

2019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
並びに
国際ワークショップ開催報告

2020 年 3 月

一般財団法人 日本特許情報機構

目次

1. はじめに	1
辻井 潤一 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	
2. 機械翻訳および関連技術	
2.1 係り受け構造に基づくアテンション制約を用いたニューラル機械翻訳	4
出口 祥之 愛媛大学 田村 晃裕 愛媛大学 二宮 崇 愛媛大学	
2.2 Neural Machine Translation based on Multi-Hop Attention	13
Shohei Iida Graduate School of Systems and Information Engineering, University of Tsukuba, Japan Hongyi Cui Graduate School of Systems and Information Engineering, University of Tsukuba, Japan Po-Hsuan Hung Graduate School of Systems and Information Engineering, University of Tsukuba, Japan Yizhen Wei Graduate School of Systems and Information Engineering, University of Tsukuba, Japan Takehito Utsuro Graduate School of Systems and Information Engineering, University of Tsukuba, Japan Masaaki Nagata NTT Communication Science Laboratories, NTT Corporation, Japan	
2.3 特許請求項定型パターンの分類と解析	19
横山 晶一 山形大学 名誉教授	
2.4 BERT によるテキスト分類における折り返し翻訳によるデータ拡張の効果の検討	26
福田 治輝 静岡大学 綱川 隆司 静岡大学 大島 純 静岡大学 大島 律子 静岡大学 西田 昌史 静岡大学 西村 雅史 静岡大学	
3. 機械翻訳評価手法	
3.1 WMT2017 データにおける単語の位置情報と分散表現に基づく自動評価法の有効性について	34
越前谷 博 北海学園大学	
3.2 新コーパスを用いた中日テストセット試験文の採集と分析	44
江原 暉将 元・山梨英和大学 長瀬 友樹 (株)富士通研究所	
3.3 独自テストセットによる中日機械翻訳の評価	52
須藤 克仁 奈良先端科学技術大学院大学 王 向莉 株式会社ディープランゲージ	
4. 国際ワークショップ開催報告：PSLT 2019	56
須藤 克仁 奈良先端科学技術大学院大学	

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	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	筑波大学大学院 教授
	須藤 克仁	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	北海学園大学大学院 教授
	江原 暉将	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	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	静岡大学大学院 助教
	中澤 敏明	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
	岡 俊行	(有) アジア産業
	高 京徹	(株) 高電社
	園尾 聡	東芝デジタルソリューションズ株式会社
	長瀬 友樹	(株) 富士通研究所 人工知能研究所
	王 向莉	株式会社ディープランゲージ

オブザーバー（（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
木下 聡	特許情報研究所 調査研究部 統括研究主幹
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	情報活用支援部 情報活用支援課 係長
船戸さやか	特許情報研究所 調査研究部 研究企画課 主任
星山 直人	情報システム部 システム企画課 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

事務局

株式会社インターグループ

平成 31 年 / 令和元年度 AAMT/Japio 特許翻訳研究会・活動履歴

令和元(2019)年 5 月 17 日

第 1 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)

令和元(2019)年 6 月 21 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)

令和元(2019)年 8 月 19 日-23 日

第 17 回機械翻訳サミット

※第 8 回特許・技術文書翻訳ワークショップ (the 8th Workshop on Patent and Scientific Literature Translation (PSLT 2019)) を開催

(於アイルランド (ダブリン) the Helix Theatre at Dublin City University)

令和元(2019)年 9 月 13 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)

令和元(2019)年 11 月 15 日

第 4 回 AAMT/Japio 特許翻訳研究会

(於東京大学)

令和元(2019)年 12 月 20 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)

令和 2(2020)年 2 月 21 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)

※新型コロナウイルス感染拡大に伴い中止となった。

令和 2(2020)年 3 月 31 日

『2019 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並びに
国際ワークショップ開催報告』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

最近、ニューラルネットを使った機械翻訳の進歩に、あらためて、感心したことがいくつか続いた。一つは、10月末にパリで開催された人工知能関係での会議で、プログラム委員会の名前で共同声明を出した時のこと。フランス側がまとめた声明案はフランス語であったが、フランス以外の委員に配布された英語版は、機械翻訳されたもので、人手による修正は入っていないとのこと。この英語での共同声明は、読んでいて不自然な部分のない、ほぼ完ぺきなものになっていた。また、ドイツのグループとの共同研究を準備していた時のこと。彼らが機械翻訳で英訳して送ってきた提案書、ドイツ国内での資金獲得のためのもので原言語はドイツ語の提案書も、人手での修正なしで、我々にとって違和感なく読めるものになっていた。この2つの例は、フランス語やドイツ語から英語への機械翻訳は、非常に完成度の高いものになっていることを示している。

海外からの研究者が多い我々のセンターでは、特に研究者に周知すべきことは、ウェブでの機械翻訳を使って英訳し、人手の修正をすることで配信している。フランス語から英語などへの翻訳に比べると、不完全で人手の修正が必要であるが、最初から人手で翻訳するよりは、はるかに早く、コストも下げられる。

最近、英語での研究者募集をするために、研究内容の説明の部分を日本語の提案書の一部を英訳したことがあった。英語訳を見ると、それなりに理解できる。ただ、これは、原文の日本語があって、それを下敷きにして英語を読み、理解できているのだと気が付いた。そういう目で、英語訳を見ると、かなりの修正をしないと、研究者募集には使えない。また、私が修正した英文を、研究分野の知識があるネイティブの研究者に見せると、また、大幅に修正された。

この経験から、(1)日本語から英語への翻訳が、フランス語・ドイツ語から英語に比べるとはるかに難度が高いこと、(2)研究内容に関する記述には専門用語の翻訳など、現在のニューラル翻訳が苦手な部分が多いこと、(3)求人広告のように簡潔で、かつ、魅力的なテキストを作る技術には言語に依存する部分があり、現在の機械翻訳の技術では、まだ解決できていないこと、ということを痛感した。

