日翻訳の system-level と segment-level の相関係数である。相関係数を求める際に使用したスコアの数は前述したとおり、system-level で 5、segment-level で 500 である。表 3 は機械翻訳システムごとに adequacy の segment-level について求めた、Spearman の順位相関係数である。使用した自動評価法は IMPACT、RIBES、そして、先述した改良版 RIBES である。しかし、IMPACT と RIBES 及び改良版 RIBES では、異なる動作環境で実験を行っているためメタ評価実験の結果という位置づけではなく、あくまでの一つの指標として掲載している [i]。

表 4 と表 5 はそれぞれ WMT14 Metrics Task データにおける多言語から英語への翻訳と英語から多言語への翻訳の system-level の Spearman の順位相関係数である。 ()” 内の数値は相関係数を求める際に使用したスコアの数である。また、表 6 と表 7 はそれぞれ WMT14 Metrics Task データにおける多言語から英語への翻訳と英語から多言語への翻訳の segment-level の Kendall の順位相関係数である。 ()” 内の数値は自動評価法のスコアと人手評価のスコアの大小比較をした際のペアの数である。更に、表中の“wmt13”は WMT13[20]に準拠し、自動評価法のスコアと人手評価のスコアが等しい場合にスコアに反映させない方法で Kendall の τ を求めたものである。即ち、スコアが等しい場合が多数出現すると相関係数が高くなる。太字の数値は自動評価法の中で最も相関係数が高いことを示している。

表 1　NTCIR-7 データにおける英語から日本語翻訳の system-level の相関係数

	Pearson		Spearman		Kendall	
	adequacy	fluency	adequacy	fluency	adequacy	fluency
IMPACT	0. 178	0. 449	0. 300	0. 500	0. 200	0. 400
RIBES	0. 914	0. 742	0. 900	0. 800	0. 800	0. 600

表 2　NTCIR-7 データにおける英語から日本語翻訳の segment-level の相関係数

	Pearson		Spearman		Kendall	
	adequacy	fluency	adequacy	fluency	adequacy	fluency
IMPACT	0.661	0.587	0.632	0.583	0.468	0.429
RIBES	0.600	0.493	0.608	0.507	0.448	0.368

表 3　NTCIR-7 データにおける機械翻訳システムごとの adequacy の segment-level での Spearman の相関係数

	tsbmt	moses	NTT	NICT-ATR	kuro	Avg.
IMPACT	0.460	0.718	0.736	0.716	0.539	0.634
RIBES	0.439	0.607	0.692	0.582	0.515	0.567
改良版 RIBES	0.452	0.670	0.727	0.608	0.550	0.601

表 4　WMT14 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.
IMPACT	0.963	0.831	0.921	0.962	0.817	**0.899**
BLEU	0.952	0.832	**0.956**	0.909	0.789	0.888
APAC	0.963	0.817	0.790	0.982	0.816	0.874
CDER	0.964	0.834	0.786	0.940	**0.820**	0.869
NIST	0.955	0.811	0.784	0.983	0.800	0.867
RIBES	**0.965**	0.821	0.706	0.989	**0.820**	0.860
TER	0.957	0.762	0.753	0.986	0.818	0.855
WER	0.957	0.759	0.745	**0.990**	0.818	0.854
PER	0.951	**0.848**	0.659	0.969	0.799	0.845

表 5　WMT14 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-de(18)	en-hi(12)	en-cs(10)	en-ru(9)	Avg.
CDER	0.948	0.428	0.953	0.973	0.932	**0.847**
RIBES	**0.965**	0.366	0.941	0.976	**0.952**	0.840
IMPACT	0.948	0.373	0.958	0.976	0.929	0.837
APAC	0.950	0.346	0.940	0.973	0.929	0.828
PER	0.940	0.273	0.985	0.979	0.927	0.821
NIST	0.941	0.200	**0.981**	**0.985**	0.927	0.807
BLEU	0.937	0.216	0.973	0.976	0.915	0.803
TER	0.954	0.432	0.640	0.972	0.935	0.787
WER	0.957	**0.438**	0.543	0.971	0.937	0.769

表 6　WMT14 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.	wmt13
IMPACT	0.363	**0.275**	0.287	**0.205**	**0.278**	**0.282**	**0.292**
APAC	**0.364**	0.271	**0.288**	0.198	0.276	0.279	0.290
RIBES	0.359	0.258	0.244	0.196	0.258	0.263	0.280
sentBLEU	0.352	0.261	0.257	0.189	0.249	0.262	0.272

表 7　WMT14 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.	wmt13
IMPACT	**0.254**	**0.211**	**0.213**	0.291	**0.390**	**0.272**	**0.288**
APAC	0.253	0.210	0.203	**0.292**	0.388	0.269	0.285
sentBLEU	0.239	0.193	0.197	0.272	0.368	0.254	0.268
RIBES	0.237	0.191	0.168	0.252	0.362	0.242	0.268

5.2.4.3　考察

表 1 より NTCIR-7 PATMT データの system-level では、RIBES の相関係数は Kendall の順位相関係数を除き、比較的高い相関係数を示している。それに対して、IMPACT の相関係数は低く、特に adequacy の相関係数が非常に低い。system-level においては、機械翻訳システムの数が 5 つと少ないため、1 つの自動評価法のスコアと人手評価のスコアとの間で評価にずれが生じると、大きく相関係数が低下すると考えられる。しかし、少ない機械翻訳システムにおいてこそ、確実に高い相関係数が得られなければならない。RIBES に関しては文献[9]より、adequacy においては Spearman の順位相関係数と Kendall の順位相関係数共に 1.000 の値となる。表 1 との違いは、MeCab による分かち書きにおいて、数字や句読点などに対する正規化の差が原因と考えられる。また、表 2 より segment-level においては、IMPACT の相関係数が RIBES の相関係数よりも高い値を示している。adequacy の相関係数に比べ、fluency の相関係数の方が IMPACT と RIBES との間で差が大きい。RIBES は adequacy に重点を置いた手法であることが影響している可能性がある。

表 3 は厳密なメタ評価結果ではないが、機械翻訳システムごとに見ると、いずれの自動評価法も同様の傾向を示していると言える。具体的には、いずれの自動評価法においても、NTT、moses、NICT-ATR、kuro、tsbmt の順で相関係数が高い値から低い値へと推移している。これらの機械翻訳システムのアーキテクチャは NTT、moses、NICT-ATR が統計翻訳、kuro が用例翻訳、tsbmt がルールベース翻訳である。したがって、いずれの自動評価法においても、統計翻訳の評価精度は比較的高いが、ルールベース翻訳の評価精度は低いといえる。このような傾向はベースラインとして広く使用されている BLEU や NIST の問題点として従来より指摘されているが、今回の結果を見る限りにおいては、他の自動評価法も同様の問題点を抱えていると考えられる。

WMT14 Metrics Task データでは、表 4 より多言語から英語の system-level の相関係数は IMPACT が最も高かった。また、表 5 より英語から多言語の system-level では CDER が最も高い相関係数を示した。他の自動評価法についても system-level においては比較的高い相関係数を示しているといえる。しかし、表 5 において、英語からドイツ語の相関係数は 0.5 をいずれの自動評価法も下回っており、他の言語間に比べ非常に低い。直感的には、機械翻訳システムの数が多く、かつ、性能が拮抗している場合、それらに対する評価は困難になると考えられる。そのような観点で言えば、英語からドイツ語の機械翻訳システムの数 18 であり、他の言語間の機械翻訳システム数より多い。次いで、機械翻訳システムの性能がどの程度拮抗しているのかを確認するために system-level における人手評価の評価値に対する標準偏差を求めた。その結果を以下の表 8 に示す。

表 8　WMT14 データにおける英語−ドイツ語機械翻訳システムに対する人手評価の標準偏差

	en-fr(13)	en-de(18)	en-hi(12)	en-cs(10)	en-ru(9)
標準偏差	0.068	0.042	0.148	0.138	0.219

表 8 より、英語からドイツ語への翻訳における機械翻訳システムの標準偏差が最も小さい。即ち、人手評価の評価値のばらつきが小さく、最も機械翻訳システムの性能が拮抗していると考えられる。これらのことが要因となり、英語-ドイツ語の相関係数が低くなった可能性は高い。しかし、英語からフランス語への翻訳においても表 8 の標準偏差は同様に小さく、かつ、機械翻訳システムの数も 13 と英語-ドイツ語に続き多い。それにもかかわらず、表 5 において、英語-フランス語の相関係数は英語-ドイツ語を除く他の言語の相関係数と変わらず 0.9 以上となっている。したがって、機械翻訳システムの数と性能の近さだけではなく、他の要因も影響していると考えられる。この点については更なる精査が必要である。

segment-level の評価精度においては、表 6, 7 より IMPACT が最も高い相関係数を示した。しかし、文献[3]より多言語から英語では DISCOTK-PARTY-TUNED[21]が 0.386、英語から多言語では BEER[22]が 0.319 と最も高い相関係数を示しており、IMPACT の 0.281 と 0.272 を大きく上回っている。したがって、相対的に IMPACT の評価精度は不十分である。今回のメタ評価の結果、segment-level においては、いずれの自動評価法も相関係数は低く、今後取り組まなければならない非常に重要な課題である。

