平成28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 29 年 3 月

一般財団法人　日本特許情報機構

目　　次

AAMT/Japio 特許翻訳研究会委員名簿

（敬称略・順不同）

委員長	辻井　潤一（※2）	国立研究開発法人 産業技術総合研究所 人工知能研究センター センター長 / 東京大学大学院 名誉教授
副委員長	宇津呂武仁（※2）	筑波大学大学院 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 主任研究員
委員	磯崎　秀樹（※1）	岡山県立大学 教授
	今村　賢治	国立研究開発法人 情報通信研究機構 先進的音声翻訳研究開発推進センター 専門研究員
	越前谷　博（※2）	北海学園大学大学院 教授
	江原　暉将（※2）	元・山梨英和大学 教授
	熊野　　明	東芝ソリューション株式会社 プラットフォームセンター ソフトウェア開発部
	黒橋　禎夫	京都大学大学院 教授
	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
	下畑　さより	沖電気工業株式会社 情報通信事業本部 ソフトウェアセンターサービス業務管理部
	綱川　隆司	静岡大学学術院 助教
	中澤　敏明（※2）	国立研究開発法人 科学技術振興機構 情報企画部 研究員 / 京都大学 大学院情報学研究科 知能情報学専攻 研究員
	二宮　　崇	愛媛大学大学院 准教授
	横山　晶一	山形大学 名誉教授
オブザーバー	潮田　　明	国立研究開発法人 産業技術総合研究所 人工知能研究センター
	高　　京徹	株式会社高電社 経営企画部 部長
	園尾　　聡（※2）	株式会社東芝 インダストリアル ICT ソリューション社 商品統括部 メディアインテリジェンス商品推進部グループ
	中川　裕志	東京大学 情報基盤センター 教授
	長瀬　友樹（※2）	株式会社富士通研究所 メディア処理研究所 主管研究員
	範　　暁蓉	東京大学大学院 中川研究室
	王　　向莉（※2）	株式会社ディープランゲージ

	守屋　敏道	一般財団法人日本特許情報機構 専務理事
	小林　　明	一般財団法人日本特許情報機構 特許情報研究所 所長
	横井　巨人	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
	木下　　聡	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 研究主幹
	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 研究管理課 課長
	白土　博之	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
	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
	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 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
	船戸　さやか	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 副主任

（※１：拡大評価部会部会長、※２：拡大評価部会メンバー）

事　務　局	小松　浩平	株式会社インターグループ
	佐藤　伶奈	株式会社インターグループ

平成 28 年度 AAMT/Japio 特許翻訳研究会・活動履歴

平成 28(2016)年 5 月 13 日
第 1 回 AAMT/Japio 特許翻訳研究会、第 1 回拡大評価部会
（於キャンパス・イノベーションセンター東京）

平成 28(2016)年 6 月 24 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 28(2016)年 7 月 29 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 28(2016)年 9 月 9 日
第 4 回 AAMT/Japio 特許翻訳研究会
（於キャンパス・イノベーションセンター東京）

平成 28(2016)年 10 月 14 日
第 5 回 AAMT/Japio 特許翻訳研究会、第 2 回拡大評価部会
（於キャンパス・イノベーションセンター東京）

平成 28(2016)年 11 月 25 日
第 4 回特許情報シンポジウム
（於東京・グランパークカンファレンス　401 ホール）

平成 28(2016)年 12 月 9 日
第 6 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 29(2017)年 2 月 10 日
第 7 回 AAMT/Japio 特許翻訳研究会、第 3 回拡大評価部会
（於キャンパス・イノベーションセンター東京）

平成 29(2017)年 3 月 31 日
『平成 28 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

AAMT/Japio 特許翻訳研究会委員長
産業技術総合研究所人工知能研究センター　センター長
辻井　潤一

機械翻訳、言語処理、言語理解に新たな研究の波が押し寄せている。音声認識、画像処理で成功を収めた深層学習の技術が、言語を取り扱う技術分野でも革新をもたらしつつある。２０１６年１１月にグーグルの翻訳サービスが従来の統計的な翻訳システムからニューラルネットによる翻訳システムに切り替えられた。これまでの統計的な機械翻訳においても、その改良は地道に続けられ、翻訳の質は徐々に良くなってきていたが、性能の改善は極めて緩やかで、ある種の技術的なプラトーに差し掛かっていた。そこに現れたのが、このニューラルネット翻訳であった。マイクロソフトも、すぐに翻訳サービスをニューラルネット翻訳に切り替えるなど、競争が激化している。

１９９０年代の初頭から始まった規則による機械翻訳から統計翻訳への切り替えが１５年以上かかったのに比べると、今回の枠組みの革新は、わずか数年間で起こった。統計翻訳のために蓄積された対訳コーパスの集積、および、画像や音声での深層学習技術のために整備されてきた計算リソース、この２つのリソースの集積がニューラルネット翻訳でもそのまま使えたことが、この枠組みの切り替えの驚くべき速さにつながった。

ニューラルネット翻訳で際立つのは、翻訳結果の流暢さであろう。統計翻訳でも、相手言語の膨大なコーパスから学習された単一言語モデルにより、翻訳文の流暢さは増していたが、ニューラルネット翻訳で出力文の流暢さは格段に向上した。この流暢さの向上により、日常的なテキストでの翻訳の誤りは格段に見つけやすくなった。ただ、専門性の高い翻訳では、専門分野の知識がない場合には、表面上の読みやすさのために逆に翻訳誤りは見つけにくくなっている。言い換えると、ニューラルネット翻訳の後編集には、これまで以上に、分野の専門知識が不可欠になっている。機械翻訳の後編集のあり方も大きく変わる。また、機械翻訳の性能評価に使われてきた様々な指標も、統計翻訳からニューラル翻訳に切り替わると再考しなければならない。機械翻訳をとりまく周辺技術も、今後、急速に変化しよう。

このように、ニューラルネット翻訳は、機械翻訳の研究を大きく変革しつつある。ただ、現在のニューラルネット翻訳の枠組みは、統計翻訳の場合と同じように、文やテキストの意味を理解しているわけではない。また、人間の翻訳家とは違って、確信のない翻訳箇所についての自覚がない。このような機械翻訳をいかにうまく使いこなして翻訳のコストを全体として低下させるか、人間との望ましい協働作業が不可欠となる。本委員会では、機械翻訳をいかに特許翻訳に活用していくかに関して、活動を続けてきた。本報告書は、我々の１年間の活動をまとめたものである。読者諸賢の参考になれば幸いである。

2. 機械翻訳および関連技術

2.1 ニューラル機械翻訳における敵対性生成モデルの応用

愛媛大学　白井 圭佑
二宮　崇

2.1.1　はじめに

機械翻訳の研究分野では、ルールベース機械翻訳、用例ベース機械翻訳、フレーズベース統計的機械翻訳、統計的同期文法など様々な手法が提案されてきたが、近年ではニューラルネットを用いたニューラル機械翻訳(Neural Machine Translation; NMT)(Kalchbrenner&Blunsom, 2013; Sutskever et al., 2014)が、従来手法を上回る高い精度を実現し注目を集めている。ニューラル機械翻訳モデルの中でも、入力系列(原言語文)を全て 1 つのベクトルにエンコードしてから出力系列(翻訳先言語文)を生成するエンコーダー・デコーダーモデル (Cho et al., 2014; Sutskever et al., 2014)や、入力系列をエンコードする過程において生成された過去の内部状態に重み付けをして翻訳を行う注意型ニューラル機械翻訳(Attention NMT; ANMT)(Bahdanau et al., 2015; Luong et al., 2015)の研究が特に盛んに行われている。

本研究は、敵対性生成モデル(Generative Adversarial Nets; GAN)(Goodfellow et al., 2014; Radford et al., 2016; Mirza&Osindero, 2014)を ANMT に応用した新しい機械翻訳モデルを提案する。GAN は、画像生成の分野において近年注目を集めている画像生成モデルの一つであり、画像を生成する生成モデルと、その画像が生成モデルによって作られたのかどうかを判別する識別モデルを用い、これらの 2 つのモデルが競い合うように学習を行うことで、より良い生成モデルを学習する手法である。識別モデルは与えられた画像が生成モデルから作られた画像か、データセット中に存在する本物の画像かを正しく見分けるように学習を行い、生成モデルは識別モデルが判別を誤るように学習を行う。本研究では、従来の ANMT の構造を GAN における生成モデルだと捉え、これに新たに識別モデルを追加する。識別モデルは ANMT が入力系列から予測した出力系列と、入力系列に対応する正解データを正しく見分けるように学習を行い、ANMT は従来の学習に加え、識別モデルが判別を誤るように学習を行う。

本稿の構成は以下のようになっている。2.1.2 節では、背景にあたるニューラル機械翻訳について説明する。2.1.3 節は、敵対性生成モデル(GAN)について説明し、GAN をニューラル機械翻訳に応用した提案手法について述べる。2.1.4 節は、提案手法を評価するための実験について報告する。2.1.5 節で本稿の主旨をまとめ、今後の課題について述べる。

2.1.2　ニューラル機械翻訳

本節は、ニューラル機械翻訳と注意型ニューラル機械翻訳について説明する。

2.1.2.1　エンコーダー・デコーダーモデル

エンコーダー・デコーダーモデルは、入力系列 $s = s_1 s_2 \cdots s_n$ を受け取り中間表現を生成するエンコーダー、エンコーダーから受け取った中間表現から出力系列 $t = t_1 t_2 \cdots t_m$ を生成するデコーダ

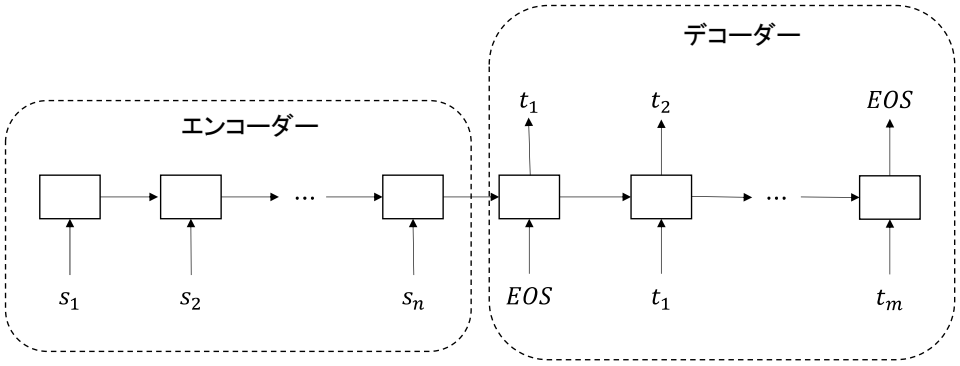


図 1: エンコーダー・デコーダーモデル

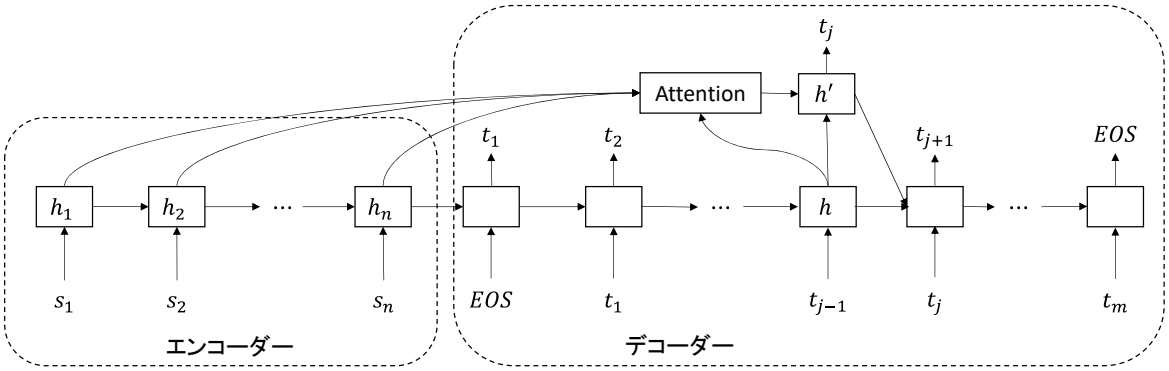


図 2: 注意型ニューラル機械翻訳

ーをそれぞれニューラルネットによって実現する。図 1 はエンコーダー・デコーダーモデルを表す。エンコーダー・デコーダーモデルの学習は、エンコーダー・デコーダーモデルに対し、下式を用いて入力系列が与えられたときの出力系列の対数尤度を最大化することで行われる。

$$p(t|s) = \sum_{i=1}^{m} \log p(t_i|\, t_{j<i}, s) \qquad \cdots (1)$$

エンコーダー、デコーダーには時系列データを解析するために LSTM を用いる。全ての入力系列を受け取った後のエンコーダー側 LSTM の内部状態をデコーダー側 LSTM の初期内部状態とする。

2.1.2.2 注意型ニューラル機械翻訳(ANMT)

注意型ニューラル機械翻訳(ANMT)は、エンコーダー・デコーダーモデルにおける出力系列の予測時の各ステップにおいて、エンコード時の各ステップにおける LSTM の内部状態の履歴を参照して出力単語を予測するモデルである。エンコーダーの内部状態の履歴を $h_1 h_2 \cdots h_n$ としたとき、ANMT は、各 h_i に対する重み α_i を、デコーダー側の LSTM の内部状態 h を用いて計算する。図 2 は t_j を出力する時の ANMT の計算の流れを表している。エンコーダーあるいはデコーダーを多

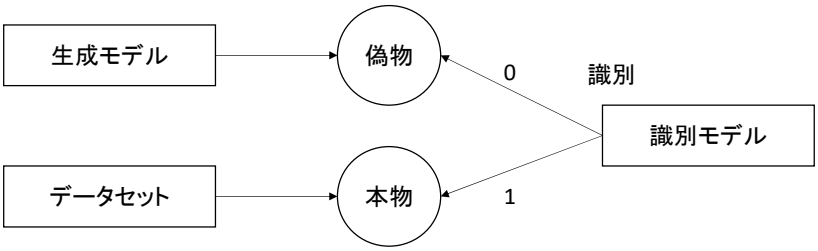


図 3: 敵対性生成モデル

層 LSTM によって実現している場合、LSTM の内部状態とは最上位層の LSTM の内部状態を指す。

Luong ら(2015)が提案したグローバルアテンションの手法では、デコード時の各ステップにおけるデコーダー側の内部状態 h と、エンコーダー側の各内部状態 h_i に対する重み $\alpha_i(h)$を次式により計算する。

$$\alpha_i(h) = \frac{\exp\{score(h, h_i)\}}{\sum_{j=1}^{n} \exp\{score(h, h_j)\}} \quad \cdots (2)$$

2.1.3 ニューラル機械翻訳における敵対性学習

本節は、敵対性生成モデルについて説明し、続いて提案手法について述べる。

2.1.3.1 敵対性生成モデル

敵対性生成モデル(GAN)(Goodfellow et al., 2014)は、ある確率分布 p_z からサンプリングしたサンプル $z = (z_1, z_2, \cdots, z_n)$ を入力として受け取り、画像 $G(z)$ を生成する。生成モデル G は識別モデル D が本物と間違うような画像を生成することが目的であり、その目的関数は以下のように表される。

$$\min_G \mathbb{E}_{z \sim p_z(z)} \left[\log\left(1 - D\left(G(z)\right)\right) \right] \quad \cdots (3)$$

識別モデルは生成モデルから生成された系列と正解データの系列を正しく見分けることが目的であるため(図 3)、その目的関数は以下のように表される。

$$\max_D \mathbb{E}_{x \sim p_{data}(x)} \left[\log\left(D(x)\right) \right] + \mathbb{E}_{z \sim p_z(z)} \left[\log\left(1 - D\left(G(z)\right)\right) \right] \quad \cdots (4)$$

敵対性生成モデルでは 2 つのモデルに関して次式で最適化を行う。

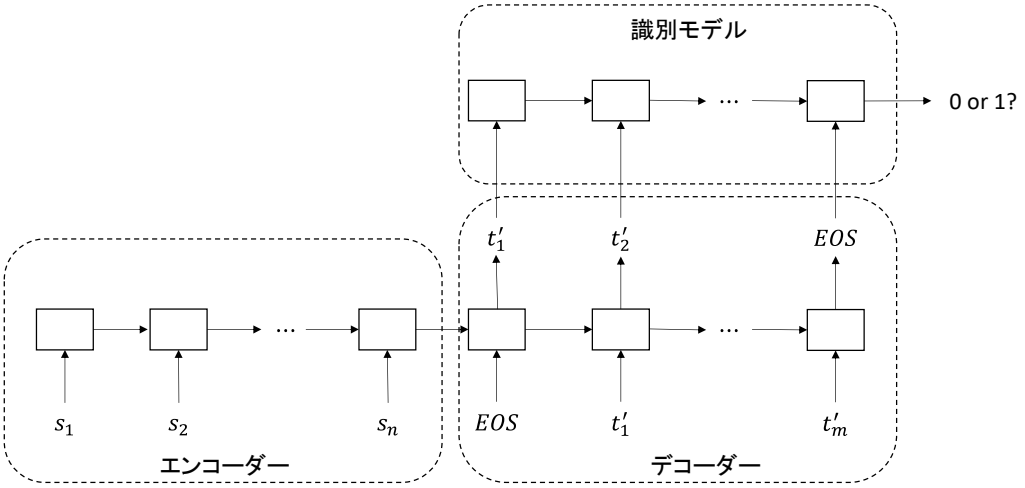


図 4: 提案モデル

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)}\left[\log\left(D(x)\right)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad \cdots (5)$$

2.1.3.2 提案手法

画像生成における GAN では、識別モデルは全結合層や畳込みニューラルネットワーク(Convolutional Neural Network; CNN)で実現される。提案手法では識別モデルに入力として画像ではなく文字列を与えるため、時系列モデルである再帰型ニューラルネットワークの一種である LSTM を用いて識別モデルを実現する。

提案モデルの全体像を図 4 に示す。提案モデルは翻訳の生成を行う ANMT と識別モデルから成り、入力として原言語の系列データ $s_1 \cdots s_n$ をエンコーダーに与え、デコーダーは翻訳先言語の系列データ $t'_1 \cdots t'_m$ を生成し、識別モデルは系列データ $t'_1 \cdots t'_m$ を受け取り、その入力が正解データか生成されたデータか識別する。

識別モデルは入力として翻訳先言語の系列データを受け取り、それが入力系列と対の正解データ $t_1 \cdots t_m$ なら 1 を、ANMT が入力系列から予測した出力系列 $t'_1 \cdots t'_m$ であれば 0 を返すように学習する(図 5)。

生成モデルに相当する ANMT の学習は入力系列と対になる正解データを用いた通常の学習(図 1)と識別モデルを用いた学習(図 6)から成る。識別モデルを用いた学習では ANMT が生成した出力系列を識別モデルが本物と間違えるように ANMT の学習を行う。つまり、識別モデルの出力を 1 として ANMT の学習を行う。

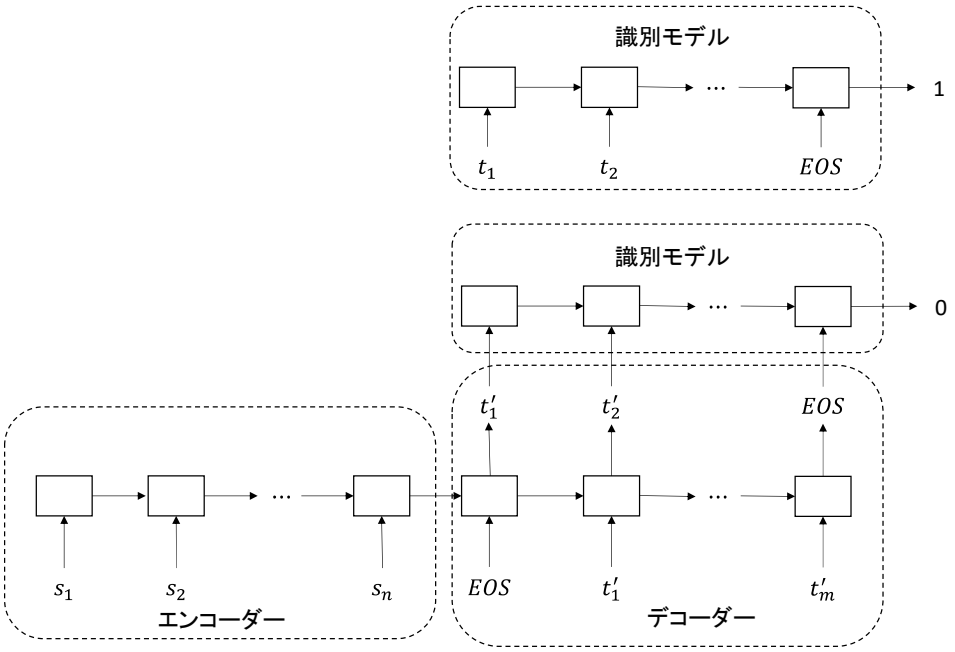


図 5: 識別モデルの学習

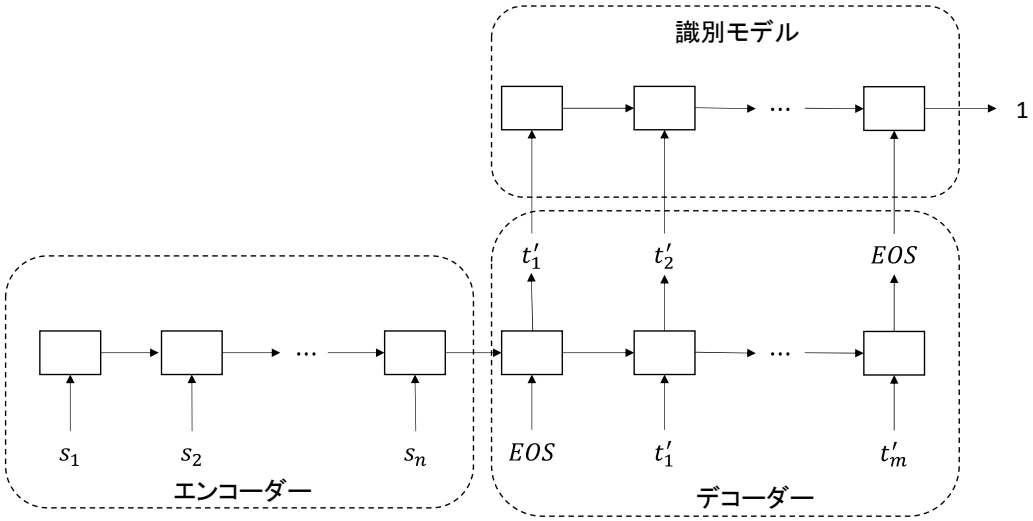


図 6: 生成モデル(ANMT)の学習

ANMT では語彙の生成にソフトマックス関数を用いるが、全ての語彙に対するノードをもつネットワークを学習することは非常に大きな計算コストを要する。そのため本研究では正解データを用いた ANMT の学習にはソフトマックス関数を用い、識別モデルを用いた学習には ANMT により選択された語彙に対してのみ、次のシグモイド関数を用いて近似的にスコアを計算する。

$$\sigma(t) = \frac{1}{1 + e^{-score(t)}}$$

ただし、$score(t)$ はデコード時の各ステップにおいて単語 t を予測したときのスコアである。

ミニバッチ内で次の手順を行うことで提案モデルのパラメータ最適化を行う。

(i) ANMT の学習 式(1)に従って、入力系列から生成した出力系列の対数尤度を最大化するように学習する(図 1)。

(ii) 識別モデルの学習 次に、式(4)に従って、生成された出力系列と正解データの系列を正しく識別する期待値を最大化するように学習する(図 5)。

(iii) 識別モデルを通した ANMT の学習 最後に、式(3)に従って、生成した出力系列を識別モデルが本物と間違える期待値を最大化するように学習する(図 6)。

2.1.4 実験

2.1.4.1 実験設定

本実験においては、英日の対訳コーパスとして Asian Scientific Paper Excerpt Corpus (ASPEC)[1]を用いて英日翻訳を行った。コーパスに対する処理としては、和文に対しては京都テキスト解析ツールキット(KyTea)[2]を用い、英文に対しては mosesdecoder[3]を用いて単語分割を行った後、mosesdecoder を用いて小文字化を行った。また、長さが 1 未満あるいは 50 以上の文を取り除いた。学習用データとしては 100 万文対(train-1.txt)に上記の処理を行い、上位 50,000 文対を抽出し用いたところ、英語、日本語の語彙は文末記号と未知語のシンボルを含めてそれぞれ 16,416 語、15,051 語であった。開発データと評価データに対しても上記の処理を行った結果、それぞれ 1,658 文対、1,812 文対であった。

ANMT は、エンコーダー、デコーダーがそれぞれ 4 層の LSTM から成るエンコーダー・デコーダーモデルを用いた。注意型のモデルは Luong ら(2015)による“Global Attention(dot) ”を採用した。また、識別モデルも同様に 4 層の LSTM を用いた。単語埋め込み層、隠れ層の大きさはともに 256 次元とし、LSTM のパラメータは $[-0.8, 0.8]$ の一様乱数で初期化した。学習には Adam (Kingma&Ba, 2015)を使用し、パラメータの初期状態は、$\alpha = 0.01$、$\beta 1 = 0.9$、$\beta 2 = 0.999$とし、ミニバッチサイズは 64 とした。また、本実験でのモデルは全て Chainer (Tokui et al., 2015)を用いて実装した。

2.1.4.2 実験結果

英日翻訳の学習を上記の設定で、従来手法(ANMT)と提案手法(ANMT + GAN)で実験を行った。最終的なモデルとしては各エポック終了後に、開発データに対して計算したパープレキシティが最も低かったものを採用した。

評価データに対してビーム幅 5 のビーム探索を行い BLEU と RIBES で評価した。表 1 はその実験結果を表す。表 1 より、提案手法のモデルでは BLEU、RIBES 共に従来手法の ANMT と比

[1] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[2] http://www.phontron.com/kytea/index-ja.html
[3] https://github.com/moses-smt/mosesdecoder

表 1: 50,000 文対を学習に用いた場合の実験結果

	BLEU(%)	RIBES
従来手法(ANMT)	15.86	0.6854
提案手法(ANMT+GAN)	17.24	0.7036

較して高い値が得られたことがわかる。

2.1.5 まとめ今後の課題

本研究は、従来手法である ANMT に GAN の学習方法を取り入れた機械翻訳のモデルを提案した。本稿で行った実験より従来手法(ANMT)に対する提案手法の有効性を示すことができた。今後は、本稿の実験で用いていない ASPEC の残りの文対を用いて実験を行い、従来手法と提案手法との比較を行う。

参考文献

N. Kalchbrenner and P. Blunsom (2013) Recurrent Continuous Translation Models, Proceedings of EMNLP 2013, pp. 1700-1709.

I. Sutskever, O. Vinyals and Q. V. Le (2014) Sequence to Sequence Learning with Neural Networks, Proceedings of NIPS 2014, pp. 3104-3112.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of EMNLP 2014, pp. 1724-1734.