機械翻訳に関係する技術の研究には、まだまだ、解決すべき課題が多く残されているとことであろう。もとにある日本語を下敷きにすると、翻訳結果の英語が理解できるという錯覚もある。この錯覚のために、意味が分からない不自然な英語のテキストが流通してしまうという危険性もあろう。

本報告書は、我々の1年間の活動をまとめたものである。進展が著しい機械翻訳技術をさらに優れたものにしていくための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 係り受け構造に基づくアテンション制約を用いたニューラル機械翻訳

愛媛大学 出口 祥之

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（ニューラル機械翻訳）が高い精度と自然な翻訳を実現することから盛んに研究されている．特に、同一文内の単語間の関係を捉える自己アテンション（Self Attention）という構造を用いた Transformer (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017) に基づくモデルが最も高い精度を実現することから現在大きく注目されている．従来のニューラル機械翻訳は、大きく分けると、各言語の中間表現を求める計算機構として、畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）を用いるニューラル機械翻訳 (Gehring, Auli, Grangier, Yarats, and Dauphin 2017) と再帰型ニューラルネットワーク（Recurrent Neural Network; RNN）を用いるニューラル機械翻訳 (Sutskever, Vinyals, and Le 2014; Luong, Pham, and Manning 2015) に分けられていたが、Transformer はいずれのモデルとも異なり、各言語の中間表現を計算するために、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アテンション（自己アテンション）を計算して中間表現を求める点を大きな特徴とする．Transformer では、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、位置エンコーディング（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 や自己アテンションを改善するための手法がいくつか提案されており、Strubell ら (Strubell, Verga, Andor, Weiss, and McCallum 2018) は、意味役割付与（Semantic Role Labeling）において、構文情報（係り受け構造）を用いて自己アテンションの重み付けを学習するマルチタスク学習や、構文情報を直接自己アテンションに用いる手法を提案している．これまで、統計的機械翻訳やニューラル機械翻訳では、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．

本稿は、原言語文、目的言語文の係り受け構造に基づいた制約を Transformer の自己アテンションに与えて訓練する新しいニューラル機械翻訳を提案する．本研究は、意味役割付与で用いられる Strubell ら (Strubell et al. 2018) の手法を機械翻訳に応用し、係り受け構造に基づく自己アテンションと機械翻訳のモデルを同時に学習するマルチタスク学習を行う研究 (Deguchi, Tamura, and Ninomiya 2019) である．構文解析器により得られる係り受け構造を自己アテンションの正解データと捉え、自己アテンションの出力と係り受け構造の差分を小さくする制約を目

的関数の一つとして導入することで、マルチタスク学習を実現する。

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) を用いて日英翻訳を行った。その結果、原言語文、目的言語文両方の係り受け構造を自己アテンションで考慮する提案モデルは従来の係り受け構造を考慮しない Transformer (Vaswani et al. 2017) より高い翻訳精度を実現した。

2.1.2 Transformer

Transformer (Vaswani et al. 2017) は、自己アテンション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ニューラル機械翻訳モデルである。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、1 単語ずつ逐次的に出力する処理が行われる。

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）から埋込み表現を表す行列に変換する。その後、Transformer は RNN に基づくニューラル機械翻訳とは異なり再帰的な構造を持たないため、位置エンコーディングにより単語の系列位置情報を付与する。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。

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

ここで、 d 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。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自己アテンション、位置ごとのフィードフォ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自己アテンション、原言語文の単語と目的言語文の単語間の関係を捉えるアテンション（言語間アテンション）、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) を行った後にレイヤ正規化 (Ba, Kiros, and Hinton 2016) が適用される。レイヤ正規化を適用する関数を $LayerNorm$ 、下位のサブレイヤからの出力を x 、現在のサブレイヤの処理を行う関数を $SubLayer$ とすると、 $LayerNorm(x + SubLayer(x))$ がサブレイヤの出力となる。