5.2.5　まとめ

本報告では、NTCIR-7 PATMT データの日本文データと WMT14 Metrics Task データを用いてメタ評価実験を実施し、その結果について述べた。評価対象を従来の英語だけではなく、日本語やヨーロッパ言語を用いてメタ評価を行った。その結果、英語を対象としたメタ評価と同様に、system-level の評価精度は比較的高いが、segment-level の評価精度は不十分であることを確認した。また、NTCIR-7 PATMT データにおける日本語を対象とした segment-level のメタ評価より、ルールベース翻訳に対する評価精度が統計翻訳に対する評価精度に比べ、低いことを確認した。

更に、WMT14 Metrics Task データより、IMPACT、RIBES 共に segment-level での相関係数が相対的にも絶対値としても不十分であることを確認した。

今後は、より多くのデータ及び自動評価法を用いたメタ評価実験を行うことで、自動評価法の現状と課題を明らかにする予定である。

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] 越前谷博, 須藤克仁, 磯崎秀樹, 江原暉将(2014) “NTCIR PATMT データと WMT METRICS TASK データにおける英文データを対象とした自動評価法のメタ評価,” 平成 25 年度 AAMT/Japio 特許翻訳研究会報告書, pp.62-70.

[2] 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.

[3] Matouš Macháček, Ondřej Bojar(2014) “Results of the WMT14 Metrics Shared Task,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.293-301.

[4] Hiroshi Echizen-ya and 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.

[5] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada(2010) “Automatic Evaluation of Translation Quality for Distant Language Pairs,” Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 944–952.

[6] 平尾努,磯崎秀樹,Kevin Duh,須藤克仁,塚田元,永田昌明(2011) “RIBES: 順位相関に基づく翻訳の自動評価法,” 言語処理学会 第 17 回年次大会発表論文集, pp.1115-1118.

[7] 平尾努, 磯崎秀樹, 須藤克仁, Duh Kevin, 塚田元, 永田昌明(2014) “語順の相関に基づく機械翻訳の自動評価法,” 自然言語処理, Vol.21, No.3, pp.421-444.

[8] 高地なつめ, 磯崎秀樹(2014) “スクランブリングを考慮した和訳の自動評価法の NTCIR-9 データによる検証,” 情報処理学会研究報告, Vol.2014-NL-219, No.2, pp.1-5.

[9] Hideki Isozaki, Natsume Kouchi, Tsutomu Hirano(2014), “Dependency-based Automatic Enumeration of Semantically Equivalent Word Orders for Evaluating Japanese Translations,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.287-292.

[10] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,” http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[11] 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), pp. 311-318.
[12] 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
[13] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul(2006) "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas, pp.223-231.
[14] Stanley F. Chen, Douglas Beeferman, Roni Rosenfield(1998) “Evaluation Metrics for Language Models,” In DARPA Broadcast News Transcription and Understanding Workshop.
[15] Christoph Tillmann, Stephan Vogel, Hermann Ney(1997) “Accelerated DP based search for statistical translation,” Proceedings of the Fifth European Conference on Speech Communication and Technology, pp.2667-2670.
[16] Gregor Leusch, Nicola Ueffing, Hermann Nry(2006) “CDER: Efficient MT Evaluation Using Block Movements,” Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp.241-248.
[17] 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," Proceedings of the ACL 2007 Demo and Poster, pp.177-180.
[18] Hiroshi Echizen’ya, Kenji Araki and Eduard Hovy(2014) “Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.381-386.
[19] “TrueSkill,” http://en.wikipedia.org/wiki/TrueSkill
[20] Matouš Macháček, Ondřej Bojar(2013), “Results of the WMT13 Metrics Shared Task,” Proceedings of the Eighth Workshop on Statistical Machine Translation, pp.45-51.
[21] Shafiq Joty, Francisco Guzmán, Lluís Màrquez, Preslav Nakov(2014) “DiscoTK: Using Discourse Structure for Machine Translation Evaluation,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.402-408.
[22] Miloš Stanojević, Khalil Sima’an(2014) “BEER: Better Evaluation as Ranking,” Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.414-419.

---

i) 表 1 は文献[5]の結果と一部一致していない。動作環境の違いが影響している可能性がある。

５．３　クラウドソーシングを利用した特許翻訳評価の可能性の検討

京都大学　中澤敏明
NHK 放送技術研究所　後藤功雄
（株）東芝研究開発センター　園尾聡

5.3.1　はじめに

機械翻訳結果の精度を正確に測るためには人手評価は必要不可欠だが、人手評価には多くの費用と時間がかかる。さらに特許文のような専門性の高いドメインを対象とする場合には、評価の正確性を保証するために専門家を用意する必要もある。本稿では特許ドメインの機械翻訳評価を、クラウドソーシングを利用して安価かつ高速に、ある程度の信頼性を保ちつつ行うことが可能かを検討する。調査対象として、NTCIR-9 特許翻訳タスク[1]において日英、英日翻訳に参加したチームの翻訳結果およびその人手評価結果を利用する。

5.3.2　データセット

NTCIR-9 特許翻訳タスクでは人手評価の指標として Adequacy と Acceptability の 2 つを利用した。前者は翻訳文の正確さを 1 から 5 の 5 段階で評価するものであり、後者はさらに目的言語文としての自然さも評価するもので、こちらも AA、A、B、C、F の 5 段階で評価している。各システムの翻訳結果は、3 人の評価者によりそれぞれ 100 文ずつ、計 300 文ずつが評価された。Acceptability 評価は Adequacy 評価が行われたシステムの一部にしか行われていないため、本稿では Acceptability 評価が行われたシステム（日英 11 システム、英日 8 システム）の翻訳結果および評価結果を調査対象データセットとして利用した。

5.3.3　クラウドソーシングを利用した翻訳結果の一対比較

5.3.3.1　評価方法

クラウドソーシングとは、たくさんの小さな仕事（タスク）を不特定多数の人（ワーカー）に少しずつ行ってもらい、作業完了までにかかる時間と費用を低く抑える枠組みのことで、現在様々なプラットフォームが利用可能である。本稿ではクラウドソーシングサービスとして Lancers（http://www.lancers.jp）を利用した。ランサーズの特徴としてタスクのカテゴリーを設定することができるため、そのカテゴリーに興味のある人や専門知識のある人がタスクを行ってくれる可能性が高いと考えられる。今回はカテゴリーとして英語翻訳・英文翻訳を設定した。

クラウドソーシングを利用した翻訳評価は、WAT2014[2]と同じ方法で行った。簡単に説明すると、ベースラインとなる翻訳結果を 1 つ設定し、これと各システムの翻訳結果を文ごとにどちらが翻訳としてより適切か、もしくは同程度かを判定し、その勝敗をスコア化してシステムをランキングするものである。本稿ではベースラインとして、NTCIR-9 のオーガナイザーが用意した BASELINE1（Moses を使ったフレーズベース SMT）を利用した。

クラウドソーシングのワーカーは特許翻訳の専門家ではないため、一人だけの評価ではその信頼性に疑問が残る。そこで同じ評価を複数のワーカーに行ってもらい、その判定を集約すること

で、信頼性を向上することが一般的に行われる。本稿では 1 つの文ペアの判断を 5 人の異なるワーカーに行ってもらった。5 人の結果をどのように集約し、最終判断をどのように決定するかは検討の余地がある。本稿でもいくつかの設定で実験を行ったが、より詳細な検討は今後の課題である。

各文ペアの判断が決定したら、ベースラインに対する勝ち数を W、負け数を L、引き分けの数を Tとして以下の式で各システムの評価をスコア化する。

$$Score = 100 \times \frac{W - L}{W + L + T}$$

5.3.3.2 タスクの難易度およびワーカーの信頼性を考慮した判定

Whitehill ら[3]は各ワーカーによる判定を観測変数として、各ワーカーの能力やタスクの難易度および各タスクの真の正解を隠れ変数としてモデル化し、これを EM アルゴリズムにより推定する方法を提案した。本稿でもこの方法を利用することで、5 人の判定を集約することを試みた。この方法を利用すると、各タスクの判定が確率値で推定されるため、この値をそのまま勝敗として利用した。たとえばある文の評価がベースラインと比較して良い確率、悪い確率、同程度の確率が順に 0.7、0.2、0.1 と推定された場合、この文の勝敗は 0.7 勝 0.2 敗 0.1 分として計算する。

5.3.3.3 クラウドソーシング費用

本稿では一人のワーカーが一つの文ペアを判断するための費用を 5 円と設定した。各文ペアは 5 人の異なるワーカーが評価するため、1 文ペアの評価にかかる費用は 25 円である。評価対象文は 1 システムにつき 300 文（実際にはとても長い 1 つの文を除いた 299 文）であるので、1 つのシステムの翻訳結果の評価にかかる費用は 7500 円である。評価対象システムはベースラインを除いて日英で 10 システム、英日で 7 システムであるので、本稿でのクラウドソーシング評価にかかった総費用は約 13 万円であった。これは専門家に評価を依頼するよりもはるかに低コストであり、また評価も 2 週間程度で完了したため、非常に高速である。

