2018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 31 年 3 月

一般財団法人　日本特許情報機構

目　次

AAMT/Japio 特許翻訳研究会委員名簿

（敬称略・順不同）

委員長	辻井　潤一（※2）	国立研究開発法人 産業技術総合研究所 人工知能研究センター 研究センター長 / 東京大学 名誉教授
副委員長	宇津呂武仁（※2）	筑波大学大学院 教授
	須藤　克仁（※1）	奈良先端科学技術大学院大学 准教授
委員	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 主任研究員
	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
	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 スマートプロダクション研究部
	綱川　隆司	静岡大学学術院 助教
	中澤　敏明（※2）	東京大学大学院情報理工学系研究科 NEDO プロジェクト 実データで学ぶ人工知能講座 AI フロンティアコース 特任講師
	二宮　　崇	愛媛大学大学院 教授
	横山　晶一	山形大学 名誉教授
オブザーバ	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
	岡　　俊行	有限会社アジア産業 研究開発部長
	高　　京徹	株式会社高電社 経営企画部 部長
	園尾　　聡（※2）	東芝デジタルソリューションズ株式会社
	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
	王　　向莉（※2）	株式会社ディープランゲージ

オブザーバ（(一財）日本特許情報機構）：

石川雄太郎	特許情報研究所 調査研究部 研究企画課 課長代理
石附　直弥	特許情報研究所 調査研究部 研究企画課 課長
大塩　只明	特許情報研究所 調査研究部 研究企画課
小川　直彦	特許情報研究所 研究管理部 研究管理課 係長
木下　　聡	特許情報研究所 調査研究部 総括研究主幹
清藤　弘晃	特許情報研究所 調査研究部 部長
小林　　明	専務理事 / 特許情報研究所 所長
土屋　雅史	情報運用部 情報運用課 主任
船戸さやか	特許情報研究所 調査研究部 研究企画課 副主任
星山　直人	情報運用部 情報整備課 係長
三橋　朋晴	特許情報研究所 調査研究部 研究管理課 課長

（※1：拡大評価部会部会長、※2：拡大評価部会メンバー）

事務局	株式会社インターグループ

2018 年度 AAMT/Japio 特許翻訳研究会・活動履歴

平成 30(2018)年 5 月 11 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 30(2018)年 6 月 22 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 30(2018)年 7 月 27 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 30(2018)年 10 月 12 日
第 4 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 30(2018)年 12 月 7 日
第 5 回特許情報シンポジウム
（於ビジョンセンター浜松町）

平成 30(2018)年 12 月 21 日
第 5 回 AAMT/Japio 特許翻訳研究会
（於キャンパス・イノベーションセンター東京）

平成 31(2019)年 2 月 8 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 31(2019)年 3 月 8 日
第 7 回 AAMT/Japio 特許翻訳研究会
（於キャンパス・イノベーションセンター東京）

平成 31(2019)年 3 月 29 日
『2018 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　研究センター長
辻井　潤一

ニューラルネットワークの機械翻訳が急速に進展し、日常的に使われるようになってきた。ソフトウェアの技術者がオープンソースの英語説明を読む場合、あるいは、ビジネスマンが英文レターを書く際に適切な句や表現を見つける場合、機械翻訳は便利なツールとして活用されている。これらの使用では、いずれも、書き手や読み手がその分野のテキストを理解する能力を持っていて、多少、誤りがあっても、内容を読み取ることができることを前提にしている。オープンソースのプログラムを使おうとしているエンジニアは、そのソフトの背景や技術を知っているので、多少の誤訳があっても理解できる。英語よりも、誤訳があっても日本語の方が理解できるエンジニアの数は、非常に大きい。

ただ、今のニューラル翻訳には、当然ながら限界も多い。ある自治体が機械翻訳を使って駅の名前を英訳してそれを公開したら、吹き出してしまうような滑稽な英訳があると指摘されて、慌ててウェッブサイトを閉じたという、笑い話のような記事があった。駅名という固有名詞を、その固有名詞を構成する語ごとに翻訳し、東京を East Capital と訳すような誤りをした、というわけである。

地名のような固有名詞はその典型であるが、長い複合語がそれ自体で一つの専門概念を表している場合がある。このような場合には、それを部分に分けて翻訳し、その翻訳結果を単純につなげるだけでは意味をなさない翻訳になる。特に、日本語の場合には、漢字のように一文字だけで意味があり、翻訳できてしまう場合には、この問題は深刻となる。東京―＞East Capital といった誤訳が起こりえる。

深刻なのは、専門用語の場合には、その分野の知識がない人間は、このような誤訳に気が付かないことである。オープンソースの内容を理解するために日本語訳を読む場合のような、いわば使い捨ての翻訳であって、そのテキストを読む人の専門知識を期待できる場合には、この種の誤訳は深刻ではない。これが会社の製品説明書として印刷物やウェッブで公開される翻訳の場合では、前述の駅名翻訳と同様に、笑い話では済まなくなる。

専門用語の翻訳に関する課題は、機械翻訳の重要な研究テーマであるとともに、人間の翻訳家にとっても深刻な課題となる。機械翻訳が下訳作成に広く使われるに伴い、それを使う人間翻訳家には、対象分野に関する理解が要求される。機械翻訳が自然な翻訳を出せるようになればなるほど、誤訳を見つけるには分野の知識が必要となる。

今後は、専門用語の翻訳問題を解決する機械翻訳技術だけでなく、機械翻訳を道具として使う人間翻訳家のための支援ツールにどのような機能が必要かを考える必要がある。

以上のように、ニューラル翻訳という新たな技術が出てきたことで、これまでとは別の技術課題が生まれてきている。本委員会での活動が、今後ますます重要になると考えている次第である。1 年間の活動をまとめた本報告書が、激しく変貌を遂げつつある機械翻訳、特許翻訳の新たな技術的課題、また、本委員会がこれらの課題にどのように取り組もうとしているかを知っていただく一助になれば幸いである。

2. 機械翻訳および関連技術

2.1 係り受け構造に対する相対的位置を用いた Transformer モデル

愛媛大学　表 悠太郎
愛媛大学　田村 晃裕
愛媛大学　二宮　崇

2.1.1 はじめに

機械翻訳は自然言語処理の初期から盛んに研究され，様々な手法が提案されてきているが，近年では，ニューラルネットワークを用いた機械翻訳（Neural Machine Translation; NMT）が高い精度を実現しており，主流となっている．NMT の中でも，特に，同一文内の単語間の関係を捉える Self Attention という構造を用いた Transformer（Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017）が state-of-the-art の精度を達成し，注目を集めている．Transformer は，従来の畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）に基づく NMT (Gehring, Auli, Grangier, Yarats, and Dauphin 2017）や再帰型ニューラルネットワーク（Recurrent Neural Network; RNN）に基づく NMT（Sutskever, Vinyals, and Le 2014; Luong, Pham, and Manning 2015）と異なり，原言語文（翻訳元言語の文）や目的言語文（翻訳先言語の文）のすべての単語の組み合わせに対する Attention (Self Attention) を計算して中間表現を求める点で大きく異なる．そのため，語順や前後関係など各単語の文中における位置に関する情報は，Positional Encoding を用いて各単語の埋め込み表現に付随させている．

Transformer や Self Attention を改善するための手法がいくつかすでに提案されている．Shaw ら（Shaw, Uszkoreit, and Vaswani 2018）は，単語の絶対的な位置情報に加えて，単語間の文中における相対的な位置関係情報を Self Attention において考慮することで Transformer の精度改善を行っている．Strubell ら（Strubell, Verga, Andor, Weiss, and McCallum 2018）は，機械翻訳ではなく意味役割付与（Semantic Role Labeling; SRL）において、構文情報（係り受け構造）を用いて Self Attention の重み付けを学習するマルチタスク学習や，構文情報を直接 Self Attention に用いる手法を提案している．これまで，統計的機械翻訳や NMT では，原言語文や目的言語文，あるいはその両方の構文情報（句構造や係り受け構造など）を活用することで翻訳精度が改善されることが知られているため（Ding and Palmer 2005; Chen, Wang, Utiyama, Liu, Tamura, Sumita, and Zhao 2017; Eriguchi, Tsuruoka, and Cho 2017; Wu, Zhang, Zhang, Yang, Li, and Zhou 2018），Transformer においても構文情報を用いることで精度が改善されることが期待される．しかしながら，これまで構文情報を陽に活用した機械翻訳のための Transformer モデルはまだ提案されていない．

本稿は，原言語側の係り受け構造の情報を Transformer の相対的位置表現に用いた新しいニューラル機械翻訳を提案する．本研究は，Shaw ら（Shaw et al. 2018）が用いている相対的位置関係に注目し，語順に対する相対的位置関係だけを用いるのではなく，原言語文を係り受け解析し，得られた係り受け構造における単語間の相対的位置関係を埋め込んだベクトルを単語埋め込みベ

クトルに付随させることを行う．提案手法は係り受け情報を単語埋め込みベクトルに付随させるだけなので, Transformer 全体の仕組みや目的関数を変更する必要がなく, その他の Transformer の拡張モデルに適応しやすく，拡張性が高い．Strubell らの手法（Strubell et al. 2018）は，タスクの対象が SRL であること，また，単語間の Attention を直接係り受け構造から学習するため Transformer のモデルや目的関数を大きく変更している点で我々の手法と大きく異なる．

本実験では，科学技術論文の概要から作成された対訳文集合である ASPEC（Asian Scientific Paper Excerpt Corpus）(Nakazawa, Yaguchi, Uchimoto, Utiyama, Sumita, Kurohashi, and Isahara 2016）を用いて英日および日英翻訳を行った．その結果，英日翻訳タスクにおいては，提案モデルによる翻訳精度の向上が見られなかったが，日英翻訳タスクにおいては，原言語文の係り受け構造を相対的位置表現で考慮する提案モデルは従来の係り受け情報を考慮しない Transformer (Vaswani et al. 2017; Shaw et al. 2018）より高い翻訳精度を実現した．

2.1.2 Transformer

Transformer は，Self Attention という構造を持ったエンコーダとデコーダから構成される NMT モデルである．Transformer の概要図を図 1 に示す．Transformer は，エンコーダレイヤとデコーダレイヤがそれぞれ複数層スタックされたエンコーダ・デコーダ構造を持つ．エンコーダでは，入力された原言語文から中間表現を獲得する処理が行われる．デコーダでは，中間表現から目的言語文を予測し，出力する処理が行われる．デコーダは，文頭の単語から順に逐次的に目的言語の単語を予測して目的言語文を生成する．具体的には，最初に目的言語文の 1 番目の文頭単語を予測し，続いて，予測された文頭単語をデコーダに入力して 2 単語目を予測する．続いて, 予測された文頭から 2 単語目までを結合した解析途中の目的言語文をデコーダに入力（shifted right に目的言語文を入力）して 3 単語目を予測する，といった処理を繰り返すことで目的言語文の予測を行う．

Transformer のエンコーダとデコーダでは，まず，埋込み層で入力単語列（エンコーダ側は原言語文の単語列，デコーダ側は目的言語文の単語列）を埋込み表現を表す行列に変換する．その後，Transformer は RNN に基づく NMT とは異なり再帰的な構造を持たないため，Positional Encoding により単語の系列情報を付与する. 具体的には, 入力単語列の埋込み表現行列に対して, 各単語の文における絶対的な位置情報をエンコードした行列 PE を加える. PE の各成分は異なる周波数の sin，cos 関数を用いて次式により算出したものである．

$$PE(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right)$$
$$PE(pos, 2i+1) = \cos\left(pos/10000^{2i/d_{model}}\right)$$

ここで，d_{model} は入力単語の埋込み次元，pos は単語の位置，i は各成分の次元を表す．単語埋込み表現行列に PE を加えたものが, 第 1 層目のエンコーダレイヤやデコーダレイヤの入力となる．

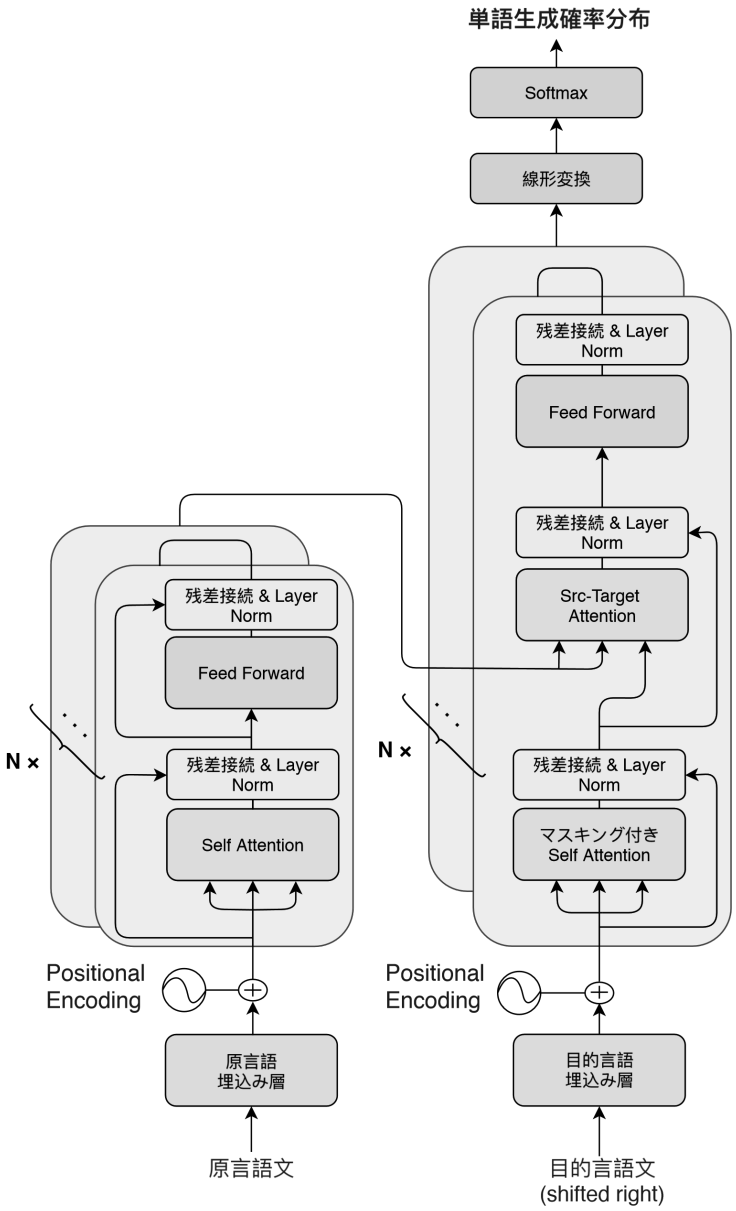


図 1: Transformer モデルの概要図

エンコーダレイヤは，下位のサブレイヤから順に，原言語文中の単語間の関係を捉える Self Attention，位置ごとのフィードフォーワードネットワーク（FFN）の２つのサブレイヤで構成されている．デコーダレイヤは，下位のサブレイヤから順に，目的言語文中の単語間の関係を捉えるマスキング付き Self Attention，原言語文の単語と目的言語文の単語間の関係を捉える Attention (Source-Target Attention), 位置ごとの FFN の 3 つのサブレイヤで構成されている．

各サブレイヤ間では，残差接続 (He, Zhang, Ren, and Sun 2016) を行った後に Layer Normalization (Ba, Kiros, and Hinton 2016) が適用される．Layer Normalization を適用する関数を$LayerNorm$，下位のサブレイヤからの出力を $\boldsymbol{x}$，現在のサブレイヤの処理を行う関数を $SubLayer$とすると，$LayerNorm\big(\boldsymbol{x} + SubLayer(\boldsymbol{x})\big)$が現在のサブレイヤの出力となる．

Self Attention と Source-Target Attention は Multi-Head Attention を用いて実現されている．Multi-Head Attention では，まず，3 つの入力ベクトル $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v} \in \mathbb{R}^{1\times d_{model}}$（本稿では特に断りがない限り，Transformer の原論文（Vaswani et al. 2017）に倣い，ベクトルを行ベクトルとして扱う）を重み行列 $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{model}\times d_k}$ $(i = 1, \cdots, h)$ により，d_{model}次元からd_k次元に線形写像した後，h 個の内積 Attention を計算する．ここで，d_{model}は元々の入力ベクトルの埋込み次元であり，$d_k = d_{model}/h$である．また，それぞれの内積 Attention をヘッド（$Head_i (i = 1, \cdots, h)$）と呼ぶ．

$$Head_i = Attention(\boldsymbol{q}', \boldsymbol{k}', \boldsymbol{v}')$$

$$Attention(\boldsymbol{q}', \boldsymbol{k}', \boldsymbol{v}') = \mathrm{softmax}\left(\frac{\boldsymbol{q}'\boldsymbol{k}'^T}{\sqrt{d_k}}\right)\boldsymbol{v}'$$

$$\boldsymbol{q}' = \boldsymbol{q}W_i^Q, \boldsymbol{k}' = \boldsymbol{k}W_i^K, \boldsymbol{v}' = \boldsymbol{v}W_i^V,$$

各ヘッドを連結した後，重み行列 $W^O \in \mathbb{R}^{d_{model}\times d_{model}}$ で線形写像する機構が Multi-Head Attention である．

$$MultiHead(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = Concat(Head_1, \cdots, Head_h)W^O$$

エンコーダの Self Attention では，上式の $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}$ に，エンコーダの内部状態系列 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ が代入される．具体的には，各ヘッドは次の出力系列 $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$ を計算する．

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij}\boldsymbol{x}_j W^V$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n}\exp(e_{ik})}$$

$$e_{ij} = \frac{(\boldsymbol{x}_i W^Q)(\boldsymbol{x}_j W^K)^T}{\sqrt{d_k}}$$

デコーダの Self Attention では，デコーダの内部状態系列を用いて計算を行う．ただし，推論時には，予測する単語より後で生成される単語を知ることはできない．そのため，デコーダの Self Attention では，予測する単語とそれより後方に位置する単語の間の関係性を考慮しないようにマスクしたマスキング付き Self Attention を用いる．

デコーダは，最終のデコーダレイヤの出力$\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n$に対して線形変換を施し，その後ソフトマックス関数を施すことによって，目的言語の各単語の生成確率分布を得る．

2.1.3 文内の相対的位置表現を用いた Transformer

Shaw ら（Shaw et al. 2018）は，2 単語間の文における相対的な位置関係を Transformer エンコーダおよびデコーダ内の Self Attention で捉える手法を提案した．Shaw らの手法では，入力文中の各単語の中間表現$\boldsymbol{x}_i, \boldsymbol{x}_j$間の関係はベクトル$\boldsymbol{a}_{ij}^V, \boldsymbol{a}_{ij}^K \in \mathbb{R}^{d_k}$で表現する．そして，サブレイヤの出力に単語間の相対的位置情報を付加して次の層への入力とする．具体的には，次式を用いて Self Attention の出力系列 $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$ を求める．

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij}(\boldsymbol{x}_j W^V + \boldsymbol{a}_{ij}^V)$$

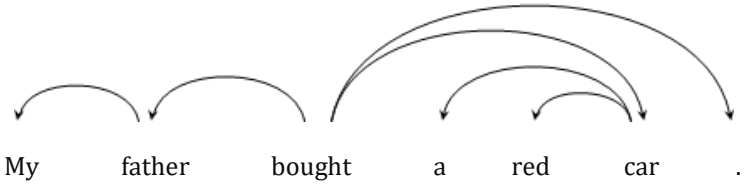


図 2: 係り受け構造の例

また，Self Attention 計算過程の e_{ij} も単語間の相対的位置情報を考慮するため，次式を用いる．

$$e_{ij} = \frac{(\boldsymbol{x}_i W^Q)\left(\boldsymbol{x}_j W^K + \boldsymbol{a}_{ij}^K\right)^T}{\sqrt{d_k}}$$

ここで Shaw らは，単語間が一定距離以上離れると離れ具合の影響は少ないと仮定し，相対的位置の距離の最大値を定数 k と定め，それより離れた相対的位置は最大値 k とした．また，文中のある単語から後ろを正の方向，前を負の方向と考え，二単語間の相対的位置関係は，以下の通り，$2k+1$個のユニークなラベル $(-k, -k+1, \cdots, 0, \cdots, k-1, k)$ で与える．

$$\boldsymbol{a}_{ij}^K = \boldsymbol{w}_{clip(j-i,k)}^K$$
$$\boldsymbol{a}_{ij}^V = \boldsymbol{w}_{clip(j-i,k)}^V$$
$$clip(x,k) = \max(-k, \min(k,x))$$

ここで，最終的に学習されるパラメータは，$\boldsymbol{w}^K = (\boldsymbol{w}_{-k}^K, \cdots, \boldsymbol{w}_k^K)$ と $\boldsymbol{w}^V = (\boldsymbol{w}_{-k}^V, \cdots, \boldsymbol{w}_k^V)$ $(\boldsymbol{w}_k^K, \boldsymbol{w}_k^V \in \mathbb{R}^{d_k})$ である．

2.1.4 提案手法

本節では，最初に本研究で用いる係り受け構造について説明し，続いて係り受け構造に対する相対的位置表現を用いた Transformer を提案する．

係り受け関係とは，単語間の「修飾」「被修飾」の関係のことであり，方向性を持つ．一例として，“My father bought a red car. ”という文の係り受け構造を図 2 に示す．ここで，矢印の向きは被修飾語から修飾語を指し示している．つまり，単語 A が単語 B を修飾するとき，単語 A は単語 B の子ノードになる．係り受け構造は単語をノードとする全域木になっている．

提案手法では，Shaw ら（Shaw et al. 2018）に倣い，原言語文の係り受け構造における相対的位置ラベルを埋込みベクトルで表現し，原言語文中の 2 単語間の係り受け構造における相対的位置の情報を Transformer エンコーダ内の Self Attention に導入する．原言語文中の 2 単語 w_i, w_j に対して，係り受け構造における相対的位置関係を表す位置ラベル $label_{i,j}$ を，次の通り与える．

- 単語 w_i と単語 w_j に対応するノード n_i とノード n_j が祖先子孫関係の場合，$label_{ij} = depth(n_j) - depth(n_i)$ とする．ここで，$depth(n)$ はノード n の深さを表す．例えば，図 2 において，“My”（$= w_1$）を基準にした“bought”（$= w_3$）との位置関係を表す位置ラベルは $label_{1,3} = 0 - 2 = -2$ である．この定義より，係り受け構造においてある単語の親方向は負，

表 1: 係り受け構造に対する相対的位置ラベルの例

	My	father	bought	a	red	car	.
My	$self$	-1	-2	non_dep	non_dep	non_dep	non_dep
father	1	$self$	-1	non_dep	non_dep	sib	sib
bought	2	1	$self$	2	2	1	1
a	non_dep	non_dep	-2	$self$	sib	-1	non_dep
red	non_dep	non_dep	-2	sib	$self$	-1	non_dep
car	non_dep	sib	-1	1	1	$self$	sib
.	non_dep	sib	-1	non_dep	non_dep	sib	$self$

子方向は正の値で相対的位置関係が表される.