D. Bahdanau, K. Cho and Y. Bengio (2015) Neural Machine Translation by Jointly Learning to Align and Translate, Proceedings of ICLR 2015.

M.-T. Luong, H. Pham and C. D. Manning (2015) Effective Approaches to Attention-based Neural Machine Translation, Proceedings of EMNLP 2015, pp. 1412-1421.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio (2014) Generative Adversarial Nets, Proceedings of NIPS 2014, pp. 2672-2680.

A. Radford, L. Metz and S. Chintala (2016) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, Proceedings of ICLR 2016.

M. Mirza and S. Osindero (2014) Conditional Generative Adversarial Nets, arXiv preprint arXiv:1411.1784.

D. P. Kingma and J. Ba (2015) Adam: A Method for Stochastic Optimization, Proceedings of ICLR 2015.

S. Tokui, K. Oono, S. Hido and J. Clayton (2015) Chainer: a Next-Generation Open Source Framework for Deep Learning, Proceedings of Workshop on Machine Learning Systems (LearningSys) at NIPS.

2.2 Neural Machine Translation of Patent Sentences with Large Vocabulary Technical Terms

University of Tsukuba Zi Long
Takehito Utsuro
University of Tsukuba Mikio Yamamoto

2.2.1 Introduction

Neural machine translation (NMT), a new approach to solving machine translation, has achieved promising results [1][2][7][8][11][12][17]. An NMT system builds a simple large neural network that reads the entire input source sentence and generates an output translation. The entire neural network is jointly trained to maximize the conditional probability of a correct translation of a source sentence with a bilingual corpus. Although NMT offers many advantages over traditional phrase-based approaches, such as a small memory footprint and simple decoder implementation, conventional NMT is limited when it comes to larger vocabularies. This is because the training complexity and decoding complexity proportionally increase with the number of target words. Words that are out of vocabulary are represented by a single unknown token in translations, as illustrated in Figure Figure 1. The problem becomes more serious when translating patent documents, which contain several newly introduced technical terms.

There have been a number of related studies that address the vocabulary limitation of NMT systems. Jean el al. [7] provided an efficient approximation to the softmax to accommodate a very large vocabulary in an NMT system. Luong et al. [13] proposed annotating the occurrences of a target unknown word token with positional information to track its alignments, after which they replace the tokens with their translations using simple word dictionary lookup or identity copy. Li et al. [10] proposed to re- place out-of-vocabulary words with similar in-vocabulary words based on a similarity model learnt from monolingual data. Sennrich et al.[16] introduced an effective approach based on encoding rare and unknown words as sequences of subword units. Luong and Manning [11] provided a character-level and word-level hybrid NMT model to achieve an open vocabulary, and Costa-jussà and Fonollosa[3] proposed a NMT system based on character-based embeddings.

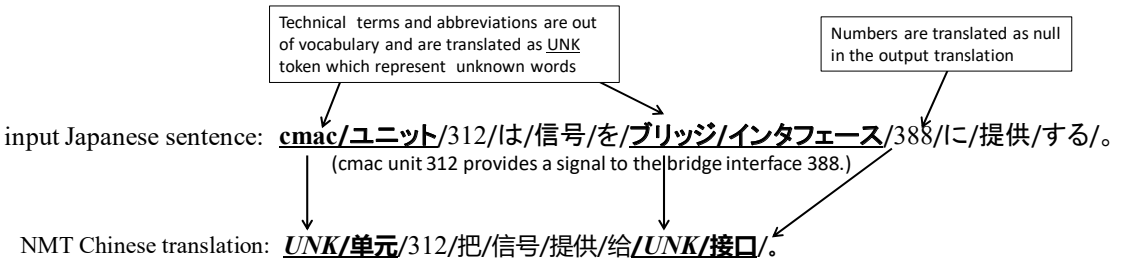


Figure 1 Example of translation errors when translating patent sentences with technical terms using NMT

However, these previous approaches have limitations when translating patent sentences. This is because their methods only focus on addressing the problem of unknown words even though the words are parts of technical terms. It is obvious that a technical term should be considered as one word that comprises components that always have different meanings and translations when they are used alone. An example is shown in Figure Figure 1, wherein Japanese word "ブリッジ"(bridge) should be translated to Chinese word "桥架" when included in technical term "bridge interface"; however, it is always translated as "桥".

In this paper, we propose a method that enables NMT to translate patent sentences with a large vocabulary of technical terms. We use an NMT model similar to that used by Sutskever et al. [17], which uses a deep long short-term memories (LSTM) [5] to encode the input sentence and a separate

deep LSTM to output the translation. We train the NMT model on a bilingual corpus in which the technical terms are replaced with technical term tokens; this allows it to translate most of the source sentences except technical terms. Similar to Sutskever et al.[17], we use it as a decoder to translate source sentences with technical term tokens and replace the tokens with technical term translations using statistical machine translation (SMT). We also use it to rerank the 1,000-best SMT translations on the basis of the average of the SMT and NMT scores of the translated sentences that have been rescored with the technical term tokens. Our experiments on Japanese-Chinese patent sentences show that our proposed NMT system achieves a substantial improvement of up to 3.1 BLEU points and 2.3 RIBES points over a traditional SMT system and an improvement of approximately 0.6 BLEU points and 0.8 RIBES points over an equivalent NMT system without our proposed technique.

2.2.2 Japanese-Chinese Patent Documents

Japanese-Chinese parallel patent documents were collected from the Japanese patent documents published by the Japanese Patent Office (JPO) during 2004-2012 and the Chinese patent documents published by the State Intellectual Property Office of the People's Republic of China (SIPO) during 2005-2010. From the collected documents, we extracted 312,492 patent families, and the method of Utiyama and Isahara[19] was applied[1] to the text of the extracted patent families to align the Japanese and Chinese sentences. The Japanese sentences were segmented into a sequence of morphemes using the Japanese morphological analyzer MeCab[2] with the morpheme lexicon IPAdic,[3] and the Chinese sentences were segmented into a sequence of words using the Chinese morphological analyzer Stanford Word Segment [18] trained using the Chinese Penn Treebank. In this study, Japanese-Chinese parallel patent sentence pairs were ordered in descending order of sentence-alignment score and we used the topmost 2.8M pairs, whose Japanese sentences contain fewer than 40 morphemes and Chinese sentences contain fewer than 40 words.[4]

2.2.3 Neural Machine Translation (NMT)

NMT uses a single neural network trained jointly to maximize the translation performance [1][2][7][8][11][12][17]. Given a source sentence $\boldsymbol{x} = (x_1, \dots, x_N)$ and target sentence $\boldsymbol{y} = (y_1, \dots, y_M)$, an NMT system uses a neural network to parameterize the conditional distributions

$$p(y_l | y_{<l}, \boldsymbol{x})$$

for $1 \leq l \leq M$. Consequently, it becomes possible to compute and maximize the log probability of the target sentence given the source sentence

$$\log p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{l=1}^{M} \log p\,(y_l | y_{<l}, \boldsymbol{x}) \quad (1)$$

In this paper, we use an NMT model similar to that used by Sutskever et al.[17]. It uses two separate deep LSTMs to encode the input sequence and output the translation. The encoder, which is implemented as a recurrent neural network, reads the source sentence one word at a time and then encodes it into a large vector that represents the entire source sentence. The decoder, another recurrent neural network, generates a translation on the basis of the encoded vector one word at a time.

One important difference between our NMT model and the one used by Sutskever et al. [17] is that we added an attention mechanism. Recently, Vinyals et al. [20]proposed an attention mechanism, a form of random access memory, to help NMT cope with long input sequences. Luong et al. [12] proposed an attention mechanism for different scoring functions in order to compare the source and target hidden

[1] Herein, we used a Japanese-Chinese translation lexicon comprising around 170,000 Chinese entries.

[2] http://mecab.sourceforge.net/

[3] http://sourceforge.jp/projects/ipadic/

[4] In this paper, we focus on the task of translating patent sentences with a large vocabulary of technical terms using the NMT system, where we ignore the translation task of patent sentences that are longer than 40 morphemes in Japanese side or longer than 40 words in Chinese side.

states as well as different strategies for placing the attention. In this paper, we utilize the attention mechanism proposed by Vinyals et al. [20], wherein each output target word is predicted on the basis of not only a recurrent hidden state and the previously predicted word but also a context vector computed as the weighted sum of the hidden states.

2.2.4 NMT with a Large Technical Term Vocabulary

2.2.4.1 NMT Training after Replacing Technical Term Pairs with Tokens

Figure 2 illustrates the procedure of the training model with parallel patent sentence pairs, wherein technical terms are replaced with technical term tokens "TT_1," "TT_2," and so on.

In the step 1 of Figure Figure, we align the Japanese technical terms, which are automatically extracted from
the Japanese sentences, with their Chinese translations in the Chinese sentences.[5] Here, we introduce the following two steps to identify technical term pairs in the bilingual Japanese-Chinese corpus:

1. According to the approach proposed by Dong et al.[4], we identify Japanese-Chinese technical term pairs using an SMT phrase translation table. Given a parallel sentence pair (S_J, S_C) containing a Japanese technical term t_J, the Chinese translation candidates collected from the phrase translation table are matched against the Chinese sentence S_C of the parallel sentence pair. Of those found in S_C, t_C with the largest translation probability $p(t_C|t_J)$ is selected, and the bilingual technical term pair (t_J, t_C) is identified.

2. For the Japanese technical terms whose Chinese translations are not included in the results of Step 1, we then use an approach based on SMT word alignment. Given a parallel sentence pair (S_J, S_C) containing a Japanese technical term t_J, a sequence of Chinese words is selected using SMT word alignment, and we use the Chinese translation t_C for the Japanese technical term t_J.[6]

As shown in the step 2 of Figure Figure, in each of Japanese-Chinese parallel patent sentence pairs, occurrences of technical term pairs (t_J^1, t_c^1), $(t_J^2, t_c^2), \cdots$, (t_J^k, t_c^k) are then replaced with technical term tokens (TT_1, TT_1), $(TT_2, TT_2), \cdots, (TT_k, TT_k)$. Technical term pairs (t_J^1, t_c^1), $(t_J^2, t_c^2), \cdots$, (t_J^k, t_c^k) are numbered in the order of occurrence of Japanese technical terms t_J^i $(i = 1,2, \cdots, k)$ in each Japanese sentence S_J. Here, note that in all the parallel sentence pairs (S_J, S_C), technical term tokens "TT_1," TT_2," and so on that are identical throughout all the parallel sentence pairs are used in this procedure. Therefore, for example, in all the Japanese patent sentences S_J, the Japanese technical term t_J^1 which appears earlier than other Japanese technical terms in S_J is replaced with TT_1. We then train the NMT system on a bilingual corpus, in which the technical term pairs is replaced by "TT_i" $(i = 1,2, \cdots)$ tokens, and obtain an NMT model in which the technical terms are represented as technical term tokens.[7]

2.2.4.2 NMT Decoding and SMT Technical Term Translation

Figure Figure illustrates the procedure for producing Chinese translations via decoding the Japanese sentence using the method proposed in this paper. In the step 1 of Figure Figure, when given an input Japanese sentence, we first automatically extract the technical terms and replace them with the technical term tokens "TT_i"

[5] In this work, we approximately regard all the Japanese compound nouns as Japanese technical terms. These Japanese compound nouns are automatically extracted by simply concatenating a sequence of morphemes whose parts of speech are either nouns, prefixes, suffixes, unknown words, numbers, or alphabetical characters. Here, morpheme sequences starting or ending with certain prefixes are inappropriate as Japanese technical terms and are excluded. The sequences that include symbols or numbers are also excluded. In Chinese side, on the other hand, we regard Chinese translations of extracted Japanese compound nouns as Chinese technical terms, where we do not regard other Chinese phrases as technical terms.

[6] We discard discontinuous sequences and only use continuous ones.

[7] We treat the NMT system as a black box, and the strategy we present in this paper could be applied to any NMT system [1][2][7][8][11][12][17].

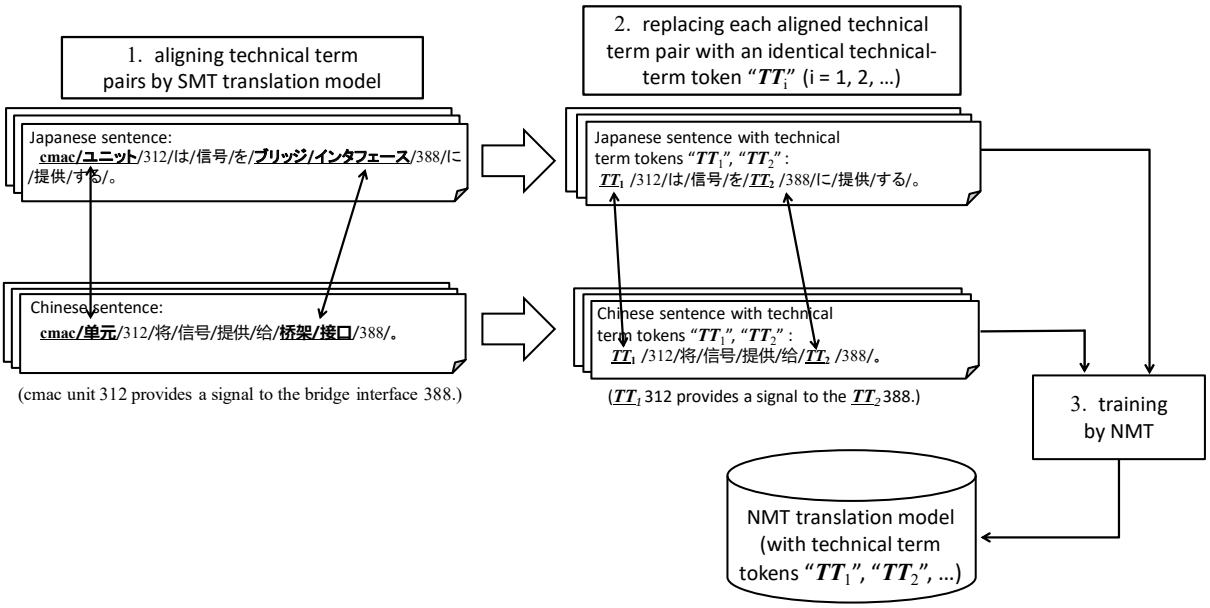


Figure 2 NMT training after replacing technical term pairs with technical term tokens "TT_i" ($i = 1, 2, \cdots$)

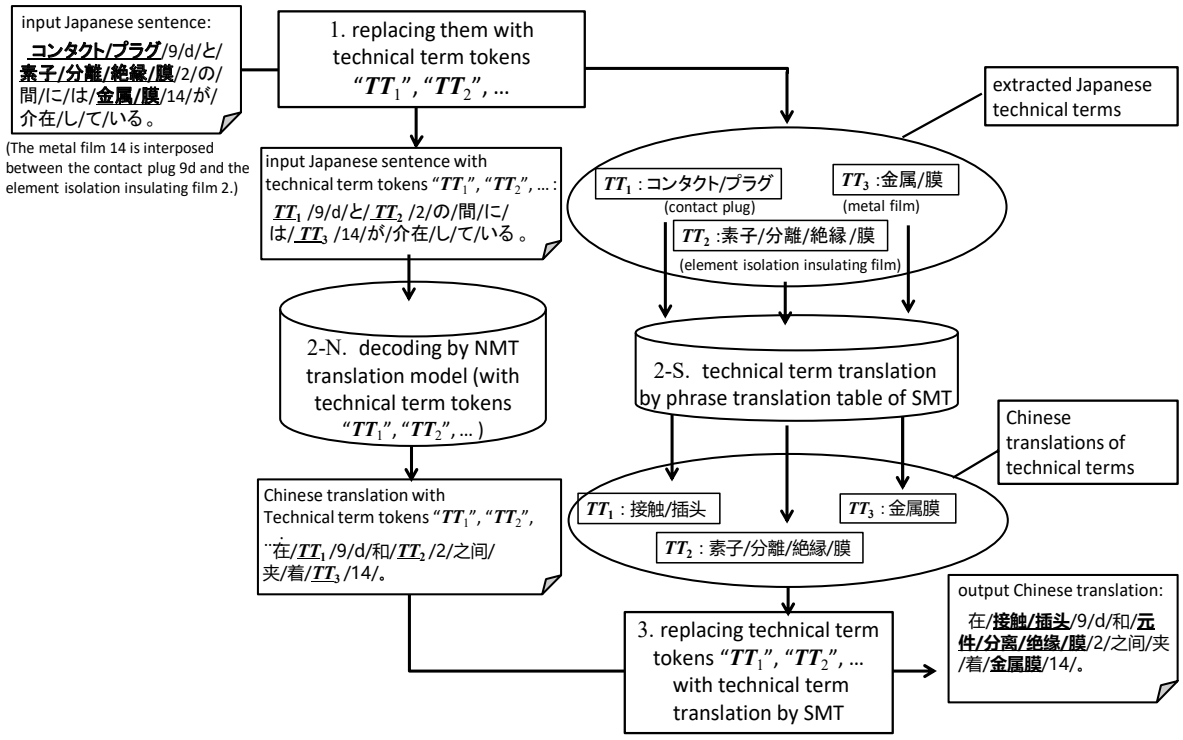


Figure 3 NMT decoding with technical term tokens "TT_i" ($i = 1, 2, \cdots$) and SMT technical term translation

($i = 1,2, \cdots$). Consequently, we have an input sentence in which the technical term tokens "TT_i" ($i = 1,2, \cdots$) represent the positions of the technical terms and a list of extracted Japanese technical

[8] We use the translation with the highest probability in the phrase translation table. When an input Japanese technical term has multiple translations with the same highest probability or has no translation in the phrase translation table, we apply a compositional translation generation approach, wherein Chinese translation is generated compositionally from the constituents of Japanese technical terms.

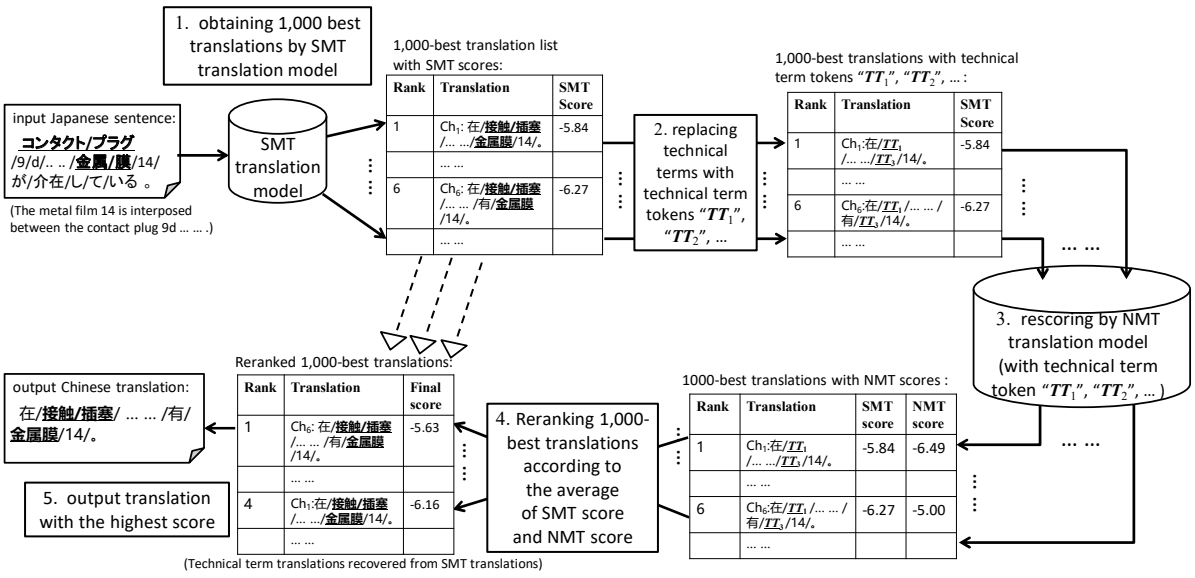


Figure 4 NMT rescoring of 1,000-best SMT translations with technical term tokens "TT_i" ($i = 1, 2, \cdots$)

terms. Next, as shown in the step 2-N of Figure Figure, the source Japanese sentence with technical term tokens is translated using the NMT model trained according to the procedure described in Section 2.2.4.1, whereas the extracted Japanese technical terms are translated using an SMT phrase translation table in the step 2-S of Figure Figure.[8] Finally, in the step 3, we replace the technical term tokens "TT_i" ($i = 1,2, \cdots$) of the sentence translation with SMT the technical term translations.

2.2.4.3 NMT Rescoring of 1,000-best SMT Translations

As shown in the step 1 of Figure Figure , similar to the approach of NMT rescoring provided in Sutskever et al.[17], we first obtain 1,000-best translation list of the given Japanese sentence using the SMT system. Next, in the step 2, we then replace the technical terms in the translation sentences with technical term tokens "TT_i" ($i = 1,2, \cdots$), which must be the same with the tokens of their source Japanese technical terms in the input Japanese sentence. The technique used for aligning Japanese technical terms with their Chinese translations is the same as that described in Section 2.2.4.1. In the step 3 of Figure Figure , the 1,000- best translations, in which technical terms are represented as tokens, are rescored using the NMT model trained according to the procedure described in Section 2.2.4.1. Given a Japanese sentence S_J and its 1,000-best Chinese translations S_C^{n} ($\mathrm{n} = 1,2, \cdots, 1{,}000$) translated by the SMT system, NMT score of each translation sentence pair (S_J, S_C^{n}) is computed as the log probability $\log p(S_C^{\mathrm{n}} | S_J)$ of Equation (1). Finally, we rerank the 1,000-best translation list on the basis of the average SMT and NMT scores and output the translation with the highest final score.