5.3.4 クラウドソーシング評価結果

図 1 に日英方向の、図 2 に英日方向のクラウドソーシング評価結果と自動評価結果である BLEU および RIBES との相関を示す。なおこの表では 5.3.3.2 章で述べた方法でクラウドソーシング評価スコアを計算している。なお、各ワーカーの判定を勝ちなら+1、負けなら-1、同程度なら 0 とし、5 人の判定を単純に足し合わせた結果、+3 以上なら最終判断を勝ち、-3 以下なら負け、それ以外を 0 としてクラウドソーシング評価スコアを計算する実験も行ったが、自動評価結果との相関に大きな違いは見られなかった。

図 3 はクラウドソーシング評価による各システムのランキングと NTCIR-9 の Adequacy および Acceptability によるランキングとの相関を示している。この結果から、翻訳精度の高いシステムはクラウドソーシング評価でも正確にシステムのランク付けが行えているが、翻訳精度の悪いシステムでは一致度が低いことがわかる。また日英翻訳結果よりも、英日翻訳結果のほうが一致

度が高い傾向にある。この原因は以下のようにさまざま考えられる：

- そもそもの翻訳精度が日英よりも英日のほうが高い。人間が翻訳結果を比較する場合には、精度の悪いもの同士を判断するより、良いものと悪いものや良い者同士を判断するほうが容易である。
- ワーカーの多くが日本語を母国語としているため、日英よりも英日のほうが正確に判定出来ている。
- NTCIR-9 では各文を評価したのは一人だけであり、その一人の主観で判断されているが、クラウドソーシング評価では複数人の意見を集約しているため、傾向が異なる。
- NTCIR-9 では絶対評価を行ったが、クラウドソーシング評価はベースラインに対する相対評価であり、傾向が異なる。
- Acceptability に関しては、多くの文が最低評価のＦと判定されており、たとえ判定Ｆの文同士に差があったとしても Acceptability 評価には反映されない。これに対し、クラウドソーシング評価ではＦ同士の文でも精度に差があれば優劣をつけており、傾向が異なる。

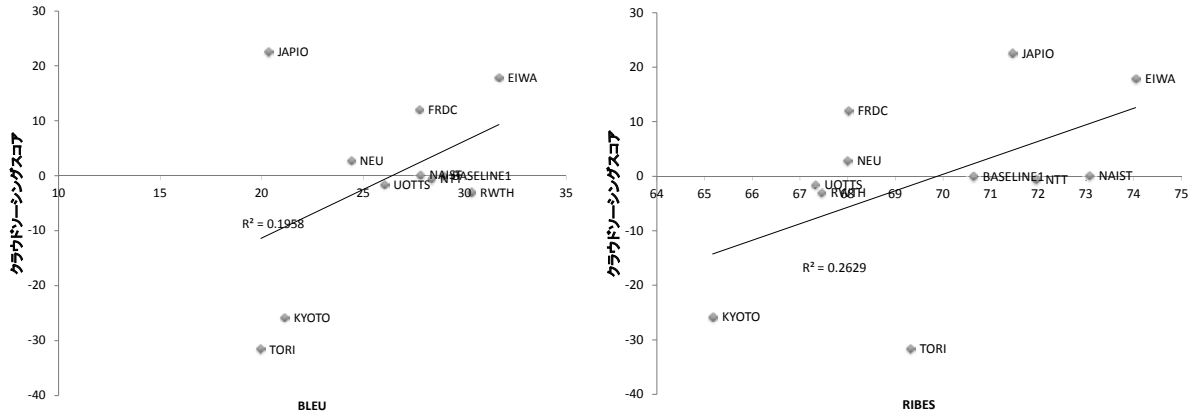

図 1：日英翻訳結果のクラウドソーシング評価と自動評価との相関

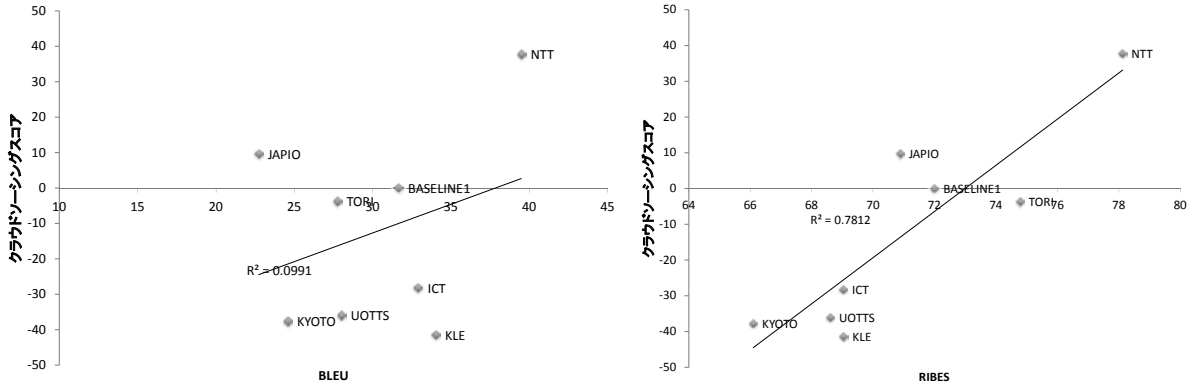

図 2：英日翻訳結果のクラウドソーシング評価と自動評価との相関

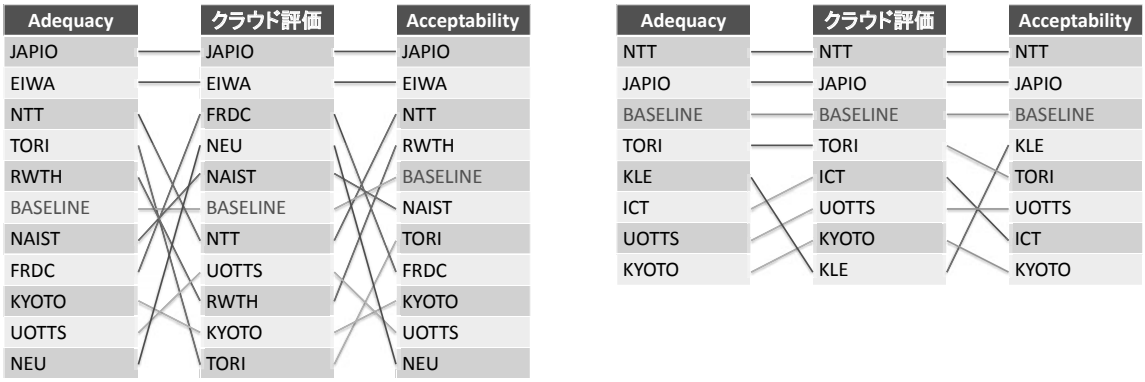

図 3：クラウドソーシング評価と NTCIR-9 での評価との相関（左：日英、右：英日）

5.3.5 クラウドソーシング評価の信頼性に対する検証、分析

クラウドソーシング評価と NTCIR-9 の人手評価を様々な観点から分析することで、クラウドソーシング評価の信頼性や有用性の検証を行った。

5.3.5.1 クラウドソーシング評価の分解能の検証

前節の結果から、全体的にみると、クラウドソーシング評価によるランキングと Adequacy や Acceptability によるランキングは必ずしも一致しないことがわかった。しかしながら、Adequacy で評価 5 の翻訳と評価 1 の翻訳や、Acceptability で評価 F の翻訳とそれ以外の評価の翻訳などは明らかに差があるはずで、クラウドソーシング評価でもこのような文ペアの判断は正確に行えると考えられる。そこで絶対評価の差とクラウドソーシング評価との関係を調査した。

各システムごとの分析結果を別紙に示す。Adequacy については、ベースライン翻訳の絶対評価に対する各システムの翻訳の絶対評価の相対値ごとに、クラウドソーシング評価の結果の分布を示した。Acceptability については、評価 F とそれ以外の評価との差が最も大きいと考え、各評価を評価 F かそうでないかの二値に分類し、ベースラインが評価 F で各システムの評価が F 以外なら+1、その逆なら-1、どちらも評価 F もしくはどちらも評価 F 以外ならば 0 として、やはりクラウドソーシング評価の結果の分布を示した。分析結果から、以下の 2 点が言える。

- 絶対評価の差が大きいものについては、クラウドソーシング評価でも正しい判断ができている場合が多い。つまり、明らかに質に差がある翻訳ペアについては、クラウドソーシング評価の信頼性は高い。逆にいえば、差が小さい翻訳ペアは人によってどの部分を評価するかが異なるため、評価が不安定になりやすい。
- 翻訳精度の低いシステムについては、明らかに差がある翻訳ペアのそもそもの数が少なくなってしまうため、クラウドソーシング評価の信頼性を検証するのに不適切であると考えられる。