- 単語 w_i と単語 w_j に対応するノード n_i とノード n_j が兄弟関係にある場合，兄弟関係ラベル sib（$label_{ij} = sib$）とする．ここで，2 ノードが兄弟関係の場合，祖先子孫関係とは異なり，方向性や距離の情報を相対的位置ラベルに与えない.
- 単語自身との相対的位置ラベルは自分自身を表すラベル $self$ とする．つまり，任意の単語 w_i に対して，$label_{ij} = self$ である.
- 上記 3 パタン以外の 2 単語間の相対的位置ラベルは，依存関係なしを表すラベル non_dep とする.

図 2 の係り受け構造における 2 単語間の相対的位置ラベルを表 1 に示す. 表 1 では, 行が単語 w_i , 列が単語 w_j に対応している.

相対的位置ラベルをもとに, 原言語文の各単語の中間表現 $\boldsymbol{x}_i, \boldsymbol{x}_j$ 間の係り受け構造に対する相対的位置関係をベクトル $\boldsymbol{b}_{ij}^V, \boldsymbol{b}_{ij}^K \in \mathbb{R}^{d_k}$ で表現し，次式を用いて Self Attention の出力系列 $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$ を求める.

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij} \left(\boldsymbol{x}_j W^V + \boldsymbol{b}_{ij}^V \right)$$

また，Self Attention 計算過程の e_{ij} も単語間の相対的位置情報を考慮するため，次式を用いる.

$$e_{ij} = \frac{(\boldsymbol{x}_i W^Q)\left(x_j W^K + \boldsymbol{b}_{ij}^K\right)^T}{\sqrt{d_k}}$$

係り受け構造における相対的位置関係においても，一定以上距離が大きいと離れ具合の影響は少なくなると仮定し，最大距離を定数 k に制限する.

上述の手法を提案手法 1 とする．提案手法 1 に加えて，Shaw ら（Shaw et al. 2018）の提案する文内における相対的位置表現の両方の情報を考慮する手法を提案手法 2 とする．具体的には，$\boldsymbol{a}_{ij}^V, \boldsymbol{b}_{ij}^V$ と $\boldsymbol{a}_{ij}^K, \boldsymbol{b}_{ij}^K$ をそれぞれ結合し，重み行列 $W_{rel}^V, W_{rel}^K \in \mathbb{R}^{2d_k \times d_k}$ を用いて線形変換を施したベクトル $\boldsymbol{c}_{ij}^V, \boldsymbol{c}_{ij}^K \in \mathbb{R}^{d_k}$ を 2 単語間の相対的位置情報として用いる．つまり，提案手法 2 は次式を用いて Self Attention の出力系列 $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$ を求める.

$$\boldsymbol{c}_{ij}^V = Concat\left(\boldsymbol{a}_{ij}^V, \boldsymbol{b}_{ij}^V\right) W_{rel}^V$$

$$\boldsymbol{c}_{ij}^{K} = Concat\left(\boldsymbol{a}_{ij}^{K}, \boldsymbol{b}_{ij}^{K}\right) W_{rel}^{K}$$

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij}\left(\boldsymbol{x}_j W^{V} + \boldsymbol{c}_{ij}^{V}\right)$$

$$e_{ij} = \frac{\left(\boldsymbol{x}_i W^{Q}\right)\left(\boldsymbol{x}_j W^{K} + \boldsymbol{c}_{ij}^{K}\right)^{T}}{\sqrt{d_k}}$$

2.1.5 実験

本実験では，科学技術論文の概要から作成された対訳文集合である ASPEC（Asian Scientific Paper Excerpt Corpus）(Nakazawa et al. 2016) を用いて英日および日英翻訳を行った．英語文は Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, and McClosky 2014) を用いて単語分割および係り受け解析を行った．日本語文は KyTea (Neubig, Nakata, and Mori 2011) を用いて単語分割した．また，EDA[1]を用いて係り受け解析を行った．モデルの学習では，学習データ（train-1.txt）から抽出した英語文・日本語文ともに文長 50 単語以下の 100,000 対訳文対を使用した．また学習データに出現した単語のうち出現頻度が 2 回以上の単語のみを語彙として用い，出現頻度が 1 回の単語は語彙に登録されていない未知語を表す$\langle UNK \rangle$タグに置き換えた．検証データとして 1,790 文対（dev.txt），テストデータとして 1,812 文対（text.txt）を用いた．

実験では，2 種類の提案手法を，従来の絶対的位置表現を考慮する Transformer ($Transformer_{abs}$) (Vaswani et al. 2017)と文中の相対的位置表現を考慮する Transformer ($Transformer_{rel}$) (Shaw et al. 2018)と比較する．評価対象の全ての Transformer モデルのハイパーパラメータは Vaswani ら（Vaswani et al. 2017）の設定に倣い，エンコーダ及びデコーダレイヤのスタック数を 6，ヘッド数を 8，埋込み次元を 512 次元とした．optimizer は Adam を用い，$\beta_1 = 0.9, \beta_2 = 0.98, \varepsilon = 10^{-9}$ と設定した．学習率の更新スケジューリングは Vaswani らの方法 (Vaswani et al. 2017) と同様にした．また，提案手法 1,2 及び $Transformer_{rel}$ において，考慮する相対的位置の最大距離は $k = 2$ とした．ミニバッチサイズは100，エポック数は50とし，検証データに対して最も精度が良かったエポックのモデルをテストデータに適用して翻訳精度を評価した．本実験では，greedy アルゴリズムにより目的言語文を生成した．

評価結果を表 2 に示す．なお，翻訳性能の評価指標は BLEU を用いた．表 2 より，英日翻訳では，提案手法 1 と $Transformer_{abs}$ の性能差は 0.03 ポイントであり，同等の性能であったことが分かる．また，提案手法 2 は，$Transformer_{abs}$ と比較して 1.11 ポイント BLEU が上回ったが，$Transformer_{rel}$ と比較して 0.24 ポイント下回る結果となった．一方で，日英翻訳では，提案手法 1 と $Transformer_{abs}$ の性能差は 0.05 ポイントであり同等の性能であったが，提案手法 2 は $Transformer_{abs}$ と比較して 0.70 ポイント BLEU が上回り，$Transformer_{rel}$ と比較して 0.40 ポイント上回った．これらの結果より，英日翻訳タスクにおいては提案手法の有効性は確認できなかったが，日英翻訳タスクにおいては，原言語文の係り受け構造を相対的位置表現で考慮することで Transformer モデルの翻訳精度を改善できることが分かった．

[1] http://www.ar.media.kyoto-u.ac.jp/tool/EDA/

表 2: 実験結果

	BLEU (%)	
	英→日	日→英
$Transformer_{abs}$	30.21	22.26
$Transformer_{rel}$	**31.56**	22.56
提案手法 1	30.18	22.21
提案手法 2	31.32	**22.96**

2.1.6 まとめ

本研究では，Transformer において原言語文の係り受け構造を活用するため，原言語の係り受け構造における単語間の相対的位置関係を Transformer エンコーダの Self Attention の中の相対的位置表現に導入する手法を提案した．ASPEC (Nakazawa et al. 2016) データを用いた評価実験を通じて，日英翻訳タスクにおいては，原言語文の係り受け構造に対する相対的位置表現を考慮することで Transformer モデルの精度改善を達成できることを確認した．

本研究の提案モデルでは目的言語側の係り受け構造を考慮しないが，今後は，目的言語側の係り受け構造を考慮できるモデルに改良したい．例えば，Wu ら (Wu et al. 2018) が用いていた依存構造解析を行う RNN を Transformer デコーダに統合することにより，デコーダ内の Self Attention で目的言語側の係り受け構造における単語間の相対的位置関係を考慮するモデルに改良できる可能性がある．また，本研究の実験で用いたデータ以外のデータセットや言語対での評価も行っていきたい．

参考文献

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin. (2017) Attention is all you need. Advances in Neural Information Processing Systems 30, pp. 5998—6008.

J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. (2017) Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning, pp. 1243—1252.

I. Sutskever, O. Vinyals, and Q. V. Le. (2014) Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems 27, pp. 3104—3112.

T. Luong, H. Pham, and C. D. Manning. (2015) Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural

Language Processing, pp. 1412—1421.

P. Shaw, J. Uszkoreit, and A. Vaswani. (2018) Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 464—468.

E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum. (2018) Linguistically-informed self-attention for semantic role labeling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5027—5038.

Y. Ding and M. Palmer. (2005) Machine translation using probabilistic synchronous dependency insertion grammars. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 541—548.

K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita, and T. Zhao. (2017) Neural machine translation with source dependency representation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2846—2852.

A. Eriguchi, Y. Tsuruoka, and K. Cho. (2017) Learning to parse and translate improves neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 72—78.

S. Wu, D. Zhang, Z. Zhang, N. Yang, M. Li, and M. Zhou. (2018) Dependency-to-dependency neural machine translation. IEEE/ACM Trans. Audio, Speech and Lang. Proc., Vol. 26, No. 11, pp. 2132—2141.

T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. (2016) ASPEC: Asian scientific paper excerpt corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), pp. 2204—2208.

K. He, X. Zhang, S. Ren, and J. Sun. (2016) Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770—778.

J. L. Ba, J. R. Kiros, and G. E. Hinton. (2016) Layer normalization. arXiv preprint arXiv:1607.06450.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. (2014) The Stanford CoreNLP natural language processing toolkit. In Association for Computational Linguistics (ACL) System Demonstrations, pp. 55—60.

G. Neubig, Y. Nakata, and S. Mori. (2011) Pointwise prediction for robust, adaptable Japanese morphological analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 529—533.

2.2 文脈を考慮するニューラル機械翻訳における最適文脈文選択法

木村 龍一郎, 飯田 頌平, 崔 鴻翌, 洪 博軒,
(筑波大学大学院 システム情報工学研究科)
宇津呂 武仁,
(筑波大学大学院 システム情報工学研究科)
永田 昌明
(NTT コミュニケーション科学基礎研究所)

2.2.1 はじめに

ニューラル機械翻訳 (NMT)モデルは近年の研究によって大きく進歩している[13, 9, 16]．しかし，これら標準的な NMT モデルは原言語一文を目的言語一文に翻訳するよう設計されているため，複数文にわたる文脈を考慮した翻訳は不可能である．この問題に対処するため，文脈として追加で文を入力できる文脈考慮型の翻訳モデルがいくつか提案された[14, 7, 10, 11, 1, 17, 15]．これらのモデルは入力できる文脈の長さによって大別することが可能である．最も典型的なモデルは，直前の文を文脈とみなして入力するモデルで，すべての入力を一つのエンコーダで処理するモデル [14]，および，文脈用に別のエンコーダを用意するモデル [7, 1, 17]に分けられる．より広い文脈を考慮するモデルとしては，直前 3 文を考慮するモデル [11]，文書中の先行文全体を考慮するモデル [15]，および，文書全体を考慮するモデル [10]に分けられる．いずれのモデルも，原言語一文のみを考慮するモデル(以下，1-to-1 翻訳モデルと呼ぶ)と比較してより高い翻訳精度を達成している．その中で，Tiedemann らの文脈考慮型 NMT モデルは，直前の文を現言語文に連結した対訳対を用いて翻訳モデルを訓練するという簡易なもので，2-to-2 翻訳モデルと呼ばれる [14]．2-to-2 翻訳モデルは 1-to-1 翻訳モデルとほぼ同等の簡易なモデルである点において他の文脈考慮型モデルより優れるものの，GPU メモリサイズの制約のために 2 文より広い範囲を文脈とすることが難しく，文脈として考慮できる範囲が限定されるという欠点がある．

本論文では，前後 5 文までを文脈文とした 2to-2 翻訳モデルによるオラクル翻訳の BLEU スコアを分析し，2-to-2 翻訳モデルにおいて直前の一文よりも広い文脈を考慮する文脈考慮型 NMT モデルの有効性を示す．そして，2-to-2 翻訳モデルを拡張するため前後 5 文までの中から適切な文脈を選択する方法を提案する．評価実験の結果として，オラクル翻訳の BLEU スコアには及ばないものの，直前のみを考慮する 2-to-2 翻訳モデルと比較して，強制逆翻訳確率の最大化によって有意に BLEU が改善することを示す．

2.2.2 文脈型 NMT のオラクル翻訳

2.2.2.1 拡張文脈

2-to-2 翻訳モデルにおいては，原言語文の直前の文を文脈文とみなして，連結記号<CONCAT>を介して文脈文と原言語文と連結して翻訳する．2-to-2 翻訳モデルの BLEU スコアは 1-to-1 翻訳モデルを上回るものの，文脈として直前 1 文までしか考慮できない点が問題である．そこで，本論文では，より広い範囲を文脈として考慮することにより，2-to-2 翻訳モデルを拡張する．具体的には，ベースラインである 1-to-1 翻訳モデルによる翻訳$\boldsymbol{y_{11}}$に加え，直前 5 文を文脈文とする 2-to-2 翻訳モデルによる訳文$\boldsymbol{y_{22}^{-1}, \ldots, y_{22}^{-5}}$，および，直後 5 文を文脈文とする 2-to-2 翻訳モデルによる訳文$\boldsymbol{y_{22}^{+1}, \ldots, y_{22}^{+5}}$を生成し，これらを訳文候補集合とする [1]．以上の 11 個の訳文候補集合の中から最適な訳文を選択する．提案手法による訳文候補集合における翻訳精度の上限を示すため，訳文候補集合中で BLEU スコア(文単位の翻訳精度を表す指標で

[1]各$\boldsymbol{y_{22}^{i}}$ (i = ±1, … , ±5) は，原言語文を起点として，前後5 文中のi 文目$\mathbf{c}^{i}$ (i = ±1, … , ±5) を文脈文とする2-to-2 翻訳モデルによる訳文を表す.

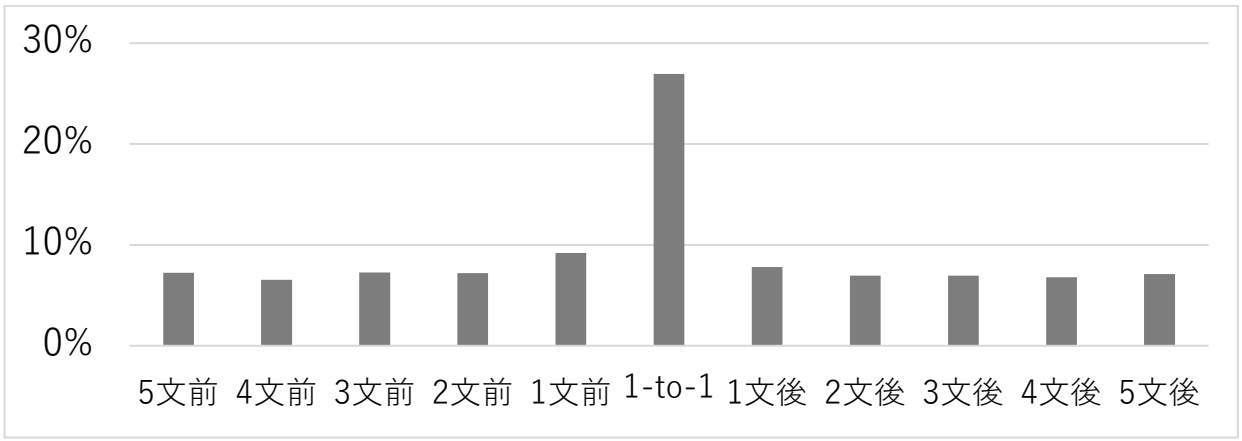
ある sentence-BLEU)最大となる訳文(オラクル翻訳と呼ぶ)を選定し，オラクル翻訳の BLEU スコア，および，文脈文位置の分布を分析する．

2.2.2.2 データセットと実験条件

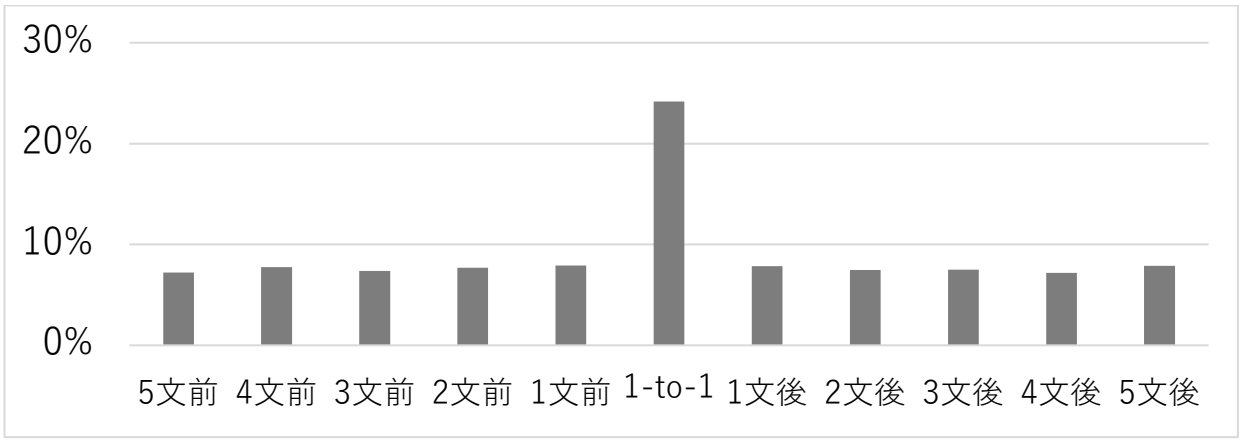
映画字幕の公開データである Opensubtitles2018 [8]を使用し，文脈付き対訳文の作成法[14]に従い，合計で 2,083,576 文の日英対訳を作成した．映画等の作品単位で 90%を訓練用， 5%を開発用，5%を評価用として無作為に分割し，合計で 1,876,624 対訳文を訓練用，104,379 対訳文を開発用，102,573 対訳文を評価用とした．オラクル翻訳の作成および BLEU スコア評価においては，評価文のうち 10,000 文を使用した [23].

[2]翻訳モデルの訓練においては，2-to-2 翻訳時の文脈文連結後の原言語文および目的言語文のいずれかが50 単語を超える文については除外した．

[3]詳細な実験条件は次の通り: 英語のtokenization にはMoses tokenizer 4)，日本語の形態素解析にはMeCab (http://taku910.github.io/mecab/) を使用．翻訳モデル作成にはOpenNMT-py を使用．語彙は頻度上位5 万語を使用．単語分散表現は512 次元，エンコーダ・デコーダとも各6 層，バッチサイズを4,096，drop out rate を0.3 として，100,000 エポックの訓練を行う．Adam opti-mizer を使用．ハードウェアはNVIDIA Tesla P100 16GBGPU 1 枚を使用．BLEU の測定及び有意差検定にはMTE-val Toolkit (https://github.com/odashi/mteval)を使用，sentence-BLEU の測定にはMoses decoder のsentence-bleu.cpp を使用．