2.2.5 Evaluation

2.2.5.1 Training and Test Sets

We evaluated the effectiveness of the proposed NMT system in translating the Japanese-Chinese parallel patent sentences described in Section 2.2.2. Among the 2.8M parallel sentence pairs, we randomly extracted 1,000 sentence pairs for the test set and 1,000 sentence pairs for the development set; the remaining sentence pairs were used for the training set.

According to the procedure of Section 2.2.4.1, from the Japanese-Chinese sentence pairs of the training set, we collected 6.5M occurrences of technical term pairs, which are 1.3M types of technical term pairs with 800K unique types of Japanese technical terms and 1.0M unique types of Chinese technical terms Out of the total 6.5M occurrences of technical term pairs, 6.2M were replaced with technical term tokens using the phrase translation table, while the remaining 300K were replaced with technical term tokens

Table 1 Automatic evaluation results

System	NMT decoding and SMT technical term translation		NMT rescoring of 1,000-best SMT translations	
	BLEU	RIBES	BLEU	RIBES
Baseline SMT [9]	52.5	88.5	-	-
Baseline NMT	53.5	90.0	55.0	89.1
NMT with technical term translation by SMT	55.3	**90.8**	**55.6**	89.4
NMT with PosUnk model [13]	54.0	90.3	55.5	89.1

using the word alignment.9 We limited both the Japanese vocabulary (the source language) and the Chinese vocabulary (the target language) to 40K most frequently used words.

Within the total 1,000 Japanese patent sentences in the test set, 2,244 occurrences of Japanese technical terms were identified, which correspond to 1,857 types.

2.2.5.2 Training Details

For the training of the SMT model, including the word alignment and the phrase translation table, we used Moses [9], a toolkit for a phrase-based SMT models.

For the training of the NMT model, our training procedure and hyperparameter choices were similar to those of Sutskever et al. [17]. We used a deep LSTM neural network comprising three layers, with 512 cells in each layer, and a 512-dimensional word embedding. Similar to Sutskever et al. [17] we reversed the words in the source sentences and ensure that all sentences in a minibatch are roughly the same length. Further training details are given below:

- All of the LSTM's parameter were initialized with a uniform distribution ranging between -0.06 and 0.06.
- We set the size of a minibatch to 128.
- We used the stochastic gradient descent, beginning at a learning rate of 0.5. We computed the perplexity of the development set using the currently produced NMT model after every 1,500 mini-batches were trained and multiplied the learning rate by 0.99 when the perplexity did not decrease with respect to the last three perplexities. We trained our model for a total of 10 epoches.
- Similar to Sutskever et al. [17], we rescaled the normalized gradient to ensure that its norm does not exceed 5.

We implement the NMT system using TensorFlow,[10] an open source library for numerical computation. The training time was around two days when using the described parameters on an 1-GPU machine.

2.2.5.3 Evaluation Results

We calculated automatic evaluation scores for the translation results using two popular metrics: BLEU [14] and RIBES [6]. As shown in Table Table 1, we report the evaluation scores, on the basis of the translations by Moses [9], as the baseline SMT[11] and the scores based on translations produced by the equivalent NMT system without our proposed approach as the baseline NMT. As shown in Table Table 1, the two versions of the proposed NMT systems clearly improve the translation quality when compared with the baselines. When compared with the baseline SMT, the performance gain of the proposed system is

[9]There are also Japanese technical terms (3% of all the extracted terms) for which Chinese translations can be identified using neither the SMT phrase translation table nor the SMT word alignment.

[10]https://www.tensorflow.org/

Table 2 Human evaluation results (the score of pairwise evaluation ranges from −100 to 100 and thecore of JPO adequacy evaluation ranges from 1 to 5)

System	NMT decoding and SMT technical term translation		NMT rescoring of 1,000-best SMT translations	
	pairwise evaluation	JPO adequacy evaluation	pairwise evaluation	JPO adequacy evaluation
Baseline SMT [9]	-	3.5	-	-
Baseline NMT	5.0	3.8	28.5	4.1
NMT with technical term translation by SMT	**36.5**	**4.3**	31.0	4.1

approximately 3.1 BLEU points if translations are produced by the proposed NMT system of Section 2.2.4.3 or 2.3 RIBES points if translations are produced by the proposed NMT system of Section 2.2.4.2. When compared with the result of decoding with the base line NMT, the proposed NMT system of Section 2.2.4.2 achieved performance gains of 0.8 RIBES points. When compared with the result of reranking with the baseline NMT, the proposed NMT system of Section 2.2.4.3 can still achieve performance gains of 0.6 BLEU points. Moreover, when the output translations produced by NMT decoding and SMT technical term translation described in Section 2.2.4.2 with the output translations produced by decoding with the baseline NMT, the number of unknown tokens included in output translations reduced from 191 to 92. About 90% of remaining unknown tokens correspond to numbers, English words, abbreviations, and symbols.12

Furthermore, we quantitatively compared our study with the work of Luong et al [13]. As the result shown in Table Table 1, compared with the NMT system with PosUnk model that is proposed as the best model by Luong et al. [13], the proposed NMT system achieves performance gains of 1.3 BLEU points and 0.4 RIBES points when the output translations are produced by NMT decoding and SMT technical term translation described in Section 2.2.3.2.

In this study, we also conducted two types of human evaluation according to the work of Nakazawa et al. [14]: pairwise evaluation and JPO adequacy evaluation. During the procedure of pairwise evaluation, we compare each of translations produced by the baseline SMT with that produced by the two versions of the proposed NMT systems, and judge which translation is better, or whether they are with comparable quality. The score of pairwise evaluation is defined by the following formula, where W is the number of better translations compared to the baseline SMT, L the number of worse translations compared to the baseline SMT, and T the number of translations having their quality comparable to those produced by the baseline SMT:

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

The score of pairwise evaluation ranges from −100 to 100. In the JPO adequacy evaluation, Chinese translations are evaluated according to the quality evaluation criterion for translated patent documents proposed by the Japanese Patent Office (JPO).[13] The JPO adequacy criterion judges whether or not the technical factors and their relationships included in Japanese patent sentences are correctly translated into Chinese, and score Chinese translations on the basis of the percentage of correctly translated information, where the score of 5 means all of those information are translated correctly, while that of 1 means most

[11] We train the SMT system on the same training set and tune it with development set.

[12] In addition to the two versions of the proposed NMT systems presented in Section 2.2.4, we evaluated a modified version of the proposed NMT system, where we introduce another type of token corresponding to unknown compound nouns and integrate this type of token with the technical term token in the procedure of training the NMT model. We achieved a slightly improved translation performance, BLEU/RIBES scores of 55.6/90.9 for the proposed NMT system of Section 2.2.4.2 and those of 55.7/89.5 for the proposed NMT system of Section 2.2.4.3.

[13] https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/tokkyohonyaku_hyouka/01.pdf (in Japanese)

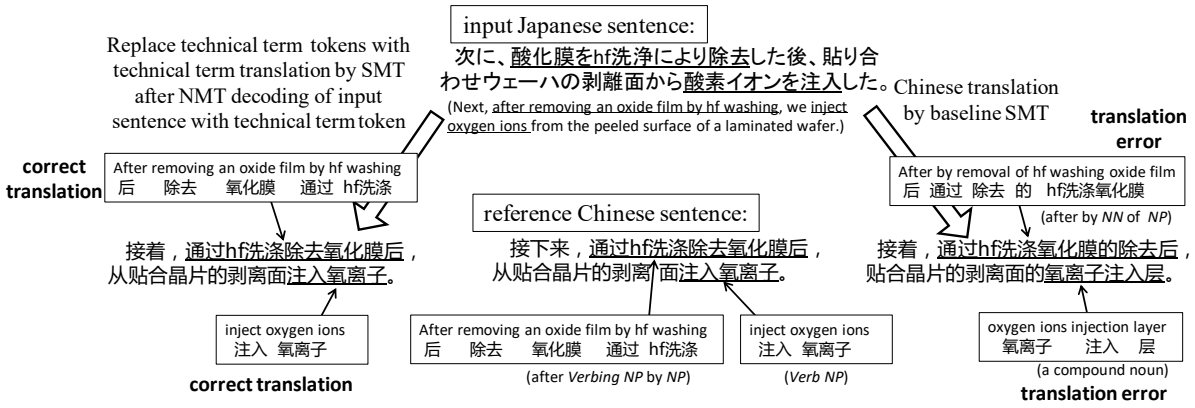


Figure 5 Example of correct translations produced by the proposed NMT system with SMT technical term translation (compared with baseline SMT)

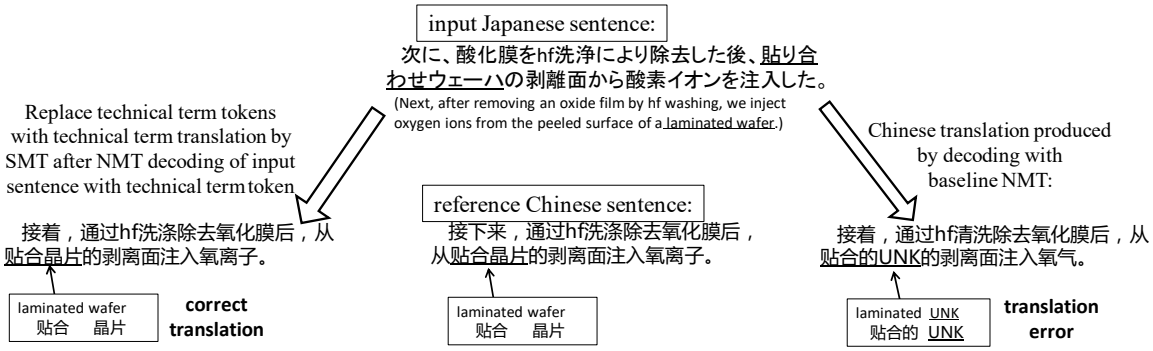


Figure 6 Example of correct translations produced by the proposed NMT system with SMT technical term translation (compared to decoding with the baseline NMT)

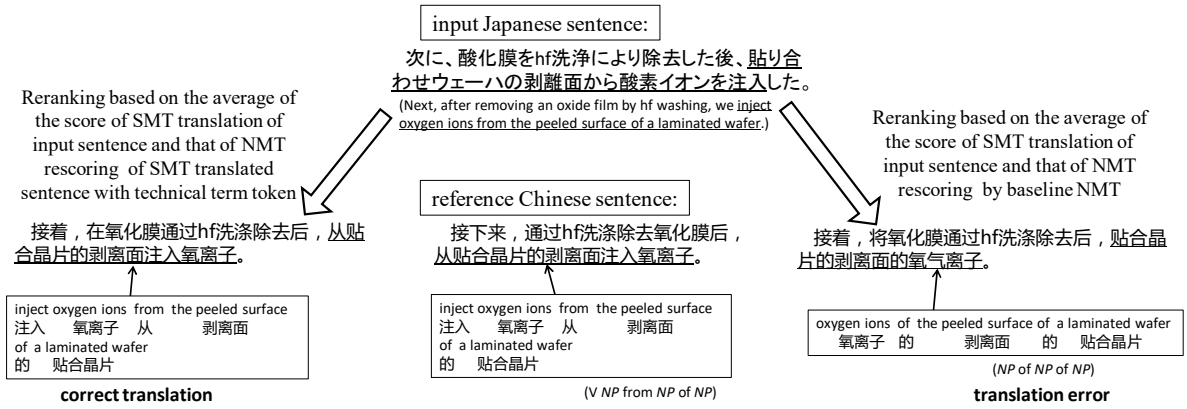


Figure 7 Example of correct translations produced by reranking the 1,000-best SMT translations with the proposed NMT system (compared to reranking with the baseline NMT)

randomly selected 200 sentence pairs from the test set for human evaluation, and both human evaluations were conducted using only one judgement. Table Table 2 shows the results of the human evaluation for the baseline SMT, the baseline NMT, and the proposed NMT system. We observed that the proposed system achieved the best performance for both pairwise evaluation and JPO adequacy evaluation when we replaced technical term tokens with SMT technical term translations after decoding the source sentence with technical term tokens.

Throughout Figure Figure ~Figure Figure , we show an identical source Japanese sentence and each of its produced by the two versions of the proposed NMT systems, compared with translations produced by the three baselines, respectively. Figure Figure shows an example of correct translation produced by the proposed system in comparison to that produced by the baseline SMT. In this example, our model correctly

translates the Japanese sentence into Chinese, whereas the translation by the baseline SMT is a translation error with several erroneous syntactic structures. As shown in Figure Figure, the second example highlights that the proposed NMT system of Section 2.2.4.2 can correctly translate the Japanese technical term “貼り合わせウェーハ”(laminated wafer) to the Chinese technical term “贴合晶片”. The translation by the baseline NMT is a translation error because of not only the erroneously translated unknown token but also the Chinese word “贴合的”, which is not appropriate as a component of a Chinese technical term. Another example is shown in Figure Figure , where we compare the translation of a reranking SMT 1,000-best translation produced by the proposed NMT system with that produced by reranking with the baseline NMT. It is interesting to observe that compared with the baseline NMT, we obtain a better translation when we rerank the 1,000-best SMT translations using the proposed NMT system, in which technical term tokens represent technical terms. It is mainly because the correct Chinese translation “晶片”(wafer) of Japanese word “ウェーハ” is out of the 40K NMT vocabulary (Chinese), causing reranking with the baseline NMT to produce the translation with an erroneous construction of “noun phrase of noun phrase of noun phrase”. As shown in Figure Figure , the proposed NMT system of Section 2.2.4.3 produced the translation with a correct construction, mainly because Chinese word “晶片”(wafer) is a part of Chinese technical term “贴合晶片”(laminated wafter) and is replaced with a technical term token and then rescored by the NMT model (with technical term tokens “TT_1,” “TT_2, ” and so on.

2.2.6 Conclusion

In this paper, we proposed an NMT method capable of translating patent sentences with a large vocabulary of technical terms. We trained an NMT system on a bilingual corpus, wherein technical terms are replaced with technical term tokens; this allows it to translate most of the source sentences except the technical terms. Similar to Sutskever et al. [17], we used it as a decoder to translate the source sentences with technical term tokens and replace the tokens with technical terms translated using SMT. We also used it to rerank the 1,000-best SMT translations on the basis of the average of the SMT score and that of NMT rescoring of translated sentences with technical term tokens. For the translation of Japanese patent sentences, we observed that our proposed NMT system performs better than the phrase-based SMT system as well as the equivalent NMT system without our proposed approach.

One of our important future works is to evaluate our proposed method in the NMT system proposed by Bahdanau et al. [1], which introduced a bidirectional recurrent neural network as encoder and is the state-of-the-art of pure NMT system recently. However, the NMT system proposed by Bahdanau et al. [1] also has a limitation in addressing out-of-vocabulary words. Our proposed NMT system is expected to improve the translation performance of patent sentences by applying approach of Bahdanau et al. [1]. We will also evaluate the present study by reranking the n-best translations produced by the proposed NMT system on the basis of their SMT rescoring. Next, we will rerank translations from both the n-best SMT translations and n-best NMT translations. As shown in Section 2.2.5.3, the decoding approach of our proposed NMT system achieved the best RIBES performance and human evaluation scores in our experiments, whereas the reranking approach achieved the best performance with respect to BLEU. A translation with the highest average SMT and NMT scores of the n-best translations produced by NMT and SMT, respectively, is expected to be an effective translation.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. 3rd ICLR*.

[2] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proc. EMNLP*.

[3] M. R. Costa-jussà and J. A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proc. 54th ACL*, pages 357–361.

[4] L. Dong, Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. 2015. Collecting bilingual technical terms from Japanese-Chinese patent families by SVM. In *Proc. PACLING*, pages 71–79.

[5] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
[6] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proc. EMNLP*, pages 944–952.
[7] S. Jean, K. Cho, Y. Bengio, and R. Memisevic. 2014. On using very large target vocabulary for neural machine translation. In *Proc. 28th NIPS*, pages 1–10.
[8] N. Kalchbrenner and P. Blunsom. 2013. Recurrent continous translation models. In *Proc. EMNLP*, pages 1700–1709.
[9] 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. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pages 177–180.
[10] X. Li, J. Zhang, and C. Zong. 2016. Towards zero unknown word in neural machine translation. In *Proc. 25th IJCAI*, pages 2852–2858.
[11] M. Luong and C. D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word- character models. In *Proc. 54th ACL*, pages 1054–1063.
[12] M. Luong, H. Pham, and C. D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*.
[13] M. Luong, I. Sutskever, O. Vinyals, Q. V. Le, and W. Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *Proc. 53rd ACL*, pages 11–19.
[14] T. Nakazawa, H. Mino, I. Goto, G. Neubig, S. Kurohashi, and E. Sumita. 2015. Overview of the 2nd workshop on asian translation. In *Proc. 2nd WAT*, pages 1–28.
[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pages 311–318.
[16] R. Sennrich, B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. 54th ACL*, pages 1715–1725.
[17] I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural machine translation. In *Proc. 28th NIPS*.
[18] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pages 168–171.
[19] M. Utiyama and H. Isahara. 2007. A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pages 475–482.
[20] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. 2015. Grammar as a foreign Language, in *Proc. NIPS(2015)*

2. 3 機械翻訳評価のための項目反応理論に基づく一対比較結果の統合

京都大学　大谷　直樹
中澤　敏明
黒橋　禎夫

2.3.1 Introduction

Manual evaluation is a primary means of interpreting the performance of machine translation (MT) systems and evaluating the accuracy of automatic evaluation metrics. It is also essential for natural language processing tasks such as summarization and dialogue systems, where (1) the number of correct outputs is unlimited, and (2) naïve text matching cannot judge the correctness, that is, an evaluator must consider syntactic and semantic information.

Recent work has used crowdsourcing to reduce costs of manual evaluations. However, the judgments of crowd workers are often noisy and unreliable because they are not experts.

To maintain quality, evaluation tasks implemented using crowdsourcing should be simple. Thus, many previous studies focused on pairwise comparisons instead of absolute evaluations. The same task is given to multiple workers, and their responses are aggregated to obtain a reliable answer.

We must, therefore, develop methods that robustly estimate the MT performance based on many pairwise comparisons.

Some aggregation methods have been proposed for MT competitions hosted by the Workshop on Statistical Machine Translation (WMT) (Bojar et al., 2015), where a ranking of the submitted systems is produced by aggregating many manual judgments of pairwise comparisons of system outputs.

However, existing methods do not consider the following important issues.

Interpretability of the estimates: For the purpose of evaluation, their results must be interpretable so that we could use the results to improve MT systems and the next MT evaluation campaigns. Existing methods, however, only yield system-level scores.

Judge sensitivity: Some judges can examine the quality of translations with consistent standards, but others cannot~¥cite{Graham2015Can}. Sensitivities to the translation quality and judges' own standards are important factors.

Evaluation of a newly submitted system: Previous approaches considered all pairwise combinations of systems and must compare a newly submitted system with all the submitted systems. This made it difficult to allow participants to submit their systems after starting the evaluation step.

To address these issues, we use a model from item response theory (IRT). This theory was originally developed for psychometrics, and has applications to academic tests. IRT models are highly interpretable and are supported by theoretical and empirical studies. For example, we can

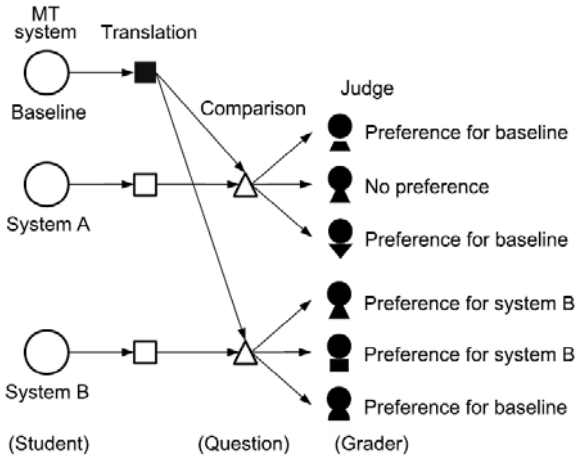


Figure 1: Illustration of manual pairwise comparison. Each system yields translations. Judges compare them with a baseline translation and report their preferences. Our goal is to aggregate the judgments to determine the performance of each system.

estimate the informativeness of a question in a test based on the responses of examinees.

We focused on aggregating many pairwise comparisons with a baseline translation so that we could use the analogy of standard academic tests. Figure 1 shows our problem setting. Each system of interest yields translations, and the translations are compared with a baseline translation by multiple human judges. Each judge produces a preference judgment.

The pairwise comparisons correspond to questions in academic tests, a judge's sensitivity to the translation quality is mapped to discrimination of questions, and the relative difficulty of winning the pairwise comparison is mapped to the difficulty of questions. MT systems correspond to students that take academic tests, and IRT models can be naturally applied to estimate the latent performance (ability) of MT systems (students).

Additionally, our approach, fixing baseline translations, can easily evaluate a newly submitted system. We only need to compare the new system with the baseline instead of testing all pairwise combinations of the submitted systems.

Our contributions are summarized as follows.[1]

1. We propose an IRT-based aggregation model of pairwise comparisons with highly interpretable parameters.
2. We simulated noisy judges on the WMT13 dataset and demonstrated that our model is less affected by the noisy judges than previously proposed methods.

[1] We also show that our method accurately replicated the WMT13 official system scores using a few comparisons. However, this is not the main focus of this paper.

2.3.2 Problem Setting

We first describe the problem setting, as shown in Figure 1.

Assume that there is a group of systems I indexed by i, a set of segments J indexed by j, and a set of judges K indexed by k.

Before a manual evaluation, we fix an arbitrary baseline system and use it to translate the segments J. Then, each system $i \in I$ produces a translation on segment $j \in J$. One of the judges $k \in K$ compares it with the baseline translation. The judge produces a preference judgment.

Let $\mathrm{u_{i,j,k}}$ be the observed judgment that judge k assigns to a translation by system i on segment j, that is,

$$\mathrm{u_{i,j,k}} = \begin{cases} 1 \text{ (preference for baseline)} \\ 2 \text{ (no preference)} \\ 3 \text{ (preference for system } i) \end{cases},$$

and let $\mathrm{c} \in \{1, 2, 3\}$ be the judgment label.

Each system i has its own latent performance $\theta_\mathrm{i} \in \mathbb{R}$. Our goal is to estimate $¥theta$ by using the observed judgments $U = \{\mathrm{u_{i,j,k}}\}_{i \in I, j \in J, k \in K}$.

2.3.3 Generative Judgment Model

We describe a statistical model for pairwise comparisons based on an IRT model.

Modified Graded Response Model

Based on the **graded response model (GRM)** proposed by (Samejima, 1968), we define a generative model of judgments. GRM deals with responses on ordered categories including ratings such as A+, A, B+ and B, and partial credits in tests. In our problem setting, judgments can be seen as partial credits. When a system beats a baseline translation, the system receives $c = 3$ credit. In the case of a tie, the system receives $c = 2$ credit. The system receives $c = 1$ credit when it lose to the baseline.

Let $\mathrm{P^*_{jkc}}(\theta_\mathrm{i})$ be the probability that judge k assigns judgment $\pi > c$ to a comparison on segment j between system i and a baseline.

$$\mathrm{P^*_{jkc}}(\theta_\mathrm{i}) = \frac{1}{1 + \exp(-a_\mathrm{k}(\theta_\mathrm{i} - b_\mathrm{jc}))}$$

where $\mathrm{P^*_{jk0}}(\theta_i) = 1$, $\mathrm{P^*_{jk3}}(\theta_\mathrm{i}) = 0$. Parameters a and b are called *discrimination* and *difficulty* parameters, respectively. a represents the discriminablity or sensitivity of the judge, and b represents a segment-specific difficulty parameter. The discrimination parameter *a* is positive, and the difficulty parameter b satisfies $\mathrm{b_1} < \mathrm{b_2}$, where $\mathrm{b_1}$ corresponds to the difficulty of not losing to the baseline ($\mathrm{c} > 1$, and $\mathrm{b_2}$ corresponds to the difficulty of beating the baseline $\mathrm{c} > 2$).

The generative probability of judgment $u_{i,j,k}$ is defined as the difference in the probabilities defined above, that is,

$$P_{jkc}(\theta_i) = P(u_{i,j,k} = c \mid \theta_i, b_j, a_k) = P^*_{jkc-1}(\theta_i) - P^*_{jkc}(\theta_i).$$

The model described above is different from the original GRM, which assumed that the values of a are independent from question to question, and that each a belongs to exactly one question. However, in our problem setting, the judges evaluate multiple segments, and discrimination parameter a is independent from segment j. This modification means that the GRM can capture the judge's sensitivity.