5.3.5.2 クラウド評価における評価品質の検証

NTCIR-9 の adequacy の評価値を、評価対象システムの訳文の評価値とベースラインシステムの訳文の評価値との相対評価値に変換し、この相対評価値（Gold 評価値）を正解としてクラウド評価の評価品質を検証した。なおここでは、クラウドの複数の評価値を集約することを行わず、個別の評価値をそのまま用いる。Gold 評価値は、評価対象システムの訳文がベースラインシステムの訳文に対して、Lose（-1）、Tie（0）、Win（+1）の 3 値のいずれかである。

5.3.5.2.1 評価者毎の評価品質に対する検証

Gold 評価値で Tie になったデータを除いて Lose または Win になったデータのみを選択して検証に用いた。ここで、同じデータに対する２つの評価値において、一方は Lose でもう一方が Win の場合に、評価が相反すると言う。そして、評価相反率と正解率を以下のように定義する。

評価相反率＝（Gold 評価とクラウド評価が相反した評価数）／（Lose または Win の評価になったクラウド評価数）

正解率＝（Gold 評価とクラウド評価が一致した評価数）／（クラウド評価数）

評価相反率は、2 値の選択問題なので、ランダム選択の期待値は 0.5 である。選択したデータにおいて、10 以上の評価値があるクラウドの評価者について、評価者毎の評価相反率と正解率を計算した。

まず、Gold 評価値の信頼性について検証する。英日評価で最も正解率が高かった評価者の正解率は 0.958 で評価数は 166 であった。また、日英評価で最も正解率が高かった評価者の正解率は 0.859 で評価数は 305 であった。もし Gold 評価の信頼性が低く、間違った評価の割合が多いもの（例えばランダムな評価値など）であればこれだけの数のデータを評価してこれだけ高い正解率が得られることはほとんどあり得ない。このことは、Gold 評価は信頼性が高いことを示していると言える。

次に、クラウドの評価者毎の評価の品質を検証する。評価者毎の結果を図 4 に示す。２つの翻訳結果の訳質が Tie であるかそうでないかの基準は評価者によって異なると考えられるため、Gold 評価値は Win または Lose でクラウドの評価値は Tie になる場合は品質に問題があるとは考えない。一方で、Gold 評価値とクラウド評価値で評価が相反する場合は、クラウド評価の品質に問題があると考える。そのため、ここでは、特に評価相反率に着目して検証する。図 4 では、横軸に評価相反率、縦軸に正解率を示す。図中の各点は 1 人の評価者を示す。評価相反率は 0.5 近辺がほぼランダムな評価であることを意味する。図より、一部で品質の高い評価をしている評価者がいる一方で、ランダムに近い評価値を付与している評価者が存在することが確認された。特に英語を評価する日英翻訳では、ランダムに近い評価値を付与した評価者が多かった。この結果から、評価の信頼性をあげるためには、単純に 1 文あたりの評価者を増やすということよりも、正解データを用意しておいてまず評価者を評価し、高品質な評価をする評価者（例えば図中で背景が灰色の部分）を選んで、選んだ評価者のみに評価してもらうことが有効だと思われる。

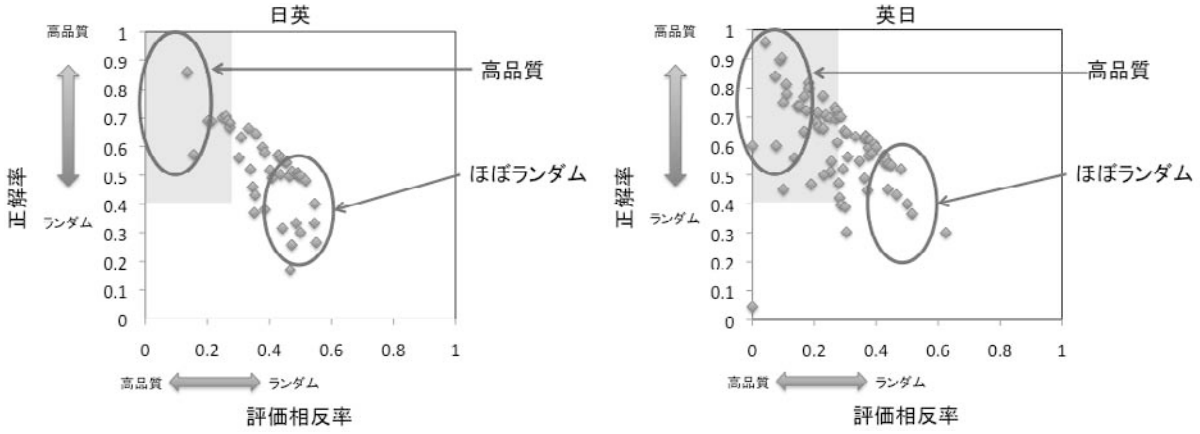

図 4：評価者毎の評価品質（左：日英、右：英日）

5.3.5.2.2 クラウド評価全体に対する検証

前節と同様に、Gold 評価値で Tie になったものを除いて Lose または Win になったデータのみを選択して検証に用いた。前節ではクラウドの評価者毎にデータを分けたが、本節ではクラウドの評価者全員のデータに対して、下記の数値を計算した。

（1）評価相反率

（2）評価不整合率

ここで、評価不整合率は、評価対象の翻訳結果の番号を i、i 番目の翻訳結果に対する評価値 v の数を $C_i(v)$ として、以下のように定義する。

$$\text{評価不整合率}=\frac{\sum_i \min\left(C_i(-1),C_i(+1)\right)}{\sum_i \left(C_i(-1)+C_i(+1)\right)}$$

評価相反率は Gold 評価値とクラウドの評価値とを比較しているのに対し、評価不整合率は、同じ翻訳結果に対するクラウドでの複数の評価者による評価値間を比較している。

結果を表 1 に示す。表中の評価相反率の括弧内の分母は Gold との一致した数＋Gold と相反した数であり、評価不整合率の括弧内の分母は Win の数＋Lose の数である。英日、日英いずれも、同じ翻訳結果に対して複数人によるクラウド評価で評価が相反した割合（評価不整合率）は、Gold 評価値と評価が相反した割合（評価相反率）はより低い値であった。すなわち、クラウド評価において、正しく訳質の優劣を判断できた割合よりも評価者間の一致率の方が高かった。このことより、単に評価者数を増やして多数派の評価結果を得ることが必ずしも信頼性を向上させるわけではないということが分かる。前節の評価者毎の評価の信頼性が大きく異なることと合わせて考えると、複数の評価者による評価結果がある際には、評価者毎に評価の信頼性を考慮して、最終的な評価値を決める必要があると言える。

なお、本節では adequacy の値の差の大きさが 1 以上のデータを一律に用いて検証したが、5.3.5.1 節の検証から adequacy の値の差が 3 以上ある場合は、差の大きさが 1 以上の場合より Gold

評価値とクラウド評価値は一致している。

表 1：評価相反率と評価不整合率

	英日	日英
評価相反率	0.285 (1593/(3987+1593))	0.389 (2995/(4714+2995))
評価不整合率	0.179 (999/(2376+3204))	0.254 (1957/(3900+3809))

5.3.5.3 評価文の特徴量によるクラウドソーシング評価の分析

5.3.5.3.1 評価者全体の評価品質に対する検証

クラウドソーシング評価における 5 人の評価値の合計値が+3 以上を Win(+1)、-3 以下を Lose(-1)、それ以外を Tie(0)とし、人手評価値（相対値）との正解率を算出した。結果を表 2 に示す。日英・英日方向ともに正解率は 50%程度とそれほど高くなかったが、相反する評価(Win->Lose または Lose->Win)となった数は少なかった。これは、全体的にベースラインシステムと同等(Tie)と評価された訳文の割合が多く、評価者間のばらつきが影響したためだと思われる。

表 2：各評価におけるクラウドソーシング評価と人手評価の関係

Adequacy-JE		クラウドソーシング評価		
		Win	Tie	Lose
人手評価	Win	308	555	87
	Tie	198	803	223
	Lose	72	493	251

一致率= 45.55%

Acceptability-JE		クラウドソーシング評価		
		Win	Tie	Lose
人手評価	Win	208	316	45
	Tie	343	1290	352
	Lose	27	245	164

一致率= 55.59%

Adequacy-EJ		クラウドソーシング評価		
		Win	Tie	Lose
人手評価	Win	245	351	84
	Tie	94	468	221
	Lose	21	226	383

一致率= 52.37%

Acceptability-EJ		クラウドソーシング評価		
		Win	Tie	Lose
人手評価	Win	170	228	42
	Tie	182	694	411
	Lose	8	123	235

一致率= 52.51%

5.3.5.3.2 評価文の特徴量に関する評価品質の検証

各評価文に対して、クラウドソーシング評価と人手評価の品質を調べるために、正解（評価が一致）したものと相反（評価が不一致）したものの二群に分割し、原文と訳文に関する以下の特徴量について t 検定を行った。