(a) 日英

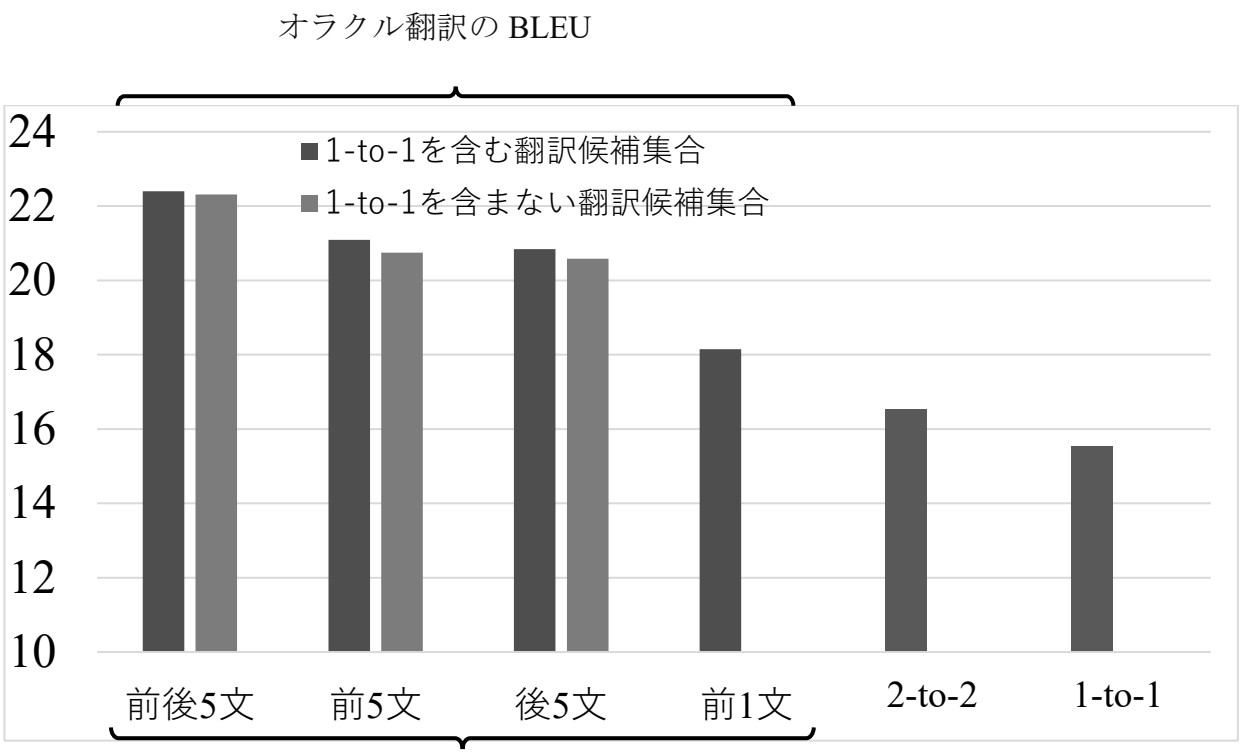


(b) 英日

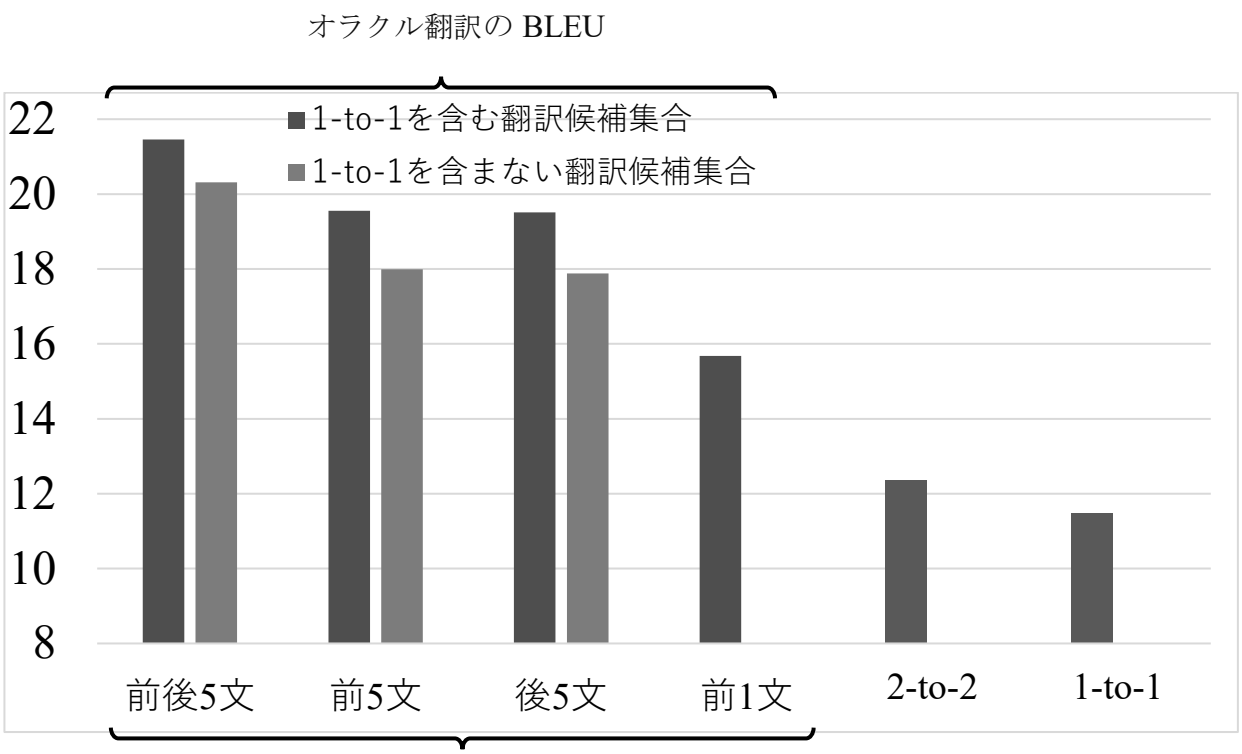
図 1　オラクル翻訳における最適文脈文位置の分布

2.2.2.3 オラクル翻訳の分布

オラクル翻訳における訳文候補選択結果の分布を図 1 に示す．オラクル翻訳 10,000 文のうち，訳文候補集合の中から sentence-BLEU 最大である訳文が一意に定まったものは日英で 39%，英日で 46%であった．sentence-BLEU 最大の訳文が一意に定まらない評価文を集計から除いて最適文脈文位置の分布を求めた．1-to-1 翻訳モデルによる訳文が sentence-BLEU 最大である割合は日英で 27%，英日で 24%であった．また，2-to-2 翻訳が sentence-BLEU 最大となる文脈位置が 1 文前である割合は日英で 9%，英日で 8%，2 文前から 5 文前である割合は日英で 28%，英日で 30%，1 文後から 5 文後である割合は日英で 36%，英日で 38%であった．この結果より，考慮する文脈として後方を追加する方式の有効性が示された．同様に，2-to-2 翻訳において 2 文前から 5 文前，または，1 文後から 5 文後のいずれかが最適文脈文となる割合は日英で 64%，英日で 68%であった．この結果より，2-to-2 翻訳モデルにおいて直前より広い文脈を考慮する方式の有効性が示された．1-to-1 翻訳モデルによる訳文($\boldsymbol{y_{11}}$)，1 文前を文脈文とする 2-to-2 翻訳モデルによる訳文 ($\boldsymbol{y_{22}^{-1}}$)，および，オラクル翻訳の BLEU スコアを図 2 に示す．オラクル翻訳においては，訳文候補の数を増やすほど BLEU スコアが改善した．具体的には，$\boldsymbol{y_{11}}$ および$\boldsymbol{y_{22}^{-1}}$に加えて$\boldsymbol{y_{22}^{-2}, \ldots, y_{22}^{-5}}$，$\boldsymbol{y_{22}^{+1}, \ldots, y_{22}^{+5}}$を訳文候補とすることにより，オラクル翻訳の BLEU スコアは英日で 6 ポイント，日英で 4 ポイント改善した (表 1)．この結果より，2-to-2 翻訳モデルにおいては，より広い範囲を文脈として考慮することで，翻訳精度の上限が改善すると言える．



(a) 日英



(b) 英日

図 2: オラクル翻訳・2-to-2 翻訳(1 文前を文脈文とする)・1-to-1 翻訳の BLEU

2.2.3 最適文脈文選択法

2節では訳文候補 $\boldsymbol{y_{11}}, \boldsymbol{y_{22}^{\pm 1}}, \ldots, \boldsymbol{y_{22}^{\pm 5}}$の中から適切な訳文を選ぶことにより，BLEUスコアが大きく改善可能であることを示した．そこで本節では，この訳文候補の中から最適な訳文を選択する手法について述べる[4][5].

2.2.3.1 強制逆翻訳確率最大化

強制逆翻訳確率は，順方向の翻訳結果を原言語文に翻訳し直すという制約付きで逆方向に翻訳したときの生成確率として定義される．単後 長 n の原言語文を $\boldsymbol{x}$，文脈文を $\boldsymbol{c}$ としたときの 順方向の訳文を $\boldsymbol{y(x,c)}$ とする．このときの原言語文の単語$\boldsymbol{x_j}$ $\boldsymbol{(1 \leq j \leq n)}$ の強制逆翻訳確率は次式となる．

$$\boldsymbol{b_j} = -\log \mathrm{p}\left(\boldsymbol{x_j} \,|\boldsymbol{x_{<j}}, \boldsymbol{y(x,c)}\right)$$

訳文$\boldsymbol{y(x,c)}$の強制逆翻訳確率はb_jの和

$$\boldsymbol{B(x, y(x,c))} = \sum_j \boldsymbol{b_j}$$

となる．強制逆翻訳確率最大化法においては，訳文の強制逆翻訳確率が高いほど意味的に原言語文に近くなるという仮説に基づき，強制逆翻訳確率を最大化する訳文を選択する[6].

2.2.3.2 逆翻訳 sentence-BLEU 最大化

Rappらは逆翻訳後の自動評価結果を尺度として順方向の翻訳結果を評価する手法を提案した[12]．逆翻訳 sentence-BLEU 最大化法においては，この手法に基づき，訳文候補の逆翻訳 $\boldsymbol{x_{11}}, \boldsymbol{x_{22}^{\pm 5}}, \ldots, \boldsymbol{x_{22}^{\pm 5}}$のうち，原言語文 $\boldsymbol{x}$ との間の sentence-BLEU を最大化する訳文を選択する[6].

2.2.4 評価

強制逆翻訳確率最大化，および，逆翻訳の sentence-BLEU 最大化による文脈文選択後の BLEU スコアを図3，図4，および， 表1に示す．文脈文選択の対象となる訳文 候補集合としては，(1) 1-to-1 翻訳$\boldsymbol{y_{11}}$，および，1文前を文脈文とする 2-to-2 翻訳$\boldsymbol{y_{22}^{-1}}$ の組，(2) $\boldsymbol{y_{11}}$，および，1文前から5文前を文脈文とする 2-to-2 翻訳$y_{22}^{-1}, \ldots, y_{22}^{-5}$の集合，(3) y_{11}，および，1文後から5文後を文脈文とする 2-to-2 翻訳$\boldsymbol{y_{22}^{+1}}, \ldots, \boldsymbol{y_{22}^{+5}}$の集合，(4) $\boldsymbol{y_{11}}$，および，前後5文を文脈文とする 2-to-2 翻訳$\boldsymbol{y_{22}^{\pm 1}}, \ldots, \boldsymbol{y_{22}^{\pm 5}}$の集合，の4通りを評価した． 1文前を文脈文とするベースライン 2-to-2 翻 訳$\boldsymbol{y_{22}^{-1}}$ に対して，強制逆翻訳確率最大化によって英日・日英方向で BLEU スコアが有意に改善した．一方，逆翻訳の sentence-BLEU 最大化によって英日方向のみ BLEU スコアが有意に改善した．これらの結果から，強制逆翻訳確率を最大化する文脈文を選択することで翻訳精度が改善することがわかった．しかし，前後5文の範囲でこれらを文脈文とする翻訳候補を追加したとき，オラクル翻訳の BLEU は大きく改善するにも関わらず，強制逆翻訳確率最大化では大きな改善が見られなかった．この結果は提案手法が広い文脈を十分考慮できていないことを示しており，さらなる改善を要する．強制逆翻訳確率最大位置の分布を図5，逆翻訳

---

[4]本節で述べる手法以外の関連研究として，Liらは，順方向翻訳時の生成確率，強制逆翻訳時の生成確率，および，目的言語側の言語モデルによる生成確率，および，訳文長の線形和尺度によって訳文候補をリランキングする手法を提案している [6]．この手法を一部導入した手法によって最適文脈文選択を行った結果のBLEU スコアは，本論文で述べる手法とほぼ同等であった．

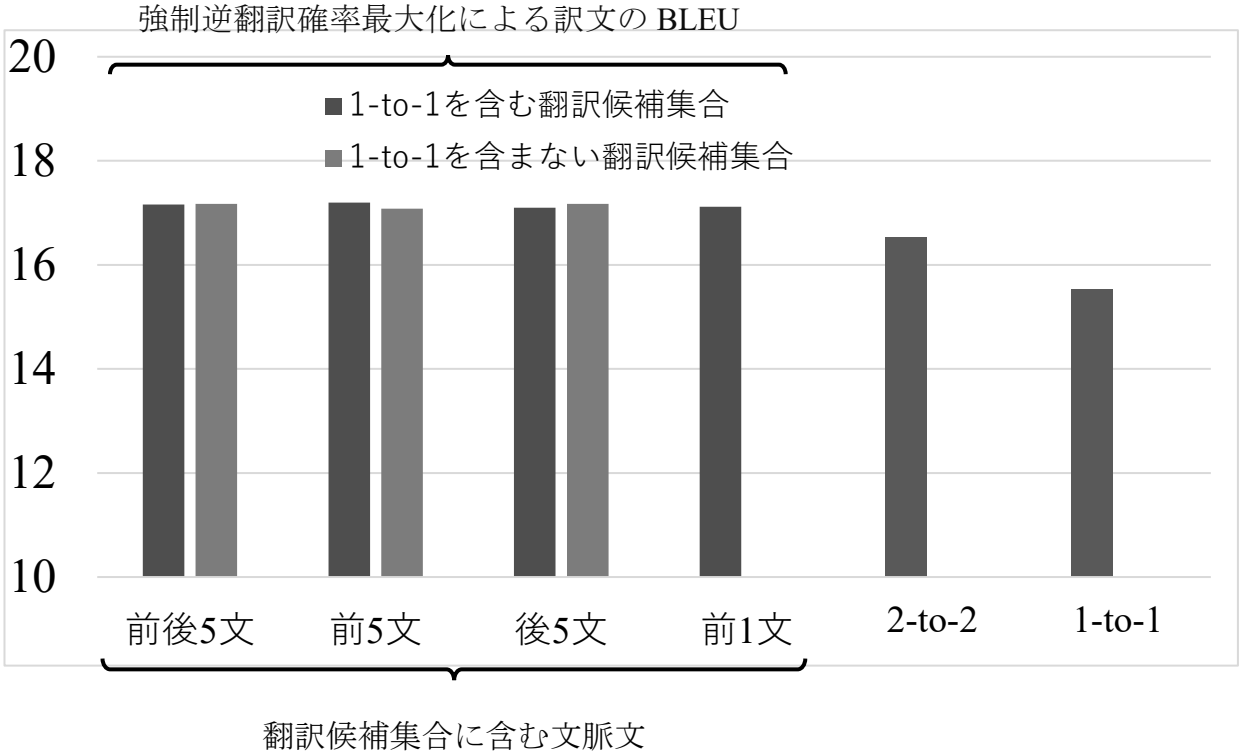
[5]本論文の手法は，文脈考慮型NMT モデルの中でも，1 文前を文脈文とする2-to-2 翻訳モデルよりも広い範囲の文脈を考慮する点においては，[11, 15, 10]に近い試みであると言える．

[6]後藤らは，NMT モデルにおける訳抜け検出において，強制逆翻訳確率比を用いる手法を提案した[3]．本論文は，文脈考慮型NMT において強制逆翻訳確率を用いて最適文脈を選択する点が[3] とは異なる．

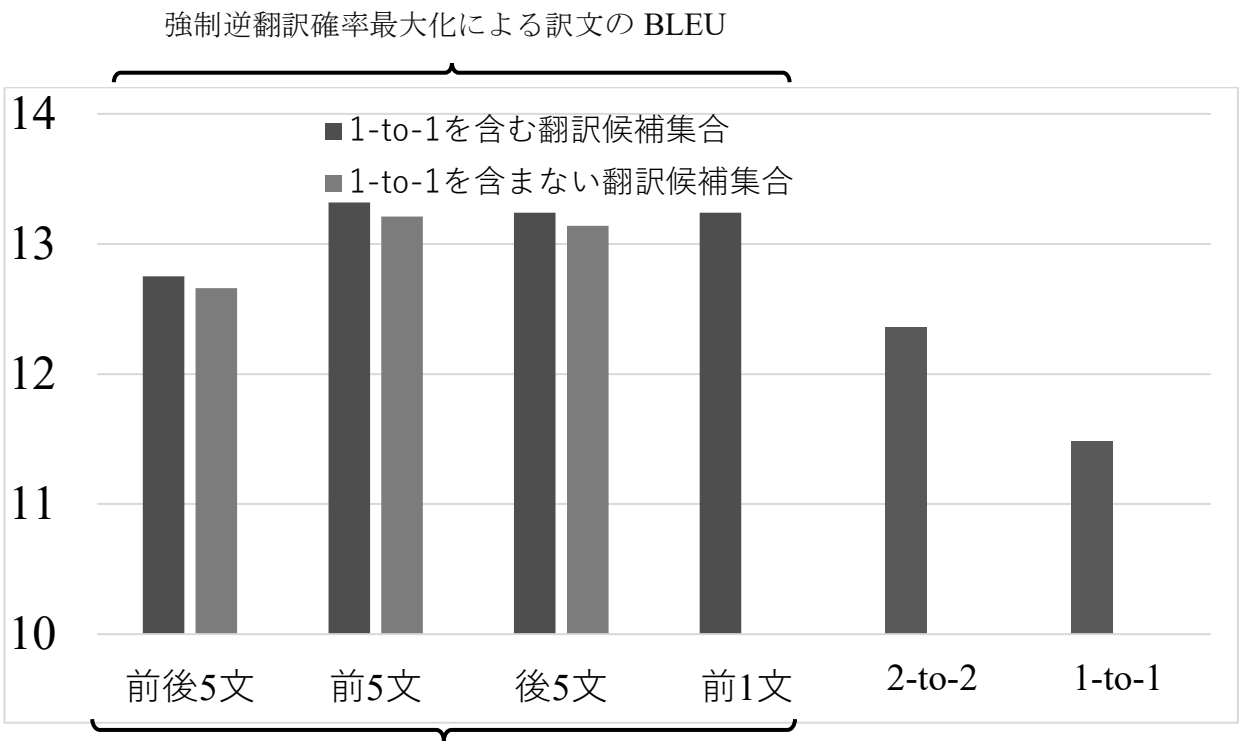
sentence-BLEU 最大位置の分布を図 6 に示す．両者の分布は，図 1 に示したオラクル翻訳の文脈位置と同様の傾向であった．

表 1: BLEU評価(強制逆翻訳確率最大化/ 逆翻訳のsentence-BLEU最大化)(**は1文前を文脈文とするベースライン2-to-2翻訳モデルに対して有意差あり($p < 0.01$)を示す)

訳文候補集合	BLEU		オラクル翻訳 BLEU	
	英日	日英	英日	日英
1-to-1	11.48	15.52	ー	ー
2-to-2 (1 文前)	12.36	16.52	ー	ー
1-to-1 + 2-to-2 (1 文前)	13.24**/12.87**	17.12**/16.61	15.61	18.15
1-to-1 + 2-to-2 (1 文前～5 文前)	13.32**/13.44**	17.20**/16.65	19.55	21.09
1-to-1 + 2-to-2 (1 文後～5 文後)	13.24**/13.20**	17.10**/16.45	19.51	20.84
1-to-1 + 2-to-2 (5 文前～5 文後)	12.75**/13.09**	17.16**/16.50	21.46	22.40