Priors

We assign prior distributions to the parameters to obtain estimates stably. We assume Gaussian distributions on θ and b, that is, $\theta \sim N(0, \tau^2)$ and $b_c \sim N(\mu_{bc}, \sigma^2_{bc})$ $(c = 1, 2)$. The discrimination parameter is positive, so we assume a log Gaussian distribution on a, i.e., $\log(a) \sim N(\mu_a, \sigma^2_a)$. Note that τ, μ and σ are hyper parameters.

2.3.4 Parameter Estimation

We find the values of the parameters to maximize the log likelihood based on the judgments U:

$$\mathcal{L}(\theta, \xi) = \log P(U; \theta, \xi).$$

We denote the parameters $a = \{a_k\}_{k \in K}$ and $b = \{b_{j1}, b_{j2}\}_{j \in J}$ to be ξ in this section.

Marginal Likelihood Maximization of Judge Sensitivity and Matchup Difficulty

Estimates are known to be inaccurate when all the parameters are optimized at once, so we first estimate the parameters ξ to maximize the marginal log likelihood w.r.t. the system performance θ.

$$m\mathcal{L}(\xi) = \log P(U, \xi) = \sum_{i \in I} \log \int_{-\infty}^{\infty} P(\theta) P(U_i | \theta, \xi) d\theta + \log P(\xi),$$

where U_i is the set of judgments given to system i The equation above can be approximated using Gauss-Hermite quadrature.

We solve the optimization problem using the gradient descent methods to maximize the approximated marginal likelihood. The inequality constraints on the parameters are handled by adding log barrier functions to the objective function.

Maximum A Posteriori (MAP) Estimation of System Performance

Given the estimates of ξ, we estimate the system performance $\theta = \{\theta_i\}_{i \in I}$ by using MAP estimation.

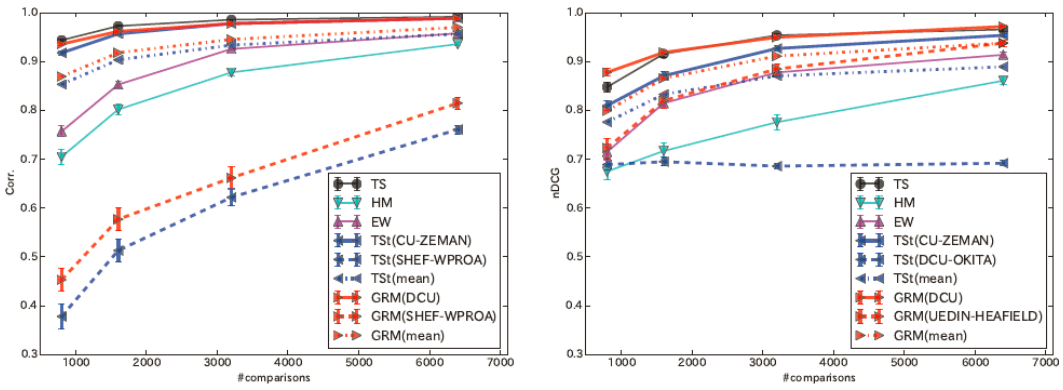
We maximize the objective function,

$$\mathcal{L}(\theta) = \log P(U, \theta; \xi) \sum_{i \in I} (\log P(\theta_i) + \log P(U_i \mid \theta_i; \xi)).$$

The estimates of θ are obtained using the gradient descent method.

2.3.5 Experiments

We conducted experiments on the WMT13 manual evaluation dataset for 10 language pairs. For details of the evaluation data, see the overview of WMT13 (Bojar et al., 2013).



(a) Correlation (b) nDCG

Figure 2: Correlation and nDCG comparing the estimated system performance and gold scores with the number of comparisons for the WMT13 Spanish–English task.

Setup

Models: Our method **(GRM)** was initialized using $a = 1.7, b = (-0.5, 0.5)$, and a θ value derived by summing up the judgments for each system and scaling θ to fit the prior distribution. For the hyper parameters, we set $\tau = \sqrt{2},\ \mu_a = \log(1.7),\ \sigma_a = 1.0,\ \mu_b = (-0.5, 0.5),\ \sigma_b = 2.0$.

To compare with our method, we trained ExpectedWins **(EW)** (Bojar et al., 2013), the model by Hopkins and May (2013), **(HM)** and the two-stage crowdsourcing model proposed by Baba and Kashima (2013) **(TSt)**. We also trained TrueSkill **(TS)** (Sakaguchi et al., 2014), which was used to produce the gold score on this experiment. We followed Sakaguchi et al. (2014), who also used the WMT13 datasets in their experiments, and initialized the HM and TS parameters. For TSt, we followed Baba and Kashima (2013).

Pairwise comparisons: The WMT dataset contains five-way partial rankings, so we converted the five-way partial rankings into pairwise comparisons. We randomly sampled 800, 1,600, 3,200 and 6,400 pairwise comparisons from the whole dataset. TS first receives all the pairwise comparisons and selects the training data based on the active learning strategy, whereas we sampled the comparisons before running the other methods.

Gold scores: We made gold scores with TS. We produced 1,000 bootstrap-resampled datasets over all of the available comparisons and collected multiple system scores by TS. The gold score is the mean of the scores.

Evaluation metrics: We evaluated the models using the Pearson correlation coefficient and the

Table 1: Correlation and nDCG between the estimated system performance and gold scores for the WMT13 Spanish--English task, based on noisy judges. The values were averaged over all the datasets. The GRM scores were averaged over all baselines. The differences from the GRM are reported for the HM and EW.

Noisy (%)	0	10	20	30	40	50
Correlation						
GRM	.929	.917	.900	.879	.849	.807
HM	+.002	-.005	-.009	-.015	-.025	-.038
EW	-.025	-.028	-.035	-.038	-.040	-.046
nDCG						
GRM	.883	.867	.847	.822	.793	.752
HM	-.024	-.130	-.137	-.144	-.152	-.168
EW	-.035	-.054	-.064	-.060	-.060	-.069

normalized discounted cumulative gain (nDCG), comparing the estimated scores and gold scores. We used nDCG because we are often interested in ranks and scores, especially in MT competitions such as the WMT translation task. These metrics were also used for experiments by Baba and Kashima (2013).

Results

Figure 2 shows the correlation and nDCG between the estimated system performance and the gold scores for the WMT13 Spanish--English task. For the GRM and TSt, the baselines used in the evaluation are shown in parentheses in the labels. Note that the main contribution of our method is not to perform better than other methods in terms of correlation and nDCG to the gold scores, but to result in highly interpretable and robust estimates discussed later.

TS resulted in the highest correlation and nDCG. It is reasonable because the gold scores themselves were produced by TS, and because it estimates the parameters using active learning, unlike the other models.

The GRM with the best baseline system (DCU) achieved almost the same scores as the TS, in terms of correlation and nDCG. Although the TSt with the best baseline resulted in accurate estimates in terms of correlation, it did not in terms of nDCG. With the worst baselines, the GRM and TSt both failed to replicate the gold scores, but the GRM was surprisingly accurate in terms of nDCG (even in the worst case). This implies that the GRM can effectively predict the top ranked systems.

Baseline Selection

It is likely that single pairwise comparisons do not work well if the baseline is very strong or

weak. As shown in Figure 2, the baseline system influences the final result. When we used SHEF-WPROA as baseline, the estimated system performance was not accurate. This is because SHEF-WPROA loses 69.4% of the pairwise comparisons and fails to discriminate between the other systems. In contrast, DCU loses 34.5% and win 34.8% of the comparisons and discriminate the other systems successfully. Thus, when we used DCU as baseline, the best correlation and nDCG were achieved. Therefore, we must determine the appropriate baseline system before the comparisons.

One possible solution is to consider the system-level scores yielded by automatic evaluation metrics such as BLEU and METEOR. We obtained relatively good results when we used a system whose system-level BLEU score and METEOR score were close to the mean of all the systems.

Analysis of Judge Sensitivity

To investigate the robustness of the GRM, we simulated *noisy* judges. We selected a subset of judges and randomly changed their decisions based on a uniform distribution. The percentage of noisy judges varied between 10% and 50% (in increments of 10%).

We trained HM and EW on the simulated datasets. We excluded TS because it assumes that we can actively request more comparisons from judges when their decisions are ambiguous.

As shown in Table 1, the accuracy of the GRM was less affected by the noisy judges than HM and EW. This is because our model estimates judge-specific sensitivities and automatically reduces the influence of the noisy judges.

Analysis of the Interpretability of the Estimated Matchup Difficulty

Our model is a natural extension of the GRM, so we can apply standard analyses for IRT models. Item information is one of the standard analysis methods and corresponds to sensitivity to a latent parameter of interest. Based on the item information, we can find which segment was difficult to be translated better than a baseline translation.

The item information is calculated using the estimated parameters ξ, that is,

$$\mathrm{I_j}(\theta) = -\mathrm{E}\left[\frac{\partial^2 \mathcal{L}(\theta;\xi)}{\partial \theta^2}\right] = \sum_{c=1}^{3}\left[\frac{\partial^2 \log \mathrm{P_{jkc}}(\theta)}{\partial \theta^2}\right]\mathrm{P_{jkc}} = \sum_{c=1}^{3}\frac{\left[\mathrm{P'_{jkc-1}}(\theta)-\mathrm{P'_{jkc}}(\theta)\right]^2}{\mathrm{P^*_{jkc-1}}(\theta)-\mathrm{P^*_{jkc}}(\theta)},$$

where $\mathrm{P'} = \partial \mathrm{P} / \partial \theta$. Because the item information is only determined by segments and is independent of the judges, we set $a_{\mathrm{k}} = 1\ (k \in K)$.

Table 2 shows translations for segments 1858 and 1818. We found that the baseline translation on segment 1818 was relatively good, whereas the baseline translation on segment 1858 contained wrong words such as “drink” and “galaxias”. Consequently, systems with low θ tended to lose to the baseline on segment 1858 due to their wrong translation (see the translation of “hawaiano de Mauna Kea”). In contrast, some of the low-ranked systems beat the baseline on segment 1818, and the segment contributed to discriminate them.

The item information is used to design academic tests that can effectively capture students' abilities. It could analogously be used to preselect segments to be translated based on the item information in the MT evaluation.

Table 2: Translation examples for the WMT13 Spanish--English task. The reference is a correct translation given by the WMT organizers and was shown to human judges. Estimates of $\boldsymbol{\theta}$ (averaged over 100 sampled datasets with 6,400 comparisons) are also reported in the table.

Segment 1858: Difficult to beat the baseline translation.

Source		Hasta 2007 los dos telescopios Keck situados en el volcán hawaiano de Mauna Kea eran considerados los más grandes del mundo.
Reference		Until 2007, the two Keck telescopes at the Hawaiian volcano, Mauna Kea, were the largest in the world.
DCU[baseline]		Until 2007, the two Keck telescopes located on the Hawaiian volcano Mauna of KEA were considered the largest in the world.
ONLINE-B	$\theta =$ 0.24	Until 2007 the two Keck telescopes located on the Hawaiian volcano Mauna Kea were considered the largest in the world.
UEDIN	0.12	Until 2007, the two Keck telescopes located on the Hawaiian volcano of Mauna Kea were considered the largest in the world.
CU-ZEMAN	-0.1	Until 2007, the two Keck telescope located in the volcano Mauna Kea hawaiano of were regarded as the world's largest.
JHU	-0.12	Until 2007, the two Telescope Keck located in the Kea volcano hawaiano of Mauna were considered the world's largest.
SHEF-WPROA	-0.92	Until 2007 the two telescope Keck located volcano hawaiano of Mauna KEA were regarded larger of world.
Segment 1818: Easy to beat the baseline translation.		
Source		Dependiendo de las tonalidades, algunas im・ágenes de galaxias espirales se convierten en verdaderas obras de arte.
Reference		Depending on the colouring, photographs of spiral galaxies can become genuine work of art.
DCU[baseline]		Depending on the drink, some images of galaxias galaxies become true works of art.
ONLINE-B	0.24	Depending on the shades, some images of spiral galaxies become true works of art.
UEDIN	0.12	(Same as ONLINE-B)
CU-ZEMAN	-0.1	Depending on the tonalidades, some images of spirals galaxies become true works of art.
JHU	-0.12	Depending on the tonalidades, some images of galaxies spirals become true works of art.
SHEF-WPROA	-0.92	Depending on the tonalidades, some images of galaxies spirals become real artwork.

2.3.6 Conclusion

We have addressed the task of manual judgment aggregation for MT evaluations. Our motivation was three folded: (1) to incorporate a judge's sensitivity to robustly measure a system's performance, (2) to maintain highly interpretable estimates, and (3) to handle with a newly submitted system.

To tackle these problems, we focused on pairwise comparisons with a fixed baseline translation so that we could apply the GRM model in IRT by using the analogy of standard academic tests. Unlike testing all pairwise combinations of systems, fixing baseline translations makes it easy to evaluate a newly submitted system. We demonstrated that our model gave robust and highly interpretable estimates on the WMT13 datasets.

References

- Yukino Baba and Hisashi Kashima. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 554–562, New York, USA, August. ACM Press.
- Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.
- Mark Hopkins and Jonathan May. 2013. Models of Translation Competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1416–1424, Sofia, Bulgaria. Association for Computational Linguistics.
- Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, pages 1–11, Baltimore, Maryland, USA.
- Fumiko Samejima. 1968. Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, 1968(1):i–169, June.

2.4 特許文請求項の構造に関する調査

山形大学名誉教授　横山晶一

2.4.1　はじめに

特許文において、課題や解決手段、請求項 1 の部分が複雑な係り受け構造を持ち、しばしば 120 字を超える長大な文または名詞句になるということは、すでに何度も言及してきた[1-4]。

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

さらに、長い修飾句の中に含まれる機能語に着目し、その性質から修飾句を形式的に分割する可能性について調査した結果についても述べた[10]。

統計的機械翻訳(SMT)においては、請求項の構造を sublanguage と捉えて組み込むことによって、翻訳品質が著しく改善されることが報告されている[11, 12]。Sublanguage としては、並列助詞「と」や機能語が一部扱われているが、意味構造を的確に把握するためには、なお解析が必要と考えられる。

本稿では、請求項 1 に含まれる構造にどのようなものがあるかを改めて考察し、さらに詳細な係り受け構造や意味構造を得るための可能性について言及する。なお、本稿は、すでに発表した調査結果[13]に、さらに新しいデータを付け加えて考察しなおしたものである。

2.4.2　資料・調査方法

2003 年の特許公開情報約 300 文献（特開 2003-180203～180492）の中から、請求項 1 の一文が 120 字を超えるものを 150 選び出した（二文以上から成るが、一文が 120 字を超えるものも含む)。文字数の内訳を表 1 に示す。

表 1　調査特許文請求項 1 の文字数の内訳

文字数	120～199	200～299	300～399	400 以上
特許数	73	56	12	9

この表から分かるように、約半数が 200 字未満である。最も短い文は 121 文字、最も長い文は 665 文字、平均で約 226 字であった。対象とした文献の中に、一文で 1807 文字という

特異的に長大な文が含まれていたが、他の文献と比較して突出しているので、解析対象からは除外した。ただし、構造的には、他の文献と同様に、いくつかの長い名詞句が並列し、それが入れ子の形になって全体として長大な名詞句を構成していることに変わりはなく、ところどころに改行を入れていることから、他の文献と同様の扱いは可能である。

これらについて、次のような観点から分析を行った。

(1) 機能語、改行による文の分割

機能語とは、以前の報告でも述べたように、ここでは、日本語の複数の形態素から成る複合語の中で、いわゆる「つなぎ言葉」的な役割をになうものと定義し、「〜において」、「〜であって」、「〜に関して」などを示す。

長い文からなる請求項 1 には、文自体に改行を入れることによって、文の構造や係り受けを明確化しているものもある。ここではそれについても調査した。

(2) 並立助詞、動詞による並列構造

典型的には「と」による並列構造の構築があげられる。すでに述べたことと重複する部分もある[5〜7]が、それらについて調査した。

(3) 照応的な語による階層構造の形成

「後」、「とともに」(「と共に」と書かれる場合もある)、「前記」、「該」など、照応的に言及する語を用いて階層構造を作ることがしばしば見られる。これらについて調査を行う。

以下では、これらの調査結果について述べる。

2.4.3 機能語・改行による文分割と係り受けの明確化

機能語については、すでに述べたように、典型的な形として、「〜[名詞句 A] [機能語]、…した[名詞句 A]」となり、その場合には多くが機能語を境として分割できることが分かっている[10]。今回調査した特許文中に出現した機能語の内訳を表 2 に示す。

表 2　特許文中の機能語の内訳

	典型	非典型	その他	
			読点あり	読点なし
であって	41	3	1	0
において	34	6	4	9

表 2 では、上記のような形で文が分離できる典型例が、「であって」では 41、「において」では 34 例あり、分離できるが典型例ではない(つまり名詞句の形が上記と異なる)ものが、それぞれ 3, 6 あることを示している。その他は、入れ子構造の下部、すなわち修飾句の一部になっていて、分離できないもので、「であって」の場合には、典型例とともに出現したものが 1 例のみある。「において」では、「において、」と読点がつくものが 4、つかないも

のが 9 あることを示している。

次に、改行であるが、長文をなるべく分かりやすくするために、文中に改行を入れているものが、今回の調査では 49 文あった。このうち、文の構造を考慮せずに改行を入れている（むしろ解析には妨げとなる）と考えられる 1 例を除いて、48 例は、改行を入れることによって、係り受け関係が比較的明瞭になっている。

また、機能語と改行を併用している例も比較的多く見られた。表 2 の「であって」の典型例 41 のうち、改行を含むものが 19、「において」の典型例 34 のうち、改行を含むものが 14 あった。この他に「において」の非典型例にも改行を伴った文が 2 例あった。

図 1 に、「であって」の典型例と、改行をともに含む例を示す。

バンド駒をピンによって連結し構成されるバンドのバンド構造であって、
ピンを固着せず、ピンを挿通させるためのバンド駒のピン穴は、ピンを被覆するパイプを備え、
ピンを固着するバンド駒のピン穴は、ピンの表面又は／及びピン穴の内壁に施されたメッキによって当該ピンと溶接されていることを特徴とするバンド構造。

図 1　機能語「であって」と改行を含む例（特開 2003-180414）

この図では、最初の 1 行は残りの行とは独立しており、その後の改行が、一部次の改行との並列句となるとともに階層構造も形成し、結局最後の「バンド構造」に係る構造となっている。文自体はそれほど長くはないが、改行によって、係り受け構造が比較的明瞭に捉えられて分かりやすくなっている。上にも述べたように、典型例では、機能語「であって、」の前の名詞句（ここでは「バンド構造」）が、最後の名詞句と同じになることが特徴である。

同様に、「において」の例を図 2 に示す。

シートの左右側に配置されるシートリクライニング装置を相互に連結するシャフトと、
前記各シートリクライニング装置のロック機構を構成し前記シャフトの少なくとも一方側に対して所定の隙間を有して係合するカムと、
該カムを回転して前記ロック機構を解除するように前記シャフトに固定される操作レバーとを有するシートバック角度調整システムにおいて、
前記シャフトと前記カムの間の前記隙間に弾性体を挿入して構成していることを特徴とするシートバック角度調整システム。

図 2　機能語「において」と改行を含む例（特開 2003-180478）

この例では、並立助詞「と」による名詞句の並列構造と、改行、機能語「において」で長

い名詞句が分離されることから、比較的構造がとらえやすいと考えられる。

2.4.4　並立助詞および動詞による並列構造

前節にも述べたように、助詞「と」で並列構造を作る場合が多く見られる。今回もこのような例が 27 見出された。また、動詞や形容詞の連用形を繰り返すことによる並列も 43 あった。このうち 7 例は、両方が用いられており、階層構造を形成する場合もあるので解析には注意が必要である。

並立助詞「と」と、動詞の連用形による並列構造が用いられた例を図 3 に示す。

開口部が形成された外容器本体と、前記開口部に着脱可能に係合する肩部材と、を有する外容器と、
前記外容器本体の開口部から外容器本体内部に着脱自在に収納されるレフィル容器と、
ポンプの押圧ボタンとなる頭部と、該頭部より径の大きい筒状部とを有し、前記レフィル容器に螺合するディスペンサーヘッドと、を有するレフィル式ディスペンサー容器において、
前記肩部材は前記ディスペンサーヘッドの頭部が突出する開口部を有し、該肩部材を外容器本体に取り付けた際は、前記ディスペンサーヘッドの頭部が肩部材の開口部から突出した状態で、前記肩部材により前記ディスペンサーヘッドの筒状部が押さえられることを特徴とするレフィル式ディスペンサー容器。

図 3　助詞、動詞等による並列構造の例（特開 2003-180445）

この図では、「と」による並列構造、「において」による名詞句の分離（典型例）、動詞の連用形（「有し」など）、改行がすべて含まれており、これも比較的解析が容易な例となっている。

また、箇条書きを用いた並列構造も 9 例確認された。改行と併用することによって、構造が比較的分かりやすくなっている例を図 4 に示す。

細胞を三次元的に、かつ高密度に培養するためのモジュールであって、少なくとも
(a)中空糸膜と、該中空糸膜の一端に連結した培養液導入口、該中空糸膜の他方の一端に連結した培養液排出口、および開閉可能な細胞接種用口を有する中空糸膜モジュールと、
(b)細胞を培養するための担体であって中空糸膜モジュール中に設けられた多孔性担体と、(c)通気手段および攪拌手段を有する培養液貯槽と、(d)該培養液貯槽から該多孔性担体を有する中空糸膜モジュールの培養液導入口に培養液を導入するための送液手段と、
(e)培養液排出口から排出された培養液を培養液貯槽に戻すための送液手段を有することを特徴とする三次元高密度細胞培養用モジュール。

図 4　箇条書きを含む例（特開 2003-180334）

図 4 では、改行の後に、(a)～(e)の箇条書きが続き、全体として最後の「モジュール」に係っているが、文によっては箇条書きの中に、以前の項目に言及するものもあるので注意が必要である。

2.4.5 照応的な語による階層構造の形成

いくつかの例においては、「該」、「前記」、「上記」など、前の部分に言及するような照応的な語が含まれている。図 4 などにも箇条書きの中にそれらの語が見出される。これらの語は、長い修飾語の中に組み込まれて名詞句の構成要素になっている場合や、前の語を受けることによって並列構造を形成する場合など、さまざまなケースがある。

また、機能語「とともに」(「と共に」と記述される場合もある) なども上記と同様の働きをしていると考えられる。

これらの語を含む例を図 5 に示す。

基板に容器の底部を画成するよう折線を形成し、該折線を介して前面部、後面部を設け、該折線を折曲げた際上記底部の上方に上記前面部と後面部が間隔をあけて対向すると共に上方に開口する開口部が形成されるよう上記前面部と後面部を折畳可能に連結し、該前面部の内側及び又は後面部の内側に捕虫用粘着剤層を形成した害虫捕獲容器。

図 5　「該」、「上記」などを含む例（特開 2003-180221）

「該」、「前記」、「上記」などは、図 5 に示すように、直前の名詞を受け、「その」といったニュアンスで記述されることが多い。その前に読点があると、前の修飾句と並列になる場合も多い。これらの語については、「とともに」も含めてなお今後検討する必要がある。

2.4.6 問題点と今後の検討

機能語については、前に主張した文の分割の可能性が、今回も裏付けられた。今後はこの観点からさらに調査対象を広げていきたい。すでに sublanguage という形で組み込むことによって、翻訳精度が上がることが示されている[11, 12]が、パターンを用いることによって、さらに精度を上げられることが考えられる。

今回、並列や機能語などの複数要因が絡み合った諸相については、やや解析が不十分なところがある。今後、これら複数要因がどのように関係しているかについてさらに調査を進め、知見を深めていく予定である。また、細かい解析には、やはり意味をきちんと把握することが必要であるし、照応的な解析（特に「前記」、「該」といった前の文脈への言及）も必要である。今後は、これらについても解析を深めていく予定である。

参考文献

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

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

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

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

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

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

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

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

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

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

[11] Masaru Fuji, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita, Yuji Matsumoto: Patent Claim Translation based on Sublanguage-specific Sentence Structure, Proceedings of MT Summit XV, vol.1, (2015) pp.1-16

[12] 富士秀、藤田篤、内山将夫、隅田英一郎、松本裕治：特許請求項に特有の文構造に基づく英中日特許請求項翻訳、自然言語処理 Vol.23, No.5, pp.407-435

[13] 横山晶一：特許文請求項の構造に関する調査、Japio Yearbook (2016) pp.242-245

2.5 F タームと特許文献中の重要語を用いた特許分類の推定

静岡大学　綱川 隆司
佐々木 深
西田 昌史
西村 雅史

2.5.1 はじめに

先行技術調査のために特許文献を検索する際には，キーワード検索だけでなく，特許分類に付与される特許分類を指定することによる検索が用いられる．特許分類は人間が各特許文献を精読して付与されているため，分類による検索はキーワード検索に比べテキストには表れない，あるいは間接的に表現されているような文献を発見できる．また，先行技術調査においては検索結果に網羅性があることが望ましく，分類による検索ではその分類に関係する文献をすべて収集できることが期待される．日本においては特許分類として IPC（国際特許分類）のほか、FI や F タームが用いられる（INPIT, 2016）.