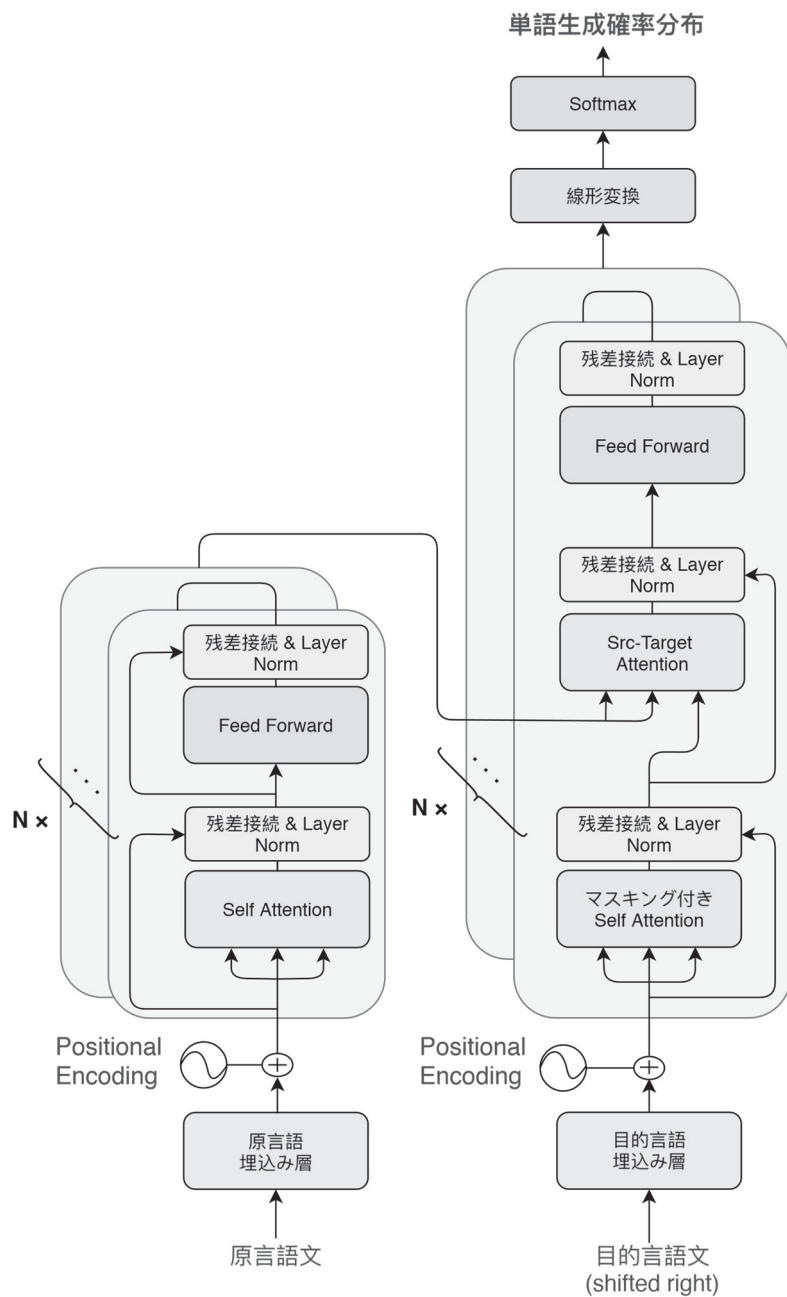


図 1: Transformer モデルの概要図

自己アテンションと言語間アテンションはマルチヘッドアテンションを用いて実現されている。マルチヘッドを用いることは、複数のアテンション機構を持つことに相当し、マルチヘッドの数を h としたとき、入力を線形変換により d 次元ベクトルを d/h 次元に縮退させ、 d/h 次元のベクトルを各ヘッドに渡す。アテンション機構への入力列を $\mathbf{x}_1, \dots, \mathbf{x}_n$ としたとき、各ヘッドは次の $\mathbf{z}_1, \dots, \mathbf{z}_n$ を計算する。

$$\begin{aligned} \mathbf{z}_i &= \sum_{j=1}^n \alpha_{ij} \mathbf{x}_j W^V \\ \alpha_{ij} &= \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})} \\ e_{ij} &= \frac{(\mathbf{x}_i W^Q)(\mathbf{x}_j W^K)^T}{\sqrt{d/h}} \end{aligned}$$

ただし, W^Q, W^K, W^V はそれぞれヘッド毎に定義される重み行列である. 各ヘッドの出力は連結され, 重み行列 W^Q で線形写像する機構がマルチヘッドアテンションである.

デコーダの自己アテンションでは, 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ができない. そのため, デコーダの自己アテンションでは, 予測する単語とそれより後方に位置する単語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自己アテンションを用いる.

デコーダは, 最終のデコーダレイヤの出力 $\mathbf{h}_1, \dots, \mathbf{h}_n$ 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 係り受け構造に基づく自己アテンションを用いた Transformer モデル

意味役割付与で用いられる Strubell ら (Strubell et al. 2018) の手法を機械翻訳に応用し, 係り受け構造に基づく自己アテンションの重み付けと機械翻訳のモデルを同時に学習するマルチタスク学習を行う.

Strubell ら (Strubell et al. 2018) は, 意味役割付与のタスクにおいて, 係り受け構造に基づく自己アテンションの重み付けと機械翻訳のモデルを同時に学習するマルチタスク学習を提案している. 本研究は, 彼らの手法に基づき, 機械翻訳のための Transformer エンコーダとデコーダの両方に彼らの手法を適用し, さらにサブワードに対応することを行う.

Transformer の自己アテンション α_{ij} の計算過程において e_{ij} は 2 単語の埋め込み表現に対する内積で計算されていたが, 内積の計算の代わりに次式で表されるバイアフィン計算を行い, 2 単語間の関係を直接学習する.

$$\begin{aligned} \alpha_{ij} &= \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})} \\ e_{ij} &= \frac{(\mathbf{x}_i W^Q) U (\mathbf{x}_j W^K)^T}{\sqrt{d/h}} \end{aligned}$$

ここで, U は 2 単語間の関係を表す重み行列となる.

次に, 係り受け解析器により得られる係り受け構造をアテンションの教師データとして捉え, 係り受け解析の学習を行う. 係り受け構造に対し, 単語 w_i が w_j に係る場合, $\hat{\alpha}_{ij} = 1$ とし, それ以外の場合は $\hat{\alpha}_{ij} = 0$ とする. 例えば, 図 2 の係り受け構造に対して, $\hat{\alpha}_{ij}$ は図 3 (b) のような行列となる.

図 3 (a) は自己アテンション α_{ij} の大きさを色の濃さで表現したヒートマップであり, 図 3 (a) のヒートマップが図 3 (b) の係り受け構造に近づくように自己アテンションの学習が行われる.