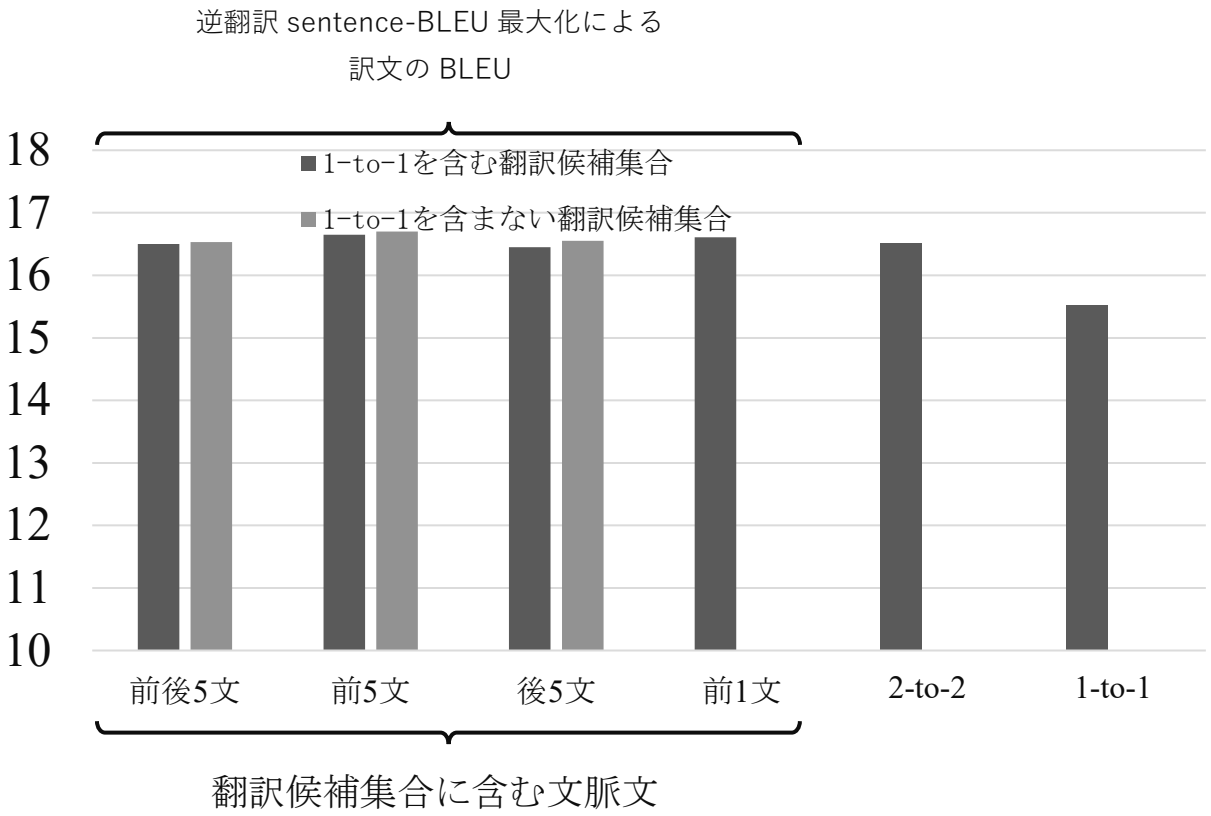


(a) 日英

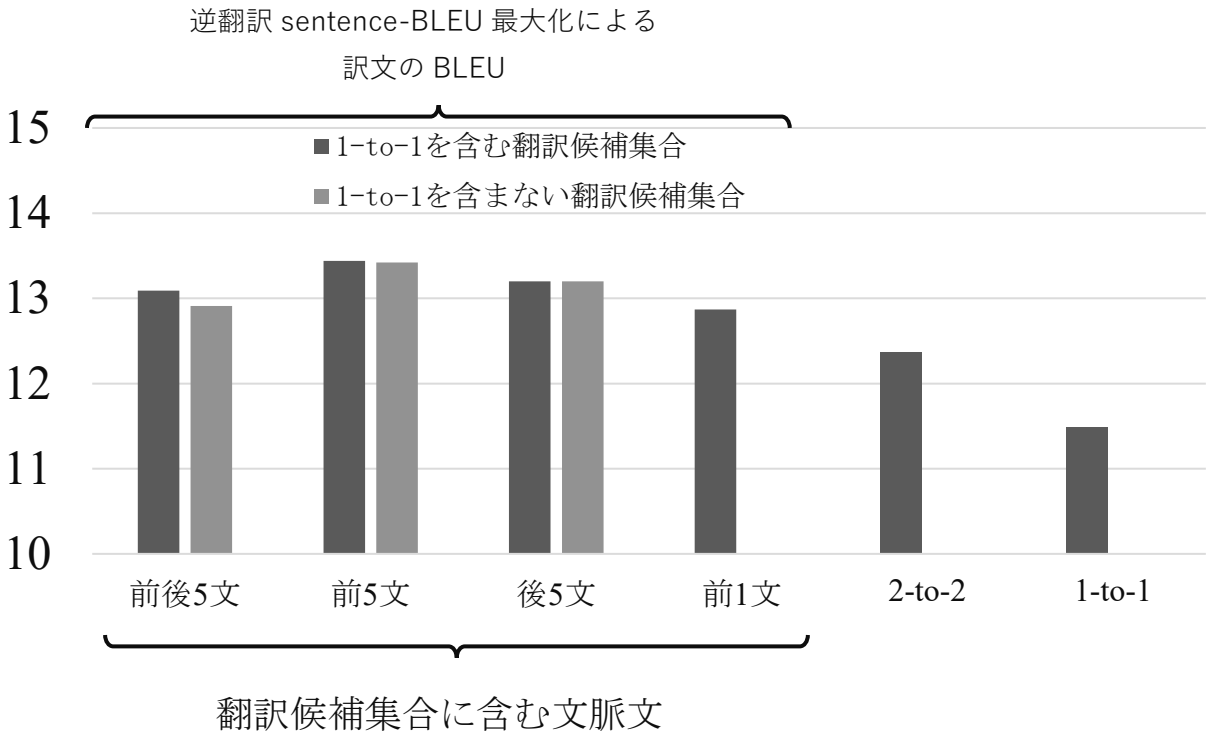


(b) 英日

図 3: 強制逆翻訳確率最大化による訳文・2-to-2 翻訳(1 文前を文脈文とする)・1-to-1 翻訳の BLEU

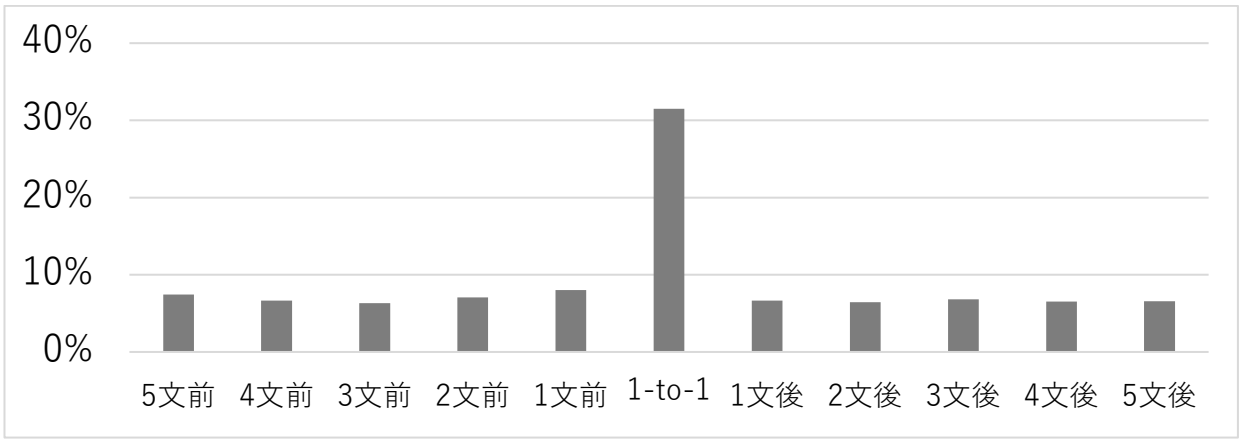


(a) 日英

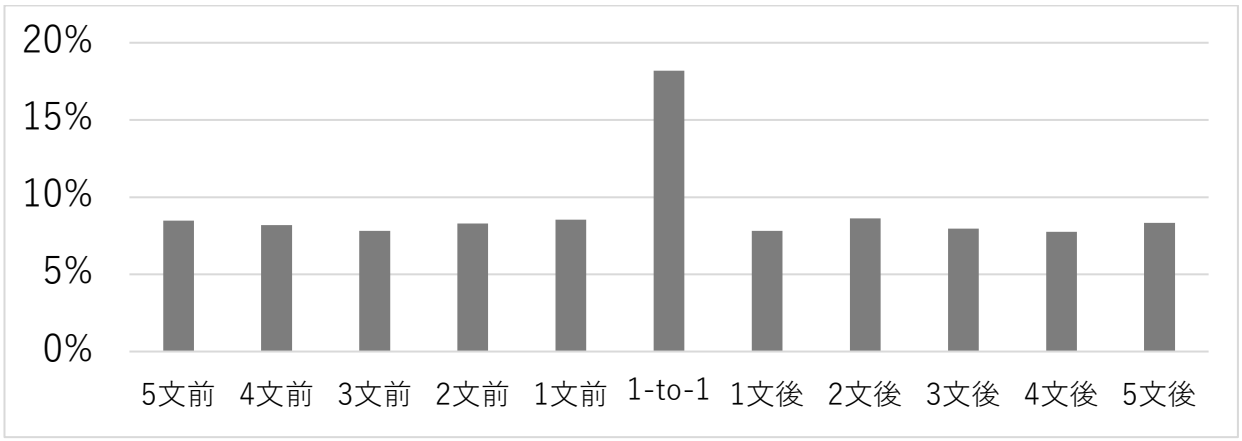


(a) 英日

図 4: 逆翻訳 sentence-BLEU 最大化による訳文・2-to-2 翻訳(1 文前を文脈文とする)・1-to-1 翻訳の BLEU

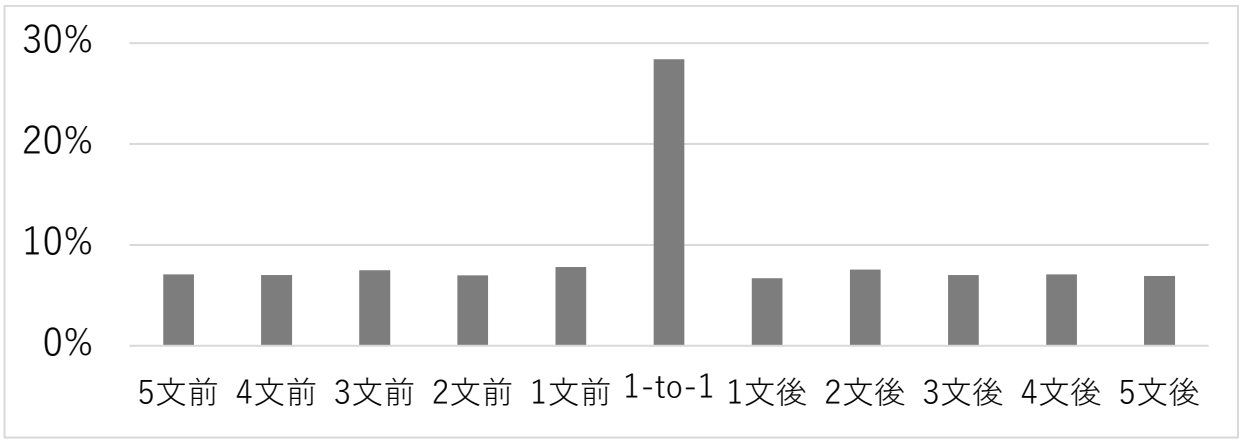


(a) 日英

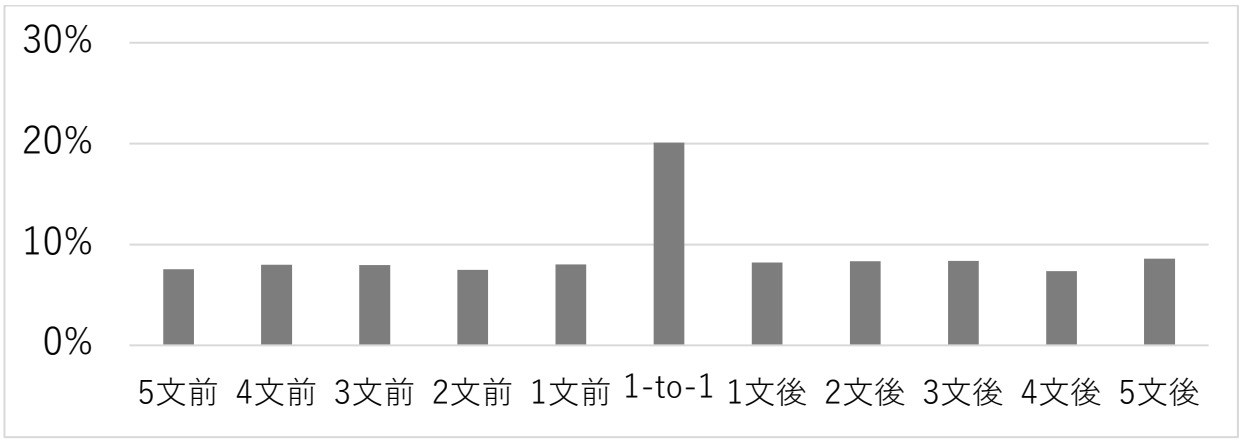


(b) 英日

図 5: 強制逆翻訳確率最大位置の分布



(a) 日英



(b) 英日

図 6: 逆翻訳 sentence-BLEU 最大位置の分布

2.2.5 おわりに

本論文では，2-to-2 翻訳モデルにおいても，文脈文の選択において従来手法よりも広い文脈を考慮することにより BLEU が改善することを明らかにした．また，前後 5 文の範囲で最適な訳文を求める文脈文の選択手法として，強制逆翻訳確率最大化法，および，逆翻訳の sentence-BLEU 最大化法を提案した．1-to-1 翻訳モデル，および， 1 文前を文脈とする 2-to-2 翻訳モデルと比較して，強制逆翻訳確率最大化によって英日・日英方向ともで BLEU スコアが有意に改善した．今後の課題として，原言語文と文脈文との間の意味的つながりの強さを利用する手法として，BERT [2]等の言語モデルの生成確率を組み合わせる手法，および，共参照関係 [5]を文脈文選択に導入する手法を検討する．

参考文献

[1] R. Bawden, R. Sennrich, A. Birch, and B. Haddow. Evaluating discourse phenomena in neural machine translation. In Proc.NAACL-HLT, pp. 1304-1313, 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In CoRR, Vol. abs/1810.04805, 2018.

[3] 後藤功雄, 田中英輝. ニューラル機械翻訳での訳抜けした内容の検出. 自然言語処理, Vol. 25, No. 6, pp. 577-597, 2018.

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

[5] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In Proc. EMNLP, pp. 188-197, 2017.

[6] J. Li and D. Jurafsky. Mutual information and diverse decoding improve neural machine translation. In CoRR, Vol. abs/1601.00372, 2016.

[7] J. Libovick´y and J. Helcl. Attention strategies for multi-source sequence-to-sequence learning. In Proc. 55th ACL, pp. 196-202, 2017.

[8] P. Lison, J. Tiedemann, and M. Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In Proc. 11th LREC, pp. 1742-1748, May 7-12, 2018 2018.

[9] M. Luong, I. Sutskever, O. Vinyals, Q. V. Le, and W. Zaremba. Addressing the rare word problem in neural machine translation. In Proc. 53rd ACL, pp. 11-19, 2015.

[10] S. Maruf and G. Haffari. Document context neural machine translation with memory networks. In Proc. 56th ACL, pp. 1275-1284, 2018.

[11] L. Miculicich, D. Ram, N. Pappas, and J. Henderson. Document-level neural machine translation with hierarchical attention networks. In Proc. EMNLP, pp. 2947-2954, 2018.

[12] R. Rapp. The back-translation score: Automatic mt evaluation at the sentence level without reference translations. In Proc. 47th ACL and 4th IJCNLP, pp. 133-136, 2009.

[13] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural machine translation. In Proc. 27th NIPS, pp.3104-3112, 2014.

[14] J. Tiedemann and Y. Scherrer. Neural machine translation with extended context. In Proc. 3rd DiscoMT, pp. 82-92, 2017.

[15] Z. Tu, Y. Liu, S. Shi, and T. Zhang. Learning to remember translation history with a continuous cache. Transactions of ACL, Vol. 6, pp.407-420, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Proc. 30th NIPS, pp. 5998-6008, 2017.

[17] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov. Context-aware neural machine translation learns anaphora resolution. In Proc.56th ACL, pp. 1264-1274, 2018.

2.3　特許文請求項からの定型パターンの抽出と調査

山形大学名誉教授　横山晶一

2.3.1 はじめに

特許の請求項 1 の文が長くて複雑な構造を持つことはよく知られており、特許文書の解析においてネックになっている[1, 2]。

特許を出願する側も、文レベル、節・句レベル、語レベルなど、さまざまなレベルで、請求項のみならず、特許文書全体を、分かりやすく書く指針を打ち出している[3]。

また、請求項の書き方について、米国特許法のスタイルを取り入れて、個々の構成要件が終了する記号としてセミコロン；や改行を用いたり、欧州や米国特許法の書き方にならって、Jepson claim（公知の事実を請求項の前半部分に記載し、改良点（発明の特徴部分）を後半部分に記載する）の形で記載する提案がなされている。これは、本稿でも述べている「特徴」という用語を構成要件が終了する記号と認識するものである[4]。

統計的機械翻訳(SMT)においては、請求項の構造を sublanguage ととらえて組み込むことによって、翻訳品質が改善されることが報告されている[5]が、意味構造を正確にとらえるためには、なお解析が必要であると考えられる。

また、ニューラルネットを用いた機械翻訳(NMT)では、学習に大量の二言語コーパス(bilingual corpus)を必要とし、そこで得られた翻訳結果の意味づけや、訳語の欠如や誤謬、もともと存在しない情報の挿入などを修正するのが困難であるということが指摘されている。当初の学習時にある程度バイアスのかかったデータを学習に用いることによって、方向性の異なるデータを同時に学習することによる不明確化を避けられるかもしれないという期待がある。実際にうまくゆくかどうかはきちんと確認する必要があるが、やってみる価値はありそうである。ただし、そのためには、データの言語学的な性質を、人間の眼でもう少し深く解明しておく必要がある。

本稿では、請求項の中の定型文として使用できそうなパターンを抜き出して、その可能性を探る。定型的な文のパターンがある程度分かることによって、そのパターンに近いものをバイアス的に学習すれば、ランダムに大量のデータを学習するのに比べて、機械翻訳である程度改善された結果が出てくる可能性があると考えている。すでに、簡単な発表を行っている[2]が、本稿では、その後に追加したデータを加えて新たに考察した結果について述べる。資料として用いた出願特許は 2003 年のもので、いささか古いが、定性的に見て、最近の特許の請求項の書き方と形式や書き方に著しい相違は見られなかったので、この資料を引き続き使用する。

2.3.2 資料・調査方法

2003 年の特許公開情報（特開 2003-180625～180880）の中から、請求項 1 の一文が 120 字を超えるものを 200 文選び出した（このうち 180738 までの 100 文は既報[2]。180739 からの 100 文が今回付け加えたもの）。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	180738 まで	37	34	18	11
	180739 以降	45	21	20	14
	合計	82	55	38	25

この表から分かるように、約 4 割が 200 字未満である。最も短い文は 120 文字、最も長い文は 760 文字、平均で約 261 字であった。この比率や平均値はこれまで調査した別の公開特許情報での分布[1]とほぼ同じである。

調査も、従来の手法を踏襲し、一部新しい知見を加えて、次のような項目について行った。

(a) 機能語「であって、」、「において、」による文の分離

これまでの報告で述べてきたように、特許文の請求項では、「[名詞句 A] [機能語（であって、においてなど）]、…した（する）[名詞句 A]」という形をとるものが非常に多い。しかも、「であって、」や「において、」の部分で、全体を 2 つに区切れる例がほとんどであることが、従来の調査で判明している。これについて、今回も確認した。

(b) 「特徴とする」の調査

上記の名詞句で、最後の部分の直前の「した（する）」が、「特徴とする」、「特徴とした」となっていることが多いこともすでに報告した。これが Jepson claim を反映しているとの発表も確認している[4]。今回もこれに関する調査を行った。

(c) 「と」による並立構造

長い名詞句にかかる修飾句は、名詞を読点で区切って並べる場合も少数見られるが、その多くは、[名詞句]+「と、」という形になる。

(d) 動詞による並列構造

「と、」という助詞ではなく、「～備え、」、「～有する」といった動詞による並列構造も多い。

(e) 改行による係り受けの明確化

主要な修飾句の部分に改行を入れることによって、係り受け関係を明確化した請求項も比較的多く見られる。ここではそれについても調査した。後半の 100 文については、外国特許の影響を調べるために、改行を含む文に関わる出願人の所属国についても調査した。

(f) 「前記」という記述による前の名詞句の省略

「前記」という名詞は、英語の“the”や“its”などに相当し、前の名詞や名詞句を受ける働きをする。「該」も同じような役割を果たす。これについても予備的な調査を行った。

これらの結果について、次節で述べる。

2.3.3 調査結果

(a) 機能語「であって、」、「において、」による文の分離

これまでの報告と同じように、表 1 に示す 200 文について、調査した結果を表 2 に示す。この表で、前、後というのは、前に報告済の 100 文[2]と、今回報告する 100 文との区別である。

ここで、典型というのは、「[名詞句 A] [機能語（であって、においてなど）]、…した（する）[名詞句 A]」の形になっていて、機能語の部分で句が前後に分割されるものを示し、非典型とは、前後の名詞句が修飾語等を含めて少し異なっているものを示す。その他とは、修飾句のもっと下位にあって、そこでは一つ上の名詞に係る形になっているが、句の大きな分割には寄与しないものを示している。

表 2　「であって」と「において」の内訳

		典型	非典型	その他	
				読点あり	読点なし
であって	前	31	4	1	0
	後	20	7	6	0
	計	51	11	7	0
であり	前	1	0	3	1
	後	0	0	8	1
	計	1	0	11	2
において	前	35	3	2	3
	後	30	6	3	15
	計	65	9	5	18

表 2 に示すように、「であって」や「において」を含む句の多くは、その部分の前後で句が大きく分割される。200 文の約 6 割が、このような形で句の長さを短くして分析できる。

ただし、句の分割のしかたはさまざまで、必ずしも前後の句が均等に分かれるわけではない。比較的均等に分かれた例を図 1 に示す。この図では、ほぼ均等に句が分割され、しかもそこに改行も入り、解析しやすくなっている。「特徴とする」という句も語末に入っている（これらについては後述）。ここで示される名詞句も「薬液処理装置」という、比較的

長いものとなっている。

常設型の薬液処理機器と、薬液搬送用の溝のあるカセットと、上記薬液処理機器と上記薬液搬送用の溝との間に配置された弾性マットとからなる薬液処理装置**において、**
上記弾性マットには、上記カセットの上記薬液搬送用溝に沿った経路を持つマット溝が形成されており、
上記マット溝には、該マット溝から上記カセットと向き合う面へ延びるスリットが設けられていることを**特徴とする**薬液処理装置。

図 1　「において」で分離される典型例（特開 2003-180834）

(b) 「特徴とする」の調査

前の報告[1, 2]にも示したが、「特徴とする」を含む例は非常に多い。今回の調査でも、「特徴とする」が前半 77 例、後半 85 例の計 162 例、「特徴を有する」が前半 1 例、「特長とする」が前半と後半それぞれ 1 例の計 165 例あり、そのいずれもが、語末の名詞句の直前に置かれたものであった。これと(a)とのかかわりについては、次節で述べるが、請求項に見られる典型的なパターンを形作るものとして重要である。なお、今回は、以前に示したような「特徴とした」というパターンは見られなかった。図 1 はその典型的な例となっている。(c)に示す図 2 もその例である。

(c) 「と」による並立構造

名詞句を列挙して、「～と、…と、」という形で修飾句を並列させる構造も、請求項に特徴的なパターンである。

液透過性のトップシートと、液不透過性のバックシートと、これら両シートの間に設けられる吸収体とを備え、**前記**吸収体の長手方向両側縁にサイドフラップが設けられた使いすておむつ**において、**
前記サイドフラップの少なくとも一部分に前記使いすておむつの長手方向に沿って連続した凹型変形部が設けられており、
また、**前記**使いすておむつの一方の胴周り領域においては、**前記**使いすておむつが股下領域を介して二つ折りにされ、着用者に装着固定するために、他方の胴周り領域に対する固定手段を有するファスニングテープが設けられ、
前記ファスニングテープの摘持部には**前記**使いすておむつの長手方向に沿って連続した凹型変形部が設けられていることを**特徴とする**使いすておむつ。

図 2　並立助詞等を含む例(特開 2003-180752)

今回調査した中でも、6 割近くが、助詞「と」を繰り返すことによって、修飾句の並列を形成している。図 2 は、「において、」、「と、」による並立、動詞による並列、「前記」、「特徴とする」を含む典型例であり、改行も含んだ長い句を示している。

(d) 動詞による並列構造

動詞は、以前の報告で述べた「備える」や「有する」といった例（活用して、「備え」、「備えた」、「有し」といった形で（読点とともに）用いられる場合も多い）もあるが、サ変動詞の並列など、さまざまなバリエーションがある。今回は、あまり動詞の形を特定せずに、動詞一般として調査を行った。全体の約半数に、この形が見られる。つまり、おおざっぱに言って、助詞、動詞両方の並列的な構造を含むものが 2〜3 割近くを占めていると言える。図 2 もその一例である。

(e) 改行による係り受けの明確化

請求項 1 の長い句の中に改行を入れている例は比較的多く、今回も前半 64 例、後半 51 例、計 115 例見られた。図 1 のように、主要な区切りのところに 1 か所だけ改行を入れているものと、図 2 のように、修飾句ごとに区切りを入れているものとがあるが、全体的には後者の方が多い。数は少ないが、あまり的確でない場所に改行を入れ、かえって解析の妨げになるものも見られる。

改行について、米国特許のセミコロン方式を踏襲（日本の特許でセミコロンを用いているものは極めて少ない）して、改行につなげている[4]可能性を探る予備的な調査として、後半の 51 文の特許について、出願人を調査した。その結果、日本人による出願が 44 件と大半を占め、残りの 7 件の内訳は、米 5 件、英 1 件、独 1 件であった。ここでは、特に海外からの特許申請が改行につながっている傾向は見られなかった。

(f) 「前記」という記述による前の名詞句の省略

「前記」という語によって、前の句とのつながりが明確になる場合が多い（「上記」という例も少数見られる）。今回は、前半部の約半数の 51 例、後半部の 6 割に当たる 64 例、計 115 例でこの記述が見られた。図 2 は比較的明確な例である。「サイドフラップ」、「使い捨ておむつ」といった語句が、前とのつながりを明確化する形で記述されている。ただし、修飾句のどのレベルでこれが使われているかは、句内部の意味記述を明らかにしないと解析できない場合も多く、今後の課題である。

(g) その他

「前記」と似ているが、「該」という語を含む例も比較的多く見られた。後半部の 100 特許の中の 31 特許（1 文中の重複も含む）に、この語が含まれていた。この語は、「前記」よ

りも曖昧な使われ方をする場合が多く、今後さらに検討が必要である（図 5 に例を示す)。

箇条書きも、しばしば見られるが、今回調査した範囲には前後半ともそれぞれ 2, 3 例しかなく、十分な知見が得られていない。また、「とともに」（「と共に」と記述される場合もある）などの機能語についても、今回は十分には調査していない。今後これらにも手を広げる予定である。

2.3.4 機能語と「特徴」との組み合わせパターン

表 3　機能語と「特徴」の組み合わせパターンのまとめ

	であって		において		であり		特徴
	典	非	典	非	典	非	含まず
前	20	3	32	2	1	0	21
後	17	6	28	6	0	0	14
計	37	9	60	8	1	0	35

被施療者にマッサージを施すための施療子と、該施療子をマッサージ動作させる駆動手段と、該駆動手段を介して上記施療子の動作をマッサージ指令に応じて制御する制御手段とを備えたマッサージ機**において、**
上記制御手段は、上記施療子の位置に対する制御である位置制御と上記施療子の被施療者への押圧力に対する制御である力制御とを切り換えて上記制御を行うものであることを**特徴とする**マッサージ機。

図 3　典型的なパターンの一例(特開 2003-180771)

皮膚面に接触又は近接させて用いられる放電管と、
前記放電管の周辺にオゾンが生成されるように前記放電管に高電圧を印加し放電を起こさせる高電圧印加手段と、
前記放電管に対し着脱自在に形成され、前記高電圧印加手段によって放電させた前記放電管の表面に帯電する電荷を集中させる電荷集中部を備えたアタッチメントとを具備することを**特徴とする**トリートメント装置。

図 4　機能語を含まない例(特開 2003-180783)

前節に述べたいくつかのパターンを組み合わせると、請求項に特有のパターンが浮かび上がってくる。

表 3 は、機能語「であって」、「において」と「特徴とする」を組み合わせた時の結果を示している。「であって」、「典」とは、

(a) [名詞句 A] であって、～を特徴とする[名詞句 A]

を示し、「において」、「典」のカラムは、

(b) [名詞句 A] において、～を特徴とする[名詞句 A]

を示している。(a)には、「特徴とする」の代わりに「特徴を有する」という形が前半部に 1、(b)には「特長とする」の 1 を前半部に加えてある。(a)の例を図 3 に、(b)の例は、図 1, 2 に示してある。後半部の「特長とする」は、機能語を伴わない例なので、この表には記載されていない。