特許分類の付与作業には大きな労力を要する．各分野を担当する作業者が分類付与を行う特許文献をふるい分けするために，自動的な分類システムが用いられている (古屋野, 2007). 現状ではふるい分けする分野が粗いため精度よく分類が可能である．自動分類がより細かい分類に及ぶほど人手による労力は軽減されることが期待されるが，精度は低下する．すでに，特許文献とそれに付与された分類を学習データとして用いた教師あり機械学習による自動分類手法が提案されている (Li et al., 2007; Fujino and Isozaki, 2007; Murata et al., 2007; 笹野, 2012; 小林, 2015). これらの研究は日本において最も細かい特許分類である F タームの自動分類に取り組んでいるが，人手による分類作業の補助として用いるには分類性能をさらに高める必要がある．また，細かい分類においては各分類に属する特許文献数が少なくなり，学習性能が低下する．さらに，分類の改正が実施された場合は新たに再分類が必要になるが，新しい分類が付与された文献がないために学習による方法は単純に適用できない.

本研究では，日本の特許分類において最も細かい分類である F タームに焦点を当て，特許文献に対する F タームの自動付与を試みる．従来の機械学習手法においては特許文献中の単語の出現頻度を主としているが，提案方法ではそれらに加え，前後の単語との連接関係を考慮した重要度 (中川ら, 2003) に基づいた複合語を重要語として抽出し，素性に加える．また，学習データの不足に対しては，特許文献と F ターム間の類似度を求める方法を提案する.

2.5.2 F ターム

F タームは，技術の複合化，融合化，製品の多様化に対応し，先行技術調査を迅速に行うために検索用に開発された特許分類であり，日本国内の特許文献にのみ適用される (INPIT, 2016). F タームは，IPC（国際特許分類）や，その細分類である FI のように特許文献についてその発明の内容や特徴を分類するのではなく，FI を所定分野ごとに複数の技術的観点から細分類したものであり，複数の観点を組み合わせることで関連特許をより効率的に絞り込むことを目的に定められている.

F タームによる分類では，まずあらかじめ区分された技術範囲を「テーマ」とし，各テーマは，

表 1　F タームリストの例

5B034	ハードウェアの冗長性						
観点	**F ターム**						
AA	AA00	AA01	AA02	AA03	AA04	AA05	…
	受動的冗長	・二重化	・・照合	・・・圧縮照合	・多重化	・・多数決	…
BB	BB00	BB01	BB02	BB03	BB04	BB05	…
	能動的冗長	・切替	・・予備切替	・・・共通予備	・・選択	・・・信頼度	…
		BB11	BB12	BB13		BB15	…
		・再構成	・・緊急制御回路	・・機番変更		・切離し	…
…	…	…	…	…	…	…	…

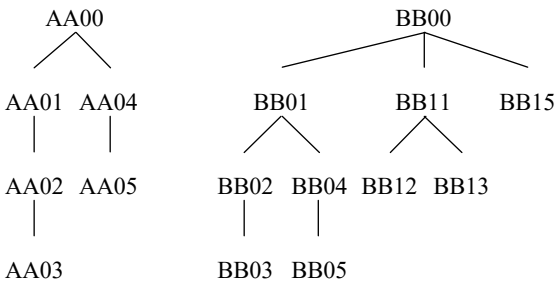


図 1　F タームリスト階層の例

表 2　F ターム解説文の例

F ターム	説明
AA00	受動的冗長 同機能の複数の装置の処理結果を比較することなどにより，システムの構成を変えることなく，エラーの検出，訂正，障害回復などを行なう技術である.
AA01	・二重化 同機能の２系統の装置を並列接続して，１系統が故障しても，そのエラーをマスクして処理を続行することができるようにしたもの.
AA02	・・照合 同機能の２系統の装置の処理結果を比較して, エラー検出を行なうもの.
…	…

対応する FI の技術範囲を示す「FI カバー範囲」，その技術分野を端的に表す「テーマ名」，英数字 5 桁からなる「テーマコード」からなる．たとえば，FI カバー範囲“G06F 11/16～11/20, 697”は，テーマ名“ハードウェアの冗長性”，テーマコード“5B034”に対応する．F タームは，テーマごとに定められる複数の「観点」に対して割り当てられる．ここで言う「観点」とはそのテーマが持つ特徴を表し，目的・用途・適用対象などのテーマに依存しない観点もあれば，そのテーマに特化した観点もあり，多種多様である．F タームは，テーマの持つ各観点から展開され，全テーマ約 2600 のうちおよそ 7 割にあたる約 1800 において作成されており, テーマコード (英数字 5 桁)，観点（英字 2 桁），数字（2 桁）の表示記号で構成される．すなわち，「テーマ」，「観点」，「F ターム」の順に展開される．たとえば，テーマ名“ハードウェアの冗長性”は 4 つの観点“受動的冗長”，“能動的冗長”，“冗長回路”，“機能・構成”を持つ．各観点について展開される F タームを表 1 に示す．F タームは表示記号に対応して端的に説明した名称があり，名称の前のドットにより階層の深さを表している．表 1 で形成される階層構造を図 1 に示す．また，各 F タームは名称を補足する説明として解説文を持つ．表 1 における F タームの解説文を表 2 に示す．F タームは組合せて検索されることが想定されており，1 つの特許文献に対して同一テーマコードから展開される複数の F タームが付与される．

2.5.3 関連研究

文書分類の研究は，1990年代以降，大量のテキストデータが利用可能になったことや，コンピュータの性能が大幅に向上したことから，機械学習による分類手法が用いられることが多くなった (Sebastiani, 2002). ナイーブベイズ (McCallum and Nigam, 1998), SVM (Joachims, 1997)といったさまざまな機械学習アルゴリズムが適用されており，近年では深層学習を適用した手法も提案されている (Lai et al., 2015). 以下，特許分類，特にFタームに焦点を当てた特許文献の自動分類に関する研究を紹介する.

NTCIR-5 および NTCIR-6 の特許検索タスクにおいて，F タームに焦点を当てた特許文献の自動分類タスクが設けられ, テストコレクションが公開されている (Iwayama et al., 2005; Iwayama et al., 2007). NTCIR-6 分類タスクは，1993～1997 年に公開された日本公開特許公報の特許文献全文を学習データとし，1998～1999 年の特許文献 21606 件に対して F タームを付与する課題であり，6 グループが参加した．Li et al. (2007) は，特許文献の bag-of-words を素性とした SVM による識別器を用いた手法を提案し，完全一致による評価で F 値 0.4125 の精度で最高性能を得た．Fujino and Isozaki (2007) は，特許文献の各要素（発明の名称，出願人・発明者，要約書，特許範囲，明細書）について bag-of-words によるナイーブベイズ識別器を作成し，これらを最大エントロピー法に基づき組み合わせる方法を提案した．Murata et al. (2007) は k-NN 法に基づく方法を提案し，特許文献間の類似度として SMART (Singhal et al., 1996) や BM25 (Robertson et al., 1994) を用いて評価を行った.

笹野 (2012) は，F タームの付与根拠データ[1]を用いた機械学習手法を提案した．この結果を踏まえ, 小林 (2015) は付与根拠データが少ない分野に対して機械学習に基づき精度を向上させるため，tf-idf および分類の階層構造を利用した分類推定方法を提案した.

これらの手法では，特許文献を形態素解析にかけ，素性ベクトルを作成し識別器を学習している．特許文献中にはその文献が属する分野の専門用語が含まれており，専門用語の多くは複合語であることが多い (中川ら, 2003). そのため，形態素解析によって単語に区切ることで，専門用語の本来の意味を適切に扱えないことが想定される．本研究では，tf-idf による単語の重みづけに加え，tf-idf とは異なる観点から得た重要語を用いる．また，小林 (2015) は，新設 F タームのように付与実績の少ない F タームは識別器を学習するための文献が不足する課題について言及している．本研究では，特許文献と各 F タームとの類似度を推定し，学習データを必要としない方法を提案する.

2.5.4 提案方法

本節では，特許文献へ特許分類を付与するための二つの提案方法である，重要語を用いた教師あり学習手法, および特許文献と F ターム間の類似度に基づく手法のそれぞれについて説明する.

[1] 特許文献への分類付与者が，付与の根拠となった文献中の箇所を特許明細書等から抽出したもの.

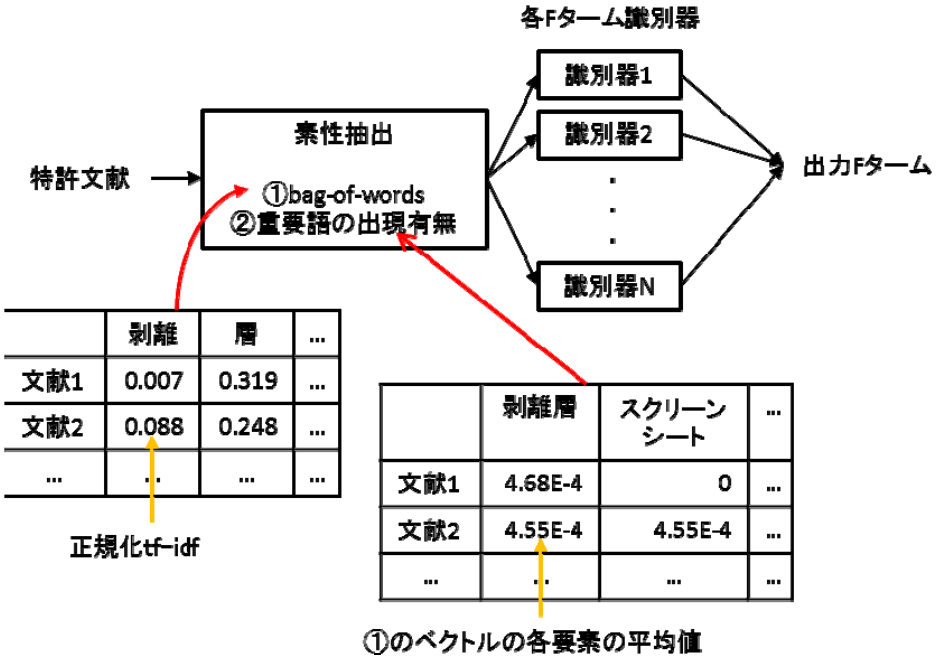


図 2　重要語を用いた教師あり学習手法の概要

2.5.4.1　重要語を用いた教師あり学習手法

図 2 に重要語を用いた教師あり学習手法の概要を示す．まず，特許文献から出現語の bag-of-words を tf-idf 値で重み付けしたもの，および重要語の出現有無を素性値として抽出し，素性ベクトルを作成する．識別対象とする個々の F タームごとに，この素性ベクトルとその F タームの付与の有無を入力し，その F タームを出力するか否かの二値識別器を学習する．識別器による F ターム付与時には，特許文献を同様に素性ベクトル化して各 F タームの識別器に入力し，付与すると判定された F タームをすべて付与する．各 F タームに対する識別器の学習はそれぞれ独立に行う．識別器には，先行研究において有効であった SVM（サポートベクタマシン）を用いる (Iwayama et al., 2007).

素性ベクトルの各要素は特許文献に出現する単語の tf-idf 値と，抽出された重要語の出現有無からなる．対象とする単語の品詞は名詞（非自立名詞，固有名詞，数を除く），自立動詞，自立形容詞，未知語とし，出現頻度が 3 回未満の単語は除く．これらの単語から tf-idf 値を要素とするベクトルを求め，ベクトルを正規化する．

次に，特許文書を特徴付ける重要語を抽出するため，専門用語自動抽出モジュール[2]TermExtract (中川ら，2003) を使用する．抽出対象とする重要語は TermExtract が出力する重要度が閾値 θ 以上のものとする．これらの重要語を素性ベクトルの要素として追加する．ここで，各重要語の要素を正規化した各 tf-idf 値の平均値[3]とする．出現しない重要語の要素は 0 とする．

ある F タームに対する識別器の学習は，その F タームが付与されている特許文献を正例，その F タームは付与されていないが，それが属するテーマコードの範囲内の他の F タームが付与され

[2] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
[3] 各 tf-idf 値の最大値，最小値，平均値をそれぞれ用いた場合の予備実験を実施し，最も精度の良かった平均値を採用した．

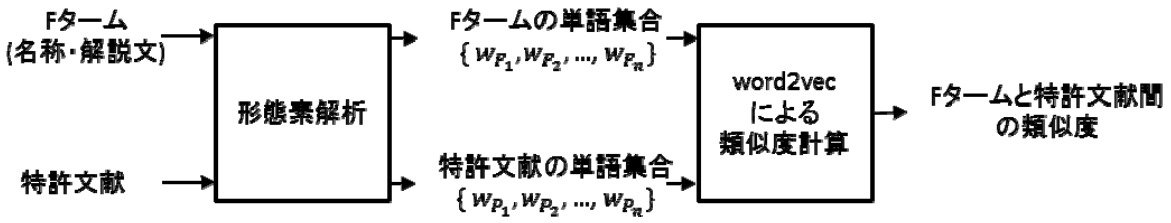


図 3　特許文献と F タームの間の類似度に基づく手法の概要

ている特許文献を負例として入力する．このとき，負例が正例に比べて圧倒的に多く，負例をすべて用いるとアンバランスな学習データになるため，負例は最大で正例数の 2 倍の数だけ無作為に選択する (Li et al., 2007).

2.5.4.2　特許文献と F ターム間の類似度に基づく手法

新設 F タームや付与実績の少ない F タームのように教師あり学習では付与が難しい F タームに対応するため，特許文献と F タームを直接比較する．図 3 に本手法の概要を示す．前節と同様に，特許文献に出現する単語集合 $\{w_{P_1}, \ldots, w_{P_m}\}$ を求める．また，識別対象となる各 F タームについては，F タームの名称と表 2 にある解説文を使用して同様に単語集合 $\{w_{F_1}, \ldots, w_{F_n}\}$ を求める．これらの単語集合間の類似度を求めるために，二つの単語 w, w' 間の類似度 $\mathrm{sim}(w, w')$ を，word2vec (Mikolov et al., 2013) に特許文献を入力し，得られる単語ベクトル間のコサイン類似度として求める．

特許文献の単語集合と F タームの単語集合の類似度 $S(\{w_{F_1}, \ldots, w_{F_n}\}, \{w_{P_1}, \ldots, w_{P_m}\})$ は，まず特許文献の各単語に対し，最も類似度の高い F タームの単語を求め，それらの類似度の平均値とする．すなわち，

$$S(\{w_{F_1}, \ldots, w_{F_n}\}, \{w_{P_1}, \ldots, w_{P_m}\}) = \frac{1}{n}\sum_{i=1}^{n} \max_{j=1,\ldots,m} \mathrm{sim}\left(w_{F_i}, w_{P_j}\right).$$

ある特許文献に対して識別対象となる F タームすべてについて類似度を求め，類似度の高い上位 n 個の F タームを付与する．

2.5.5　評価実験

2.5.5.1　実験設定

前節で述べた二つの提案方法について，NTCIR-6 (Iwayama et al., 2007) のデータコレクションを用いて評価実験を行い，有効性を検証する．NTCIR-6 のデータコレクションは，学習データが 1993～1997 年の特許文献，テストデータが 1998～1999 年の特許文献からなる．テストデータに付与されるテーマコードは 108 種類あり，これらは IPC の最も粗い分類であるセクションレベルから無作為に選択されている．本実験では，これらの中から 20 種類のテーマコードを無作為に選択し，それらのテーマコードを持つ F タームが付与された特許文献を実験対象とする．本実験に使用した特許文献数，および 1 文献に付与される F ターム数を表 3 に示す．なお，word2vec に入力する特許文献は，表 3 に示した学習データおよびテストデータを合わせた 49643 件を用いた．

教師つき学習の識別器の SVM のカーネル関数にはシグモイドカーネルを用いた．特許文書から

単語集合を得るための形態素解析器には MeCab[4] を用いた．また，重要語抽出の重要度閾値 θ は 20 とした．

F ターム付与の評価指標には，適合率，再現率およびそれらの調和平均である F 値を用いた．

表 3　実験に用いたデータ

訓練データの文献数		45631
テストデータの文献数		4012
1 文献に付与された F ターム数	平均	8.37
	最大	43
	最小	1

表 4　テーマコード付与結果

適合率	再現率	F 値
0.880	0.951	0.914

表 5　教師あり学習による F ターム付与結果

素性	適合率	再現率	F 値
tf-idf のみ	0.340	0.305	0.322
提案手法	0.335	0.358	0.346

表 6　テーマコード別の F ターム付与結果

テーマコード	素性	適合率	再現率	F 値
3F054	tf-idf のみ	0.554	0.264	0.358
	提案手法	0.537	0.296	0.381
3G023	tf-idf のみ	0.465	0.383	0.420
	提案手法	0.466	0.483	0.474
2C088	tf-idf のみ	0.124	0.517	0.201
	提案手法	0.114	0.590	0.192
5E082	tf-idf のみ	0.390	0.471	0.427
	提案手法	0.371	0.495	0.424

2.5.5.2　重要語を用いた教師あり学習によるテーマコード・F ターム付与実験結果

2.5.4.1 節で述べた提案方法は, F タームだけでなく任意の特許分類に適用可能である．そこで, F タームに比べ粗い分類であるテーマコードについての付与実験を行った．表 4 にテーマコード付与の実験結果を示す．実験に使用した 20 種類のテーマコードには，5B064（文字認識）と 4L045（紡糸方法及び装置）のように共通点に乏しいものが多いため識別が比較的容易であると考えられるものの，高精度で自動付与ができることを示している．本実験では，テーマコードの同定ができることを仮定し，一つのテーマコードに属する F タームを対象に F タームの識別器の学習を行い, F タームの付与を行う．

表 5 に，教師つき機械学習による F ターム付与の実験結果を示す．比較対象として素性ベクトルに重要語を用いない場合を示した. 重要語の導入により, F 値が 2.46 ポイント改善した．また，実験に使用した 20 種類のテーマコードのうち，18 種類で F 値の改善がみられた．表 6 に 4 つのテーマコードに関する結果を示す．すべてのテーマコードについて，適合率については同程度またはわずかに低下し，再現率は上昇する傾向がみられた．

2.5.5.3　特許文献と F ターム間の類似度に基づく F ターム付与実験結果

表 7 に特許文献と F ターム間の類似度を用いた F ターム付与実験結果を示す．表 7 における n は特許文献に対して付与する F ターム数を示す．多くの特許文献において，実際に付与された F タームは類似度の上位にはあまり含まれず，教師つき機械学習の結果に比べ精度は低かった．また，表 7 (b) は特許文献の単語集合に TermExtract によって抽出した重要語のみを用いた場合を示

[4] http://taku910.github.io/mecab/

表 7　特許文献と F タームの類似度に基づく F ターム付与結果

(a) 特許文献全体による単語集合

n	適合率	再現率	F 値
5	0.100	0.065	0.079
10	0.096	0.125	0.108
15	0.090	0.176	0.119
20	0.086	0.224	0.125
25	0.083	0.270	0.127
30	0.081	0.314	0.128
35	0.079	0.357	0.129
40	0.076	0.397	0.128

(b) 重要語のみによる単語集合

n	適合率	再現率	F 値
5	0.105	0.068	0.083
10	0.099	0.129	0.112
15	0.095	0.184	0.125
20	0.091	0.237	0.132
25	0.088	0.286	0.135
30	0.084	0.329	0.134
35	0.081	0.367	0.132
40	0.078	0.404	0.130

表 8　1 文献あたりに出力された F ターム数

	tf-idf のみ	提案方法
平均	8.32	9.93
最大	75	77
最小	0	0

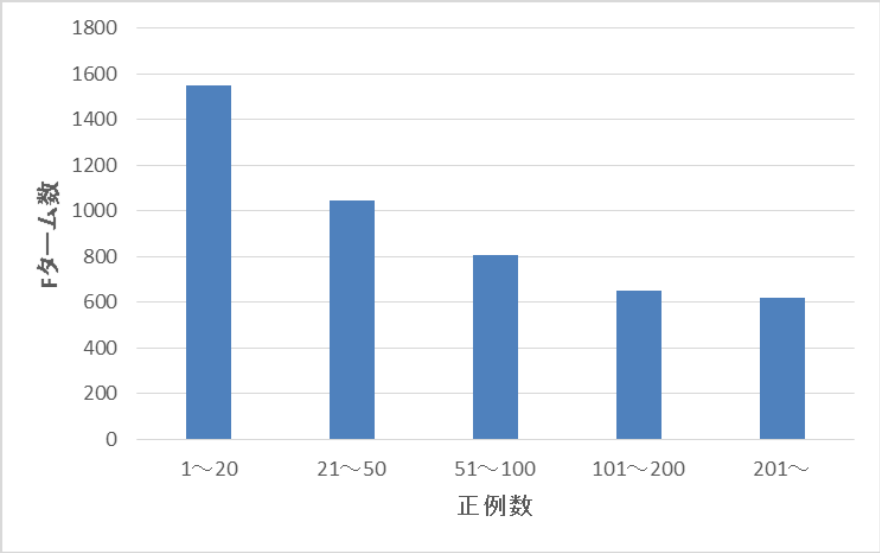


図 4　付与特許文献数による F タームの分布

しており，特許文献の単語集合全体を用いる場合と比べわずかに良い結果が得られた．

2.5.5.4　考察

2.5.5.2 節の教師つき学習による F ターム付与において，1 つの特許文献に対して付与された F ターム数を表 8 に示す．重要語の導入前は，テストコレクションにおける 1 文献当たりの平均 F ターム数とほぼ同じであるが，重要語の導入により平均 1.6 程度 F タームを多く出力しており，これにより再現率の上昇につながっている．特許分類を用いた先行技術調査では検索漏れが少ないことが望ましく，再現率はより重要であると考えられる．

図 4 は，各 F タームが付与された特許文献数（正例数）によって区分した F タームの分布を示

表 9　特許文献と F タームの類似度ランキング例

(a) 特許文献全体による単語集合

順位	F ターム	類似度	正誤	正例数
1	AB02	0.763	誤	60
2	AD13	0.742	誤	119
3	**AB05**	**0.729**	**正**	**754**
4	**AD02**	**0.726**	**正**	**466**
5	AD07	0.725	誤	407
6	AD03	0.724	誤	456
7	AA00	0.721	誤	115
8	AD05	0.702	誤	446
9	AC02	0.697	誤	211
10	AD29	0.697	誤	235
…	…	…	…	…
57	**AF00**	**0.629**	**正**	**40**
…	…	…	…	…
67	**AF04**	**0.584**	**正**	**5**
…	…	…	…	…

(b) 重要語のみによる単語集合

順位	F ターム	類似度	正誤	正例数
1	**AD02**	**0.674**	**正**	**466**
2	AC02	0.664	誤	211
3	**AB05**	**0.663**	**正**	**754**
4	AD13	0.660	誤	119
5	AD09	0.658	誤	373
6	AD05	0.658	誤	446
7	**AC01**	**0.644**	**正**	**171**
8	AD03	0.644	誤	367
9	AF02	0.641	誤	235
10	AD22	0.640	誤	219
…	…	…	…	…
41	**AF04**	**0.584**	**正**	**5**
…	…	…	…	…
72	**AF00**	**0.539**	**正**	**40**
…	…	…	…	…

す．正例数が 50 以下の F タームが全体の半数以上を占めており，このような F タームは，テストデータにおいて付与される文献が少ないかあるいはまったく存在しない．

教師つき学習においては，このような F タームはテストデータの特許文献すべてに対して付与されないことが多い．たとえば，テーマコード 3C045（旋削加工）に属する F タームは 197 種類あり，そのうち 3C045DA20（旋削加工の形態＞・非円形断面＞・・多角形）が付与された特許文献は，学習データに 3 件，テストデータに 1 件のみ存在した．教師つき学習による付与の結果，テーマコード 3C045 を持つテストデータ 193 件すべてについて当該 F タームは付与されなかった．学習データ 3 件には，多角形を示す“ポリゴン”という語が出現するが，テストデータの 1 件には“非円形”，“突起”の 2 語で多角形であることが示唆されており，語彙が一致しなかった．また，学習データの文献数が少ないため，他の出現語彙の類似性からこれらの文献間の共通性を見出して当該 F タームが付与されることも期待できない．また，テーマコード 3G023（内燃機関燃焼法）において，F ターム 3G023AF04（対象とする機関＞・側弁式機関）についても，学習データ 5 件，テストデータ 2 件にのみ付与されており，テストデータのいずれにも当該 F タームは付与されなかった．学習データの文献のいくつかには“側弁”，“サイドバルブ”といった語句が含まれているが，テストデータの文献には含まれておらず，形状は文章の説明および図によって示されている．このケースも学習データが不十分であり，この F タームに特徴的な重要語の共通性もみられないため，本手法の適用が困難な例である．

F ターム 3G023AF04 が付与されたテストデータの特許文献 1 件について，類似度に基づいて算出した上位の F タームを表 9 に示す．この特許文献には 9 種類の F タームが付与されているが，重要語のみを類似度計算に用いることで，正解となる F タームの順位が上昇し，上位 10 個には正解が 3 つ含まれるようになった．実際に付与されている F タームには正例数が 50 以下の F ターム

が 2 種類（AF00, AF04）あるが，いずれに対しても類似度は低く，この例においては提案方法の有効性は示されなかった．

2.5.6 おわりに

本研究は，特許分類 F タームの自動付与精度向上のため，従来の教師あり学習手法に特許文献の重要語を素性として用いる方法を提案した．重要語の出現有無を考慮することにより，複合語からなる専門用語の意味を反映することができた．また，付与例が少ない F タームや新設 F タームに対しても適用が可能な，特許文献と F タ ーム間の類似度に基づく方法を提案した．NTCIR-6 テストコレクションを用いた評価実験では，重要語を用いない場合と比較して教師あり学習の付与の再現率を向上させ，F 値を 2.46 ポイント改善した．特許文献と F ターム間の類似度に基づく方法は，得られた類似度上位の F タームには正解のうち付与例の少ないものがあまり含まれず，十分な精度が得られなかった．

今後の課題として，テーマが持つ観点数の差の考慮が挙げられる．特許文献に付与されるテーマについて，そのテーマが持つ観点数に関わらず特許文献に付与される観点数はある程度限定される．そのため，比較的精度の良いと思われる観点レベルでの識別を先に行い，付与すると判定した観点に対してのみそれに属する F タームの識別を行うことが考えられる．また，単語の類似度計算においては複合語を考慮していないため，形態素解析に使用する辞書を拡張することも挙げられる．

謝辞

本研究を遂行するにあたり NTCIR-6 テストコレクションをご提供いただいた国立情報学研究所，ならびに F タームを含む整理標準化データをご提供いただいた特許庁に感謝致します．

参考文献

Fujino, A. and Isozaki, H. (2007). Multi-label patent classification at NTT Communication Science Laboratories. In *Proc. of NTCIR-6 Workshop Meeting*, pp.381-384.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Proc. of the 29th AAAI Conference on Artificial Intelligence*, pp. 2267–2273.