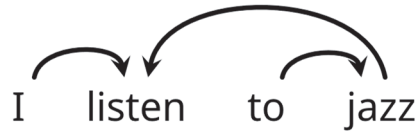


図 2: 係り受け構造

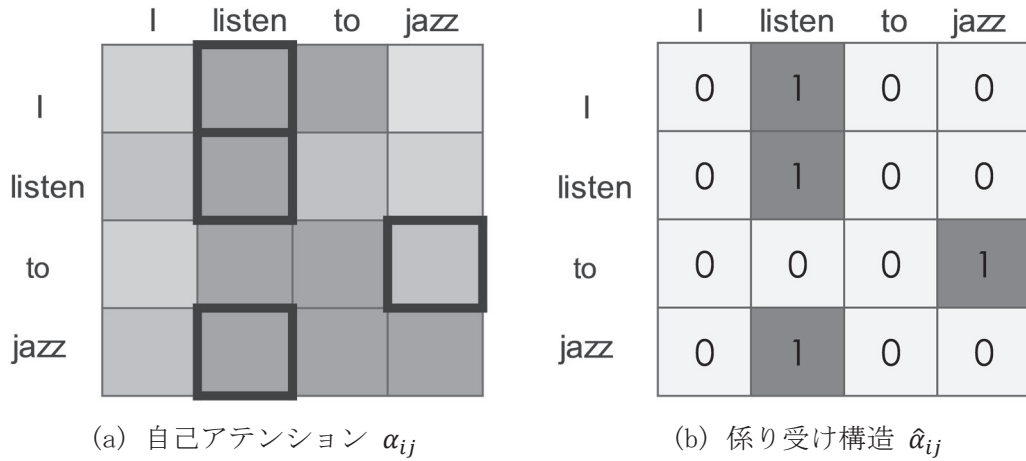


図 3: 自己アテンションと係り受け構造

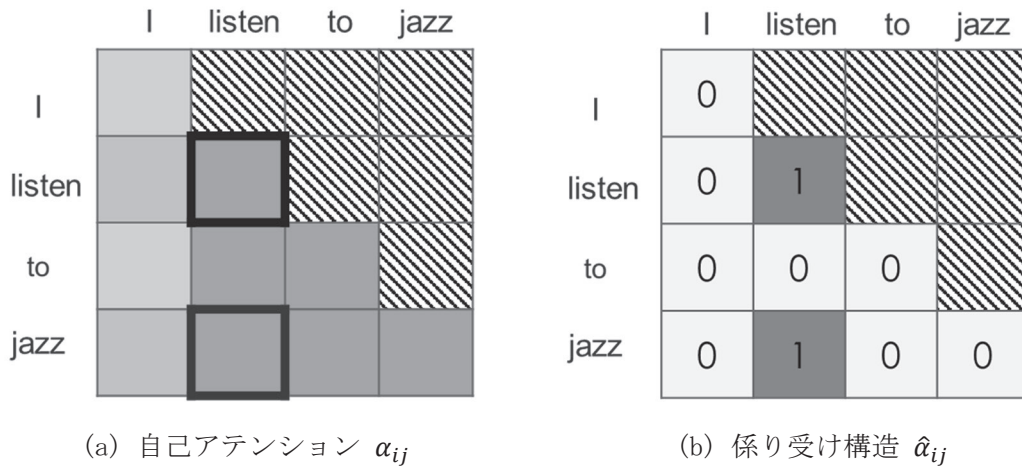


図 4: 自己アテンションと係り受け構造

デコード時には、Transformer デコーダにおいて、未来に生成される単語の自己アテンションを計算することができないため、ヒートマップおよび係り受け構造のマスキング処理を行う。図 4 はデコーダのマスキング処理を表す。斜線で表された要素がマスクされており、マスクされた要素の自己アテンションの計算および係り受け構造との差分の計算は行わない。



図 5: 係り受け構造のサブワード化

現在の多くのニューラル機械翻訳はサブワード列 (Sennrich, Haddow, and Birch 2016) を入力とするサブワード化が行われているが、係り受け解析は単語単位で行われるため、提案手法に対して直接サブワード化を適用することはできない。本研究では、サブワード化に対応するため、係り受け関係にある単語に対し、(1) 他の単語からの係り先は一番左端のサブワードとし、(2) 単語中のサブワードの係り先は右隣のサブワードとし、(3) 他の単語に係る係り元は右端のサブワードとすることで対応した。図 5 は “listen” が “li@@”, “s@@”, “ten” というサブワードに分割された場合の係り受け構造の例を示している。図左側が単語単位の係り受け構造を表し、図右側がサブワード単位の係り受け構造を示している。

提案手法の学習は、機械翻訳の学習、エンコーダにおける係り受け構造の学習、デコーダにおける係り受け構造の学習に対する目的関数を線形補間した以下の式で表される損失関数 $\mathcal{L}$ を目的関数としたマルチタスク学習により行う。

$$\mathcal{L} = \mathcal{L}_{token} + \lambda_1 \mathcal{L}_{encparse} + \lambda_2 \mathcal{L}_{decparse}$$

ただし、 $\mathcal{L}_{token}$ はラベル平滑化交差エントロピーによって求められる翻訳の損失、 $\mathcal{L}_{encparse}$, $\mathcal{L}_{decparse}$ はそれぞれ交差エントロピーによって求められるエンコーダ側、デコーダ側の自己アテンションと係り受け構造の間の損失、 λ_1, λ_2 はそれぞれ $\mathcal{L}_{encparse}$, $\mathcal{L}_{decparse}$ が目的関数に影響する度合いをコントロールするためのハイパーパラメータである。

2.1.4 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) を用いて日英翻訳を行った。英語文の単語分割は Moses (Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst 2007) を用い、係り受け解析は Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, and McClosky 2014) を用いて行った。日本語文は KyTea (Neubig, Nakata, and Mori 2011) を用いて単語分割を行い、EDA を用いて係り受け解析を行った。サブワード化には BPE (Sennrich et al. 2016) を用い、BPE の語彙数は 100,000 とした。モデルの学習では、訓練データの上位 1,500,000 文対に対して BPE を適用した際のサブワード長が 250 以下となる 1,198,149 文対を使用した。検証データとして 1,790 文対、テストデータとして 1,812 文対を用いた。

実験では、提案モデルと従来の Transformer (Vaswani et al. 2017) を比較する。提案モデルと従来モデルのハイパーパラメータは base モデル (Vaswani et al. 2017) に倣い、エンコーダとデコーダのレイヤ数をそれぞれ 6、アテンションのヘッド数を 8、埋込み次元を 512 次元とした。

表 1: 実験結果

	BLEU (%)
Transformer	27.29
提案モデル	28.29

モデルの学習には Adam を用い、 $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$ とした。学習率の更新スケジューリングは Vaswani ら (Vaswani et al. 2017) の方法と同様にした。ミニバッチサイズは 224 文とし、エポック数は 20 とした。従来モデルと提案モデルの翻訳の損失関数であるラベル平滑化交差エントロピーの ϵ (Szegedy, Vanhoucke, Ioffe, Shlens and Wojna 2016) は 0.1 とした。提案手法はエンコーダ、デコーダそれぞれ 4 層目の自己アテンションに対して、ヘッド 1 つを置き換えることによって適用した。提案モデルの目的関数のハイパーパラメータ λ_1, λ_2 はどちらも 1.0 とした。従来モデルと提案モデル両方において、モデルの評価は検証データに対して最も精度が高いエポックのモデルをテストデータに適用して翻訳性能を評価した。翻訳時はビーム探索を用い、ビーム幅は 4 とした。文長正規化パラメータ (Wu, Schuster, Chen, Le, Norouzi, Macherey, Krikun, Cao, Gao, Mache et al. 2016) は 0.6 とした。