200 の例を調査しただけで、全体像を示すのはやや乱暴であるが、請求項の約半数が、この形のパターンを持っていることが期待できる。

このような機能語がない例を図 4 に示す。「特徴とする」は、ここで調査した限りでは、必ず最後の名詞句の直前に位置しているので、機能語が無い場合でも、大きな構造を捉えるのに役立つ可能性がある。

図 5 には、「であって、」、「特徴とする」を含む非典型例を示す。「特徴とする」の後に、「吸収性物品の」という修飾語が入って、最初の名詞句とのマッチングが取れなくなっている。この例では、「該」という語も多く用いられている。

面ファスナーフック部材に係合する不織布製ターゲットテープ**であって、**
該不織布製ターゲットテープが、多数の熱融着性短繊維からなるウエッブに、高温エアーを吹き付けて該繊維同士を高密度に交絡させて熱融着してなるエアスルー不織布であり、
該不織布の坪量が１７～５０ｇ／m<SP>２</SP>であり、
該不織布の短繊維の繊度が０．５～８デニールであることを**特徴とする**吸収性物品の不織布製ターゲットテープ。

図 5　「であって、」と「特徴とする」を含む非典型例（特開 2003-180748）

2.3.5 問題点と今後の検討

NMT 全盛の時代に、文の構造を調査するのは、ある意味無駄な作業に思えるかもしれない。しかしながら、照応など、一文を超える現象などに関しては（最近は簡単な照応現象をニューラルネットで学習する試みも行われているが)、まだ、ある程度の基本調査が必要と考えられる。

理想的には、たとえば、昔話の解析のように、いくつかの定型的なあらすじをさらに一般化するところまで広がって行ければ面白い[6]が、特許文の請求項に対してこのような試

みが成功するかどうかは不明である。

今後は、さらにデータを増加させるとともに、ここでは触れられなかった諸現象にも眼を配り、できる限り談話的な部分を含んだ一般化を進めていきたいと考えている。また、さまざまな現象がどのように組み合わさっているかについてもさらに研究を進めたい。

また、系統的に調査したわけではないが、アメリカやドイツの特許の請求項（英：claim、独：Anspruch）もやはり長い名詞句であることが多いようである（たとえば、ドイツ語では、関係副詞“wobei”（英語 wherein、その場合）などを使って、長々とした名詞句にする例などが見られる）。今後これら外国特許とも比較してみたい。

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。公開特許には、類似のものを同時期にまとめて出願するために、ある程度の偏りも見られるので、幅広い調査を行った上で知見を深める必要がある。

動詞については、「～を備え、」というパターンは確認できたが、その他の動詞において、明確なパターンがあるかどうかの確認が必要である。今後さらにデータを増加させて調査する予定である。

また、前回調査の対象とした「該」、「前記」といった照応的な語については、今回も、さまざまなバリエーションがあって、十分な調査が行えなかった。これらについても今後さらに検討していく予定である。

出願人や出願国についても、今後調査を進めていく予定であるが、現時点ではどの程度特許内容と関連するか不明である。

参考文献

[1] 横山晶一：パターンを用いた特許文請求項の構造解析、平成 29 年度 AAMT/Japio 特許翻訳研究会報告書 (2017) pp.32-37,
http://aamtjapio.com/kenkyu/files/kenkyu05/AAMT_Japio_20180423.pdf

[2] 横山晶一：特許文請求項からの定型パターンの抽出、Japio Yearbook (2018) pp.296-299
http://www.japio.or.jp/00yearbook/files/2018book/18_4_07.pdf

[3] (財)日本特許情報機構 特許情報研究所：特許ライティングマニュアル（第 2 版）「産業日本語」(2018)

[4] 小池誠：請求項の記載を構造化する提案、第 5 回特許情報シンポジウム(2018) pp.81-88
http://aamtjapio.com/kenkyu/files/symposium2018/AAMT_symposium_20181207.pdf

[5] 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 (2016) pp.407-435

[6] マックス・リュティ、小澤俊夫訳：ヨーロッパの昔話、岩波文庫(2017)

3. 機械翻訳評価手法

3.1 拡大評価部会の活動概要

奈良先端科学技術大学院大学　須藤　克仁

AAMT/Japio 特許翻訳研究会では、2012 年度より機械翻訳の評価に関する議論を深めるための下部組織として「拡大評価部会」を設置し、研究会委員以外の識者も加えた活動を展開している。

拡大評価部会では、翻訳評価を様々な視点から捉えるため、以下の３つのサブグループに分かれ個別の課題についての検討を行い、部会会合において部会全体で議論している。

・自動評価サブグループ（越前谷 博、須藤 克仁）
　自動評価尺度の改善の検討、およびメタ評価に関する議論を行う
・テストセットサブグループ（江原 暉将、長瀬 友樹、王 向莉）
　機械翻訳における典型的な課題を含む評価用データセットの設計・作成・評価を行う
・人手評価サブグループ（中澤 敏明、園尾 聡、後藤 功雄）
　種々の人手評価方法の検討、および実際の評価データの分析・議論を行う

2018 年度は 3 回の部会会合を開催した。

・2018 年 5 月 11 日　　年度活動計画の策定
・2018 年 10 月 12 日　中間報告および今後の活動内容についての議論
・2019 年 3 月 8 日　　年度活動報告および報告書内容の確認

自動評価サブグループは、昨年度検討した単語分散表現に基づく手法を、チャンクに基づく手法と組み合わせる拡張について検討を行い、さらに人手評価との相関を改善できることを示した。

テストセットサブグループでは、昨年度までに構築した中日評価テストセットを利用した評価サイトを利用した評価結果の分析を行うとともに、機械翻訳の文法能力評価に着目した新しいテストセットの設計・構築を行った。

人手評価サブグループでは、WAT2018 の評価結果の分析を行い、人手評価による機械翻訳の優劣が自動評価では十分判別できないレベルまで機械翻訳精度が向上していることを示唆する結果が得られた。

なお、7 年に渡り続けてきた拡大評価部会の活動は、2018 年度をもって休止とし、翻訳評価に関する課題の検討は研究会の活動に戻す形で 2019 年度以降継続する予定である。これまで拡大評価部会の活動にご協力いただいた部会員の皆様に、特に部会長として発足にご尽力くださった江原暉将先生、翻訳評価に人一倍強い情熱を持って部会長を務めてくださった磯﨑秀樹先生（故人）に御礼申し上げる。

3.2 MT の文法能力を評価するための中日文テストセットの設計と構築

（株）ディープランゲージ　王　向莉

3.2.1 はじめに

機械翻訳(MT)の対象となる言語対は、文法の近い同語族言語対（例えば日本語と韓国語）と文法の大きく異なる異語族言語対（例えば日本語と中国語）に分けられる。人間翻訳と同じであるが、同語族言語対 MT と異語族言語対 MT は難易度が大きく異なる。同語族言語対 MT は、翻訳対象となる源言語と目的言語は文法が似ているため、単語さえ正しく対応すればよい。その一方、異語族言語対を機械翻訳する場合、単語レベルだけでなく、言語間の文法上の違いをうまく対処しなければならない。

従来の MT 評価手法においては、難易度の違う同語族言語対と異語族言語対の MT を分けずに、同じ尺度または基準で評価するのが一般的であった。その一方、王と辻井（2017）が、異語族言語対を対象とする MT は、翻訳を仕事とする人間と同じように、言語対の文法上の違いを対処する能力が重要で、それを評価する手法が不可欠であると指摘した。具体的には、１）同語族言語対 MT と異語族言語対 MT を分けて評価する；２）異語族言語対 MT を文法と語彙の 2 次元に分けて評価する；３）異語族言語対 MT の文法能力を評価すべきである；４）MT の文法能力を評価するために、中国語深層理解のための文法資源 CSSG（Chinese Sentence Structure Grammar）[2]を活用することを提案した。

今年度の評価部会では、「MT の文法能力を評価する」というアイディア[1]に基づいて、科学技術文献の中日 MT（特にニューラルネットワーク MT）を一層有効に評価することを目的として、中国語深層理解のための文法資源 CSSG を活用することで、中日 MT 評価用テストセットを設計し、最初版を構築した。本稿では、それについて報告する。

3.2.2 CSSG の視点による中国語と日本語の文法上の違い

CSSG は Sentence Structure Grammar (SSG)という形式文法枠組みにおけるアイディアに基づいて設計・構築した中国語深層理解のための文法知識資源である[2]。CSSG では、文は述語動詞とその深層格要素および話者の主観的な表現等からなるものとみなす。ここでは、中国語文と日本語文における深層格要素、述語動詞、主観的な表現を比較し、それぞれの違いをまとめる。

3.2.2.1 深層格要素と格助詞

・日本語文でも中国語文でも深層格要素に格助詞が付くことがある

言語	例文
日本語	シャーレ内の濾紙の上**に**種子**を**放置する
中国語	**将**种子放置**在**浅底盘内的滤纸上

例えば上記の例では、中国語の「将」と「在」は日本語の「を」と「に」に対応する格助詞である。格助詞「将」は深層格要素「种子（種子）」に、格助詞「在」は深層格要素「浅底盘内的滤纸

上（シャーレ内の濾紙の上)」に付く。

・日本語文では、格助詞が深層格要素の後ろにあるが、中国語文の場合では、格助詞が深層格の前に置かれる

言語	例文
日本語	シャーレ内の濾紙の上**に**
中国語	**在**浅底盘内的滤纸上

この例では、中国語の「在」は日本語の「に」と対応する格助詞だが、深層格要素「浅底盘内的滤纸上（シャーレ内の濾紙の上)」の直前に位置する。

・日本語文では、格助詞が深層格要素に付くが、中国語文では、格助詞が必ずしも深層格要素に付くとは限らない

言語	例文
日本語	**シャーレ内の濾紙の上に**置いた
中国語	放置**在**了**浅底盘内的滤纸上**

上記の例のように、中国語格助詞「在」と深層格要素「浅底盘内的滤纸上（シャーレ内の濾紙の上)」の間に「了」という単語がある。

・日本語文では、深層格要素に格助詞が付くのが普通であるが、中国語文では、深層格要素は位置が重要で、それに必ずしも格助詞がつくとは限らない

言語	例文
日本語	棒状着磁磁石と超電導体の間**に**３ｍｍの厚さのスペーサ**を**置く
中国語	棒状磁化磁铁与超导体之间放置３ｍｍ厚的隔板

この例の場合、中国語文では、日本語の格助詞「に」と「を」に対応する格助詞がない。

・日本語文では、深層格要素の位置が比較的自由であるが、中国語文では、深層格要素の位置における制約が厳しい

言語	例文	正誤
日本語	棒状着磁磁石と超電導体の間**に**３ｍｍの厚さのスペーサ**を**置く	○
中国語	棒状磁化磁铁与超导体之间放置３ｍｍ厚的隔板	○

例えば上記の例では、中国語文も日本語文も文法は正しいが、下記の例のように深層格の位置が変わるとそうでなくなる。下記の例では、日本語文における「３ｍｍの厚さのスペーサ」という深層格要素とそれに付く格助詞「を」は文頭に移動しても文法的正しいが、中国語文の場合、それに対応する深層格要素「３ｍｍ厚的隔板（３ｍｍの厚さのスペーサ)」は文頭に移動すると、文法の正しくない非文になってしまう。

言語	例文	正誤
日本語	**３ｍｍの厚さのスペーサを**棒状着磁磁石と超電導体の間に置く	○
中国語	**３ｍｍ厚的隔板**放置棒状磁化磁铁与超导体之间	×

・深層格要素の位置は格助詞がそれに付くことで換えられる

言語	例文	正誤
日本語	**３ｍｍの厚さのスペーサを**棒状着磁磁石と超電導体の間**に**置く	○
中国語	**【把】**３ｍｍ厚的隔板放置**【在】**棒状磁化磁铁与超导体之间	○

上記の例では、深層格要素「３ｍｍ厚的隔板（３ｍｍの厚さのスペーサ)」と「棒状磁化磁铁与超导体之间（棒状着磁磁石と超電導体の間)」にそれぞれ格助詞「把」と「在」が付くと、正しい文として成り立つ。

3.2.2.2　述語動詞

・日本語文では、述語動詞は文末に置かれるが、中国語文では、述語動詞は文中に置かれるのが普通である

言語	例文
日本語	シャーレ内の濾紙の上に種子を**放置する**
中国語	将种子**放置**在浅底盘内的滤纸上

例えば、上記の例では、日本語の述語動詞「放置する」は文末に置くが、中国語のそれに対応する述語動詞「放置」は文中に位置する。

・日本語文では、述語動詞には語尾変化があるが、中国語文では、述語動詞に語尾変化がない

言語	例文
日本語	種子はシャーレ内の濾紙の上に**放置される**
中国語	将种子**放置**在浅底盘内的滤纸上

例えば、上記の例に示すように、受動文の場合では、日本語の述語動詞「放置する」は「放置される」に変化するが、中国語の述語動詞は「放置」のままであり、述語動詞に語尾変化がない。

3.2.2.3　主観的な表現

・日本語文では、主観的な表現は述語動詞の後ろに現れるが、中国語文では、主観的な表現は一般的に述語動詞の前に置かれる

言語	例文
日本語	シャーレ内の濾紙の上に種子を放置する**かもしれない**
中国語	**也许**将其放置在浅底盘内的滤纸上

この例では、日本語の「かもしれない」も中国語の「也许」も「推測」という主観的な表現を表す単語であるが、「かもしれない」が述語動詞の後ろにあるのに対して、「也许」は述語動詞の前に置かれる。

・日本語文と中国語文においては、主観的な表現の順序が違う

言語	例文	正誤
日本語	シャーレ内の濾紙の上に種子を放置し**ないかもしれない**	○
中国語	**也许没**将种子放置在浅底盘内的滤纸上	○

例えば、上記の例のように、「否定」と「推測」との 2 つの主観的な表現は、日本語文では、「否

定（ない)」、「推測（かもしれない)」という順序となるが、中国語文では、「推測（也许)」、「否定（没)」という順序になる。下記の例のようにその順序を変えると、日本語文も中国語文も文法が正しくなくなり、非文になってしまう。。

言語	例文	正誤
日本語	シャーレ内の濾紙の上に種子を放置し**かもしれないない**	×
中国語	**没也许**将种子放置在浅底盘内的滤纸上	×

3.2.3 MT の文法能力を評価するための中日文テストセット

MT 研究はルールベース MT 時代、統計 MT 時代を経て、近年、ニューラルネットワーク MT が急速な進展を遂げている。時代が変わったとしても、日本語と中国語のような異語族言語対を翻訳対象とする MT にとっては、言語対の文法上の違いを対処する力、すなわち文法能力が不可欠である。文法能力は MT 評価の重要な尺度として取り扱うべきだと考えられる。評価部会では、中日科学技術文献を翻訳する MT の文法能力を評価するために、CSSG を活用することで、中日文テストセットを設計し、その最初版を構築した。

3.2.3.1 中日文テストセットの設計

１）評価用文型の選定

文法能力は源言語側における様々な文型を目的言語に正しく翻訳する能力であると考えられる。CSSG では、中国語述語動詞がその深層構文特徴によって分類され、各種類の動詞が取れる様々な文型は網羅的に取集されている。CSSG から典型的な動詞種類を選び、動詞種類ごとから代表的な動詞を取り出し、それが取れる様々な文型から科学技術文献で頻繁に用いられるものを取り出して、中日 MT の文法能力を評価すると考えている。

２）文型における評価ポイント

CSSG 文型は述語動詞、深層格要素、主観的な表現からなるので、評価は表１に示す評価ポイントに着目して評価を行う。

表１：テストセットにおける評価ポイント

番号	評価ポイント	評価点数
1	述語動詞は正しく訳されるか	3
2	述語動詞の態は正しく訳されるか	2
3	各深層格要素は正しく訳されるか	2
4	各深層格要素に付いてある格助詞は正しく訳されるか	2
5	主観的な表現は正しく訳されるか	2

3.2.3.2 中日文テストセットの構築

中日文テストセットは下記の６つのステップで構築される。表 2 に完成したテストセット文の具体例を示す。

1）CSSG から異なる深層文法特徴を持つ述語動詞グループを５つ選ぶ
2）述語動詞グループごとから、代表的な動詞を１つ選ぶ
3）各動詞について科学技術文献で頻繁に表れる深層文型を５つ選ぶ
4）「日中論文抜粋コーパス（ASPEC-JC)」という既存日中対訳文コーパスから、文型ごとに様々な表層表現文（基本的には 5 文ずつ）とそれに対応する日本語文を収集する
5）中国語文と日本語文に深層文型情報を付ける
6）各文に対して評価ポイントに評価点数を付ける

表 2：中日文テストセットの具体例

中国語文	日本語訳文	中国語文型情報	日本語文型情報	評価ポイント	総評価点数
棒状磁化磁铁与超导体之间放置了３ｍｍ厚的隔板。	棒状着磁磁石と超電導体の間に３ｍｍの厚さのスペーサを置いた。	[棒状磁化磁铁与超导体之间]Loc[放置]V [了]past[３ｍｍ厚的隔板]Obj。	[棒状着磁磁石と超電導体の間]Loc[に]mark_Loc[３ｍｍの厚さのスペーサ]Obj[を]mark_Obj[置い]V[た]past。	・述語動詞 V :3 点 ・動詞の時制 past:2 点 ・Object 格 Obj :2 点 ・格助詞 mark_obj:2 点 ・Location 格 Loc :2 点 ・格助詞 mark_loc:2 点	13 点

3.2.3.3　テストセットの構成

「日中論文抜粋コーパス（ASPEC-JC)」は科学技術振興機構（JST）の所有物である「アジア学術論文抜粋コーパス（ASPEC)」[3]の一部で、日本語学術論文を人手で中国語に翻訳したものである。ASPEC-JC は表３のように 4 つの部分からなるが、訓練データの部分だけを利用した。ASPEC-JC の訓練データから、選ばれた文型ごとにおける様々な表層表現文（基本的には 5 文ずつ）とそれに対応する日本語文を選んで、テストセット文対とした。

表 3：ASPEC-JC の中身

データ種類	ファイル名	文対数
訓練データ	train.txt	672,315
開発データ	dev.txt	2,090
開発実験データ	devtest.txt	2,148
実験データ	test.txt	2,107

ASPEC-JC から 125 文対を収集する予定であったが、動詞「降低/低下する」の５つの文型のうち、1 つについては対応する文は 4 つしか見つからなかった。ASPEC-JC は約 67 万文対あるが、カバーされていない CSSG 文型が多数存在していると考えられる。また、選ばれた文対における日本語文をチェックしたところ、ふさわしくない文対が 16 対見つかり、このような文対を取り除いて、最終的に 108 の文対を集め、テストセットの文対として用いた（表 4)。

表 4：中日文テストセットの中身

動詞番号	動詞	文型数	文対数
1	放置/放置する	5	24
2	収集/集める	5	22
3	提供/提供する	5	20
4	降低/低下する	5	21
5	添加/添加する	5	21
合計	5	25	108

テストセットとしてふさわしくない文対には、下記のようなものがある。

・中国語文が意訳されたものは最も多い

言語	例文
日本語	それら知識ベースの構築は規模の大きさに重点が置かれ,
中国語	这些知识库的构筑的重点被放置在规模大小上,

例えば、この例では、日本語文における主題「それら知識ベースの構築」と主語「重点」を一つの名詞句「这些知识库的构筑的重点（それら知識ベースの構築の重点)」として訳される。

・誤訳が入っている文対もある

言語	例文
日本語	水素ガスに変換し燃料電池コジェネレーションに供給して発電する方式
中国語	生成的氢气提供给燃料电池联合发电，进行发电的方式

この例では、「燃料電池コジェネレーション」の翻訳が難しいためか、中国語訳文には「燃料电池联合发电，进行发电」のような誤りがあった。

・日本語文側に理解しにくいところがある文対もある

言語	例文
日本語	実油汚染土壌（約１２，０００−１７，０００ｍｇ／ｋｇ程度）を油分濃度約２，４００−3，３００ｍｇ／ｋｇ程度まで，低減できる
中国語	可以将实际油污染土壤（约１２０００－１７０００ｍｇ／ｋｇ程度）的油分浓度约降低到约２４００－３３００ｍｇ／ｋｇ左右

上記の例の日本語文は、「実油汚染土壌の油分濃度（約１２，０００−１７，０００ｍｇ／ｋｇ程度）を約2，４００−3，３００ｍｇ／ｋｇ程度まで，低減できる」のように修正したほうが分かりやすいと思われる。

3.2.4 まとめ

本稿では、MT 評価用中日文テストセットの設計と構築についてまとめた。本テストセットは「MT の文法能力を評価する」というアイディアに基づいて、CSSG という中国語深層理解のた

めの文法資源を活用することより、設計・構築したものである。構築した中日文テストセットが科学技術文献における中日 MT の文法能力を有効に評価することが期待される。このテストセットをどのように生かすのか、その有効性をどのように検証するのかなどは、今後の課題となる。

参考文献

[1] 王向莉, 辻井潤一. 2017. MT の文法能力を評価する. AAMT/Japio 特許翻訳研究会評価部会内部発表資料.

[2] Xiangli Wang, Yi Zhang, Yusuke Miyao, Takuya Matsuzaki, Junichi Tsujii. 2013. Deep Context-Free Grammar for Chinese with Broad-Coverage. Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), pages 11–19.