- 原文の長さ(src_len)
- 原文の単語数(src_word)
- ベースライン訳文の長さ(base_len)
- 評価対象訳文の長さ(tgt_len)
- ベースライン訳文の単語数(base_word)
- 評価対象訳文の単語数(tgt_word)
- ベースライン訳文と評価対象訳文の Word-Error-Ratio(wer)

分析結果を表 3 に示す。日英方向では、Adequacy に関して WER との間に有意差が見られ、類似

している訳文のクラウドソーシング評価の正解率が高いという傾向が確認された。クラウドソーシング評価者は、原文と訳文の対応関係よりも評価文間の差分に注目し、より正しい翻訳結果を選択していると推察される。

一方で、英日方向では、Adequacy に関して原文、訳文の長さ及び単語数との間に有意差が見られた。クラウドソーシングでは短時間に評価作業を行わなければならず、長い評価文に対する評価の品質が悪化した可能性がある。両方向とも Acceptability との関係性は特に見られなかった。

表 3：評価文の特徴量に関する評価品質の統計的検定

Adequacy-JE	base_len	tgt_len	src_len	base_word	tgt_word	src_word	wer
平均値(一致)	169.58	171.02	62.98	30.27	29.98	39.20	0.11632
平均値(不一致)	173.57	174.00	64.01	30.98	30.45	39.79	0.12020
T値	6.628%	17.170%	18.452%	6.762%	21.567%	19.702%	0.045%
	-	-	-	-	-	-	**

Acceptability-JE	base_len	tgt_len	src_len	base_word	tgt_word	src_word	wer
平均値(一致)	173.80	174.27	64.04	31.01	30.54	39.89	0.11796
平均値(不一致)	169.28	170.68	62.93	30.22	29.88	39.08	0.11878
T値	3.752%	9.998%	15.236%	4.075%	7.723%	8.142%	45.832%
	-	-	-	-	-	-	-

Adequacy-EJ	base_len	tgt_len	src_len	base_word	tgt_word	src_word	wer
平均値(一致)	61.32	61.41	174.01	38.81	40.08	29.91	0.11328
平均値(不一致)	66.54	66.21	188.34	41.29	42.65	32.18	0.11518
T値	0.005%	0.014%	0.021%	0.167%	0.156%	0.049%	28.062%
	**	**	**	**	**	**	-

Acceptability-EJ	base_len	tgt_len	src_len	base_word	tgt_word	src_word	wer
平均値(一致)	6259.42%	6258.69%	17780.89%	3943.13%	4078.62%	3047.68%	11.25%
平均値(不一致)	6515.29%	6491.75%	18418.31%	4060.66%	4188.03%	3156.44%	11.61%
T値	4.67%	6.46%	9.98%	13.60%	17.77%	9.47%	4.02%
	-	-	-	-	-	-	-

5.3.6 まとめと今後の課題

本稿では調査対象として NTCIR-9 特許翻訳タスクの日英､英日翻訳結果およびその人手評価結果を利用して、特許ドメインの機械翻訳評価をクラウドソーシングを利用して行うことが可能かを検討した。検討の結果、専門家による評価と比べると信頼性は劣るものの、翻訳精度のある程度高いシステムに対する評価や、翻訳精度に明らかに差がある文ペアの評価は高精度に行えることが分かった。このため、まずクラウドソーシング評価を行ってシステムをランキングし、その上位のシステムに対してのみ専門家による評価を行うといった使い方は十分に可能であり、全体的な翻訳評価コストを抑えることができると考えられる。

今後の課題としては、人手評価とクラウドソーシング評価とで結果が異なった原文や訳文の特徴の更なる調査や、NTCIR-9 の結果の相対評価と反対の評価および一致する評価の割合を翻訳システムごとや文長ごとなど別の視点から調査することなどが挙げられる。また今回は各システム 300 文の翻訳結果を利用したが、翻訳精度の低いシステムについては翻訳品質に大きな差が出る

ような翻訳ペアが少なくなってしまう傾向があり、クラウドソーシング評価の信頼性検証に不適切なケースがあった。このため、NTCIR-10 のデータを使うなど、より多くのデータで検証を行うことも必要である。将来的には、より効率がよく、信頼性の高い人手評価手法の構築も検討していきたい。

References

[1] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the NTCIR-9 workshop. In Proc. of NTCIR-9, 2011.

[2] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kuro- hashi, and Eiichiro Sumita. 2014. Overview of the 1st workshop on asian translation. In Proceedings of the 1st Workshop on Asian Translation (WAT2014).

[3] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, in Advances in Neural Information Processing Systems 22 (2009)