Li, Y., Bontcheva, K., and Cunningham, H. (2007). SVM based learning system for F-term patent classification. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 396-402.

Iwayama, M., Fujii, A., and Kando, N. (2005). Overview of classification subtask at NTCIR-5 patent retrieval task. In *Proc. of NTCIR-5 Workshop Meeting*.

Iwayama, M., Fujii, A., and Kando, N. (2007). Overview of classification subtask at NTCIR-6 patent retrieval task. In *Proc. of NTCIR-6 Workshop Meeting*, pp.366-372.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant

features. *Machine Learning: ECML-98 - Proc. of 10th European Conference on Machine Learning*, pp. 137-142.

McCallum, A. and Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *Proc. of AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41-48.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Murata, M., Kanamaru, T., Shirado, T., and Isahara, H. (2007). Using the k-nearest neighbor method and SMART weighting in the patent document categorization subtask at NTCIR-6. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 407-413.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1994). Okapi at TREC-3. In *Proc. of the third Text REtrieval Conference (TREC-3)*, pp. 109-126.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), pp. 1-47.

Singhal, A., Choi, J., Hindle, D., and Pereira, F. (1997). AT&T at TREC-6. In *SDR Track in NIST Special Publication 500-226: The 6th Text REtrieval Conference (TREC6)*, pp. 227-232.

小林英司. (2015). 特許分類の自動推定に向けた取り組みー機械学習による自動分類推定の課題と今後の展開ー. *Japio YEAR BOOK 2015*, pp. 272-275.

古屋野浩史. (2007). 特許分類等の付与精度向上への取り組み. *Japio 2007 YEAR BOOK*, pp. 118-119.

笹野秀生. (2012). 特許分類の自動推定に向けた取り組みー機械学習による自動分類技術の特許文献への適用ー. *Japio YEAR BOOK 2012*, pp. 208-211.

独立行政法人工業所有権情報・研修館 (INPIT). (2016). 特許分類の概要とそれらを用いた先行技術調査～IPC, FI, F ターム編～（平成 28 年度版）.

中川裕志, 湯本紘彰, 森辰則. (2003). 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, 10(1), 27-45.

3. 機械翻訳評価手法

3.1 拡大評価部会の活動概要

岡山県立大学　磯崎 秀樹

2012 年度から、本研究会の下部組織として「拡大評価部会」を設置し、機械翻訳の評価に関する議論を深めてきた。

拡大評価部会は、以下の３つのサブグループから構成されている。

1. 自動評価サブグループ
2. 人手評価サブグループ
3. テストセットサブグループ

本年度も、昨年度同様、３回の部会を開催した。

・2016 年 5 月 13 日　今年度の活動計画の策定
・2016 年 10 月 14 日　中間報告と今後の活動内容についての議論
・2017 年 2 月 10 日　最終活動報告と年度報告書の執筆について

自動評価サブグループは、英日・日英翻訳の自動評価手法として 2010 年に提案され、標準的評価法として定着してきた RIBES について、最近分かってきた問題点を改善する取り組みについて 3.2 で報告する。3.3 では、単語の重要度を考慮した別の自動評価法について報告する。

テストセットグループでは、中日翻訳において、離れた単語で構成される頻出パターンをテストするためのテストセットについて 3.4 で報告する。

人手評価については、Workshop on Asian Translation (WAT) 2016 における特許翻訳タスクの人手評価について 3.5 で報告する。

3.2 現在の翻訳自動評価が抱える問題点

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

磯崎が 2010 年に提案した翻訳自動評価法 RIBES（ライビーズ）は、日英・英日翻訳の翻訳自動評価法として定着してきたが、以下の２つの問題が指摘されている。

1.　SMT では BLEU を目的関数としたチューニングが行われるが、この BLEU の代わりに RIBES を目的関数としたチューニングが難しい。

Kevin Duh ら[1]は以下のように指摘している。

We found that RIBES can be difficult to tune directly. It is an extremely non-smooth objective with many local optima – slight changes in word ordering causes large changes in RIBES.

2.　最近急速に広まっている Neural Machine Translation (NMT)の評価が適切にできない。NMT の出力は、Statistical Machine Translation (SMT)の出力と違い、語順の問題が少なく、訳語の問題が大きいので、語順を重視し、訳語の間違いを軽視する RIBES では、適切な評価ができない。この点は、Takayuki Sato ら[2]などが指摘している。

そこで、岡山県立大学磯崎研究室では、これらの問題の解決に取り組んだ。まず、問題 1 については、RIBES が語順の近さを測定するために用いている Kendall の τ が階段関数であることに原因があると推測した。階段関数は、いたるところで傾きがゼロであるため、どちらに向かって進めば良いかがわからず、最適化しづらい。

τ をもっと滑らかにすることで、チューニングしやすくできるのではないか、と考え、滑らかにした関数 Smooth Kendall’s tau (SKT)を考案した。図 1 に SKT のグラフを示す。SKT には滑らかさを示すパラメタ σ があり、σ を小さくするほど滑らかになる。

RIBES の τ を SKT で置き換えた MERT を実装し、チューニングの実験中である[3]。

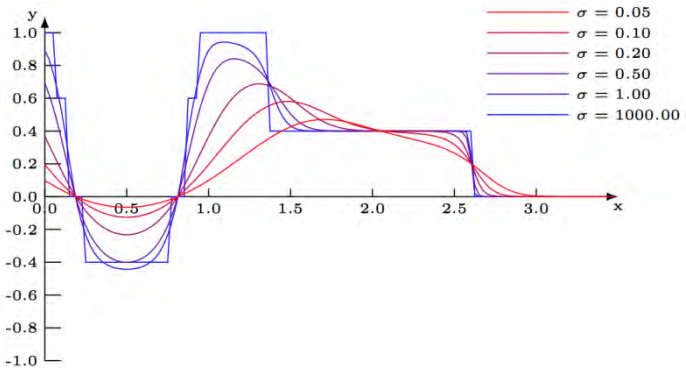


図 1　Kendall の τ を滑らかにした関数 SKT
σ が大きいほど、階段関数である τ に近づく。

問題 2 については、NMT のシステムが多い Workshop on Asian Translation (WAT) 2016 のデータを入手し、RIBES の様々な変種と人手評価の相関を調べた。その結果、WAT のデータでは、NTCIR のデータと異なり、Recall つまり機械訳中の単語が参照訳に含まれている割合が人手評価と高い相関があることが判明した[4]。

[1] Kevin Duh et al. 2012: Learning to Translate with Multiple Objectives, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 1-10, 2012.
[2] Takayuki Sato et al. 2016: Japanese-English Machine Translation of Recipe Texts, Proceedings of the 3rd Workshop on Asian Translation, pp.58-67, 2016.
[3] 森崎 2017: 統計的機械翻訳における目的関数の変更の試み、岡山県立大学修士論文、2017.
[4] 川本 2017: 翻訳自動評価法 RIBES における適合率再現率の重視、岡山県立大学卒業論文、2017.

3.3 統計的アプローチを用いた単語アライメントに基づく自動評価法

北海学園大学　越前谷　博

3.3.1　はじめに

近年、ニューラルネット翻訳[1][2]に対する注目が集まっている。その結果、ニューラルネット翻訳に対する様々な研究が行われており、大きな成果を挙げている。こうした背景より、ニューラルネット翻訳に対応した自動評価法に対するニーズも高まっている。これまで自動評価の研究は統計翻訳研究[3][4]の進展を背景として行われてきた。しかし、意味を考慮して翻訳を行うことが可能とされるニューラルネット翻訳に対応しているかどうかについては大きな疑問が残る。特に、言語に非依存の自動評価法としてデファクトスタンダートとなっている BLEU[5]を始めとする、METEOR[6]、IMPACT[7]、そして、RIBES[8]などの自動評価法は意味的な情報を十分に利用した評価尺度とはいえず、ニューラルネット翻訳においても有効な評価法として存在意義を示せるかどうかは議論の余地がある。

このような状況において、近年にはニューラルネット翻訳を意識した、即ち、単語の意味を考慮した自動評価法が提案されるようになってきた。その場合、単語の意味は word2vce[9]などの分散表現を用いることが多い。また、文の構造においては、単語アライメントを行うことで語順を考慮している。本報告では、このようなニューラルネット翻訳のための自動評価法の第 1 段階として新たな自動評価法を提案する。提案手法では、統計的なアプローチにより単語間の対応関係を求める。その結果、表層的には一致しない単語間においても翻訳文と参照訳との間の単語アライメントを適切に得ることができる。また、単語アライメントの結果に基づいて単語の位置情報を得る。そして、その位置情報により翻訳文と参照訳間の語順の違いをスコアに反映させる。したがって、提案手法は表層情報に基づく自動評価法では困難な単語アライメントを可能とし、より高い精度での翻訳文に対する評価が期待できる。

3.3.2 関連研究

単語の意味を反映させた文書間類似度計算及び自動評価法が近年提案されている。文書間類似度の計算においては、柳本[10]は word2vec による単語の分散表現と Earth Mover's Distance（以下、EMD と表記）[11]を用いて、文書間の類似度を求めている。その結果、単語の同義語や類義語を考慮した類似度計算が可能となった。EMD は 2 つの分布間の距離の計算を輸送問題として捉え、最適な輸送コストを求めるために定義される。これまでは類似画像検索の分野で広く利用されてきたが、近年、文書間類似度のために EMD を用いる手法[12]も提案されている。文献[12]では WordNet を用いて単語間の距離を定義している。また、EMD は文書検索にも利用されている[13]。更に、Kusner ら[14]は、EMD を用いて文書間類似度を求めるにあたり、ストップワード以外の単語間におけるアライメントを行ったうえで、word2vec より対応付けのコストが最も低い場合のコストの総和を類似度として求める、Word Mover's Distance（以下、WMD と記す）を提案している。

単語の意味を考慮した自動評価法としては、松尾ら[15]は、単語分散表現を用いた単語アライメントを自動評価に応用し、その有効性について検証している。WAT2015、NTCIR-8 の日本語-英語データを用いた性能評価実験の結果、意味的文類似度に基づく評価法として用いた、「One-hot 表現に基づく意味的文類似度」、「文の分散表現に基づく意味的文類似度」、「Whole Alignment Similarity」、「Maximum Alignment Similarity」、そして、「Hungarian Alignment Similarity」の中で、「Maximum Alignment Similarity」が文単位の人手評価との相関が最も高くなることを示した。「Maximum Alignment Similarity」ではすべての単語のコサイン類似度ではなく、単語間類似度が最大となる値のみの平均を用いる。また、最大値を求める際には、翻訳文から参照訳の最大値を取る場合と参照訳から翻訳文の最大値を取る場合の 2 つのパターンの平均値を用いる。

提案手法では、EMD を用いて翻訳文と参照訳間の類似度計算を行う。その際には、今回は分布の特徴量は単語とし、単語の重みには文レベルの tf・idf、単語間の距離計算には Dice 係数に基づく計算式を用いた。単語の重みに tf・idf を使用することにより機能語と内容語を差別化することが可能になる。また、単語間の距離計算に Dice 係数に基づく計算式を用いる理由は、翻訳文の単語と参照訳の単語の間で表層的に一致しない単語であっても、確率的に単語間の類似度を求めることで適切に対応関係を得ることができると考えられるためである。更には、EMD における距離行列の要素には、すべての単語間の距離計算の値を用いるのではなく、単語アライメントの結果得られた、対応関係が存在する単語間のみに距離計算の値を用いる。その結果、適切な EMD の類似度計算が期待できる。

3.3.3 提案手法

提案手法は次の 3 つの処理（単語の重み付け、単語アライメント、EMD によるスコアの計算）より構成されている。以下に、それぞれの詳細について述べる。

3.3.3.1 単語の重み付け

提案手法では、EMD を用いてスコアを求める際に、特徴量に文中の単語、その重みには文レベルの tf・idf を用いる。tf・idf を使用する目的は文中の単語において、機能語と内容語を動的に区別するためである。WMD では事前に機能語、即ち、ストップワードを除いたうえで単語アライメントを行う。それに対して、本報告では特定の言語に依存することなく機能語と内容語を得るために、tf・idf を用いる。tf・idf は任意の文書において出現頻度が高く、他の文書に出現する数が少ないほど特徴的な単語であると見なす。提案手法では、文レベルで tf・idf を用いる。即ち、任意の文において出現頻度が高く、他の文に出現する回数が少ない単語ほど特徴的な単語と見なす。文レベルの tf・idf を提案手法では、tf・isf として、以下の式(1)より求める。そして、得られた値を EMD を用いる際の単語の重みとする。

$$tf \cdot isf = (\log(tf(w,s)) + 1) \times \frac{|S|}{sf(w)} \qquad (1)$$

式(1)の tf(w, s)は任意の文 s における単語 w の出現頻度である。文中で同じ単語が複数出現する可能性は内容語よりも機能語の方が高いと考えられる。しかし、log を用いているため、機能語

においても tf は大きな値とはならず、内容語と機能語の差別化には大きく影響しない。また、log(tf(w, s))に 1 を加えているのは tf(w, s)が 1 の場合、tf が 0.0 になることを防ぐためである。

そして、|S|は翻訳文の総数及び参照訳の総数である。sf(w)は翻訳文の総数及び参照訳の総数に対する単語 w が出現する数である。したって、isf は機能語のように多くの文に出現する場合、小さくなり、内容語のように特定の文のみに出現する場合には大きくなる。その結果、tf・isf は isf の値が強く反映され、機能語の場合には小さく、内容語の場合には大きくなる。

3.3.3.2 単語アライメント

EMD を自動評価に適用する際の問題点の一つは語順が変わっても類似度の値が同じになることである。即ち、同じ単語で構成される翻訳文が複数存在する場合、語順が大きく異なるにもかかわらず類似度の値は同じになってしまい、語順が正しい翻訳文と誤った翻訳文を区別できない。したがって、EMD を自動評価に用いる際には、語順の違いを反映した距離計算を行う必要がある。提案手法では語順の違いを反映するために始めに単語アライメントを行う。

提案手法における単語アライメントでは、以下の式(2)により信頼度を求め、この信頼度に基づき翻訳文と参照訳における単語間の対応関係を決定する。

$$\text{信頼度} = \begin{cases} \dfrac{Dice\text{係数} + 1.0}{2.0} (w_c = w_r \text{の場合}) \\ \dfrac{Dice\text{係数}}{2.0} (w_c \neq w_r \text{の場合}) \end{cases} \qquad (2)$$

また、式(2)の Dice 係数は以下の式(3)より求める。

$$Dice\text{係数} = \frac{2 \times f_{cr}}{f_c + f_r} \qquad (3)$$

提案手法では、Dice 係数を翻訳文と参照訳中の単語の確率的な近さを求めるために用いる。例えば、表層的には一致していない単語“centrum”と“center”は意味的には近いため Dice 係数の値は高くなると考えられる。この Dice 係数のみを用いて信頼度とすることも考えられるが、Dice 係数は確率的なアプローチであるため使用する文数が不十分な場合、対応関係にない単語間に対しても値が高くなることがある。そこで、提案手法では、式(2)のように、2 つの単語 w_c（翻訳文中の単語）と w_r（参照訳中の単語）において、表層的に一致する場合には重みを 1.0 として、この重みと Dice 係数との平均値を信頼度とする。

この式(2)により得られる単語間の信頼度に基づき単語アライメントを行う。まず、翻訳文中の任意の単語と参照訳文中の全単語との間で信頼度を求める。そして、求めた信頼度の中で最大の信頼度を持つ単語を対応関係にある単語と見なす。ただし、Dice 係数が 0.5 未満の場合には単語間の対応関係が一意に決定できたとしても、誤っている可能性が高いとしてアライメントの対象から外す。以下の図 1 に翻訳文“基板用ブラケットの詳細構成を図 8 に示す。”と参照訳“この基板ブラケットの詳細構成を図 8 に示す。”における単語アライメントの具体例を示す。

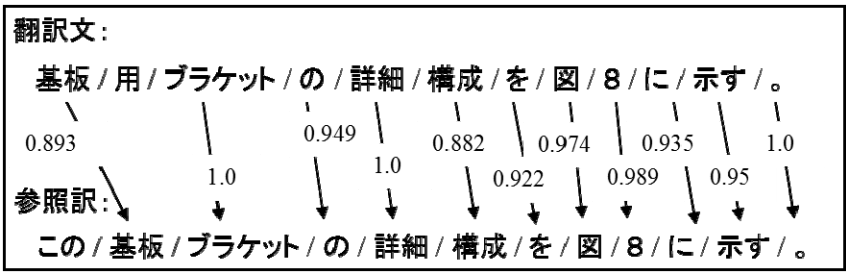


図 1　単語アライメントの具体例

図 1 の数値は信頼度である。図 1 では、翻訳文の単語“用”において最大の Dice 係数は 0.25 であり、0.5 未満であるため単語アライメントの対象外となる。提案手法では、Dice 係数による確率的な情報と表層情報の両方を用いて単語アライメントを行うことで、語順情報を EMD の距離計算により適切に反映させることが可能となる。

3.3.3.3 EMD によるスコアの計算

EMD は線形計画問題の 1 つである輸送問題の解を求めることで計算される。輸送問題とは、一定の供給量を持つ複数の供給地と同じく一定の需要量を必要とする需要地を設定し、各供給地から需要地までの輸送コストが与えられた際に、需要地の需要を満たすように供給地から需要地へ最小の輸送コストで荷物を輸送する方法を探す問題である。EMD の計算方法について述べる。まず、m 個の供給地を持つ供給地集合 P と n 個の需要地を持つ需要地集合 Q を以下のように表す。

$P=\{(p_1, w_{p1}),\ldots, (p_m, w_{pm})\}$、$Q=\{(q_1, w_{q1}),\ldots, (q_n, w_{qn})\}$

ここで、p_i を i 番目の供給地を表す特徴ベクトル、w_{pi} を i 番目の供給地の供給量とする。また、q_j を j 番目の需要地を表す特徴ベクトル、w_{qj} を j 番目の需要地が必要とする需要量とする。提案手法では、供給地 p_n を翻訳文の単語、需要地 q_n を参照訳の単語とし、供給量 w_{pn} を翻訳文の単語の tf・isf、w_{qn} を参照訳の単語の tf・isf とする。ただし、EMD を求める際には w_{p1} から w_{pm} の総和及び w_{q1} から w_{qn} の総和が 1.0 になるように調整する。

EMD では総輸送コストは以下の式(4)となる。ここで、p_i と q_j 間の距離を d_{ij} とする。また、p_i から q_j への輸送量を f_{ij}、全輸送量を F とする。EMD はこの式(4)に対して、最適な全輸送量 F* を用いて総輸送コストを最小化する。

$$WORD(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \qquad (4)$$

ただし、総輸送コストを求める際に、以下の式(5)から式(8)の制約条件を満たす必要がある。

$$f_{ij} \geq 0 \ (1 \leq i \leq m, 1 \leq j \leq n) \qquad (5)$$

式(5)は供給地から需要地の 1 方向のみに輸送することを意味する。逆方向での輸送は行われない。

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i} \ (1 \leq i \leq m) \qquad (6)$$

式(6)は供給地から輸送できる容量は供給量 w_{pi} を超えることはできないことを意味する。

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j} \ (1 \leq j \leq n) \qquad (7)$$

式(7)は需要地が受け取れる容量は w_{qj} 以下であることを意味する。

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \qquad (8)$$

そして、式(8)は供給地から移動できる最大の輸送量は供給量の総和と需要量の総和の小さい方となることを示している。これらの制約のもとで EMD は最適な輸送量 F^*を用いて以下の式(9)より得られる。

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}^*}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}^*} \qquad (9)$$

式(9)において、最小コストを輸送量 F^*で割っているのは、輸送量によって EMD の範囲が変わらないように正規化するためである。

EMD の距離計算として式(2)と式(3)の信頼度を単純にすべての単語間の距離に用いる場合、対応関係にない単語間の信頼度も高くなり適切な類似度が得られないことが考えられる。そこで、提案手法では、3.3.3.2 で述べた単語アライメントの結果に基づき、距離行列を生成する。具体的には、以下の式(10)を用いて距離を計算する。

$$d = \begin{cases} 1.0 - \text{信頼度} \times pos\_diff & (\text{対応関係あり}) \\ 1.0 & (\text{対応関係なし}) \end{cases} \qquad (10)$$

式(10)の信頼度は式(2)より得る。pos_diff は信頼度に対する重みである。その計算式を式(11)に示す。

$$pos\_diff = 1.0 - \left|\frac{pos(w_c)}{len(c)} - \frac{pos(w_r)}{len(r)}\right| \qquad (11)$$