実験結果を表 1 に示す。翻訳性能の評価指標には BLEU を用いた。表 1 より、提案手法を用いることにより BLEU (%) が 1.0 ポイント上昇していることがわかる。この結果より、係り受け構造に対する自己アテンションの学習を行うことで、Transformer モデルの翻訳精度を改善できることが分かった。

2.1.5 まとめ

本研究では、原言語および目的言語における係り受け構造を自己アテンションの学習に用いることで Transformer モデルを改善する手法を提案した。提案手法は、係り受け構造を自己アテンションの正解とみなすことにより、機械翻訳の学習と、エンコーダの自己アテンションの学習と、デコーダの自己アテンションの学習を同時に学習するマルチタスク学習を行う。デコーダに適用するにはマスキングを施すことにより、学習時とテスト時の不整合を解消する。サブワード化に対応するため、単語単位の係り受けをサブワード単位の係り受けに変換する手法を提案した。提案手法は、学習時には係り受け構造を解析する係り受け解析器が必要となるが、係り受け解析も同時に学習されるため、テスト時には係り受け解析器を必要としない。ASPEC データ (Nakazawa et al. 2016) を用いた日英翻訳の評価実験を通じて、提案手法により Transformer モデルの精度改善を達成できることを確認した。

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 *Proc. of ICML 2017*, 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 *Proc. of EMNLP 2015*, pp. 1412—1421.
- E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum. (2018) Linguistically-informed self-attention for semantic role labeling. In *Proc. of EMNLP 2018*, pp. 5027—5038.
- Y. Ding and M. Palmer. (2005) Machine translation using probabilistic synchronous dependency insertion grammars. In *Proc. of ACL 2005*, 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 *Proc. of EMNLP 2017*, pp. 2846—2852.
- A. Eriguchi, Y. Tsuruoka, and K. Cho. (2017) Learning to parse and translate improves neural machine translation. In *Proc. of ACL 2017 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.*, Vol. 26, No. 11, pp. 2132—2141.
- H. Deguchi, A. Tamura, and T. Ninomiya. (2019) Dependency-Based Self-Attention for Transformer NMT. In *Proc. of RANLP2019*.
- K. He, X. Zhang, S. Ren, and J. Sun. (2016) Deep residual learning for image recognition. In *Proc. of CVPR 2016*, pp. 770—778.

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

T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. (2016) ASPEC: Asian scientific paper excerpt corpus. In Proc. of LREC 2016, pp. 2204—2208.

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. of ACL on Interactive Poster and Demonstration Sessions, pp. 177—180.

G. Neubig, Y. Nakata, and S. Mori. (2011) Pointwise prediction for robust, adaptable Japanese morphological analysis. In Proc. of ACL 2011, pp. 529—533.

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

R. Sennrich, B. Haddow, and A. Birch. (2016) Neural machine translation of rare words with subword units. In Proc. of ACL 2016, pp. 311—318.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, (2016) Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.

2.2 Neural Machine Translation based on Multi-Hop Attention

Shohei Iida, Hongyi Cui, Po-Hsuan Hung, Yizhen Wei, Takehito Utsuro
(Graduate School of Systems and Information Engineering, University of Tsukuba, Japan)
Masaaki Nagata
(NTT Communication Science Laboratories, NTT Corporation, Japan)

2.2.1 Introduction

Multi-hop attention was first proposed in end-to-end memory networks[11] for machine comprehension. In this paper, we define a hop as a computational step which could be performed for an output symbol many times. By “multi-hop attention”, we mean that some kind of attention is calculated many times for generating an output symbol. Previous multi-hop attention can be classified into “recurrent attention”[11] and “hierarchical attention”[7]. The former repeats the calculation of attention many times to refine the attention itself while the latter integrates attentions for multiple input information sources. The proposed multi-hop attention for the Transformer is different from previous recurrent attentions because the mechanism for the first hop attention and that for the second hop attention is different. It is also different from previous hierarchical attention because it is designed to integrate attentions from different heads for the same information source.

In neural machine translation, hierarchical attention[2, 7] can be thought of a multi-hop attention because it repeats attention calculation to integrate the information from multiple source encoders. On the other hand, in the Transformer[12], the state-of-the-art model for neural machine translation, feed-forward neural network (FFNN) integrates information from multiple heads. In this paper, we propose a multi-hop attention mechanism as a possible alternative to integrate information from multi-head attention in the Transformer.

We find that the proposed Transformer with multi-hop attention converges faster than the original Transformer. This is likely because all heads learn to influence each other, through a head gate mechanism, in the second hop attention(Figure 1). Recently, many Transformer-based pre-trained language models such as BERT[5] have been proposed and take about a month for training. The speed at which the proposed model converges may be even more important than the fact that its accuracy is slightly better.

2.2.2 Multi-Hop Multi-Head Attention for the Transformer

2.2.2.1 Multi-Head Attention

One of the Transformer’s major successes is multi-head attention, which allows each head to capture different features and achieve better results compared to a single-head case.

$$a^{(h)} = \text{softmax}\left(\frac{Q^{(h)}K^{(h)T}}{\sqrt{d}}\right)V^{(h)} \quad (1)$$

$$\mathbf{m} = \text{Concat}(a^{(1)}, \dots, a^{(h)})W_o \quad (2)$$

Given the query $\mathbf{Q}$, the key $\mathbf{K}$, and the value $\mathbf{V}$, they are divided into each head. Here, $\mathbf{h}(= 1, \dots, \mathbf{H})$ denotes the index of the head, where a is the output of scaled dot-product attention, $\mathbf{W}_o$ is a parameter for a linear transformation, and d is a scaling factor. Finally, the output of multi-head attention, $\mathbf{m}$, is input to the next layer. The calculation of attention using scaled dot-product attention is defined as the first hop (Figure 1).