[3] http://orchid.kuee.kyoto-u.ac.jp/ASPEC/

3.3 中日テストセットを用いた特許文献の翻訳評価

－拡充されたテストセットの評価サイトへの組み込み－

元・山梨英和大学　江原 暉将
(株)富士通研究所　長瀬 友樹

3.3.1　はじめに

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

筆者らは、中国語特許文献の中日機械翻訳評価のためにテストセットの構築を行い、昨年度までに以下のことを実施した[4)5)6)7)8)]。

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

今年度は、昨年度までに拡充されたテストセットを評価サイトに組み込み、あわせていくつかの機械翻訳システムを評価した。

3.3.2　テストセットの拡充

これまでに収集・拡充した中日特許文テストセットの規模を表1に示す。2014年度と2015年度は連続パターンを収集し、2016年度と2017年度は分離パターンを収集した。また、2017年度は発明の名称部分の拡充も行った。

表1　中日特許文テストセットの規模の推移

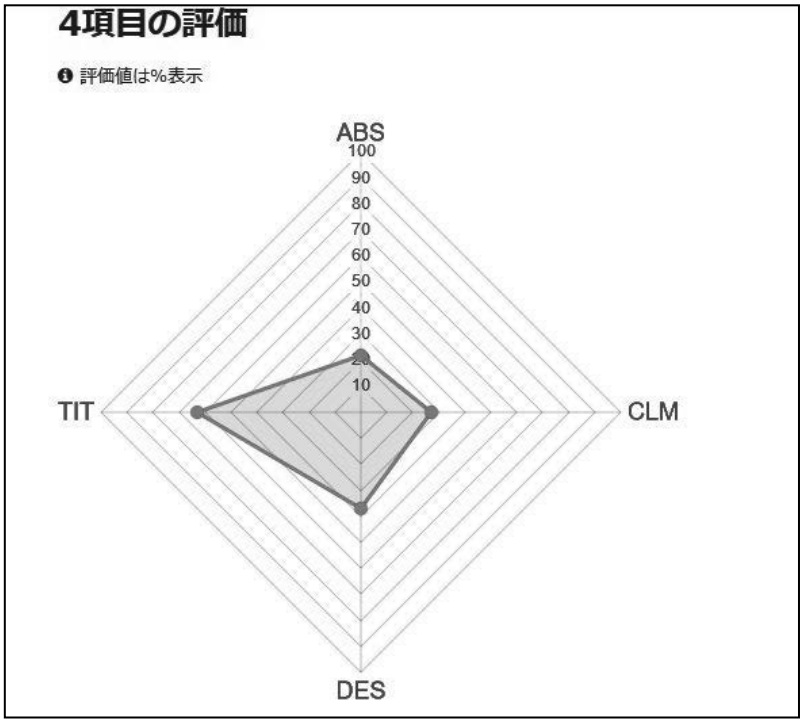
年度	テスト文数				
	発明の名称	要約	請求範囲	詳細説明	合計
2014	3	38	22	276	339
2015	4	67	53	720	844
2016	4	70	54	936	1064
2017	22	71	55	1003	1151

2018年度はテストセットの拡充を行うことはしなかった。また、日本語設問の追加も行わなかった。

[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。

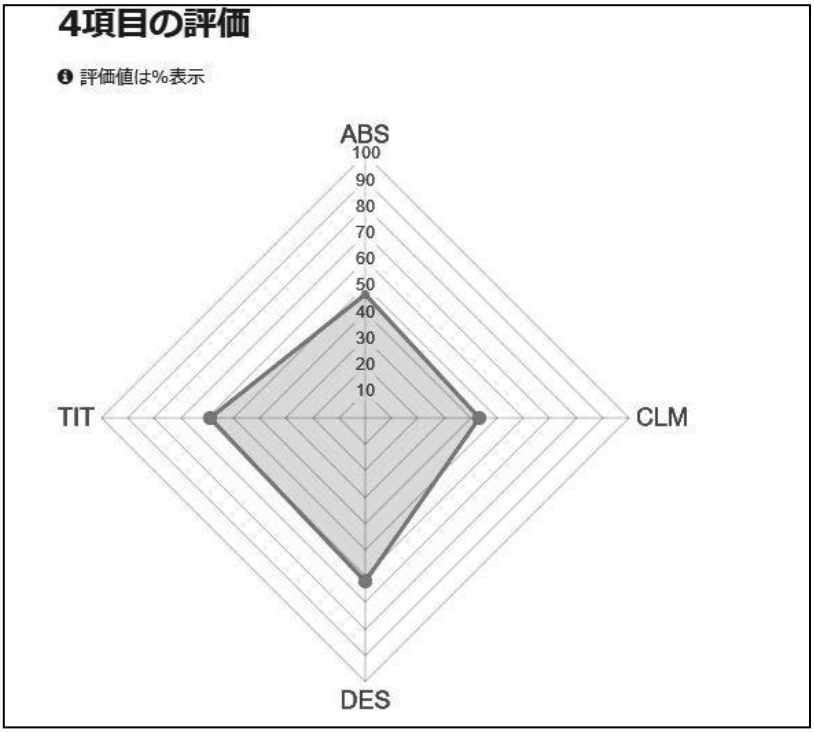
3.3.3 AAMT 自動評価サイトでの試験

2017 年度までに構築したテストセットを AAMT 自動評価サイトに組み込んだ。そのサイトを用いて、いくつかの翻訳システムについて評価を行った。結果を図 1 に示す。RBMT、SMT、オンラインサイト(2019 年 2 月 17 日アクセス)、NMT での翻訳である [2]。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。評価結果は、評価値の高い順に NMT、オンラインサイト、SMT、RBMT の順である。NMT とオンラインサイトは出典によって評価値に差異がある。TIT と DES は NMT が高く、ABS と CLM はオンラインサイトの方が高い。

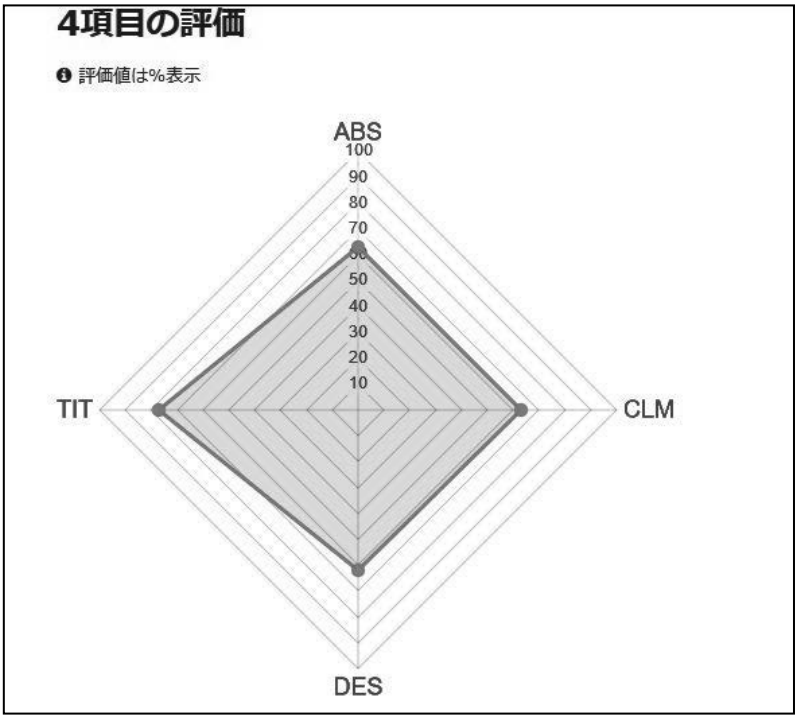


(a)　　RBMT

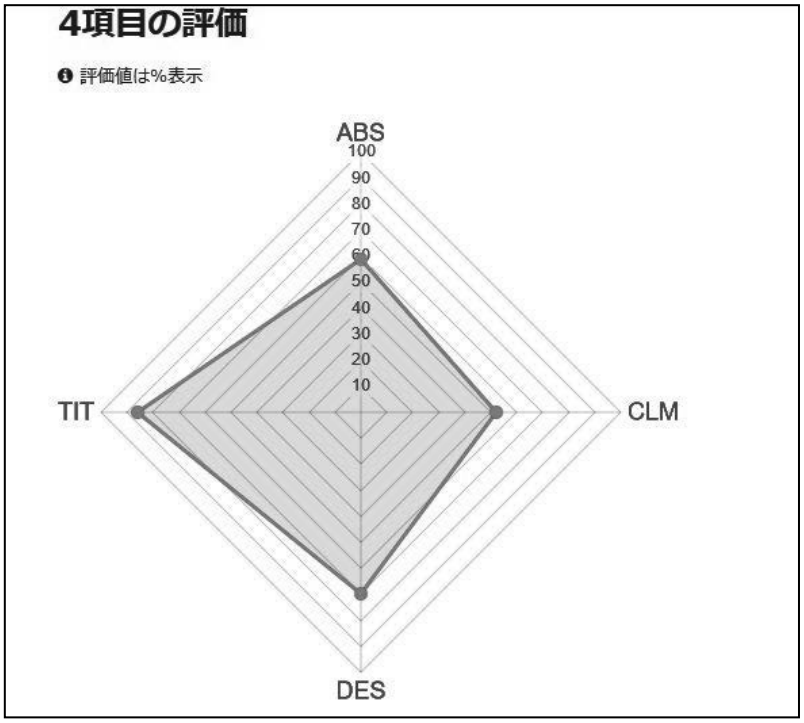
[2] RBMT は市販の機械翻訳ツール、SMT は WAT2016 の文献 9)のシステム、NMT は WAT2018 の文献 10)のシステムである。システムの選択は非網羅的であり、評価結果に一般性はない。



(b) SMT



(c) online site



(d)　　NMT

図 1 テストセットを用いた評価結果例

図 1 に示す評価結果例を出典別にまとめると表 2 が得られる。表 2 には、TIT、ABS、CLM、DES の各評価値を設問数で重みをつけて平均した値として Test set score も示してある。さらに、システムおよび試験データは異なるが WAT での Pairwise score を類似したシステムから求め[3]、表 2 に示した。図 2 には、Test set score と Pairwise score との散布図および線形回帰直線と R^2 の値を示した。R^2 は 0.9915 と高く、システム数が 4 と少ないものの、Test set score と Pairwise score には強い相関がみられる。

表 2　テストセットを用いた評価結果のまとめ

	RBMT	SMT	online	NMT
TIT	0.65	0.62	0.79	0.89
ABS	0.23	0.48	0.64	0.61
CLM	0.28	0.44	0.63	0.53
DES	0.39	0.62	0.62	0.71
Test set score	0.380	0.603	0.625	0.699
Pairwise score	-40	39	54.25	69.75

[3] Pairwise score の値は、WAT の中日特許翻訳タスク(JPCzh-ja)から求めた。具体的には RBMT は WAT2015[11]の RBMT A の値、SMT は WAT2016[12]の EHR 1 の値、online は WAT2017[13]の ONLINE A の値、NMT は WAT2017[13]の EHR 2 の値を用いた。WAT2018 ではベースラインが変更になったので単純には比較ができなかった。

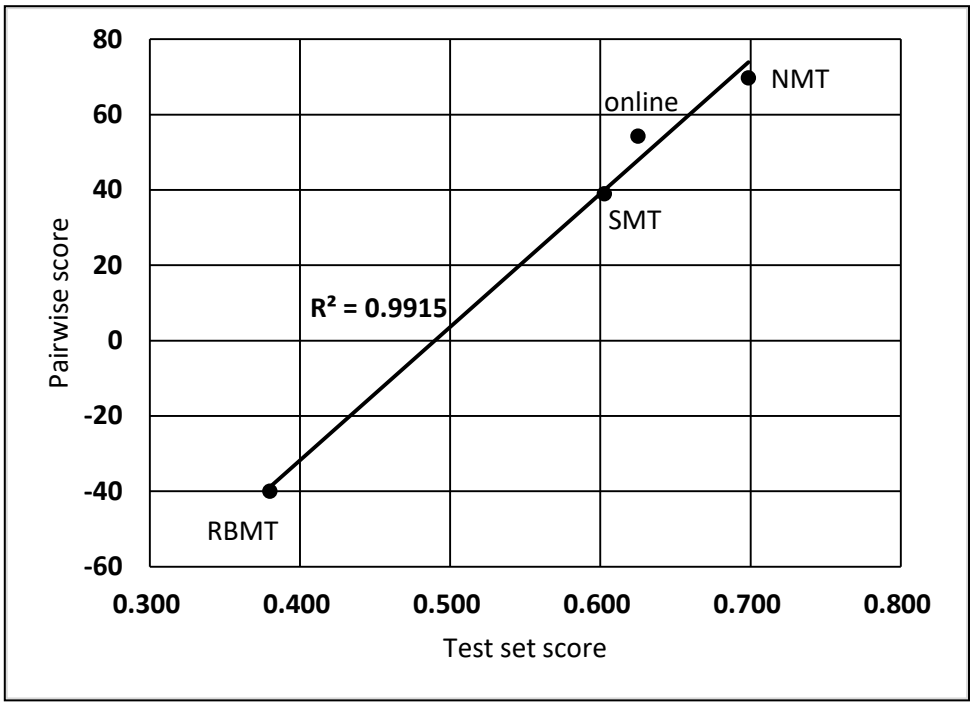


図 2　Test set score と Pairwise score の関係

3.3.4　まとめと今後の課題

拡充したテストセットを評価サイトに組み込んでいくつかのシステムについて評価を実施した。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) 江原暉将, 長瀬友樹, 王向莉. 2017. 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―, *平成 28 年度 AAMT/Japio 特許翻訳研究会報告書,* pages 62-66.

8) 江原暉将, 長瀬友樹, 王向莉. 2018. 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拡充および評価の実施―, *平成 29 年度 AAMT/Japio 特許翻訳研究会報告書,* pages 56-60.

9) 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.

10) Terumasa Ehara. 2018. SMT reranked NMT (2), *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*.

11) Toshiaki Nakazawa et. al., 2015. Overview of the 2nd Workshop on Asian Translation, *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 1-28.

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

13) Toshiaki Nakazawa et. al., 2017. Overview of the 4th Workshop on Asian Translation, *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1-54.

3.4 単語の位置情報と単語分散表現を用いたEarth Mover's Distanceに基づく自動評価法とチャンクに基づく自動評価法の組み合わせ

北海学園大学　越前谷　博

3.4.1　はじめに

ニューラル機械翻訳[1][2]は統計的機械翻訳との比較においてより良い翻訳文を出力することが知られている。特に流暢さ（fluency）においては高品質な翻訳が可能とされている。しかし、ニューラル機械翻訳では訳抜けが起こりやすいという問題点が指摘されている[3]。そして、この訳抜けを含む翻訳文に対して BLEU[4]や NIST[5]等の従来の自動評価法は十分に対応しているとはいえない。訳抜けのある翻訳文と参照訳においては、単語そのものが翻訳文に存在していないため単語間のミスマッチとはならない。したがって、訳抜けはスコアに反映されにくいと考えられる。BLEU では翻訳文と参照訳との間の構成単語数の違いに基づくペナルティにより訳抜けを含む翻訳文の評価に対応しているが、そこでは個々の単語間の意味的な近さや位置情報を考慮していないため不十分である。

そこで、本報告では、単語間の意味的な近さと位置情報を用いた Earth Mover's Distance[6][7]に基づく自動評価法として MEMD（**M**etric based on **E**arth **M**over's **D**istance）を提案する。さらにこの MEMD と従来より著者が提案しているチャンクに基づく自動評価法である IMPACT[8]を組み合わせた自動評価法 MEMD$_{\text{COMB}}$ を提案する。IMPACT は文単位において人手評価との間で比較的高い相関が得られることから、文単位での評価精度の向上が期待できる。The 4th Workshop on Asia Translation （WAT2017）[9]の日英方向と英日方向のデータを用いた性能評価実験より、提案手法 MEMD 及び MEMD$_{\text{COMB}}$ が人手評価との間で最も高い相関を示すことを確認した。

3.4.2 単語の位置情報と単語分散表現を用いたEarth Mover's Distanceに基づく自動評価法：MEMD

3.4.2.1 Earth Mover's Distanceの利用

提案手法MEMDは二つの分布間の距離を求めるための尺度であるEarth Mover's Distance(以降、EMD と記す）に基づいている。EMD を用いて距離を得るためには個々のシグネチャーを構成する特徴量と重みを定義する必要がある。そして、シグネチャー間の距離を計算するための距離式を定義する必要がある。EMD に基づく MEMD においてはこれら 3 つのパラメータを自然言語文に適合することでスコアを得る。MEMD ではシグネチャーを文中の単語とし、特徴量には単語分散表現、重みには文レベルの *tf・idf* を用いる。そして、距離計算にはコサイン類似度を使用する。さらに、MEMD では翻訳文と参照訳の単語間の位置情報を距離計算に反映させることでより高い精度でのスコア計算を実現する。

3.4.2.2 単語の重み付け

提案手法 MEMD のパラメータの一つであるシグネチャーにおける重みには *tf・idf* を用いる。*tf・idf* は以下の式（1）より得られる。

$$tf \cdot idf = tf \times \left(\log \frac{N}{df} + 1.0 \right) \qquad (1)$$

式（1）の *tf* は任意の単語が文中に出現する回数である。*N* は全文数、*df* は任意の単語が出現する文数である。式（1）より、様々な文に出現する助詞などは *tf・idf* の値が低くなり、特定の文にのみ出現する名詞などは *tf・idf* の値が高くなると考えられる。したがって、式（1）より機能語と内容語の差別化が可能になると考えられる。

また、式（1）を MEMD に適用する際には文中の全単語の *tf・idf* の総和が 1.0 になるように正規化を行う。

3.4.2.3 単語アライメント

EMD では全シグネチャー間の距離を要素とした距離行列を生成することで分布間の距離を得る。それに対して提案手法 MEMD では全単語間のコサイン距離と単語の位置情報を用いて距離を計算し、距離行列を生成する。また、距離行列を生成するにあたり、前処理として単語アライメントを行う。これは単語アライメントの結果に基づいて対応関係にある単語間の相対位置のずれを決定し、それを単語の位置情報とするためである。

単語アライメントは翻訳文を基準として、翻訳文の先頭の単語から対応する参照訳中の単語をコサイン類似度に基づいて決定する。単語アライメントの具体例を図 1 に示す。図 1 は翻訳文として“A Japanese person went to British Museum.”、参照訳として“A Japanese man went to the British Museum.”を用いた単語アライメントの例である。

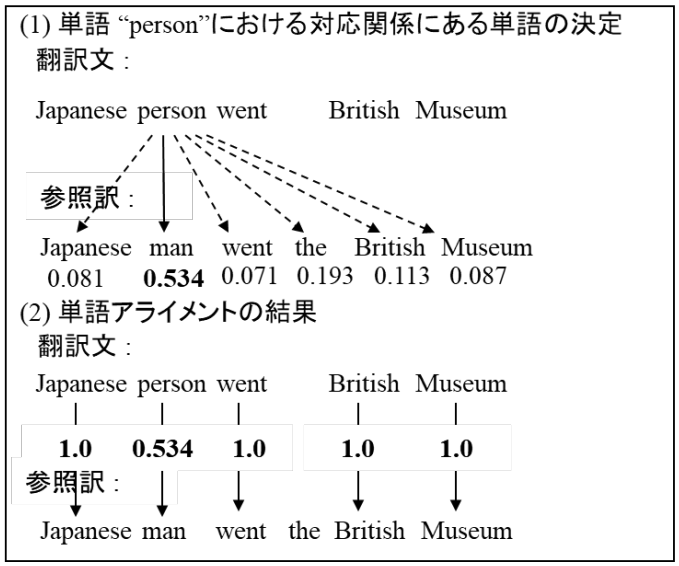


図 1　単語アライメントの具体例

図中では単語“A”、“to”、そして、“.”が省略されているが、これは単語分散表現のモデルの中にストップワードが含まれていない単語分散表現のモデルを用いた場合の例となっているためである。図１の（１）は翻訳文中の単語“person”に対応する参照訳中の単語を求める処理を示している。まず、翻訳文中の単語“person”と参照訳文中のすべての単語間のコサイン距離を求める。そして、得られたコサイン類似度が最も高い単語を対応関係にある単語と位置付ける。この場合、“man”とのコサイン類似度が 0.534 と最も高い値を示したため、“person”と“man”に対応関係があると決定される。このような処理を翻訳文中のすべての単語について行うことで図１の（２）に示すような単語アライメントの結果が得られる。ここで、参照訳文中の単語との間で最も高い類似度が複数得られた場合には、対応関係が一意に決定できないとして単語間の対応付けは行わない。また、図１では参照訳中の“the”は対応する単語が存在しない。これは翻訳文中のいずれの単語との間においてもコサイン類似度が低かったため単語間の対応付けは行われなかったためである。

3.4.2.4 距離計算

提案手法 MEMD では、距離行列を生成するために用いる単語間の距離計算を **3.3.2.3** で述べた単語アライメントの結果に基づいて行う。自然言語文に EMD を適用する場合には語順は重要であるとの観点より、単語アライメントの結果に基づいて単語間の相対位置のずれを距離計算に反映させる。距離の計算式を以下の式（２）と式（３）に示す。

$$d = \begin{cases} 1.0 - cosine \times pos\_info \\ (\text{対応関係がありかつ} wd_t = wd_r) \\ 1.0 - cosine^2 \times pos\_info \\ (\text{対応関係はあるが} wd_t \neq wd_r) \\ 1.0 \\ (\text{対応関係なし}) \end{cases} \qquad (2)$$

$$pos\_diff = 1.0 - \left| \frac{pos(wd_t)}{len(t)} - \frac{pos(wd_r)}{len(r)} \right| \qquad (3)$$

式（２）では、単語アライメントにより対応関係が存在し、かつ翻訳文中の単語 wd_t と参照訳中の単語 wd_r が表層レベルでも一致する場合、*1.0-cosine*×*pos_info* を距離計算式とする。ここで式（２）の *pos_info* は式（３）より得られ、翻訳文中の単語 wd_t と参照訳中の単語 wd_r の相対位置のずれを示している。$pos(wd_t)$ と $pos(wd_r)$ はそれぞれ翻訳文中と参照訳中における単語 wd_t と wd_r の出現位置である。また、*len(t)* と *len(r)* はそれぞれ翻訳文と参照訳の構成単語数を示している。式（３）より相対位置のずれが大きいほど *pos_diff* の値は小さくなる。そのため、コサイン類似度 *cosine* に対して大きなペナルティとなる。逆に相対位置のずれが小さい場合、*pos_diff* の値は大きくなるため小さなペナルティとなる。また、式（２）では、単語アライメントにより対応関係はあるが翻訳文中の単語 wd_t と参照訳中の単語 wd_r が表層レベルでは一致しない場合には

コサイン距離 *cosine* を 2 乗することで距離 *d* を大きな値にする。これは単語分散表現においては“man”と“woman”のような対義語のコサイン類似度が高くなるという問題を解消するためである。例えば、図 1 の参照訳に対して翻訳文として“A woman went British Museum”が存在した場合、参照訳文中の“man”と翻訳文中の“woman”のコサイン類似度が高くなってしまい、結果として高い評価スコアが出力されることになる、そこで、式（2）の距離計算式 $1.0\text{-}cosine^2 \times pos\_info$ を用いることでコサイン類似度は小さくなり、その結果、距離 *d* が大きくなる。したがって、対義語によるコサイン類似度の影響を小さくすることができると考えられる。

式（2）と式（3）を用いて生成される翻訳行列の具体例を表 1 に示す。表 1 の例は図 1 で示した翻訳文と参照訳との間で求めた翻訳行列である。

表 1　翻訳行列の具体例

		参照訳					
		Japanese	man	went	the	British	Museum
翻訳文	Japanese	**0.028**	1.0	1.0	1.0	1.0	1.0
	person	1.0	**0.727**	1.0	1.0	1.0	1.0
	went	1.0	1.0	**0.056**	1.0	1.0	1.0
	British	1.0	1.0	1.0	1.0	**0.028**	1.0
	Museum	1.0	1.0	1.0	1.0	1.0	**0.014**

表 1 の太字の値は単語アライメントにより対応関係にある単語間の距離である。ここでは具体的な計算例として、翻訳文の単語“person”と参照訳の単語“man”との間の距離 0.727 について述べる。図 1 に示すようにコサイン類似度として 0.534 が得られた。単語“person”と単語“man”は表層レベルで一致しないため式（2）の $1.0\text{-}cosine^2 \times pos\_info$ が使用される。また、式（3）の *pos_diff* については、*pos(”person”)*と *pos(“man”)*が共に 3 であり、*len(t)*と *len(r)*はそれぞれ 8 と 9 である。したがって、*pos_diff* の値は 0.958(=1.0-|3/8-3/9|)となる。その結果、式（2）より *d* の値として 0.727(=1.0-0.534²×0.958)が得られる。このような処理により翻訳行列を求めることで個々の単語に着目したより精度の高い評価スコアが得られると考えられる。