*** 英日: adequacy ***

NTT-UT
	4	3	2	1	0	-1	-2	-3	-4
1	6	40	48	44	34	8	1	0	0
0	0	2	3	6	22	5	0	0	0
-1	3	2	12	14	20	11	3	4	0

Win: 150.17, Loss: 51.39, Even: 19.44, Sum: 221.00, Score: 44.69

JAPIO
	4	3	2	1	0	-1	-2	-3	-4
1	6	34	34	32	31	9	1	0	0
0	0	2	3	7	12	3	0	0	0
-1	2	5	12	23	41	19	10	3	0

Win: 117.58, Loss: 78.31, Even: 18.12, Sum: 214.00, Score: 18.35

KYOTO
	4	3	2	1	0	-1	-2	-3	-4
1	0	2	12	18	24	12	3	0	0
0	0	1	0	3	15	8	1	1	0
-1	0	0	3	11	48	68	29	20	6

Win: 52.04, Loss: 140.98, Even: 16.97, Sum: 210.00, Score: -42.35

TORI
	4	3	2	1	0	-1	-2	-3	-4
1	0	7	23	33	44	13	4	1	0
0	0	1	3	3	12	6	1	1	0
-1	0	0	9	19	43	41	7	15	2

Win: 84.62, Loss: 96.51, Even: 17.87, Sum: 199.00, Score: -5.98

KLE
	4	3	2	1	0	-1	-2	-3	-4
1	0	4	4	14	35	5	1	0	0
0	0	0	0	5	24	7	2	0	0
-1	0	0	2	28	74	54	15	13	1

Win: 32.18, Loss: 116.97, Even: 15.85, Sum: 165.00, Score: -51.39

UOTTS
	4	3	2	1	0	-1	-2	-3	-4
1	0	5	3	15	38	10	1	1	0
0	0	1	0	2	27	6	1	0	0
-1	0	1	3	17	74	42	16	23	5

Win: 38.11, Loss: 109.72, Even: 11.17, Sum: 159.00, Score: -45.04

ICT
	4	3	2	1	0	-1	-2	-3	-4
1	0	3	7	20	52	5	0	1	0
0	0	0	0	5	21	4	0	0	0
-1	0	0	6	14	83	27	18	14	9

Win: 39.16, Loss: 92.37, Even: 10.47, Sum: 142.00, Score: -37.47

*** 日英: adequacy ***

JAPIO
	4	3	2	1	0	-1	-2	-3	-4
1	0	42	39	43	37	9	2	1	0
0	0	2	2	4	2	1	0	0	0
-1	0	10	20	26	27	19	2	0	0

Win: 138.10, Loss: 80.81, Even: 12.09, Sum: 231.00, Score: 24.80

EIWA
	4	3	2	1	0	-1	-2	-3	-4
1	0	27	41	37	41	13	3	0	0
0	0	0	3	3	10	3	0	0	0
-1	0	10	17	22	34	19	6	1	0

Win: 124.56, Loss: 77.67, Even: 10.77, Sum: 213.00, Score: 22.02

TORI
	4	3	2	1	0	-1	-2	-3	-4
1	0	7	13	20	32	14	3	0	0
0	0	0	2	3	10	5	0	0	0
-1	0	4	13	35	68	43	15	4	0

Win: 59.27, Loss: 117.04, Even: 11.70, Sum: 188.00, Score: -30.73

NAIST
	4	3	2	1	0	-1	-2	-3	-4
1	0	3	24	28	51	17	5	1	0
0	0	0	2	6	13	8	2	0	0
-1	0	0	7	17	48	35	15	5	0

Win: 81.09, Loss: 83.96, Even: 20.95, Sum: 186.00, Score: -1.54

KYOTO
	4	3	2	1	0	-1	-2	-3	-4
1	0	7	11	14	39	17	8	1	0
0	0	0	0	4	9	4	1	0	0
-1	0	0	5	13	73	52	27	3	0

Win: 61.49, Loss: 103.73, Even: 10.78, Sum: 176.00, Score: -24.00

RWTH
	4	3	2	1	0	-1	-2	-3	-4
1	0	7	20	20	60	14	5	1	0
0	0	0	0	3	17	4	0	0	0
-1	0	1	6	20	63	28	14	6	0

Win: 71.39, Loss: 77.21, Even: 9.39, Sum: 158.00, Score: -3.68

NEU
	4	3	2	1	0	-1	-2	-3	-4
1	0	0	5	29	69	27	5	2	0
0	0	0	0	5	15	4	2	0	0
-1	0	0	5	15	56	24	20	9	0

Win: 70.57, Loss: 75.05, Even: 12.37, Sum: 158.00, Score: -2.83

NTT-UT
	4	3	2	1	0	-1	-2	-3	-4
1	0	9	15	26	60	13	5	0	0
0	0	0	2	3	21	4	0	0	0
-1	0	3	8	15	60	30	11	1	0

Win: 71.14, Loss: 72.77, Even: 12.08, Sum: 156.00, Score: -1.04

UOTTS
	4	3	2	1	0	-1	-2	-3	-4
1	0	2	8	27	60	26	5	3	0
0	0	0	0	2	16	4	0	0	0
-1	0	1	1	6	70	36	15	6	0

Win: 73.42, Loss: 68.00, Even: 9.58, Sum: 151.00, Score: 3.59

FRDC
	4	3	2	1	0	-1	-2	-3	-4
1	0	3	16	23	74	24	8	2	0
0	0	0	1	2	21	4	0	0	0
-1	0	0	3	12	54	27	14	4	0

Win: 78.40, Loss: 62.00, Even: 8.60, Sum: 149.00, Score: 11.00

*** 英日: acceptability ***

JAPIO
     1   0 -1
 1 79 63   5
 0 10 19   0
-1 30 74 14
Win: 84.96, Loss: 44.85, Even: 11.19, Sum: 141.00, Score: 28.44

NTT-UT
     1   0 -1
 1 96 85   3
 0   6 32   2
-1 17 40 13
Win: 99.51, Loss: 31.05, Even: 9.44, Sum: 140.00, Score: 48.90

TORI
     1   0 -1
 1 19 98 10
 0   2 26   1
-1 11 88 40
Win: 30.13, Loss: 51.68, Even: 4.19, Sum: 86.00, Score: -25.05

KYOTO
     1   0 -1
 1 16 56   4
 0   1 28   2
-1   5 135 48
Win: 20.84, Loss: 54.13, Even: 4.03, Sum: 79.00, Score: -42.14

ICT
     1   0 -1
 1 10 75   4
 0   1 28   1
-1   9 124 41
Win: 15.54, Loss: 51.50, Even: 2.96, Sum: 70.00, Score: -51.36

UOTTS
     1   0 -1
 1   8 63   4
 0   1 35   1
-1   4 138 40
Win: 13.03, Loss: 45.87, Even: 3.10, Sum: 62.00, Score: -52.97

KLE
     1   0 -1
 1   9 55   2
 0   0 38   1
-1   7 150 33
Win: 12.07, Loss: 40.66, Even: 2.28, Sum: 55.00, Score: -51.98

*** 日英: acceptability ***

JAPIO
     1   0 -1
 1 97 73   4
 0   5   9   0
-1 34 67   6
Win: 102.51, Loss: 41.05, Even: 5.45, Sum: 149.00, Score: 41.25

EIWA
     1   0 -1
 1 71 87   7
 0   5 14   1
-1 28 70 12
Win: 78.38, Loss: 41.34, Even: 7.28, Sum: 127.00, Score: 29.17

NAIST
     1   0 -1
 1 34 89   9
 0   1 28   4
-1 12 88 31
Win: 43.43, Loss: 43.77, Even: 5.80, Sum: 93.00, Score: -0.36

NTT-UT
     1   0 -1
 1 30 92   8
 0   2 29   1
-1 11 98 23
Win: 39.19, Loss: 34.58, Even: 4.23, Sum: 78.00, Score: 5.90

TORI
     1   0 -1
 1 15 67   7
 0   2 17   2
-1 13 139 32
Win: 23.45, Loss: 46.01, Even: 4.55, Sum: 74.00, Score: -30.49

RWTH
     1   0 -1
 1 30 96   5
 0   0 25   0
-1   7 106 25
Win: 35.62, Loss: 33.94, Even: 1.44, Sum: 71.00, Score: 2.38

KYOTO
     1   0 -1
 1 19 72   7
 0   1 17   1
-1   7 138 31
Win: 27.88, Loss: 39.26, Even: 2.86, Sum: 70.00, Score: -16.26

NEU
     1   0 -1
 1 11 117 10
 0   0 25   1
-1   6 87 36
Win: 21.88, Loss: 43.44, Even: 2.67, Sum: 68.00, Score: -31.70

FRDC
     1   0 -1
 1 22 116 13
 0   0 28   0
-1   4 89 22
Win: 36.47, Loss: 27.08, Even: 1.45, Sum: 65.00, Score: 14.45

UOTTS
     1   0 -1
 1 10 110 12
 0   1 23   1
-1   5 103 29
Win: 23.26, Loss: 35.48, Even: 3.26, Sum: 62.00, Score: -19.70

5.4　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

長瀬友樹　江原暉将　王向莉

5.4.1　はじめに

機械翻訳評価の一手法として、文法項目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文法項目がうまく訳されていることのみ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、一昨年から中国語特許文献の中日機械翻訳評価のためのテストセットについて検討を行い、昨年より実際にテストセットの作成に着手している。今年度はテストセットに含まれる例文の拡充を行うとともに、計算機による自動評価を考慮したテストセットの作成を行った。このテストセットをインターネット上の評価用サイトへ実装し、中国語特許文の中日翻訳評価をネット経由で実行できる環境を整備した。本稿では、テストセット拡充のために行った作業と、テストセットを用いた自動評価の概要、そして評価用サイトの構成および実行例について説明する。

5.4.2　特許文テストセットの拡充

5.4.2.1　中日対訳文の抽出

テストセットはインターネット等で公開中の特許情報をもとに作成した。まず、中国語特許と日本語特許の patent family のリストをもとに、公開のデータベースサービスを使い、中日対訳例文を自動で抽出した。中国語と日本語の文章の対応づけは、中国語文献中の文章を中日機械翻訳した結果と日本語文献中の文章をつき合わせて、文献中の出現場所や共通の文字列を含む割合をもとに対応度合いを定量化した[4)]。タイトル、抄録、クレーム、本文をあわせて 19410 文の中日対訳文候補を抽出した。

ﾊﾟｰﾄ	C文献番号	C文数	J文献番号	J文数	対応スコア	中国語文	中日機械翻訳結果	日本語文
DES	CN102597199A	239	2013-503960	244	0.45	通过酸碱滴定法评估碱度(表3)。	酸っぱい 塩基 の 滴 定 の フランス を 通る の は アルカリ 度 の （ ３ 表す ） を 評価 する 。	アルカリ 度 を 、 酸 塩基 滴 定 を により 評価 した （ 表 ３ ） 。
ABS	CN101808947A	2	2010-533785	3	0.44	该金属硫酸盐可以是碱金属硫酸盐和/或碱土金属硫酸盐，优选为硫酸钠或硫酸镁。	この 金属 の 硫酸 塩 は アルカリ 金属 の 硫酸 塩 と ／ あるいは アルカリ 土 類 金属 の 硫酸 塩 な ことが できて 、 最適 化 して 硫酸 ナトリウム あるいは 硫酸 マグネシウム に なる 。	前記 金属 硫酸 塩 は アルカリ 金属 硫酸 塩 および ／ または アルカリ 土 類 金属 硫酸 塩 で ある ことが 好ましい 。
CLM	CN101673026A	4	2010-79298	4	0.44	5.根据权利要求1所述的照相机，其中，至少一第一光辐射频率包括至少一第一频带。	5. 権利 によって 1つ の 述べる カメラ を 求めて 、 その 中 、 少なくとも 1 第 1 光 放射 の 周波数 は 少なくとも 1 第 1 周波 帯 を 含む	上記 少なくとも 1つ の 第 1 の 周波数 が 、 少なくとも 1つ の 第 1 の 周波数 帯 を 含む こと を 特徴 と する 請求 項 1 に 記載 の カメラ 。

図１　中日対訳文候補の例

5.4.2.2　テスト文の抽出

中日対訳文候補の中から特許文に特有な中日パターンを含む文章を選択し、これをテストセット文の候補とした。特有なパターンの選択は、日本語が文の重要部分が文末にくる主要素後置型言語であることを利用して、日本語文の末尾に注目した方法が知られている[5)]。実際の作業では、日本語文を文節単位に区切り、文末の５つの文節が同一表記の例文をグルーピングして、各々のグループから１文対を候補に加えるという方法を用いた。この方法によれば、たとえば「を特徴とする。」、「定義される。」、「ことを目的とする。」、「備えている。」、「存在してもよい。」などの日本語パターンを重複することなく抽出できる。中国語パターンとの対応付けについては、今回は中国語と日本語のバリンガルによる人手作業で行った。日本語の文末表現以外のパターンについても、作業者が気づいた範囲で、テストセット文候補の追加を行った。

なお、テストセット文候補の中の中国語文または日本語文に完結した文章以外が含まれている場合や、中国語と日本語の文で情報の過不足があるものについては、テストセット文候補から除外するようにした。以上の作業により、テストセット文の候補として各パートより合計 339 文対を抽出した。