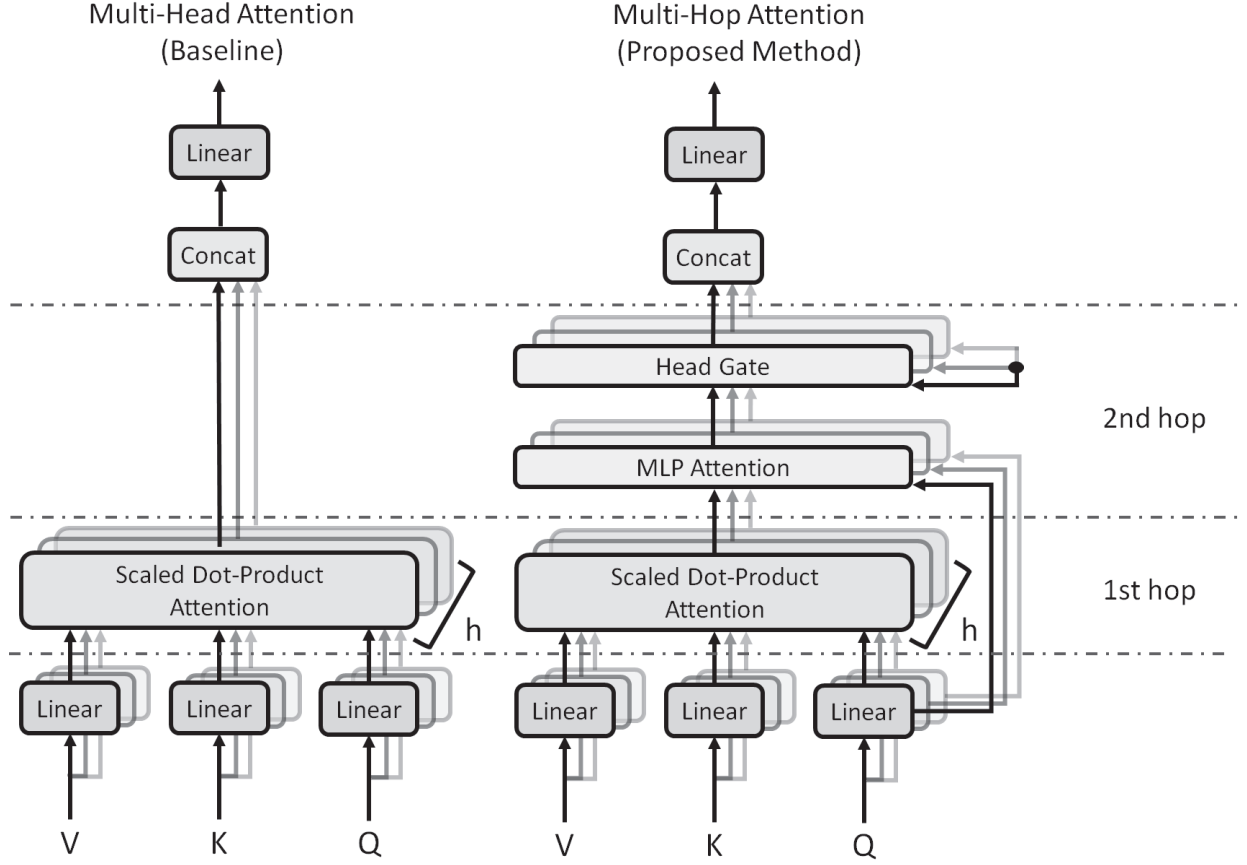


Figure 1: Multi-hop attention

2.2.2.2 Multi-Hop Attention

In the original Transformer[12], information from each head is integrated by simple concatenation followed by a linear transformation. Attention is refined by stacking the combination of self-attention sub-layer and position-wise feed-forward neural network sub-layer. However, as layers are stacked, convergence becomes unstable. Consequently, there is a limit to the iterative approach by layering. Therefore, we propose a mechanism to repeat the calculation of attention based on a mechanism other than stacking layers. The original Transformer is considered to consist of six single-hop attention layers. On the contrary, in the proposed method, some layers have a multi-hop (two-hop) attention. By experiments, we have established the appropriate position of the proposed multi-hop attention in the neural machine translation system. If the number of layers for encoders and decoders are six, then there are six self-attention layers in both the encoder and the decoder, respectively, and six source-to-target attention layers in the decoder.

The first hop attention of the multi-hop attention is equivalent to the calculation of scaled dot-product attention (Equation 1) in the original Transformer. The second hop attention consists of multi-layer perceptron (MLP) attention and head gate, as shown in Figure 1 and the following equations.

$$e_i^{(h)} = v_b^T \tanh(W_b Q^{(h)} + U_b^{(h)} a_i^{(h)}) \quad (3)$$

$$\beta_i^{(h)} = \frac{\exp(e_i^{(h)})}{\sum_{h=1}^H \exp(e_i^{(h)})} \quad (4)$$

$$a_i'^{(h)} = \beta_i^{(h)} U_c^{(h)} a_i^{(h)} \quad (5)$$

First, MLP attention between the output of the first hop, $a_i^{(h)}$, and the query Q , is calculated. Attention is considered as the calculation of a relationship between the query and the key/value. Therefore, in the second hop, attention is calculated again by using the output of the first hop, rather than the key/value.

Equations 4 and 5 are head gate in Figure 1. The head gate normalizes the attention score of each head to $\beta_i^{(h)}$, using the softmax function, where h ranges over all heads. In hierarchical attention[2], the softmax function is used to select a single source from multiple sources. Here, the proposed head gate uses the softmax function to select a head from multiple heads. Finally the head gate calculates new attention, $a_i'^{(h)}$, using the learnable parameters $U_c^{(h)}$, $\beta_i^{(h)}$, and $a_i^{(h)}$. The second hop MLP attention learns the optimal parameters for integration under the influence of the head gate. Although Vaswani et al.[12] reported that dot-product attention is superior to MLP attention, we used MLP attention in the second hop of the proposed multi-hop attention because it can learn the dependence between heads by appropriately tuning the MLP parameters. We conclude that we can increase the expressive power of the network more efficiently by adding the second hop attention layer, rather than by stacking another single-hop multi-head attention layer.