3.4.2.5 MEMD スコア

提案手法 MEMD のスコアは最終的に以下の式（4）より得られる。

$$MEMD = 1.0 - EMD \qquad (4)$$

EMD は距離を求めるための尺度であり、2 つの分布が類似しているほど値は小さくなる。自動評価法ではスコアの観点より類似しているほどスコアが高くなる方が直感的にもわかりやすいと考えられる。そのため、**3.3.2.1** から **3.3.2.3** の処理で得られる EMD の値をそのまま MEMD スコアとするのではなく、1.0 から EMD の値を引くことで、類似しているほど MEMD スコアが高くなるように変換する。

3.4.2.6 MEMD とチャンクに基づく自動評価法の組み合わせ

本報告では提案手法 MEMD と他の自動評価法を組み合わせることで文レベルの評価精度の向上を図る。自動評価法においてはシステムレベルに比べて文レベルの相関係数が低いことが問題となっている。そこで、MEMD との組み合わせに適した自動評価法を決定するために、NTCIR-7[10]における日英方向のデータより MEMD と他の自動評価法を用いて文レベルの相関係数を求めた。表 2 に NTCIR-7 の日英方向における文レベルの相関係数を示す。

表 2　NTCIR-7 の日英方向における文レベルの相関係数

自動評価法	Kendall		
	adequacy	fluency	Avg.
IMPACT	0.429	0.433	<u>0.431</u>
BLEU	0.333	0.354	0.344
RIBES[11]	0.370	0.341	0.356
WMD[12]	0.206	0.249	0.228
METEOR[13]	0.366	0.384	0.375
MEND	0.413	0.428	0.421

表 2 において、下線は adequacy と fluency の相関係数の平均が最も高かった相関係数を示している。表 2 より IMPACT が最も高い相関係数を示していることが確認できる。上述したように自動評価法においては文レベルの相関係数の低さが問題となっているため、MEMD との組み合わせに用いる自動評価法として文レベルの相関係数が最も高かった IMPACT を使用することとする。IMPACT は翻訳文と参照訳間の表層レベルの共通単語列を最長共通部分列（LCS）に基づいて求めることでスコアを得る、チャンクに基づく自動評価法である。本報告では、以下に示す式（5）を用いて MEMD と IMPACT のスコアを組み合わせる。

$$MEMD_{COMB} = \frac{\alpha \times MEMD + \beta \times IMPACT}{\alpha + \beta} \qquad (5)$$

式（5）の *MEMD* と *IMPACT* は文レベルのスコアを示している。また、α と β は *MEMD* と *IMPACT* の重みである。したがって、式（5）は MEMD と IMPACT の加重平均を求める式となっている。α と β のペア(α, β)が（1, 1）の場合には相加平均となる。

この α と β の最適な値は NTCIR-7 データを用いた相関係数を得ることで決定した。具体的には α と β のペアを（1, 1）、（1, 2）、（1, 3）、（1, 4）、（1, 5）、（1, 9）、（2, 1）、（3, 1）、（4, 1）そして、（9, 1）の 10 通りの組み合わせを用いて式（5）よりスコアを求め、相関係数の推移を調査した。その結果、英日方向のシステムレベルは、どの組み合わせにおいても IMPACT よりも低い相関係数であった。また、日英方向のシステムレベルでは MEMD を超える相関係数は得られなかったが、IMPACT よりもすべての組み合わせで高い相関係数を示した。したがって、日英方向において最も高い Pearson の相関係数 0.892 を示したペア（2, 1）、（3, 1）、そして、（4, 1）を α と β の組み合わせに用いることとした。また、文レベルの $MEMD_{COMB}$ の相関係数の推移を図 2

に示す。

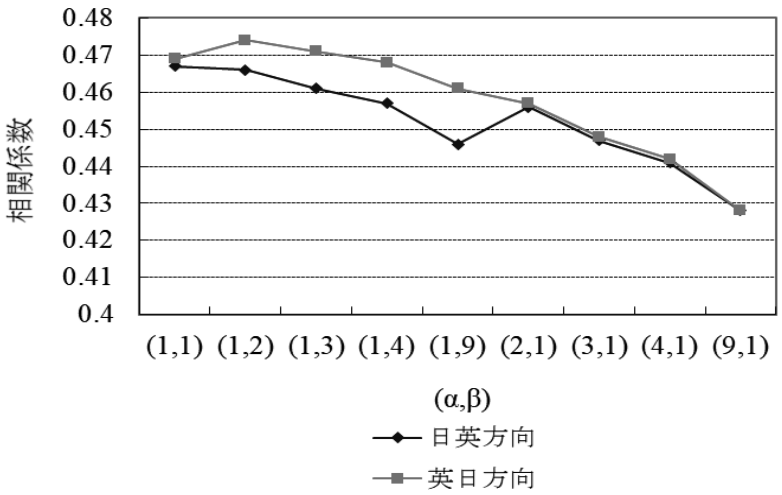


図２　NTCIR-7 データにおける $MEMD_{COMB}$ の文レベルの相関係数の推移

図２において、日英方向では（1, 1）の組み合わせの相関係数が最も高く、英日方向では（1, 2）の組み合わせの相関係数が最も高かった。そこで、日英方向と英日方向の平均を求めた。その結果、（1, 1）の組み合わせでは 0.468、（1, 2）の組み合わせでは 0.470 となった。したがって、文レベルにおいては α と β の組み合わせとして日英方向と英日方向の平均が最も高かった（1, 2）を用いることとした。本報告では α と β の組み合わせを反映させた $MEMD_{COMB}$ を $MEMD_{COMB(\alpha,\ \beta)}$ のように表記することとする。

3.4.3 性能評価実験

3.4.3.1 実験データ

本報告では、提案手法 MEND と $MEMD_{COMB}$ の有効性を検証するために the 4th Workshop on Asia Translation (WAT 2017)の英日方向と日英方向の翻訳文、参照訳、そして、人手評価を用いた。WAT2017 では、日英及び英日方向のデータとして the Asian Scientific Paper Except Corpus (ASPEC)、the JIJI Corpus、the Japan Patent Corpus (JPC)、そして、the Recipe Corpus が含まれている。また、人手評価においては、ペアワイズと adequacy の 2 つの評価方法が含まれている。ペアワイズの評価では個々の翻訳文に対して 5 名の評価者が win、loss、そして、tie のいずれかを付与している。具体的には、ベースラインより良ければ１、ベースラインより悪ければ -１、そして、ベースラインと同じであれば 0 が付与されている。一方、adequacy の評価においては １から５までのスコアが付与されている。

3.4.3.2 実験方法

本報告では提案手法を含めた複数の自動評価法によるメタ評価を実施した。使用した自動評価法は BLEU、RIBES、IMPACT、WMD、そして、提案手法 MEMD と $MEMD_{COMB}$ である。また、$MEMD_{COMB}$ については、システムレベルには $MEMD_{COMB(2,\ 1)}$、$MEMD_{COMB(3,\ 1)}$、そして、

$MEMD_{COMB(4,\ 1)}$、文レベルには $MEMD_{COMB(1,\ 2)}$を用いた。翻訳文と参照訳に対しては、英語は tokenizer.perl[14]と lowercase.perl[14]による前処理を行った。また、日本語は MeCab[15]を用いて分かち書きを行った。そして、提案手法 MEMD と $MEMD_{COMB}$、そして、WMD においては単語分散表現を得るために word2vec[16]モデルを用いた。このモデルについては、英語は GoogleNews-vectors-negative300.bin（300 次元、語彙数：3,000,000）、日本語は日本語エンティティの entity_vector.model.bin（200 次元、語彙数：1,015,474）をそれぞれ用いた。GoogleNews-vectors-negative300.bin は新聞記事を学習データとしており、ストップワードは含んでいない。entity_vector.model.bin は日本語 Wikipedia を学習データとしており、ストップワードも含んでいる。

また、自動評価法に対するメタ評価はシステムレベルでは Pearson の相関係数を用いて人手評価との相関を得た。文レベルでは Kendall の相関係数を用いて人手評価との相関を得た。

3.4.4.3 実験結果

表３にメタ評価結果として WAT2017 データを用いたシステムレベルでの adequacy における相関係数を示す。また、表４と表５には文レベルでの adequacy とペアワイズにおけるそれぞれの相関係数を示す。表３から表５まで自動評価法は全データの相関係数の平均である“Avg.”の高い順にソートされている。

表３　WAT2017 におけるシステムレベルでの adequacy の評価による Pearson の相関係数

自動評価法	ASPEC	JIJI	JPC	RECIPE			Avg.
				ING	STE	TTL	
$MEMD_{COMB(4,\ 1)}$	0.052	-0.996	0.855	1.0	1.0	1.0	0.485
MEMD	-0.137	-0.981	0.890	1.0	1.0	1.0	0.462
RIBES	0.483	-1.0	0.790	1.0	1.0	0.0	0.379
BLEU	0.494	-0.996	0.706	1.0	1.0	0.0	0.367
IMPACT	0.384	-0.998	0.769	1.0	1.0	0.0	0.359
WMD	0.394	-1.0	0.748	1.0	1.0	0.0	0.357
$MEMD_{COMB(2,\ 1)}$	0.181	-0.998	0.835	1.0	1.0	0.0	0.336
$MEMD_{COMB(3,\ 1)}$	0.105	-0.997	0.847	1.0	1.0	0.0	0.326

表４　　WAT2017 における文レベルでの adequacy の評価による Kendall の相関係数

自動評価法	ASPEC	JIJI	JPC	RECIPE			Avg.
				ING	STE	TTL	
$MEMD_{COMB(1, 2)}$	**0.371**	0.134	**0.268**	0.416	**0.446**	**0.382**	0.336
IMPACT	**0.358**	0.079	0.256	0.430	**0.438**	**0.381**	0.324
MEMD	0.303	0.220	0.225	0.420	**0.386**	0.372	0.321
sentBLEU	0.306	0.051	0.226	0.502	0.385	0.379	0.308
RIBES	0.313	0.091	0.242	0.417	0.395	0.356	0.302
WMD	0.302	0.131	0.226	0.066	0.371	0.330	0.237

表５　　WAT2017 における文レベルでのペアワイズの評価による Kendall の相関係数

自動評価法	ASPEC	JIJI	JPC	RECIPE			Avg.
				ING	STE	TTL	
$MEMD_{COMB(1, 2)}$	**0.527**	0.118	**0.410**	**0.574**	**0.533**	0.411	0.445
IMPACT	**0.556**	0.105	0.392	**0.550**	**0.542**	0.420	0.444
RIBES	**0.592**	0.186	0.468	**0.491**	**0.473**	0.327	0.435
MEMD	0.370	0.127	0.410	**0.551**	**0.467**	0.401	0.400
sentBLEU	0.384	0.041	0.234	0.503	0.457	0.410	0.364
WMD	0.154	0.081	0.138	0.349	0.441	**0.428**	0.289

表３から表５において、“Avg.”以外の相関係数は日英と英日方向の相関係数の平均である。また、それらの相関係数において太字と下線の相関係数はそれぞれ BLEU と IMPACT との間で統計的有意性があることを示している。

3.4.4.4　考察

表３より $MEMD_{COMB(4, 1)}$の“Avg.”が最も高い相関係数を示した。これは“RECIPE”と“TTL”の相関係数が 1.0 となっていることが影響している。具体的な内容としては、２つの機械翻訳システム（例えば、“A”と“B”）に対する $MEMD_{COMB(1, 4)}$のスコアはそれぞれ 0.343 と 0.342 であった。同様に、adequacy による人手評価はそれぞれ 4.200 と 4.075 と機械翻訳システム“A”の方が“B”よりも高かった。したがって、$MEMD_{COMB(4, 1)}$と人手評価共に機械翻訳システム“A”の方がわずかに“B”よりも高く、、その結果、高い相関係数が得られたと考えられる。一方、RIBES では機械翻訳システム“A”と“B”に対するスコアはそれぞれ 0.529 と 0.550 であった。したがって、人手評価とは異なり、“B”の方がスコアは高く、相関係数が低くなったと考えられる。また、“JPC”においても $MEMD_{COMB}$ と MEMD の相関係数が他の自動評価法に比べて高い相関係数を示していることも“Avg.”が高くなった原因になったと考えられる。

そして、表４と表５の文レベルの相関係数においては共に $MEMD_{COMB(1, 2)}$が最も高い“Avg.”の値を示した。これは、提案手法 MEMD とチャンクに基づく自動評価法である IMPACT の組み

合わせが有効であったことを示している。

3.4.5 まとめ

本報告では、単語間の位置情報と単語の分散表現を用いた Earth Mover's Distance に基づく自動評価法として MEMD（**M**etric based on **E**arth **M**over's **D**istance）を提案した。さらに、文レベルの人手評価との相関の向上を目的に MEMD とチャンクに基づく自動評価法である IMPACT を組み合わせた $\text{MEMD}_{\text{COMB}}$ を提案した。WAT2017 の日英方向と英日方向のデータを用いたメタ評価実験の結果、システムレベルのメタ評価では $\text{MEMD}_{\text{COMB}(4,1)}$、文レベルのメタ評価では $\text{MEMD}_{\text{COMB}(1,2)}$が最も高い相関係数を示した。これらの結果より、提案手法の有効性を確認することができた。

今後は日本語-英語間の性能評価実験だけではなく、他の言語間の翻訳文を用いた評価実験を行う予定である。また、文レベルの評価精度は相関係数が低く十分とはいえないため、文レベルの相関係数を向上させるための改良を行っていく予定である。

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] Klubička, G., Toral, A. and Sánchez-Cartagena V. M.(2017) "Fine-grained Human Evaluation of Neural Versus Phrase-based Machine Translation," The Prague Bulletin of Mathematical Linguistic, No.108, pp.121--132.

[4] 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.

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

[6] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas (1998) "A Metric for Distributions with Applications to Image Database," Proceedings of the 6th International Conference on Computer Vision, pp.59-66.

[7] 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.

[8] 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.
[9] Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig and Sadao Kurohashi (2017) "Overview of the 4th Workshop on Asian Translation," Proceedings of the 4th Workshop on Asian Translation, pp.1-54.
[10] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro(2008) "Overview of the Patent Translation Task at the NTCIR-7 Workshop," Proceedings of NTCIR-7 Workshop Meeting, pp.389-400.
[11] 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.
[12] 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.
[13] 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.
[14] "Welcome to Moses!", http://www.statmt.org/moses/
[15] "MeCab: Yet Another Part-of-Speech and Morphological Analyzer,"
http://taku910.github.io/mecab/
[16] 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.

3.5 WAT2018 人手評価結果について

東京大学　中澤　敏明
NHK 放送技術研究所　後藤　功雄
東芝デジタルソリューションズ株式会社　園尾　　聡

3.5.1　はじめに

今年度の拡大評価部会人手評価グループからは、WAT2018[1]で行われた一対比較による評価（Pairwise Evaluation）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（JPO Adequacy Evaluation）の 2 種類の人手評価結果について報告する。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。

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

前回までは全てのベースライン翻訳はフレーズベース SMT を採用していたが、今回からは十分な量の訓練データが提供されている一部の翻訳タスクにおいて、ベースラインを最もシンプルな NMT に変更した。この変更によりベースラインが強化されたため、前回までの Pairwise 評価結果と今回の結果を直接比較することはできない。

JPO Adequacy Evaluation（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」[2]）の結果についての報告を行う。JPO Adequacy Evaluation は、テストセットのうちの 200 文を対象に、2 名の評価者が以下の基準での絶対評価を行う。

表 1：JPO Adequacy Evaluation の評価基準

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%)

翻訳タスクは昨年度からの継続である科学技術論文（日⇔英、中）、特許文（日⇔英、中、韓）、新聞記事（日⇔英）、料理レシピ（日⇔英）、混合ドメイン（ヒンディー⇔英、日）に加えて、新たに混合ドメイン（ミャンマー⇔英）と映画字幕（インド諸語 7 言語⇔英）の翻訳タスクが追加

された。しかし残念ながら新聞記事および料理レシピ翻訳タスクは参加者がおらず、特許文翻訳タスクも最大 2 チームの参加にとどまった。もっとも参加チーム数が多かったのはミャンマー語翻訳タスクの 8 チームで、ついで科学技術論文タスクの 5 チーム、インド諸語タスクの 4 チームであった。

Pairwise Evaluation は JPO Adequacy Evaluation に回すためのデータの選別のために行うため、参加者数が少ないなどの理由で選別の必要のないサブタスクにおいては Pairwise Evaluation は実施していない。JPO Adequacy Evaluation はインド諸語の一部を除いて全ての翻訳タスクについてが行われたが、ここでは科学技術論文タスクの結果についてのみ報告する。

3.5.2　科学技術論文タスクの人手評価結果

図 1 から図 4 に日英、英日、日中、中日翻訳の上位チームの評価結果を示す。左側のグラフは JPO Adequacy Evaluation の詳細であり、2 名の評価者による評価結果の割合を示している。右側のグラフは 3 種類の自動評価（BLEU、RIBES、AM-FM）と 2 種類の人手評価（JPO Adequacy Evaluation と Pairwise Evaluation）の結果をまとめたものである。

左側のグラフの中で点線で囲まれているシステムは Transformer をベースとしたモデルを採用したシステムであり、それ以外は RNN ベースのシステムである。これを見ればわかるように、Transformer ベースのモデルの方が RNN ベースのモデルよりも一貫して高精度な翻訳が行えている。今回の特許文翻訳タスクにおいては Transformer ベースのモデルはなかったが、特許文においても Transformer ベースのモデルの方が高精度な翻訳が行える可能性がある。

英日翻訳においては、BLEU スコアは昨年度に比べて 2.5 ポイント向上しているが、人手評価の平均値はほとんど変わらなかった。逆に日中翻訳では人手評価結果の平均値は 0.5 ポイントほど向上したが、BLEU スコアの向上はわずかである。これは全体的な翻訳精度が向上し、良い翻訳と悪い翻訳の区別が自動評価ではつけられない領域に来ているためと思われ、自動評価での翻訳の質の判定が限界に来ていると考えられる。

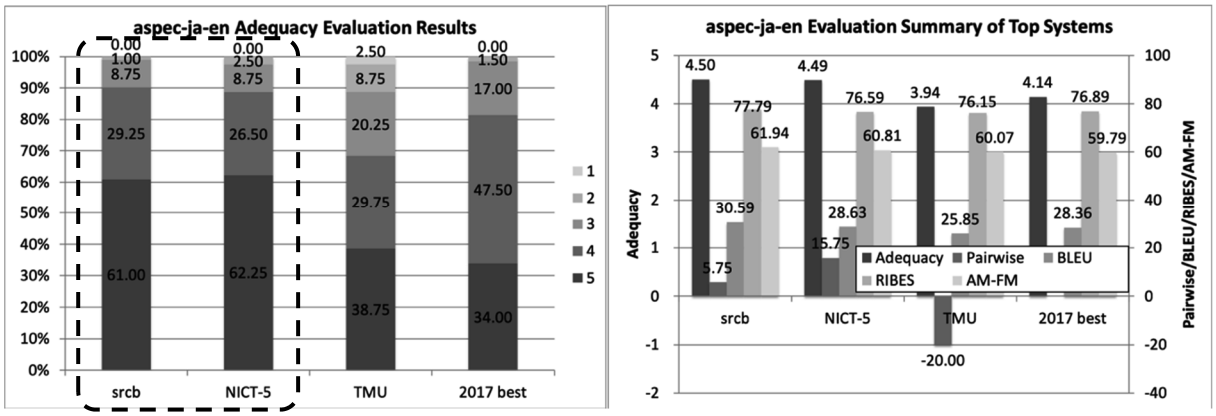


図 1: ASPEC ja-en 評価結果

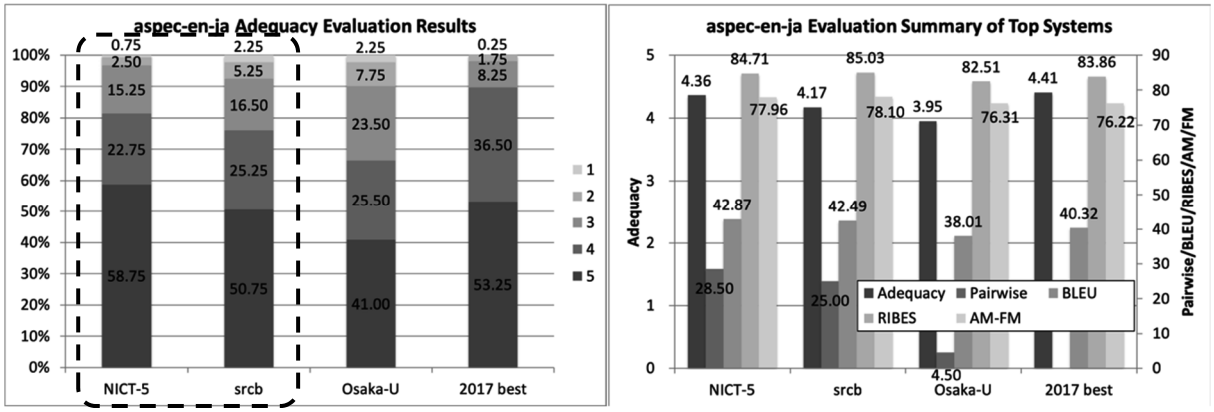


図 2: ASPEC en-ja 評価結果

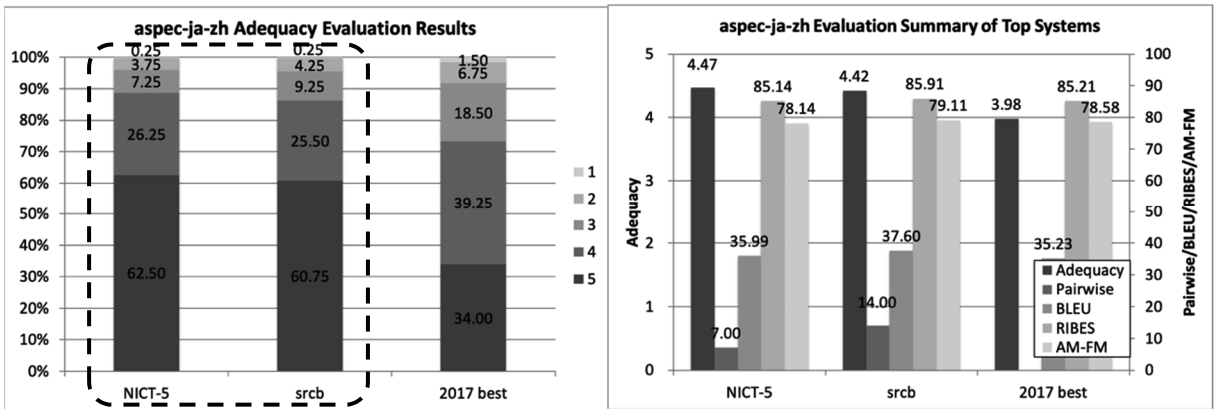


図 3: ASPEC ja-zh 評価結果

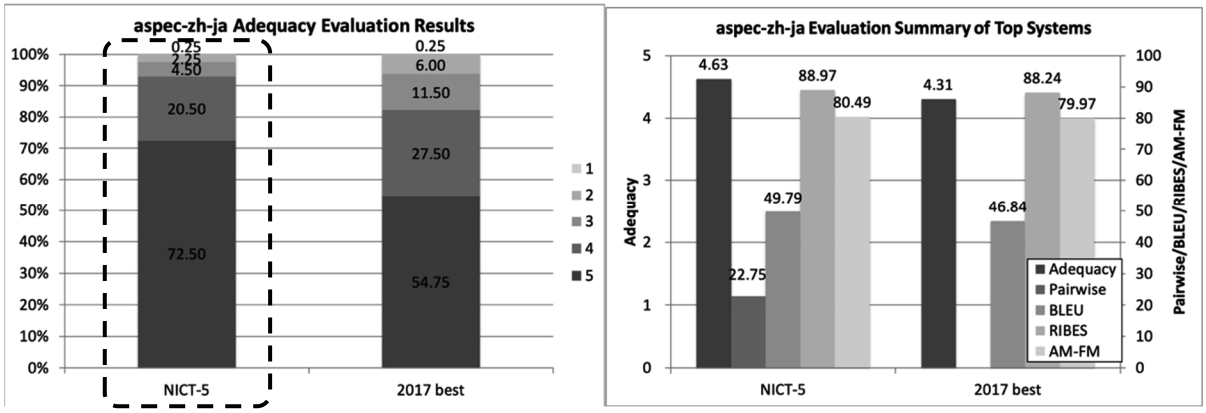


図 4: ASPEC zh-ja 評価結果

3.5.3 まとめ

本項では WAT2018 の科学技術論文翻訳タスクの人手評価結果について報告し、その結果から得られた知見をまとめた。これまでは文単位での評価しか行っていないが、今後は訳語の一貫性や適切な代名詞の補完など、文脈情報を利用するような翻訳が必要となるであろう。またそのよ

うな評価が適切に行えるようなデータセットや評価指標の整備も必要となる。これらの解決を今後の課題としたい。

参考文献

[1] Overview of the 5th Workshop on Asian Translation, In Proceedings of the 5th Workshop on Asian Translation (WAT2018).