タイトル	3 文
抄録	38 文
クレーム	22 文
本文	276 文

5.4.2.3　設問の作成

テストセットによる機械翻訳文評価では、各テスト文について設問を設定する必要がある。設問設定では、人間が回答する場合と、計算機が自動回答する場合とを分けて考える必要がある。人間が回答する場合は設問表現に特に制約はないが、計算機が回答する場合は具体的な文字列の有無を問う形にブレークダウンする必要がある。たとえば、「図の番号が正しく訳出されていますか？」という設問に計算機は答えることができないため、「訳文テキストに“図１４””が含まれていますか？」のように文字列マッチングで回答が可能な形に設問内容を置き換える必要がある。

今回は計算機による自動評価を前提とするため、訳文に含まれるべき（含まれるべきでない）部分の表記を参照訳（日本語文）から抜き出す作業が、実質的な設問設定の作業となる。図２が設問設定に使用したワークシートの例である。「JP パターン」に記載されている内容が中国語文を翻訳した結果の日本語訳に含まれていなければならない文字列を示している。たとえば、一行目の例は、「具体地参照图 2，烤炉 10 还可以具有控制器 70。」という中国語文の翻訳結果に「備えることができる」または「有することができる」という文字列が含まれていなければならないことを意味している。

中国語文	日本語文	CN_パターン	JP_パターン
具体地参照图2，烤炉10还可以具有控制器70。	特に図2においてオーブン10はまた制御器70を備えることができる。	可以具有	備えることができる。|有することができる。
参照图9，示出了上面阐述的过程的逻辑和流程图。	図9において、上述の工程の論理及びフロー図が示されている。	示出了	(示(している|すされ(てい)?る)|明示(する|しているされ(てい)?る))。
然后，如步骤(6)中所示的，使用者可以手动地解除节能模式。	使用者はそれから、ステップ(6)に示すように、手動で省エネルギーモードを作動停止することができる。	可以	ことができる。

図２　設問設定の例

JP パターンの内容は、Perl プログラムに変換されて、訳文を与えれば設問の回答を Perl プログラムが自動的に判定することが可能である。

5.4.3　AAMT の自動評価サイトへの組込み

5.4.3.1　AAMT 自動評価サイトの概要

AAMT 課題調査委員会では、一昨年より、基本文の日中機械翻訳を評価するための自動評価サイト「AAMT 機械翻訳自動評価サイト（以降、AAMT 自動評価サイト)」を整備してきている [6)]。今回、AAMT 自動評価サイトの枠組みを借りて、拡大部会で作成した中日特許文評価のためのテストセットを組み込み、インターネット経由で特許文の中日翻訳評価ができる環境を作成した。

図３は AAMT 自動評価サイトの構成を示している。システムは Web アプリケーションとして実現されており、Web ブラウザのみで翻訳結果の評価実行が可能である。テストセットは評価プログラムとともに評価サーバー（AAMT サーバー）内で管理されている。利用者はテストセットの原文をダウンロードし、評価したい翻訳エンジンで翻訳した結果を評価サーバーにアップロードするだけで評価結果を得ることができる。

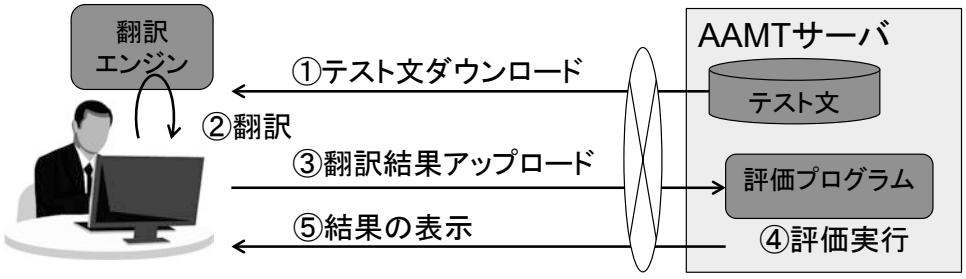

図３　システムの構成

5.4.3.2　自動評価の実行手順

AAMT 翻訳サイトから実際に機械翻訳評価を実行する操作手順は下記のとおりである。

(1)ユーザは評価の対象となる中日機械翻訳システムが翻訳を実行できる環境を手元に用意する。評価対象システムは PC にインストールするソフトウェアでもインターネット上のサービスでもよいが、複数のテキスト文章が一括で翻訳できる環境が必要である。

(2)AAMT 翻訳サイトにアクセスしてテスト文（原文：中国語）をダウンロードし、ユーザの PC 内部に保存する。

(3)評価したい翻訳システムを立ち上げ（Web サイトの場合はサイトをブラウザで表示し)、ダウンロードしたテキストを翻訳する。翻訳結果（日本語）は全文をクリップボードにコピーしたうえで PC 内に保存する。

(4)AAMT 翻訳サイトの翻訳結果送信画面にテスト文の翻訳結果全体を貼り付け、送信ボタンをクリックする。

(5)数秒間待つと、サーバ内で設問の回答を自動判定し、その結果を集計した結果が画面上に表示される。

5.4.3.3　評価結果の確認

サーバー内で評価が終了すると図４に示すような「評価結果一覧」画面が表示される。

評価結果をレーダーチャートの形で表示するのは、たとえば評価したい文法項目の情報をテスト文に持たせることによって、文法項目ごとの得点を可視化するためである。これにより、ユーザは評価対象の翻訳エンジンの課題をより具体的に知ることができる。

今回作成したテストセットは、文法項目の観点からテスト文を分類してはいないが、その代わりにテスト文の出典元のパートを属性として持たせおり、結果表示のレーダーチャートには出典パート別に得点の集計結果を表示させている。出典パートとしては、「発明の名称」、「要約」、「請求範囲」、「詳細説明」の４つに分類した。

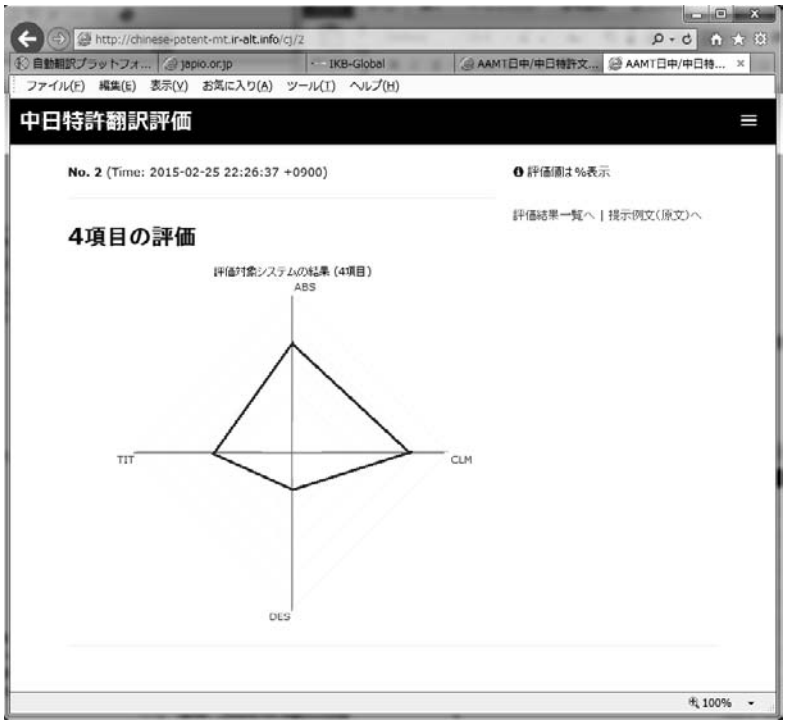

図４　評価結果の表示例

5.4.4　おわりに

昨年に引き続きパテントファミリをもとに特許文翻訳評価のためのテストセットの整備を行い、昨年度作成文と合わせて 371 文から成るテストセットを作成した。テストセットの設問を文字列の有無を問うだけのシンプルな形にすることで、テストセット評価の自動化を実現した。また、既存の自動評価サイトに今回作成したテストセットを組み込み、インターネット経由で特許翻訳の自動評価ができる環境を構築した。

今年作成したテストセットの有効性の確認については今後の課題である。従来の人間評価や既存の自動評価と比較した場合の本手法のメリット/デメリットや、信頼性の高い特許文評価ツールとして使うために収集しなければならないテスト文の数など、今後検証を行う予定である。

References

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*, 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) 江原暉将 ： 中国語特許文書から文パターンを抽出する一方法, *Japio Year Book 2013, pp.270-275*, Nov. 2013.
5) 特許庁：日本語特許出願書類の中国語への機械翻訳に関する調査報告書、2011
6) アジア太平洋機械翻訳協会：AAMT 日中/中日テストセットに基づく翻訳自動評価サイト公開、*AAMT Journal Vol.57, pp.77-83,* 2014