Table 1: Best position for multi-hop

Model	2nd hop	BLEU
		de-en
Baseline	-	36.96
Multi-hop	Encoder, 1st layer	37.53†
Multi-hop	Encoder, 2nd layer	37.35†
Multi-hop	Encoder, 3rd layer	37.36†
Multi-hop	Encoder, 4th layer	37.27†
Multi-hop	Encoder, 5th layer	37.21‡
Multi-hop	Encoder, 6th layer	37.16‡
Multi-hop	Encoder, 1st to 6th layers	36.48
Multi-hop	Encoder, 2nd to 6th layers	36.91
Multi-hop	Encoder, 3rd to 6th layers	37.10
Multi-hop	Encoder, 4th to 6th layers	36.98
Multi-hop	Encoder, 5th and 6th layers	37.06

†($p \leq 0.01$) and ‡($p \leq 0.05$) indicate that the proposed methods significantly outperform the Transformer baseline. The encoder and the decoder each had six layers, respectively.

2.2.3 Experiment

2.2.3.1 Data

We used German-English parallel data obtained from the IWSLT2017¹ and the WMT17² shared tasks. The IWSLT2017 training, validation, and test sets contain approximately 200K, 9K, and 9K sentence pairs, respectively. There are approximately 5.9M sentence pairs in the WMT17 training dataset. For the WMT17 corpus, we used newstest2013 as the validation set and newstest2014 and newstest2017 as the test sets.

For tokenization, we used the subword-nmt tool[10] to set a vocabulary size of 32,000 for both German and English.

¹ <https://sites.google.com/site/iwslt2017/>

² <http://www.statmt.org/wmt17/translation-task.html>

2.2.3.2 Experimental Setup

In our experiments, the baseline was the Transformer[12] model. We used fairseq³ toolkit[6] and the source code will be available at our github repository. For training, we used the Adam optimizer with a learning rate of 0.0003. The embedding size was 512, the hidden size was 2,048, and the number of heads was 8. The encoder and the decoder each had six layers. The number of tokens per batch was 2,000. The number of training epochs for IWSLT2017 and WMT17 were 50 and 10, respectively. In all experiments using the IWSLT2017, models were trained on an Nvidia GeForce RTX 2080 Ti GPU, while in all experiments using the WMT17, models were trained on an Nvidia Tesla P100 GPU.

2.2.3.3 Results

Table 2: Evaluation Result

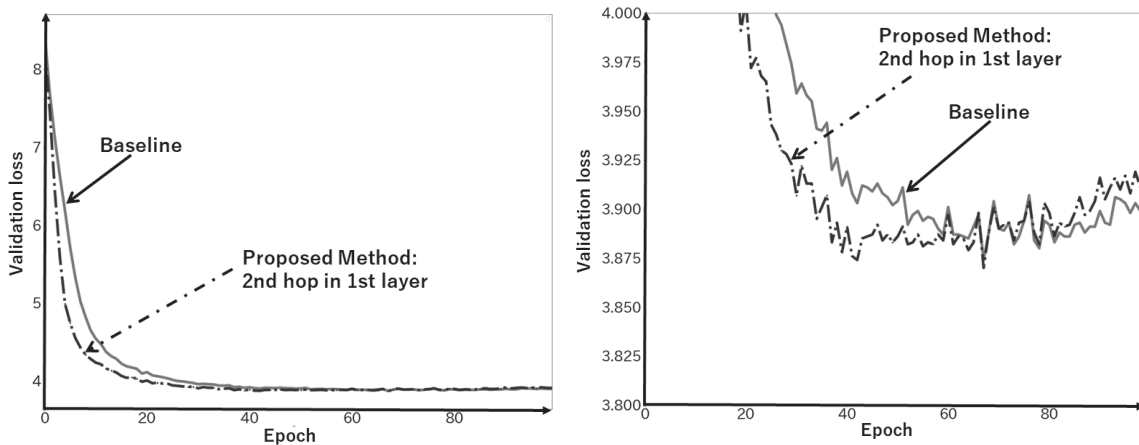
Model	IWSLT2017		WMT17
	de-en	en-de	en-de
Baseline	36.96	30.57	20.03
Multi-hop	37.53 [†]	30.80 [‡]	20.69 [‡]

[†]($p \leq 0.01$) and [‡]($p \leq 0.05$) indicate that the proposed methods significantly outperform the Transformer baseline. The encoder and the decoder each had six layers, respectively.

Results of the evaluation are presented in Tables 1 and 2. In Table 2, the proposed multi-hop attention is used only at the fourth layer in the encoder. In the evaluation of German-to-English translation for IWSLT2017, the proposed method achieved a BLEU score of 37.53, which indicates that it significantly outperforms the Transformer baseline, which returned a BLEU score of 36.96. For WMT17, the proposed method achieved a BLEU score of 20.69, indicating that it also significantly outperformed the Transformer baseline, which returned a BLEU score of 20.03.

In IWSLT2017 German-to-English and English-to-German translation tasks, various conditions were investigated, as shown in Table 1. The best models are shown in Figure 2. The baseline training time was 1143.2s per epoch in IWSLT2017 German-to-English translation, and the training time for the proposed method is 1145.6s per epoch. We found that increasing the number of parameters did not affect training time.

2.2.4 Effect on Learning Speed



³ <https://github.com/pytorch/fairseq>

(a) All learning curve view

(b) Enlarged view

Figure 2: Validation loss by each epoch for IWSLT2017 de-en

Figure 2 shows the validation loss of models for the IWSLT2017 German-to-English translation task with the second hop layers in 1st layer whose dropout rate is 30%, and both models have 6 layers. Figure 2(a) is apparent that the models with the second hop converge faster than the baseline model. Figure 2(b) is an enlarged view of Figure 2(a), focused on the lowest validation loss for different models. We find that the validation loss is lower when there are fewer second hop attentions.

2.2.5 Related Work

The mechanism of the proposed multi-hop attention for the Transformer was inspired by the hierarchical attention in multi-source sequence-to-sequence model[7]. The term “multi-hop” is borrowed from the end-to-end memory network[11] and the title “attention over heads” is inspired by Attention-over-Attention neural network[3], respectively.