[2] https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

3.6　自動評価と人手評価の比較

－WAT2018 での BLEU 値と pairwise evaluation 値の比較－

元・山梨英和大学　江原 暉将

3.6.1　はじめに

機械翻訳評価の方法としては、まず人手評価が提案された。例えば、文献 1)や 2)がその嚆矢である。特許文献については、文献 3)で、内容の伝達レベルの評価、重要技術用語の翻訳精度の評価、流暢さの評価、チェックリストによる評価、が提案されている。しかし人手評価にはコストも時間もかかるため、計算機で自動的に行える自動評価が提案された[4]。文献 4)の BLEU は、提案当初から課題が指摘されており[5]、改良を加えた自動評価基準が提案されている[6)7)]。その中で BLEU は基準となる自動評価手法として現在も多く用いられている。

アジア言語についての機械翻訳ワークショップ(WAT: Workshop on Asian Translation)では、自動評価基準として BLEU、RIBES、AMFM が用いられており、人手評価基準としては JPO Adequacy (文献 3)の内容の伝達レベルの評価に相当する)およびベースライン翻訳結果との pairwise evaluation (対比較) score が用いられている[8)]。筆者も WAT2018 に参加し[9)]、その中で BLEU 値と pairwise evaluation score に逆転があることを経験したので、その分析を行う。

3.6.2　自動評価と人手評価の逆転現象

WAT2018 での筆者のシステム(EHR)を用いて科学技術文献の英日翻訳タスク(ASPEC-en-ja)に参加した結果、BLEU 値が 37.97 であった。一方、ベースラインの BLEU 値は 36.37 であり、EHR より低い。しかしながら、EHR とベースラインの pairwise score の結果は－0.5 であり、ベースラインの方が EHR より評価値が高い[1]。この逆転現象を文毎に分析する。そのために試験データ(1,812 文)に対する EHR およびベースラインの翻訳結果について、文毎の BLEU 値を計算した。この計算には NIST のツール mteval-v13a.pl[2] を用いた。そして、文毎にそれらの差(EHR の BLEU 値マイナスベースラインの BLEU 値)を求めた。この値を diff BLEU と呼ぶ。pairwise score は 5 人の評価者が各々{-1, 0, 1}で評価する。-1 はベースラインのほうが EHR より良い、0 は同程度、1 は EHR のほうがベースラインより良い、という評価値である。したがって、5 人の評価者の評価値の合計は-5 以上 5 以下である。本文では、この合計値を pairwise score と呼ぶ。試験データ(1,812 文)に対する diff BLEU と pairwise score の散布図を図 1 に示す。第 1 象限と第 3 象限のデータ点では、両者に逆転はないが、第 2 象限と第 4 象限のデータ点には、逆転がある。同データに対する符号での分割表を表 1 に示す。

---

[1] 推定値は負値であるが、信頼区間の上端は正値であり、ベースラインの方が EHR より評価値が高いということに有意性はない。

[2] ftp://jaguar.ncsl.nist.gov/mt/resources/mtevalv13a.pl

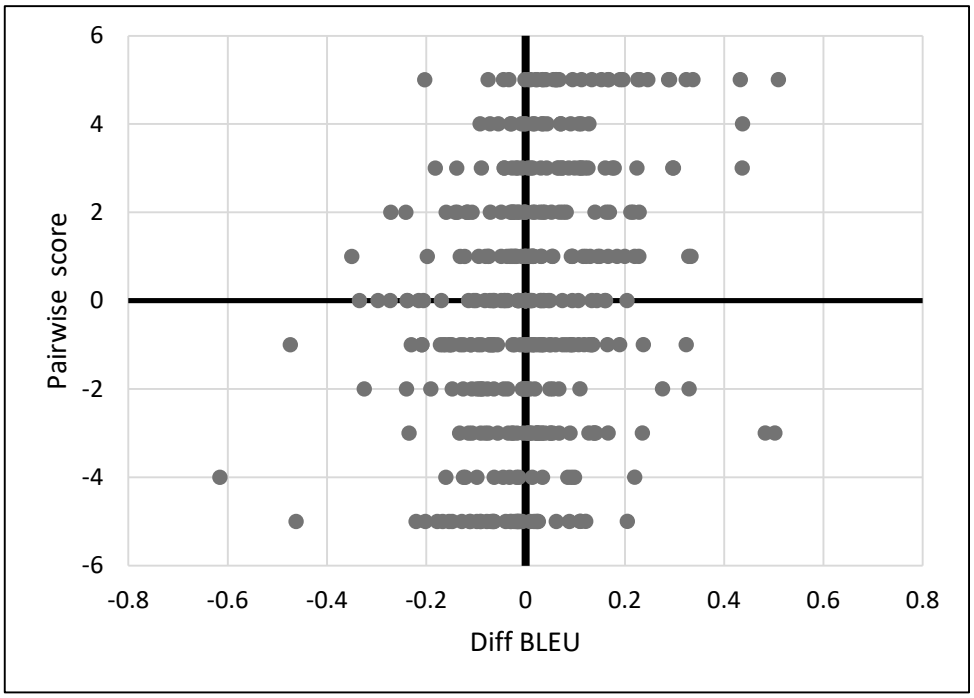


図 1　試験データ(1,812 文)に対する Diff BLEU と Pairwise score の散布図

表 1　試験データ(1,812 文)に対する Diff BLEU と Pairwise score の符号での分割表

Diff BLEU＼Pairwise	＜ 0	≧ 0
＜ 0	94	84
≧ 0	85	137

3.6.3　文毎の分析

前節で述べた diff BLEU と pairwise score で逆転の起こっているデータ点について、典型的な例をあげて文毎の分析をする。表 2 には、diff BLEU が正で pairwise score が負である例を示す。表 3 には、diff BLEU が負で pairwise score が正である例を示す。表中、BLEU_b はベースラインの BLEU 値、BLEU_r は EHR の BLEU 値である。diff BLEU= BLEU_r－BLEU_b である。

diff BLEU が正で pairwise score が負となる理由として二つのことが見いだされた。表 2 の文番号 365 と 157 では、重要情報が EHR では誤訳あるいは訳抜けとなっており、そのため、人手評価では低い評価値となった。一方、長い句において EHR と参照訳が一致しているため、BLEU_r が高くなった。その句において、ベースライン訳は語形は異なるものの、意味的には参照訳に近いため、人手評価値を下げることはなかった。これが第一の理由である。表 2 の文番号 277 と 118 では、第二の理由が読み取れる。これらの文では、ベースライン訳が直訳となっている一方、EHR と参照訳は意訳になっており、しかも語形が一致している。そのため、BLEU_r は BLEU_b より高くなる。一方、pairwise 評価の評価者は参照訳を見ていないので、より原文の表現に近いベースライン訳を EHR 訳より良い訳と判断していると推察される。

表 2　diff BLEU が正で pairwise score が負である例

文番号	365
原文	Toxic substances in heavy oil is a cause of failure.
参照訳	重油 中 に 含ま れる 有害 物質 が 障害 の 原因 と なる 。
ベースライン	重油 中 の 有害 物質 は 故障 の 原因 である 。
EHR	重 質 油 中 の 有害 物質 が 障害 の 原因 と なって いる 。
BLEU_b	0.1027
BLEU_r	0.4261
pairwise	-1
備考	EHRでは、重要情報である "heavy oil" が "重 質 油"と誤訳されている。EHRでは、長い句"有害 物質 が 障害 の 原因 と"が参照訳と一致しており、BLEU_rが高い。ベースライン訳の" 有害 物質 は 故障 の 原因"は参照訳と意味が近いが語形は異なる。

文番号	157
原文	During the follow-up period 176 people died by malignant tumor.
参照訳	追跡 期間 中 に 176 名 が 悪性 腫瘍 に より 死亡 した 。
ベースライン	追跡 期間 中 176 人 が 悪性 腫よう で 死亡 した 。
EHR	フォロー アップ 期間 中 ， 悪性 腫よう に より 死亡 した 。
BLEU_b	0.21
BLEU_r	0.3191
pairwise	-5
備考	EHRでは、重要情報である "176 people" が訳抜けしている。EHRでは、長い句"悪性 腫瘍 に より 死亡 した 。"が参照訳と一致している。ベースライン訳の "悪性 腫よう で 死亡 した 。" は参照訳と意味が近いが語形は異なる。

文番号	277
原文	In addition, the neutronic characteristics of the demonstration reactor are introduced.
参照訳	この ほか 実証 炉 の 核 特性 に ついて 紹介 した
ベースライン	さらに ， 実証 炉 の 中性子 特性 を 紹介 した 。
EHR	さらに ， 実証 炉 の 核 特性 を 紹介 した 。
BLEU_b	0.1836
BLEU_r	0.4035
pairwise	-4
備考	ペースラインでは、"neutronic characteristics" が "中性子 特性" と直訳されている。EHRと参照訳では、"核 特性" と意訳されている。

文番号	118
原文	A fractal analysis result of potentials related to events was applied to the survey of fluctuating conditions of intelligent activities.
参照訳	事象 関連 電位 の フラクタル 分析 を 行い，その 結果 を 知的な 活動 の 変動 的な 状態 の 調査 に 使った 。
ベースライン	事象 に 関連 した ポテンシャル の フラクタル 解析 結果 を，知的 活動 の 変動 条件 の 調査 に 適用 した 。
EHR	事象 関連 電位 の フラクタル 解析 結果 を，知的 活動 の 変動 条件 の 調査 に 適用 した 。
BLEU_b	0.1413
BLEU_r	0.2825
pairwise	-3
備考	ベースラインでは、"potentials related to events" が "事象 に 関連 した ポテンシャル" と直訳されている。EHRと参照訳では、"事象 関連 電位" と意訳されいる。

diff BLEU が負で pairwise score が正である理由としても、逆の場合のベースライン訳と EHR 訳を入れ替えた理由が見いだされる。表 3 の文番号 128 と 76 が、その例である。一方、表 3 の文番号 36 は、やや微妙である。参照訳、ベースライン訳、EHR 訳ともにほぼ同一の意味を表現しており、原文の意味を反映している。しかし、ベースライン訳は ”the detailed analysis of the intensity dependence of excited light”を”励起 光 の 強度 依存 性 の 詳細な 解析”と直訳であり、多義的な助詞「の」が重なって意味が若干不明確である[10]。一方、EHR 訳は”励起 光 の 強度 依存 性 を 詳細に 解析 して”と意味が明確であり、日本語としての自然性も高い。そのため、人手評価では EHR の方が良い訳と判断したものと推察される。

表 3　diff BLEU が負で pairwise score が正である例

文番号	128
原文	No side effect was noted during treatment.
参照訳	治療 中 ， 副 作用 は 認め なかった 。
ベースライン	副 作用 は 認め なかった 。
EHR	治療 中 副 作用 は 認め られ なかった 。
BLEU_b	0.6065
BLEU_r	0.4033
pairwise	5
備考	ベースラインでは、重要情報"during treatment" が訳抜けしている。ベースラインでは、長い句 "副 作用 は 認め なかった 。"が参照訳と逸しいてる。EHR訳の "副 作用 は 認め られ なかった 。" は参照訳と意味は近いが語形は異なる。

文番号	76
原文	Cutting, patterning polishing, and metalizing work of diamond films with laser beams are reviewed.
参照訳	レーザ に よる ダイヤモンド 膜 の 切断 ， パターニング ， 研摩 ， 金属 化 加工 に ついて 総 説 した 。
ベースライン	レーザ ビーム を 用いた ダイヤモンド 膜 の 切断 ， パターニング ， および メタライゼーション 作業 に ついて レビュー した 。
EHR	レーザ ビーム を 用いた ダイヤモンド 膜 の 切削 ， パターニング 研磨 ， 金属 化 に ついて レビュー した 。
BLEU_b	0.3412
BLEU_r	0.1592
pairwise	3
備考	ベースラインでは、"metalizing work" が "メタライゼーション 作業" と直訳されている。EHRと参照訳では、"金属 化" と意訳されている。ベースラインでは、長い句 "ダイヤモンド 膜 の 切断 ， パターニング" が参照訳と一致しいてる。EHR訳の "ダイヤモンド 膜 の 切削 ， パターニング" は参照訳と意味は近いが語形は異なる。

文番号	36
原文	Through the detailed analysis of the intensity dependence of excited light, this is judged to be due to photoionization.
参照訳	励起 光 強度 依存 性 の 詳細 解析 に より ， 光 イオン 化 に よる もの と 判断 した
ベースライン	励起 光 の 強度 依存 性 の 詳細な 解析 に より ， 光 イオン 化 に よる と 判断 した 。
EHR	励起 光 の 強度 依存 性 を 詳細に 解析 して ， これ は 光 イオン 化 に よる もの と 判断 した 。
BLEU_b	0.592
BLEU_r	0.4533
pairwise	3
備考	参照訳、ベースライン、EHRともにほぼ同一の意味であるが、参照訳とベースラインはやや直訳であり、さらに語形としても近い。一方、EHRは文意がベースラインより明確であり、自然性も高い。そのため人手評価では高い評価値となったと推定される。

3.6.4 まとめと今後の課題

自動評価基準の BLEU と人手評価基準の pairwise score の間で評価値に逆転が起こる場合について分析した。分析の結果、いくつかの原因が見いだされた。今後は、これらの原因を考慮に入れて自動評価基準を改良することが課題である。

参考文献

1) Automatic Language Processing Advisory Committee, National Research Council. 1966. Language and Machines, Computers in Translation and Linguistics.

2) 長尾真、辻井潤一. 1985. Mu プロジェクトにおける日英翻訳結果の評価、*情報処理学会研*

究報告、*自然言語処理 47-11*、pages 79-88.

3) 日本特許情報機構. 2014. 特許文献機械翻訳の品質評価手法に関する調査報告書.

4) Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* pages 311-318.

5) Chris Callison-Burch, Miles Osborne, Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research, *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249-256.

6) Hiroshi Echizen-ya, Kenji Araki. 2007. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, *Proceedings of the Eleventh Machine Translation Summit (MT SUMMIT XI)*, pages 151-158.

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

8) Toshiaki Nakazawa et al. 2018. Overview of the 5th Workshop on Asian Translation. *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*.

9) Terumasa Ehara. 2018. SMT reranked NMT (2), *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*.

10) 日本特許情報機構. 2018. 特許ライティングマニュアル(第 2 版), page 20.

4. シンポジウム開催報告

4. 第5回特許情報シンポジウム 開催報告

奈良先端科学技術大学院大学　須藤　克仁

4.1 開催概要

本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30年12月7日（金）午後にビジョンセンター浜松町において開催した。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5回となる。シンポジウムの企画および運営は実行委員長を宇津呂先生、副委員長を須藤とし、研究会委員で組織する実行・プログラム委員を中心に行った。一般講演の受付についてはこれまで同様綱川先生にご尽力いただき、プログラム委員で選定を行った。

今回は午後半日の開催で、招待講演3件、特別講演1件と一般講演4件の構成とし、研究会からの報告は行わなかった。参加者は93名（研究会関係者を含む）となり、大変盛況であった。

4.2 講演の概要

4.2.1 招待講演（1）: 松谷 洋平 様（特許庁 総務部総務課 特許情報室 室長補佐）

「特許庁における機械翻訳活用の現状と今後の課題」

2010年頃を境にしての中国における特許出願件数の突出した伸び、また日本から海外への出願件数の増加を受け、機械翻訳の重要性は増す一方である。WIPO や EPO ではニューラル機械翻訳による翻訳サービスを提供しているが、日本特許庁でも審査における海外特許情報の参照や審査結果の海外発信のために機械翻訳を積極的に活用しており、また機械翻訳の更なる充実へ向けて辞書やコーパスの整備、技術調査を継続的に行っている。今後はニューラル機械翻訳に基づく機械翻訳プラットフォームの提供、また審査書類の機械翻訳に関する事業を実施する予定である。

4.2.2 招待講演（2）: 井佐原 均 様（豊橋技術科学大学 情報メディア基盤センター長／情報・知能工学系 教授）

「ニューラル翻訳時代の機械翻訳システムの社会実装」

ニューラルネットワークによる機械翻訳の急速な精度向上によって機械翻訳のニーズが高まっている。豊橋技科大では機械翻訳の社会実装のため、機械翻訳エンジンのベンダー、応用システム開発のベンダー、そしてユーザとなる企業や自治体との連携による共同開発・実証実験プロジェクトを進めている。自治体における情報提供では分野特化型の高品質な翻訳が求められるため、ドメイン内の対訳データの学習に加え辞書登録機能が実装され、システムの運用・評価・効果検証を進めている。

4.2.3 招待講演（3）: 田村 晃裕 様（愛媛大学 大学院理工学研究科 助教）

「ニューラル機械翻訳の研究動向」

ニューラルネットワークによる機械翻訳の技術について初期の方法から最新の研究までを概観

する。2014 年の回帰型ニューラルネットワーク(RNN)による系列変換モデルの提案に始まり、訳文中の対応関係をモデル化する注視(attention)機構の導入、8 層にも及ぶ多層化によって高精度化を達成した Google NMT、現在標準的な手法と認識されている Transformer までのわずか 3 年ほどで翻訳精度は大きく向上した。さらにサブワードの利用、逆翻訳による単言語データの活用、さらには対訳文データを要しない教師なし機械翻訳、画像の情報を活用するマルチモーダル翻訳等、様々な新しい技術とその応用が急速に広がりつつある。

4.2.4 特別講演 : 日本知的財産翻訳協会（NIPTA）特許機械翻訳研究会（奥山 尚一 様[全体紹介]、上野 哲也 様[化学分科会]、湯浅 豊裕 様[機械工学分科会]、菊地 公一 様[電気・電子工学分科会]、宮城 三次 様[知財法務実務分科会]、新田 順也 様[ツール開発]）

「ユーザーから見た NMT の使い勝手と活用の展望」

ニューラル機械翻訳によって特許等の翻訳においても機械翻訳が高い精度でできるようになりつつあり，NIPTA では実務において現状のニューラル機械翻訳の技術水準の評価を行いその活用方法を検討するための研究会を立ち上げ，複数の分科会に分かれて活動している。各分科会において NMT を翻訳実務に利用可能であるかどうかの検討を行った。NMT では SMT と異なる訳文の自然性が大きく向上したが、知財翻訳においては NMT で頻出する訳抜けや湧き出しの問題に加え、専門用語の訳出や用語の統一等の面で多くの課題が残されているのが現状である。しかしながら様々な工夫によって NMT の能力を活かすことができ、知財翻訳実務において有益であると考えられる。

4.2.5 一般講演

一般講演は発表募集を行い、8 件の発表申込に対して以下の 4 件を採択の上、質疑込みで各 15 分の発表とした。いずれも特許の翻訳や検索に関する実用視点からの発表であり、産業における特許情報処理の重要性を感じさせるものであった。

・「特許明細書の翻訳で注意すべきこと」 吉川 潔 様（翻訳業）

・「請求項の記載を構造化する提案」 小池 誠 様（弁理士）

・「用語集による原文の一括置換と正規表現による一括並べ替えを用いた特許明細書の高品質翻訳」 杉山 範雄 様（株式会社杉山特許翻訳）

・「知的財産リスクマネージメントにおける効果的な特許検索」 宮崎 祐 様、河野 京子 様（ヤフー株式会社）

4.3 所感

前回 2016 年の開催時と比較し、ニューラル機械翻訳の技術が広く浸透しつつあること、そして翻訳業界での利用が急速に進みつつあることを改めて強く感じさせる内容であった。技術の進展に合わせ、各国特許庁による機械翻訳利用が活発化し、企業等の知財実務・知財翻訳・産業翻訳にも変化をもたらしつつある。

今回は知財行政・研究開発に関する招待講演に加え、特別講演としてユーザの立場から見たニ

ューラル機械翻訳の利点と欠点について詳細なご報告があり、機械翻訳の研究に携わる者として非常に興味深かった。実用のニーズがさらに高まることは間違いなく、研究会としてもこうしたニーズの高まりに対して学術的研究の視点からどういった価値を提供できるか、また技術と実用との橋渡しをすることができるかが重要であろう。研究会としてはこうしたニーズの高まりに対して学術的研究の視点からどういった価値を提供できるか、また実用との橋渡しを図ることができるかが重要であろうと考える。

4.4 謝辞

本シンポジウムの開催にあたり、一般社団法人 日本翻訳連盟および特定非営利活動法人 日本知的財産翻訳協会よりご協賛を賜りました。また、特別講演の実施にあたり、河野弘毅様にご協力いただきました。シンポジウム実行委員会を代表し、この場を借りて御礼申し上げます。

2018年 度 AAMT/Japio特許翻訳研究会報告書

発 行 日　平成 31 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