式(11)の $pos(w_c)$は翻訳文における単語の出現位置である。先頭の単語を 1 として何番目に位置するかを示す。$pos(w_r)$は参照訳における単語の出現位置である。そして、これらを翻訳文の単語数 len(c)と参照訳の単語数 len(r)を用いてそれぞれ 0.0 から 1.0 に正規化する。したがって、$|pos(w_c)/len(c)-pos(w_r)/len(r)|$ は単語 w_c と単語 w_r の出現位置の相対的なずれを意味する。更に、1.0 から引くことで、相対位置のずれが大きいほど式(11)の pos_diff の値が小さくなるようにしている。これは、pos_diff は信頼度に対する負の重みとして用いるためである。したがって、式(10)

より得られる値は小さいほど対応関係にあり、大きいほど対応関係にはないことを示す。

これらの式(10)と式(11)を用いて、翻訳文と参照訳間のすべての単語間の距離計算を行い、距離行列を生成する。図 2 に図 1 で用いた翻訳文“基板用ブラケットの詳細構成を図８に示す。”と参照訳“この基板ブラケットの詳細構成を図８に示す。”における距離行列の例を示す。

	この	基板	ブラケット	の	詳細	構成	を	図	8	に	示す	。
基板	1.0	0.182	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
用	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ブラケット	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
の	1.0	1.0	1.0	0.051	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
詳細	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
構成	1.0	1.0	1.0	1.0	1.0	0.118	1.0	1.0	1.0	1.0	1.0	1.0
を	1.0	1.0	1.0	1.0	1.0	1.0	0.078	1.0	1.0	1.0	1.0	1.0
図	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.026	1.0	1.0	1.0	1.0
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.011	1.0	1.0	1.0
に	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.065	1.0	1.0
示す	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.05	1.0
。	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0

図 2　距離行列の具体例

距離行列の要素の値は小さいほど距離が近いことを意味している。したがって、単語アライメントより対応関係が得られた単語間の距離は小さな値となっている。そして、このような距離行列を用いた場合、EMD では類似性が高いほど値は小さくなる。しかし、自動評価法においてはその多くが評価値とスコアは比例している。そのため、提案手法でも多くの自動評価法と同様に以下の式(12)を用いて、評価値が高い場合にはスコアが高くなるようにしている。

$$score = 1.0 - EMD(C, R) \qquad (12)$$

3.3.4 性能評価実験

3.3.4.1 実験データ

本報告では提案手法の有効性を検証するために NTCIR-7 データ[16]を用いた。NTCIR-7 データでは英日方向については 5 の機械翻訳システムがそれぞれ出力した翻訳文 500 文(=100 文×5)、日英方向については 15 の機械翻訳システムがそれぞれ出力した翻訳文 1,500 文(=100 文×15)が含まれている。また、すべての翻訳結果に対応する参照訳が 1 つずつ存在する。更に、人手評価については 3 名の評価者が Adequacy と Fluency の 2 つの観点より 5 段階評価を行っている。今回の

性能評価実験では、3 名の平均値を人手評価のスコアとして用いた。日本語の翻訳文及び参照訳についてはMeCab[17]を用いて分かち書きを行った。

3.3.4.2 評価方法

評価は提案手法により得られるスコアと人手評価のスコアの間の Pearson と Kendall τ の相関係数をシステム単位と文単位でそれぞれ求めることで行った。また、提案手法の有効性を検証するために、ベースラインとして単語アライメントを行わない場合のシステムを用いた。その際、距離行列の生成にはすべての単語間の Dice 係数の値を用いた。更に、語順情報の有効性を確認するために、式(10)と式(11)より与えられる pos_diff を用いない場合のシステムについても相関係数を求めた。

3.3.4.3 実験結果

実験結果を表 1 から表 4 に示す。表 1 は英日翻訳におけるシステム単位の相関係数、表 2 は英日翻訳における文単位の相関係数である。また、表 3 は日英翻訳におけるシステム単位の相関係数、そして、表 4 は日英翻訳における文単位の相関係数である。そして、表 1 から表 4 には従来の表層情報のみに基づく自動評価法として BLEU と IMPACT の相関係数も参考までに付与している。

表 1　EtoJ におけるシステム単位の相関係数

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	-0.313	0.365	0.000	0.200
ベースライン	-0.246	0.351	0.000	0.200
pos_diff なし	-0.597	0.066	0.000	0.200
BLEU	-0.199	0.184	0.000	0.200
IMPACT	0.254	0.489	0.200	0.400

表 2　EtoJ における文単位の相関係数

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	0.537	0.503	0.346	0.354
ベースライン	-0.124	-0.070	-0.022	0.006
pos_diff なし	0.415	0.424	0.260	0.298
sentBLEU[18]	0.440	0.447	0.267	0.306
IMPACT	0.657	0.583	0.461	0.419

表 3　JtoE におけるシステム単位の相関係数

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	0.815	0.938	0.785	0.612
ベースライン	0.752	0.864	0.632	0.536
pos_diff なし	0.701	0.877	0.287	0.191
BLEU	0.730	0.881	0.498	0.440
IMPACT	0.814	0.935	0.689	0.555

表 4　JtoE における文単位の相関係数

	Pearson		Kendall	
	adequacy	fluency	adequacy	fluency
提案手法	0.521	0.539	0.367	0.386
ベースライン	0.067	0.155	0.090	0.110
pos_diff なし	0.464	0.519	0.331	0.359
sentBLEU	0.463	0.478	0.333	0.354
IMPACT	0.631	0.646	0.466	0.467

3.3.4.4　考察

表 1 から表 4 までの実験結果より、提案手法は単語アライメントを行っていないシステム（ベースライン）及び語順情報を反映させていないシステム（pos_diff なし）に対して、高い相関係数を示した。表 1 の EtoJ におけるシステム単位の相関係数においてはいずれのシステムも非常に低い相関となっている。この原因としては、英語から日本語の翻訳システムは 5 つと非常に少ないことから、1 つの翻訳システムに対する評価精度が低いと相関係数に与える影響が高くなってしまうことが考えられる。表 2 から表 4 においては、提案手法の相関係数はベースラインと pos_diff なしのシステムの相関係数よりもすべて上回っており、これは提案手法の有効性を示している。

表層情報のみに基づく自動評価法 BLEU と IMPACT との比較においては、提案手法は表 1 を除き、すべての相関係数で BLEU よりも高い。しかし、IMPACT との比較においては、表 3 の JtoE におけるシステム単位の相関係数においては高くなっているが、文単位では表 2 の EtoJ、表 4 の JtoE のすべてにおいて低い相関係数となった。この原因については詳細な検証が必要であるが、式(1)の tf・isf による重みの付与の困難さが影響したと考えられる。その具体例を図 3 に示す。

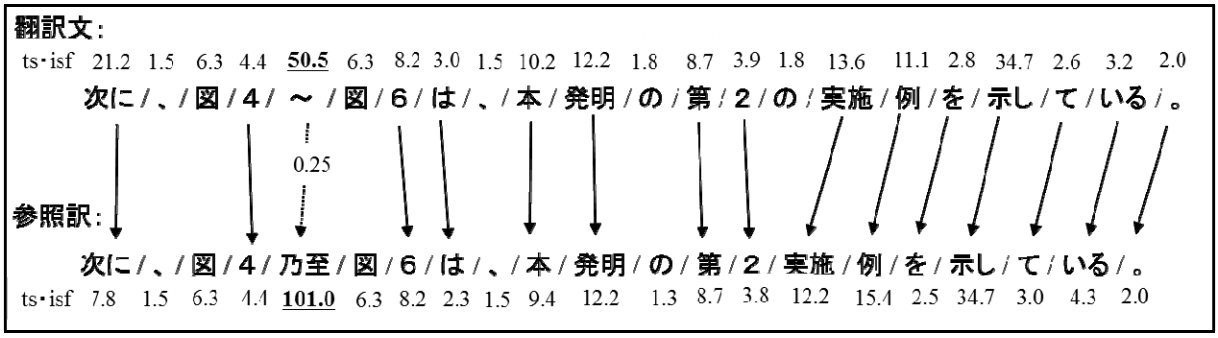


図 3　ts・isf の付与の具体例

図 3 では翻訳文の“～”と参照訳の“乃至”が対応関係にあるが、この間の Dice 係数は 0.25 であり、0.5 未満であるため対応関係は得られない。また、“～”の tf・isf の値は 50.0、“乃至”の tf・isf の値は 101.0 である。これらの値は翻訳文と参照訳それぞれの単語において、最大の値であった。今回用いた翻訳文と参照訳においては、これらの単語の出現頻度は低く、特徴的な単語として認識された。その結果、対応関係が得られた単語の tf・isf の値は相対的に小さくなり、スコアとしては 0.454 が得られた。このスコアは直感的には低いと考えられる。“～”や“乃至”のようなそれほど重要ではない単語に対しては、重みを小さくすることが理想である。統計的なアプローチを利用する以上はこのような問題はある程度避けられないが、他のコーパスも利用するなどの方法を用いることで解決できる可能性がある。なお、“、”や“の”の Dice 係数は 0.5 を超えているが、文中に複数存在し、一意に対応関係を決定できないため、単語アライメントの対象外としている。

3.3.5　まとめ

本報告では、統計的アプローチを用いた単語アライメントに基づく自動評価法を新たに提案した。提案手法では、EMD を用いてスコアを算出するが、単語アライメントを統計的アプローチで行ったうえで、その結果に基づき単語間の距離計算を行う。更に、語順の情報を距離計算に反映させることで評価精度の向上を図った。性能評価実験の結果、単語アライメントを行わなかった場合と語順情報を用いなかった場合に比べ、提案手法は人手評価に対して高い相関係数を示した。これは提案手法の有効性を示している。

今後は、ニューラルネット翻訳に対応した評価が行えるように提案手法を改良する予定である。即ち、EMD の適用において特徴ベクトルには word2vec による単語の分散表現を用いることで、単語の意味を考慮した自動評価法を構築し、性能評価実験を行う予定である。

謝辞

この研究は国立情報学研究所との共同研究に関連して行われた。

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V.L.E (2014) “Sequence to Sequence Learning with

Neural Networks," Advances in Neural Information Processing Systems, pp. 3104-3112.
[2] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning (2015) "Effective Approaches to Attention-based Neural Machine Translation," arXiv preprint arXiv:1508.04025.
[3] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation," Computational Linguistics, 19 (2), 263–311.
[4] Philipp Koehn, Franx Josef Och, and Daniel Marcu (2003) "Statistical Phrase-Based Translation," Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003), pp.48-54.
[5] Kishore Papineni, Salim Roukos, Todd Ward, and 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.
[6] 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.
[7] 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.
[8] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and 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.
[9] Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) "Efficient Estimation of Word Representations in Vector Space," Proceedings of Workshop at International Conference on Learning Representations 2013.
[10] 柳本豪一 (2015) "単語の分散表現を利用した文書類似度," Proceedings of the 29th Annual Conference of the Japanese Society for Artificial Intelligence, 4K1-1.
[11] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas (2000) "The Earth Mover's Distance as a Metric for Image Retrieval," International Journal of Computer Vision, 40(2), pp.99-121.
[12] Xiaojun Wan, and Yuxin Peng (2005) "The Earth Mover's Distance as a Semantic Measure for Document Similarity," Proceedings of the 14th ACM International Conference on Information and Knowledge management, pp.301-302.
[13] 藤江悠五，渡部広一，河岡司 (2009) "概念ベースと Earth Mover's Distance を用いた文書検索," 自然言語処理, 16(3), pp.25-49.
[14] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger (2015) "From Word

Embeddings To Document Distances," Proceedings of the 32nd International Conference on Machine Learning, pp.957-966.
[15] 松尾潤樹，小町守，須藤克仁 (2016) "単語分散表現を用いた単語アライメントによる日英機械翻訳の自動評価尺度," 情報処理学会第 227 回自然言語処理研究会, Vol.2016-NL-229 No.20, pp.1-7
[16] 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.
[17] "MeCab: Yet Another Part-of-Speech and Morphological Analyzer,"
http://taku910.github.io/mecab/
[18] Matouš Macháček, Amir Kamran, Philipp Koehn, and Ondřej Bojar (2015) "Results of the WMT15 Merics Shared Task," Proceedings of the Tenth Workshop on Statistical Machine Translation, pp.256-273.

3.4 中日テストセットを用いた特許文献の翻訳評価
－中国語分離パターンの利用－

元・山梨英和大学　江原 暉将
(株)富士通研究所　長瀬 友樹
(株)ディープランゲージ　王　　向莉

3.4.1　はじめに

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

筆者らは、中国語特許文献の中日機械翻訳評価のためにテストセットの検討を行い、昨年までに以下のことを実施した[4)5)6)]。

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

昨年度までのテストセットの作成において、844 個の中国語表現パターンとそれを含む中国語特許文の収集および中国語表現パターンに対する日本語翻訳パターン設問の作成を行った。これまでは、中国語において単語が連続する連続表現パターンのみを扱ってきたが、今年度は単語が連続しておらず、文中において、他の単語を間に挟む分離表現パターンに対するテストセットを構築した。併せて本テストセットを用いた評価サイトを研究会内限定で公開した。

3.4.2　中国語分離表現パターンの収集

100 万文対からなる中日特許文平行コーパス[2]から中国語部分を抜き出し、以下の手順で中国語分離表現パターンを抽出した。

まず、中国語文を文献 8)に示す方法で単語分割し、他の単語を間に挟む分離した単語バイグラム(分離バイグラムと呼ぶ)について出現度数をカウントする。ここで、分離間隔(間の単語数)は 4 以上、15 以下とした。分離間隔が 3 以下では、連続パターンの一部になる場合が多い、16 以上では生起が独立となると判断して分離間隔を設定した。

単語w_iとw_jの分離バイグラムカウントを$C(w_i, w_j)$とし、Vを全単語の集合とするとき

$$C(w_i) = \sum_{W_j \in V} C(w_i, w_j)$$

$$C(w_j) = \sum_{W_i \in V} C(w_i, w_j)$$

[1] 本部分は、AAMT 課題調査委員会で整備したサイトを利用させてもらっている。
[2] 本コーパスは WAT2016 の JPCzh-ja task において開示されたコーパスを利用している[7)]。

$$C = \sum_{W_i \in V} \sum_{W_j \in V} C(w_i, w_j)$$

と定義する。このとき、w_iとw_jの$t_score(w_i, w_j)$を以下のようにして求める。w_iとw_jの生起が独立とみなした時の期待される出現度数$C_{ind}(w_i, w_j)$は

$$C_{ind}(w_i, w_j) = C(w_i) \times C(w_j)/C$$

となり、

$$t\_score(w_i, w_j) = \frac{C(w_i, w_j) - C_{ind}(w_i, w_j)}{\sqrt{C(w_i, w_j) + C_{ind}(w_i, w_j)}}$$

で求められる。t_scoreの大きな単語対を表 1 に示す。表 1 では、中国語表現パターンとしては不適切な単語対(例えば、読点の対など)も含まれている。そこで、中国語表現パターンとして適切な対をt_scoreの大きい順に 500 対選択した。一部を表 2 に示す。

表 1 t_scoreの大きな単語対

前部分	後部分	t_score
、	、	271.5547
(	)	243.3445
在	中	179.6108
是	的	149.7051
图	的	132.5581
表示	的	124.4097
的	。	123.6213
在	,	118.0289
图	实施	117.9554
图	本	117.4901
图	发明	117.33
-	-	115.1469
对	进行	112.0874
当	时	109.3007
部	部	108.1184
图	方式	105.3952
在	上	100.3719
、	等	99.57987
图	实施例	97.94059
)	)	91.4757
图	图	91.43866
本	方式	87.41019
(	(	86.36381
在	下	85.77429
单元	单元	85.35363

表 2 中国語分離表現パターン(部分)

前部分	後部分	t_score
在	中	179.6108
在	上	100.3719
在	下	85.77429
在	有	77.30509
将	到	64.2659
从	到	47.15492
将	至	46.41461
在	内	43.047
使用	来	42.72821
可以	或	41.58661
从	向	41.07052
为	以下	40.05859
基	来	39.23804
与	地	38.40671
可	或	37.75489
可以	来	34.73906
以上	以下	33.71879
由于	而	31.79233
中	是否	31.09576
是	所	29.4842

500 対の中国語表現パターンに対して、それを含む中国語文をこれまでに作成した中日特許文平行コーパス[3]から収集した。その結果、全部で 19,090 文が収集できた。これらの文の中から各中国語表現パターンが主要な役割を持っている文をテストセット用の中国語文として選択した。つぎに、中国語表現パターンに対応する日本語翻訳パターン設問を設定した。設定にあたっては上述の中日特許文平行コーパスでの翻訳を参考にした。表 3 に今回収集した分離した中国語表現パターンを含むテストセットの例を示す。表中、CN は中国語テスト文、JA は日本語翻訳例、CN pattern は中国語表現パターン、JA pattern は日本語翻訳パターン設問を表す。

表 3 中国語表現パターンと日本語翻訳パターン設問例

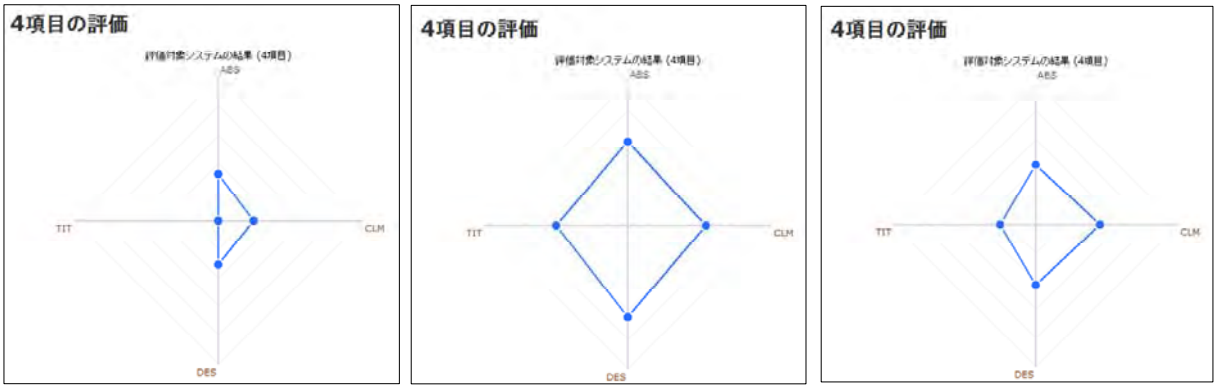
CN	JA	CN pattern	JA pattern
水分 和 甲醇 都 可 通过 以 干燥 空气 喷射 来 去除 。	水分およびメタノールは両方とも、乾燥空気による散布により、排除することができる。	以　来	によ(り|って).*(除去|排除)(する|し|でき|され)
耳朵 在 0 _ 220 MPa 的 不同 压力 下 处理 15 秒 。	耳を15秒間、0から220MPaの様々な圧力で処理した。	在　下	圧力(下)で
轴承 箱 55 依靠 四 个 螺栓 51 装配 在 支撑 平面 48 上 。	ベアリング・ハウジング55は、4つのボルト51によってサポートプレート48に取り付けられる。	在　中	(48|48)(上)に
保持 时间 从 0 秒 (对照) 到 60 秒 。	保持時間は0秒(対照区)から60秒間で変化させた。	从　到	から.*(60|60)秒(間|)(まで|で)

3.4.3 AAMT 自動評価サイトでの試験

昨年度までに作成したテストセットを AAMT 自動評価サイトにアップし、動作確認を行った。評価例を図 1 に示す。RBMT、SMT、オンラインサイトでの翻訳結果の評価例である。評価結果は、「発明の名称(TIT)」、「要約(ABS)」、「請求範囲(CLM)」、「詳細説明(DES)」 の 4 種類の出典別に 0～100%の範囲で

[3] 本コーパスは独自に収集したものであり 5)、WAT2016 で開示されたコーパスとは異なる。

設問への正解率が表示される。表示された四辺形の面積が大きいほうが評価値が高い。



(a) RBMT　　(b) SMT　　(c) online site

図 1 テストセットを用いた評価例

3.4.4　まとめと今後の課題

昨年度までのデータ作成と今年度の作成を合わせて 1064 の設問設定ができた。中国語パターンとしてはかなりの程度が収集できたのではないかと考える。今後の課題としては以下のことがあげられる。

・日本語翻訳パターンのバラエティが不足している部分があり、より適切な設問とすることが必要である。
・数式や化学式、数量表現など特許に特有な表現パターンが不足している。
・自動評価や人手評価とテストセット評価との比較を行い、双方のメリット・デメリットを明らかにする。

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

参考文献

1) Isahara, H. 1995. JEIDA's Test-Sets for Quality Evaluation of MT Systems -Technical Evaluation from the Developer's Point of View-. *Proc. of MT Summit V.*
2) Uchimoto, K., K. Kotani, Y. Zhang and H. Isahara. 2007. Automatic Evaluation of Machine Translation Based on Rate of Accomplishment of Sub-goals. *Proc. of NAACL HLT*, pages 33-40.
3) Nagase, T., H. Tsukada, K. Kotani, N. Hatanaka and Y. Sakamoto. 2011. Automatic Error Analysis Based on Grammatical Questions. *Proc. of PACLIC.*
4) 長瀬友樹, 江原暉将, 王向莉. 2014. 中日特許文評価用テストセットの作成, *平成 25 年度 AAMT/Japio 特許翻訳研究会報告書*, pages 78-82.
5) 長瀬友樹, 江原暉将, 王向莉. 2015. 中国語特許文献の中日翻訳評価のためのテストセットの改良と評価サイトの作成, *平成 26 年度 AAMT/Japio 特許翻訳研究会報告書*, pages 104-109.
6) 江原暉将、長瀬友樹、王向莉. 2016. 中国語特許文献の中日翻訳評価のためのテストセットの拡充, *平成 27 年度 AAMT/Japio 特許翻訳研究会報告書*, pages 40-42.
7) Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig and Sadao Kurohashi.

2016. Overview of the 3rd Workshop on Asian Translation. *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 1-46.

8) Terumasa Ehara. 2016. Translation systems and experimental results of the EHR group for WAT2016 tasks, *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 111-118.