6　第 3 回特許情報シンポジウム開催報告

静岡大学大学院情報学研究科　梶　博行

6.1　はじめに

AAMT/Japio 特許翻訳研究会の活動の一環として、2014 年 11 月 28 日（金）10:00～17:40、キャンパスイノベーションセンター東京の国際会議室において第 3 回特許情報シンポジウムを開催した。2010 年の第 1 回、2012 年の第 2 回と同様、研究者、開発者、利用者、政策担当者が一堂に会して議論し、特許情報処理の発展に資することを目的とした。以下にシンポジウムの概要を報告する。

研究会の辻井潤一委員長の開会挨拶のあと、3 件の招待講演、2 件の研究会報告、1 件の特別講演、6 件の一般講演が行われ、Japio の守屋敏道専務理事の閉会挨拶でシンポジウムを締めくくった。過去 2 回と異なり、外国の特許庁関係者による招待講演はなく、すべて日本語による発表であった。参加者は計 84 名で、講演者ならびに AAMT/Japio 特許翻訳研究会メンバー以外の内訳は AAMT 会員 2 名、大学等教育機関 13 名、団体・研究機関 10 名、翻訳会社（システム提供を含む）11 名、調査会社（DB 提供を含む）5 名、一般事業会社・個人等 13 名であった。

6.2　招待講演

最初に、『特許庁における機械翻訳への取組』と題して、特許庁総務部特許情報企画室の櫻井健太氏にご講演いただいた。特許審査業務の重要なステップである先行技術調査では外国文献を含めて調査しなければならないが、英語だけでなく中国語や韓国語の文献が重要になっている。審査官には中国語や韓国語の読解力がある人が少ないので、機械翻訳が必須である。中国語・韓国語の特許公報を日本語に機械翻訳し、日本語で検索できるシステムを活用していること、今後は ASEAN 言語についても同様なサービスが必要になることが述べられた。また、各国特許庁が審査情報を共有する AIPN（Advanced Industrial Property Network）が紹介され、日本の審査結果を英語で提供するために日英機械翻訳を利用していること、利用者である海外特許庁審査官からのフィードバックや未知語の収集・翻訳メモリの構築など特許庁における取り組みが述べられた。

次に、『多言語機械翻訳の研究開発動向』と題して、（独）情報通信研究機構ユニバーサルコミュニケーション研究所の隅田英一郎氏にご講演いただいた。文法規則を人間が作成する旧方式「規則翻訳（RBMT）」と翻訳例を集めたパラレルコーパスから翻訳知識を自動的に学習する新方式「統計翻訳（SMT）」を対比し、後者の翻訳精度が優ってきていることが強調された。旅行分野のような限られたドメインだけでなく、特許翻訳にも新方式が有望であるという見方が示された。また、従来は原言語と目的言語の構文解析が必要であったが、目的言語の構文解析のみを必要とし、原言語の文法を自動獲得する新しいアイデアも紹介され、新方式が多言語翻訳に適した方法であると述べられた。

さらに、『特許情報検索サービスにおける機械翻訳の活用』と題して、日本パテントデータサービス（株）企画室の早川浩平氏にご講演いただいた。特許情報と機械翻訳は切り離せない関係

になっており、同社の特許情報検索サービスでは各国の特許情報をすべて英語に翻訳し串刺し検索を可能にしている。対訳比較表示などのインタフェースや活用方法を含めてこのサービスが紹介された。また、翻訳精度に関しては改善が必要であることが指摘された。特許文書特有の言い回しで正しく翻訳されない例、文脈に沿って訳出できない多義語の例、アメリカ英語とイギリス英語の取り扱いの問題などが具体的に示された。

6.3 研究会報告・特別講演

AAMT/Japio 特許翻訳研究会の重要なテーマの一つである辞書の自動構築に関し、宇津呂武仁委員（筑波大学）が『パテントファミリーからの専門用語対訳辞書の構築』を報告した。パテントファミリーとは同一発明を各国に出願することによって得られるコンパラブルな明細書の組である。そのようなコーパスから対訳専門用語を抽出する手法に関する宇津呂委員の研究グループの研究成果が紹介された。文レベルで対訳になっている部分については、統計的機械翻訳の標準的な手法であるフレーズテーブルを作成することによって、また文レベルで対訳になっていない部分については、要素合成法を用いることによって、それぞれ高精度で対訳が抽出できることを明らかにした。また、一つの語の訳語として抽出された複数の語が同義語であるかどうかを判定する試みにも言及した。

特許翻訳研究会が力を入れているもう一つのテーマが翻訳システム／翻訳文の自動評価である。これに関連し、越前谷博委員（北海学園大学）が『自動評価法を用いた機械翻訳の定量的評価』と題して、本年 6 月にボルチモアで開催された WMT 2014 (The 9th Workshop on Statistical Machine Translation) の自動評価タスクについて報告するとともに、日本発の自動評価法である APAC と RIBES を紹介した。さまざまな自動評価法が提案されているが、ほとんどが機械翻訳システムによる翻訳文と人間が作成した参照訳との類似度を計算する方法である。自動評価法自体の評価は、自動評価法によるスコアと人手による翻訳文評価の相関を求めることによって行われる。WMT 2014 の自動評価タスクでも、参加した自動評価法がそのような方法で競い合った。機械翻訳システムに対して一つのスコアを出力する“システムレベルの評価”では人手評価との相関が 0.8 を超えるのに対し、翻訳文ごとにスコアを出力する“セグメントレベルの評価”では人手評価との相関が 0.4 程度と低かったこと、どの言語対に対しても一貫して優位な自動評価法はなかったことなどが報告された。自動評価法も発展途上にあるといえる。APAC は越前谷委員が考案した方法で、参照訳との一致単語列（チャンク）に着目した方法である。また、RIBES は磯崎秀樹委員（岡山県立大学）の提案によるもので、参照訳との語順の近さを測定する方法である。

特許翻訳においても中国語や韓国語などの重要性が増していることから、アジア言語の機械翻訳システムの研究開発に携わっている中澤敏明委員（科学技術振興機構／京都大学）に特別講演『アジア言語を中心とした機械翻訳研究』をお願いした。講演の前半は、科学技術振興機構と京都大学が中国科技技术信息研究所などと協力して推進中の日中・中日機械翻訳実用化プロジェクトの紹介であった。言語構造や語順が大きく異なる言語対であることを考慮した依存構造木のアライメント・依存構造木間の翻訳方法を開発し、従来方法より高い翻訳精度を達成している。後半は、中澤委員がオーガナイザとして本年 10 月に東京で開催した WAT 2014 (The 1st Workshop on

Asian Translation) の報告であった。今回は日・中・英 3 言語間の科学技術論文翻訳タスクを実施したが、来年はインドネシア語 - 英語の新聞記事翻訳や日本語 - 中国語の特許文献翻訳を含めることが検討されている。

6.4 一般講演

投稿ベースの一般講演では、機械翻訳に関して以下の 3 編の論文が発表された。

- 田中浩之，園尾聡，木下聡，釜谷聡史：統計的訳語選択技術による韓日機械翻訳の高精度化
  ルールベースの韓日機械翻訳に統計的な訳語選択技術を組み込むことにより、BLEU スコアが 8.5 ポイント向上した。
- 正林真之，杉浦伸夫：特許事務所における機械翻訳と人手による翻訳の Mix 事例
  翻訳メモリと翻訳辞書による機械翻訳とを組み合わせた「Hybrid 翻訳」により、翻訳品質の向上と時間短縮を両立させている。
- 吉川潔：翻訳ソフトを優れた辞書として活用する方法
  インターネットから利用できる翻訳ソフトが新語や造語の訳語を調べるのにたいへん有用であることを指摘した。

さらに、翻訳以外の特許情報処理に関しても興味深い論文 3 編が発表された。

- 福田悟志，難波英嗣，竹澤寿幸，乾孝司，岩山真，橋田浩一，藤井敦：F タームに基づいたオントロジーの構築
  特許分類コード体系である F タームをベースとしてブートストラッピングにより、上位－下位、全体－部分といった用語間の関係を獲得する。
- 池田紀子，田中一成：特許文書からの化学物質情報の抽出
  特許文書から化学物質名と化学式の対応関係を抽出し、化学物質名の命名規則を用いてそれらを部品化し、部品を組み合わせることによって新しい化学物質名に対する化学式を生成する。
- 岩本圭介：特許情報分析のためのマイニング手法と分析ツール Patent Mining eXpress
  特徴技術ネットワーク、課題－解決手段のクロスマップなどの特徴的な機能をもつ。

6.5 おわりに

以上のように、さまざまな立場からさまざまな内容が発表され、会場からも多数の質問と意見が出た。研究開発者と利用者の相互理解が深まったものと思われる。Japio の守屋専務理事の閉会挨拶では、発表者と参加者への謝意とともに 2016 年の第 4 回シンポジウム開催への期待が表明された。18:00 から別室で開かれた懇親会でも議論が続き、有意義な一日であった。

（注）AAMT/Japio 特許翻訳研究会のホームページ（http://aamtjapio.com）からシンポジウムの論文集をダウンロードすることができます。

禁 無 断 転 載

平成 26 年度 AAMT/Japio 特許翻訳研究会報告書

発 行 日　平成 27 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