3.5 WAT2016 における特許文翻訳タスクの人手評価結果の分析

科学技術振興機構　中澤　敏明
（株）東芝　インダストリアル ICT ソリューション社　園尾　　聡
NHK 放送技術研究所　後藤　功雄

3.5.1 はじめに

今年度、拡大評価部会人手評価グループでは、WAT2016[3]で行われた一対比較による評価と、特許庁が提案している「特許文献機械翻訳の品質評価手順」のうち「内容の伝達レベルの評価」に従った翻訳の専門家による評価の 2 種類の人手評価結果の分析を行った。一対比較評価では 400 文に対して、各システムとベースラインとなるシステムとの間で、1 文ずつ、どちらの翻訳の方がより良いか（もしくは同程度か）を判定し、その勝敗数をスコア化して各システムをランキングする。システムの出力がベースラインより良い場合は+1、悪い場合は-1、同程度の場合は 0 とし、5 人の異なる評価者の判断を足し合わせる。足しあわせた結果が+2 以上ならばその文ペアについては Win、-2 以下ならば Lose、それ以外ならば Tie と判定する。400 文に対してそれぞれ判定を行い、最終的に一対比較スコア（Pairwise）は以下の式で計算される。

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

内容の伝達レベルの評価は、各文について以下の基準での絶対評価を行う。なお内容の伝達レベルの評価は、一対比較評価の対象である 400 文のうち、ランダムに選択された 200 文に対して行った。また各言語対ごとに、一対比較評価の上位 3 チームに対してのみ内容の伝達レベルの評価を行った。

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

本稿ではこれらの評価結果に対して以下の 2 つの点からの分析を行ったので、報告する。

・一対評価の実施方法の違いによる信頼性の違い
・絶対評価(JPO-adequacy)と一対評価(pairwise)の関係性

3.5.2 一対評価の実施方法の違いによる信頼性の違い

3.5.2.1 仮説

WAT2015 では、クラウドソーシングにより一対評価を実施した。それに対して、WAT2016 では、価格はクラウドソーシングと同程度であるが、翻訳会社に委託して一対評価を実施した。クラウ

ドソーシングで実施した評価では、評価の質で評価者を評価する仕組みがないため、ランダムに評価値を付与している評価者がいる可能性がある。それに対して、翻訳会社では、評価者をマネージメントしているため、ランダムに評価値を付与している評価者がいる可能性は低いことが期待される。そこで、WAT2016 の一対評価の信頼性が WAT2015 よりも高いという仮説を立てる。この仮説について、データを比較することにより検証する。

3.5.2.2 用いたシステムと言語

ベースラインシステムの翻訳品質に比べて、評価対象のシステムの翻訳品質が高く、品質の差が大きい言語対とシステムについて、結果を分析した。その理由は、差が小さい場合は、ベースラインシステムと評価対象のシステムのどちらの翻訳品質が本当に高いのかが明確でないためである。さらに、WAT2015 と WAT2016 のいずれにおいても評価を実施したシステムである必要がある。我々は、上記の条件を満たすシステムとして、特許文の中日翻訳（JPC-CJ）の上位 3 システムと科学技術論文の日英翻訳（ASPEC-JE）の上位 3 システムを選択した。ここで、上位 3 システムは、WAT2015 での一対評価の上位 3 システムである。

3.5.2.3 信頼性を測る考え方

2014 年度の報告書[1]で、クラウドソーシングでの評価の信頼性について分析結果を報告している。その結果の概要は次の通りである。クラウドソーシングでは、正解と比較してランダムに近い評価値を付与しているワーカーも存在していた。正解とある程度相関している評価値を付与しているワーカーも存在していた。ほとんどの場合で正解と一致する評価値を付与しているワーカーも一部存在していた。そして、正解と強い負の相関を持つ評価値を付与しているワーカーは存在していなかった。このため、信頼性の低い評価値が含まれていても、全体を平均すると、ベースラインとの比較は、信頼できる結果になることが分かった。

上記の分析結果から、ベースラインより高性能なシステムとベースラインとを比較した評価の場合に、評価の信頼性は、ランダムに付与された評価値の割合が低いほど評価の信頼性が高いことに基づいて評価の信頼性を測ることができると考える。

3.5.2.4 具体的な指標

ここで用いる評価結果は、ベースラインシステムよりも（平均的に）高性能な評価対象システムとベースラインシステムの対であることを前提とする。

ランダムに付与された割合が高い評価結果と低い評価結果を比較すると、次の３つの統計量を指標としたときに、それぞれ以下のように考えられる。

（１）評価値平均値

全てランダムであれば、スコアの平均値はほぼ 0 となる。一方全て完全な評価であれば、スコアの平均値は 0 より大きい値になる。この値をここでは、s_{perfect} とする。ランダムなスコアの割合が高ければ、その割合に応じて、s_{perfect} の値が 0 に近づくと考えられる。すなわち、平均値が 0 に近いほどランダムに評価している割合が多いと考えることができる。

（２）Tie の割合

性能差があるシステム間で、文の翻訳結果の優劣を比較するとき、多くの場合で差があると考えられるため、評価値が 0（Tie）の割合は低いと考えられる。それに対して、ランダムに 3 択の選択肢の 1 つを選択すると、評価値が 0 の割合は 1/3 となる。すなわち、評価値が 0 の割合が 1/3 に近いほどランダムに評価している割合が多いと考えられる。

（３）相反する評価の率

同じ文の翻訳結果に対して、5 つの評価値が付与されている。ランダムに評価している割合が高ければ、同じ翻訳結果に対して、+1 と-1 が同時に付与される可能性が多くなると考えられる。すなわち、全く逆のスコアが付与された割合が多いほどランダムに評価している割合が高いと考えられる。

相反する評価の率は、同じ翻訳結果に付与された 5 つの評価値のうち、+1 の数を n_{+1} とし、-1 の数を n_{-1} とすると、

相反する評価の率= $\min\{n_{+1}, n_{-1}\} / (n_{+1} + n_{-1})$

として計算する。相反する評価の率が 20%というのは、同じ翻訳結果に付与された評価値のうち Win または Lose のものの中で少数のものの割合が 20%を意味している。相反する評価の率が 50%というのは、1 文あたりの評価の Win の数と Lose の数がちょうど同数であることを意味する。

以上の（１）、（２）、（３）の指標について分析する。なお、（１）と（２）は、スコア全体（2000 スコア）全体の傾向を対象としているのに対し、（３）は、各文での 5 つのスコア内の整合性の傾向を対象としているという違いがある。

3.5.2.5 分析結果

表 1 に（１）評価値平均値を示す。表 2 に（２）Tie の割合を示す。表 3 に（３）相反する評価の率を示す。

表 1: 評価値平均値

	SYSTEM1	SYSTEM2	SYSTEM3
WAT2015	0.165	0.179	0.166
WAT2016	0.224	0.226	0.278

(a) JPC-CJ

	SYSTEM4	SYSTEM5	SYSTEM6
WAT2015	0.254	0.227	0.194
WAT2016	0.339	0.282	0.248

(b) ASPEC-JE

表 2: Tie の割合

	SYSTEM1	SYSTEM2	SYSTEM3
WAT2015	0.376	0.241	0.360
WAT2016	0.260	0.255	0.231

(a) JPC-CJ

	SYSTEM4	SYSTEM5	SYSTEM6
WAT2015	0.187	0.167	0.193
WAT2016	0.130	0.152	0.148

(b) ASPEC-JE

表 3: 相反する評価の率

	SYSTEM1	SYSTEM2	SYSTEM3
WAT2015	0.210	0.216	0.219
WAT2016	0.208	0.195	0.200

(a) JPC-CJ

	SYSTEM4	SYSTEM5	SYSTEM6
WAT2015	0.239	0.246	0.257
WAT2016	0.233	0.250	0.249

(b) ASPEC-JE

表 1 より、WAT2015 の評価は、（1）評価値平均値が WAT2016 より 0 に近いことが分かる。これは WAT2016 での評価値が WAT2015 の評価値よりランダムに付与された割合が低い根拠になると考えられる。

表 2 より、WAT2015 の評価は、（2）Tie の割合が WAT2016 より WAT2015 のほうがほとんどの場合で、0.33=1/3 に近いことが分かる。これは WAT2016 での評価値が WAT2015 の評価値よりランダムに付与された割合が低い根拠になると考えられる。

表 3 より、WAT2015 と WAT2016 の評価は、（3）相反する評価の率がほぼ同程度であまり違いは見られなかった。この結果からはランダムに付与された割合の違いは見いだせなかった。これは（1）と（2）の結果とは傾向が異なる。しかし、相反する評価の率は 0.2〜0.25 程度であり、このことから WAT2015 と WAT2016 はいずれもランダムな選択の割合は、Tie を除く結果のうちの半数以下にとどまっており（相反する評価の率が 0.25 というのは、2 値分類の場合はその 2 倍すなわち 0.5 がランダム選択と考えられる)、Tie を除く結果のうち半数以上はランダム選択ではなく、訳質に基づいた評価結果と考えられる。

以上をまとめると、スコア全体（2000 スコア）全体の傾向の指標である（1）と（2）の結果から、WAT2016 の結果のほうが WAT2015 の結果よりもランダムに選択した評価値の割合が低く、

それゆえ信頼性が高いと言える。一方、各文でのスコア内の整合性の傾向の指標である（３）の結果からはWAT2015とWAT2016での信頼性の違いは観察されなかった。

3.5.2.6 Fleiss'κとの関係

評価の信頼性の指標として、Fleiss'κの値が利用されることがある。WAT2015とWAT2016の概要論文[2,3]にFleiss'κの値が報告されている。それらの値をWAT2015とWAT2016との間で比較する。

表4: Fleiss'κ

	SYSTEM1	SYSTEM2	SYSTEM3
WAT2015	0.087	0.117	0.111
WAT2016	0.096	0.131	0.123

(a) JPC-CJ

	SYSTEM4	SYSTEM5	SYSTEM6
WAT2015	0.104	0.070	0.076
WAT2016	0.078	0.066	0.068

(b) ASPEC-JE

表4にFleiss'κを示す。JPC-CJとASPEC-JEのいずれも、WAT2015とWAT2016の値はほぼ同じである。この結果は、前節での分析結果の（１）と（２）の指標の結果とは整合しないが、（３）の指標の結果と整合する。Fleiss'κが各文でのスコア内の整合性の傾向を測る指標であり、（３）も同様の指標であるためと考えられる。

さらに、ASPEC-JEでは、わずかにWAT2016の値がWAT2015の値より3つのシステムのいずれも低くなっている。これは（３）の指標の結果と整合しない。この主な原因として、次のことが考えられる。Fleiss'κは独立したカテゴリの分類の一致を測る指標であり、Win, Tie, Loseのようにカテゴリの順番に意味がある場合には、順番を考慮しないために必ずしも適切な指標にはなっていない。具体的には、ある翻訳結果に対する5つの値がWin３つ、Lose２つの場合と、Tie３つ、Lose２つの場合では、Win３つ、Lose２つのほうが評価のばらつきが大きいことを意味するが、これらの２つケースでFleiss'κは違いがないということである。

3.5.3 絶対評価(JPO-adequacy)と一対評価(pairwise)の関係性

WAT2016の特許翻訳タスクにおいて、絶対評価(JPO-adequacy)と一対評価(pairwise)のスコアは概ね一致する傾向となっているが、一部の翻訳方向／システムでは、JPO-adequacyのスコアが高いにもかかわらず、pairwiseのスコアが低くなるという結果が得られた。ここでは、JPO-adequacyとpairwiseのスコアの関係性について文レベルの分析を行ったので、その結果について報告する。

一対評価では、5名の評価者が、ベースラインの翻訳結果とターゲットの翻訳結果を比較し、

ベースラインよりターゲットの翻訳精度が、高い(Win)・同等(Tie)・低い(Lose)の相対評価を行っている。したがって、文レベルで見てみると、ベースライン自体の翻訳精度次第では、pairwise のスコアは低い（高い）が、JPO-adequacy のスコアが高い（低い）といった状況が起こりうる。そこで、ベースラインの翻訳精度が、JPO-adequacy と pairwise のスコアの乖離に影響しているのでは、という仮説を立て、以下のような検証を行った。

1. JPO-adequacy の評価が実施された全ての評価文について、pairwise で参照されたベースラインの翻訳結果に対する文レベルの BLEU スコア(sentence BLEU; sBLEU)を算出。
2. 各タスクの評価文を、JPO-adequacy のスコアが 4 以上[1]、かつ、pairwise のスコアが-2 以下[2]の群（incongruous LOSE）と、それ以外の群(OTHER)に分け、両群の sBLEU の平均値を比較。

結果を表 5 に示す。この結果、JPC-JE/EJ/CJ タスクにおいて、JPO-adequacy のスコアが高いにも関わらず、pairwise のスコアが低く評価された群は、ベースラインの翻訳精度(sBLEU)が有意に高いことが確認された。

表 5: 文レベルでのベースライン翻訳精度に対する分析(**: p<0.01, *: p<0.05)

Task	Incongruous LOSE	OTHER
JPC-JE	**34.981** (**)	26.554
JPC-EJ	**45.141** (**)	34.737
JPC-JC	25.240	23.729
JPC-CJ	**39.442** (*)	32.850
JPC-KJ	68.549	66.944

incongruous LOSE に含まれる評価文の一例を表 6 に示す。これらの例では、評価文(TGT)の翻訳結果は、原文の意味を十分伝えているが、ベースライン(BASE)は、ほぼ参照訳通りの訳となっており、一部の微妙な表現の差異により一対比較のスコアが低くなったと考えられる。

表 6: JPO-adequacy のスコアが高い(>=4)が、pairwise のスコアが低い(Lose)評価文の例

SRC:	FIG. 2 shows an example of the data configuration of the account registration data 1110 .
REF:	図２に、アカウント登録データ１１１０のデータ構成の一例を示す。
TGT:	図 2 は、アカウント登録データ 1110 のデータ構成の一例を示す図である。
BASE:	図２は、アカウント登録データ１１１０のデータ構成の一例を示す。
SRC:	System 10 further includes one or more sensors 28 .
REF:	システム１０はさらに、１つ以上のセンサ２８を含む。

[1] 2 名の評価者による評価結果（伝達レベル＝1（低い）～5（高い））の平均値を採用。また、日中／中日方向では、JPO-Adequacy のスコアが 3 以上とした。

[2] 5 名の評価者による評価結果（Win=+1, Tie=0, Lose=-1）の合計値を採用。

TGT:	システム 10 はさらに、一つまたは複数のセンサー28 を含む。
BASE:	システム１０は、さらに、１つ以上のセンサ 28 を含む。
SRC:	ビット線層５２は、ビット線ＢＬとして機能する。
REF:	The bit line layer 52 functions as the bit lines BL.
TGT:	Bit line layer 52 functions as a bit line BL.
BASE:	The bit line layer 52 functions as the bit line BL.
SRC:	この第一回転軸Ａｘ１は、ゴルフボール２の２つの極点Ｐｏを通過する。
REF:	The first rotation axis Ax 1 passes through the two poles Po of the golf ball 2 .
TGT:	The first rotation axis Ax 1 passes through the two poles of the golf ball 2 .
BASE:	The first rotation axis Ax 1 passes through the two poles Po of the golf ball 2.

一方、JPO-adequacy が低いにも関わらず、pairwise スコアが低く評価された文も少なからずあった。これらの文は、ベースラインの翻訳精度が相対的に低いものであると思われる。実際、未知語がそのまま訳出されている翻訳結果や、文法構造が破綻して非文となっている翻訳結果であったが、例数が僅かであったため、詳細な分析を行っていない。

また、JPO-JC/KJ では、トップシステムの JPO-adequacy のスコアが高いにも関わらず、pairwise のスコアが極端に低い傾向が見られた。図 1 に JPO-JC の、図 2 に JPO-KJ の上位システムの評価結果を示す。表 5 の分析結果では、sBLEU の有意な差は見られなかった。そこで、比較対象を Lose から Tie へ拡張し、すなわち、JPO-adequacy のスコアが 4 以上、かつ、pairwise のスコアが 0 以下の群(incongruous TIE)と、それ以外の群(OTHER)に分けて追加検証を行った。結果を表 7 に示す。この結果を見ると、JPO-adequacy のスコアが高いにも関わらず、pairwise のスコアが同等以下と評価された群は、ベースラインの翻訳精度(sBLEU)が有意に高いことが確認された。さらに、この群に属する評価文は、全体の評価文の内、JPC-JC で約 27%、JPC-KJ で約 44%を占めていることが分かった。これらの結果より、WAT2016 の JPC-JC/KJ のタスク評価において、JPO-Adequacy のスコアが高くとも、pairwise 評価において、同等以下と評価される文の割合が多いため、pairwise のスコアが低くなった（0 に近づく）ものと思われる。

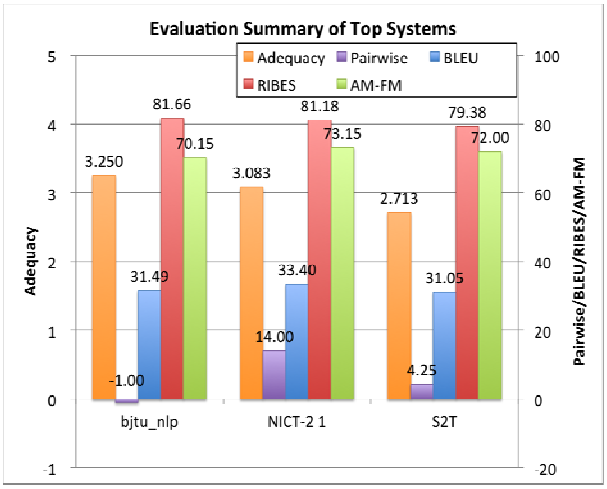


図 1 JPO-JC の上位システムの評価結果

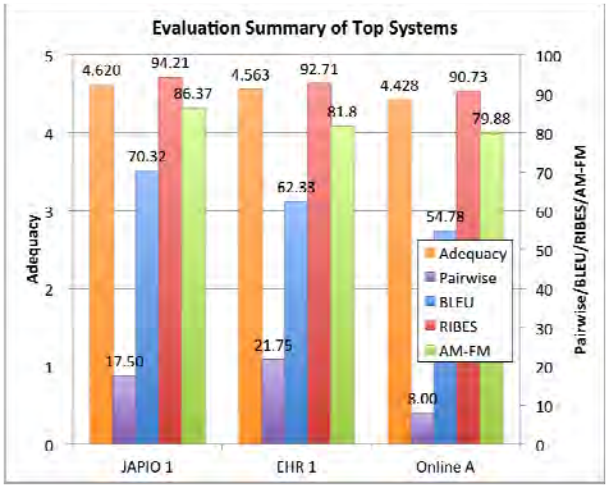


図 2 JPO-KJ の上位システムの評価結果

表 7: 文レベルでのベースライン翻訳精度に対する分析 2(**: p<0.01, *: p<0.05)

Task	Incongruous TIE	OTHER
JPC-JC	**34.981** (**)	22.531
JPC-KJ	**70.598** (**)	64.522

3.5.4 まとめ

本稿では WAT2016 の人手評価結果に対して、2 つの観点からの分析を行った。一対評価の実施方法の違いによる信頼性の違い分析の結果から、クラウドソーシングを利用するよりも翻訳会社を利用する方が全体として信頼度の高い結果が得られていることがわかった。一方で同一文の評価が乖離している割合はどちらも大差がないことがわかった。これは本質的に翻訳の評価は難しく、評価者がどこを重視して評価するかの違いが現れている可能性がある。ここの点をより詳細に分析することは今後の課題である。

また絶対評価(JPO-adequacy)と一対評価(pairwise)の関係性の分析結果からは、絶対評価と一対評価は基本的には相関があるものの、一対評価のベースラインシステムの質によって傾向が変わるということが定量的に確認された。これより、一対評価においてより信頼できる結果を得るためにはベースラインシステムを適切に選択することも重要であることが示唆される。

References

[1] 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, 平成 26 年度 AAMT/Japio 特許翻訳研究会報告書.

[2] Overview of the 2nd Workshop on Asian Translation, In Proceedings of the 2nd Workshop on Asian Translation (WAT2015).

[3] Overview of the 3rd Workshop on Asian Translation, In Proceedings of the 3rd Workshop on Asian Translation (WAT2016).

4. 第4回特許情報シンポジウム報告

4. 第4回特許情報シンポジウム報告

NTT コミュニケーション科学基礎研究所　須藤 克仁

1. 概要

AAMT/Japio 特許翻訳研究会の活動の一環として、翻訳を中心とする特許情報処理に関する情報交換と議論の場を提供するための「特許情報シンポジウム」を、平成 28 年 11 月 25 日（金）にグランパークカンファレンスにおいて開催した。本シンポジウムは 2010 年に第 1 回を開催し、以後 2 年ごとの開催で今回が第 4 回となる。

シンポジウムの企画及び運営は筑波大学宇津呂教授を委員長、NTT コミュニケーション科学基礎研究所須藤を副委員長とし、研究会委員で組織する実行・プログラム委員会を中心に行った。今回は招待講演 4 件、特別講演 1 件、一般講演 3 件と、研究会からの報告 2 件の構成であった。参加者は 93 名（研究会関係者を含む）と盛況であった。

2. セッション内容

2.1 招待講演（1）

午前のセッションは特許庁及び大学から講演者をお招きしての招待講演であった。

1 件目は特許庁特許情報室の加藤啓氏に「特許庁における機械翻訳の取り組み」と題して、特許庁で進めている外国特許公報の翻訳や審査書類の翻訳の取り組みやその課題についてご講演いただいた。近年の統計翻訳技術の進歩や対訳コーパスの整備により機械翻訳への期待は大きくなっており、今後さらに審査書類翻訳の精度向上や ASEAN 言語への対応が求められているというメッセージをいただいた。

2 件目は東京大学准教授の鶴岡慶雅氏に「ニューラルネットワークを用いた自然言語処理の最先端」と題して、近年急速に発展しているニューラルネットワークの言語処理における応用についてご講演いただいた。特に機械翻訳では従来の統計翻訳を凌駕する高性能を達成しており、折しもシンポジウム開催の直前に Google のニューラル機械翻訳が公開されたこともあり注目度は高かった。講演終了後も多くの参加者の方々が質問に来ていたのが印象的であった。

2.2 研究会報告

午後最初のセッションは AAMT/Japio 特許翻訳研究会から、翻訳評価に関する 2 件の報告を行った。

1 件目は岡山県立大学教授の磯﨑秀樹委員より、翻訳の自動評価方法についてご報告いただいた。世界的によく使われている自動評価尺度 BLEU の問題点と、それを踏まえて考案された語順の誤りに敏感な尺度 RIBES とその改良についての紹介であった。

2 件目は科学技術振興機構の中澤敏明委員より、第 3 回アジア翻訳ワークショップ(WAT)の機械翻訳共通タスクにおける人手評価結果についてご報告いただいた。WAT はシンポジウムの半月後の開催であったため本シンポジウムで先行してご紹介いただいたことになる。WAT の共通タスクは科学技術論文と特許が中心となっており、特に今回はニューラル機械翻訳を利用したチームが上位を占める結果となったことが示された。一方でニューラル翻訳特有の誤り（大規模な訳抜けや異なる内容への置換）もあり、現時点ではすべての面で従来の機械翻訳を上回るものとは言えないが、今後のさらなる発展によって従来の機械翻訳が過去の遺産となる可能性も示唆された。

2.3 招待講演 (2)

午後の招待講演セッションでは、産業界と特許事務所から講演者をお招きした。

1 件目は韓国 NAVER の Hyoung-Gyu Lee 氏に「Domain Adaptation for Machine Translation at NAVER LABS」と題して、NAVER での機械翻訳サービスや機械翻訳における分野適応の取り組みについてご講演いただいた。様々な分野の翻訳をするにあたって対訳リソースが不足する場合に、一般的な対訳リソースから当該分野の文に近いものを抜き出してきて利用することの有効性についての内容であった。

2 件目は久遠特許事務所の奥山尚一氏に「日本語の素晴らしさとユーザーの機械翻訳への大きな期待」と題してご講演いただいた。日本語の言語特性や歴史を踏まえ、思考のツールとして有効であること、機械翻訳の発展によりコミュニケーションのツールとしての英語の重要性が低下すること、特許情報が世界中で利用可能になることへの期待、といったお話であった。

2.4 特別講演及び一般講演

一般講演に先立ち、AAMT からの特別講演として、秋桜舎の山本ゆうじ氏に「文章と翻訳の品質を改善する-構造化用語データ UTX による用語管理と実務日本語ルール」と題してご講演いただいた。翻訳実務における用語集を共有・再利用しやすくするための共通規格として策定された UTX についての紹介であった。対訳用語対に関しては研究会でも様々な検討が行われており、その活用先として有望であると見込まれる。

一般講演は論文投稿を募集したもので、投稿された 3 件すべてが採択された。

1 件目は北海道大学の小野寺大輝氏の「課題とその対象を軸としたマトリクス型特許マップの自動生成方法の提案」であった。技術動向の分析のために特許マップは重要なツールの一つであり、複数の観点で技術の分類をするための拡張トピックモデル fbLDA によって、特許マップの自動生成を試みた研究の報告であった。

2 件目は翻訳者の吉川潔氏の「特許明細書の翻訳で注意すべきこと（翻訳者からのノウハウ）」であった。特許明細書の翻訳においては、単純に字面通りの翻訳をすれば十分というわけではなく、限定詞や並列構造等の正確な訳出のために文意を正確に把握しなければならないこと、現在の機械翻訳はその域には達していないことなどを述べられた。

3 件目は静岡大学の綱川隆司氏の「特許文献中の重要語に着目した特許分類の推定」であった。特許の分類のために用いられる F タームの自動付与のために、文献中の全出現語に加え、重要語を素性とした F ターム分類問題として定式化を行い改善を試みる研究の報告であった。

3. 所感

技術のグローバル化によって、特許情報処理は日本国内に閉じた問題ではなく、翻訳や言語横断処理を含めて考慮すべき問題となった。本シンポジウムでは機械翻訳の話題が多くなる傾向にあるが、ニューラル機械翻訳の急速な発展によって平均的な機械翻訳精度がかなり向上してきた現在、機械翻訳ができさえすれば、あとは今まで通りのワークフローでよい、ということではないと考える。つまり、決して完璧ではない機械翻訳、単言語向けに構築された特許の分類・分析方法が融合して、言語を問わず特許情報を扱うためのワークフローの改善に寄与することが特許情報処理の技術には求められる。そうした面では、今後のシンポジウムにおいて特許事務所、企業等の知財担当部署、あるいは知財分析業者等の特許実務業界の観点を強化することも、単に翻訳のみを切り出して考えるよりもさらに実用的な技術の実現に向けて重要なのではないだろうか。

平成 28 年度 AAMT/Japio 特許翻訳研究会報告書

発 行 日　平成 29 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