Ahmed et al.[1] proposed Weighted Transformer which replaces multi-head attention by multiple self-attention branches that learn to combine during the training process. They reported that it slightly outperformed the baseline Transformer (0.5 BLEU points on the WMT 2014 English-to-German translation task) and converges 15-40% faster. They linearly combined the multiple sources of attention, while we combined multiple attention non-linearly using softmax function in the second hop.

It is well known that the Transformer is difficult to train[9]. As it has a large number of parameters, it takes time to converge and sometimes it does not do so at all without appropriate hyper parameter tuning. Considering the experimental results of our multi-hop attention experiments, and that of the Weighted transformer, an appropriate design of the network to combine multi-head attention could result in faster and more stable convergence of the Transformer. As the Transformer is used as a building block for the recently proposed pre-trained language models such as BERT[5] which takes about a month for training, we think it is worthwhile to pursue this line of research including the proposed multi-hop attention.

Universal Transformer[4] can be thought of variable-depth recurrent attention. It obtained Turing-complete expressive power in exchange for a vast increase in the number of parameters and training time. As shown in Table 4, we have proposed an efficient method to increase the depth of recurrence in terms of the number of parameters and training time. Recently Voita et al.[13] and Michel et al.[8] independently reported that only a certain subset of the head plays an important role in the Transformer. They performed analyses by pruning heads from an already trained model, while we have proposed a method to assign weights to heads automatically. We assume our method (multi-hop attention or attention-over-heads) selects important heads in the early stage of training, which results in faster convergence than the original Transformer.

2.2.6 Conclusion

In this paper, we have proposed a multi-hop attention mechanism for a Transformer model in which all heads depend on each other repeatedly. We found that the proposed method significantly outperforms the original Transformer in accuracy and converges faster with little increase in the number of parameters. In future work, we would like to implement a multi-hop attention mechanism to the decoder side and investigate other language pairs.

References

- [1] K. Ahmed, N. S. Keskar, and R. Socher. 2018. Weighted transformer network for machine translation. arXiv preprint arXiv:1711.02132
- [2] R. Bawden, R. Sennrich, A. Birch, and B. Haddow. Evaluating discourse phenomena in neural machine translation. In *Proc. NAACL-HLT*, pp. 1304-1313, 2018.
- [3] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proc. 55th ACL*, pages 593–602.
- [4] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser. 2019. Universal transformers. In *Proc. 7th ICLR*.
- [5] 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.
- [6] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proc. 34th ICML*.
- [7] J. Libovický and J. Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proc. 55th ACL*, pp. 196-202, 2017.
- [8] P. Michel, O. Levy, and G. Neubig. 2019. Are sixteen heads really better than one? arXiv preprint arXiv:1905.10650.
- [9] M. Popel and O. Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.
- [10] R. Sennrich, B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. 54th ACL*, pages 1715–1725.
- [11] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. 2015. End-to-end memory networks. In *Proc. 28th NIPS*, pages 2440–2448.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. 30th NIPS*, pp. 5998-6008, 2017.
- [13] 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 特許請求項定型パターンの分類と解析

山形大学 名誉教授

横山 晶一

2.3.1 はじめに

特許文の請求項が一般的に複雑な長文で解釈しにくい、その中に決まったパターンが存在し、そのパターンを利用すれば、長文を比較的容易に分割したり、係り受けを明確化できるということは、これまで繰り返し示してきた[1]。

また、同様のことが、外国語の特許についても言えることもすでに述べた。

A liquid level detecting device **comprising**:
a sensor housing;
a holder which is held rotatably by the sensor housing;
a float arm which is fixed to the holder;
a float which is attached to a tip portion of the float arm and varies in position following a surface level of liquid stored in a tank;
a magnet provided in the holder; and
a Hall element which is provided in the sensor housing and detects a displacement of the magnet of the holder,
wherein the magnet is a plastic magnet in which magnetic powder and a resin are mixed;
and
wherein a surface of the plastic magnet is coated by a coating layer.

図 1 (a) 米国特許の請求項の例 (US20180348037A1)

Flüssigkeitsstand-Erfassungsvorrichtung, **umfassend**:ein Sensorgehäuse;ein Halter, der durch das Sensorgehäuse drehbar gehalten ist;einen Schwimmerarm, der an dem Halter befestigt ist;einen Schwimmer, der an einem Spitzenabschnitt des Schwimmerarms angebracht ist und seine Position ändert, indem er einem Oberflächenstand einer Flüssigkeit, die sich in einem Behälter befindet, folgt;einen Magneten, der in dem Halter vorgesehen ist; und ein Hall-Element, das in dem Sensorgehäuse vorgesehen ist und eine Verschiebung des Magneten des Halters erfasst,wobei der Magnet ein Kunststoffmagnet ist, **in dem** ein magnetisches Pulver und ein Harz vermischt sind; und **wobei** eine Oberfläche des Kunststoffmagneten mit einer Beschichtungsschicht beschichtet ist.

図 1 (b) ドイツ特許の請求項の例 (DE102018208795A1)

センサハウジングと、前記センサハウジングに回転可能に保持されたホルダと、前記ホルダに固定されたフロートアームと、前記フロートアームの先端に設けられ、タンク内に貯留された液体の液面に追従して変位するフロートと、前記ホルダに設けられたマグネットと、前記センサハウジングに設けられ、前記ホルダのマグネットの変位を検出するホール素子と、を備え、前記マグネットは、樹脂に磁粉が混入されたプラスチックマグネットであり、表面がコーティング層によって覆われていることを特徴とする液位検出装置。

図 1 (c) 日本特許の請求項の例 (JP2018205135A)

図 1 は、パテントファミリーをなす、同じ特許の英語、ドイツ語、日本語による請求項 1 を示したものである。これは、「液位検出装置」に関する特許申請の例で、一般的には、パテントファミリーをなすからといって、請求項の記述が必ずしも似たものになるとは限らないが、図 1 の(a)～(c)はほぼ同じ内容になっている。

図 1(a)で、“comprising” (含む) という現在分詞で以下を規定する部分は、(b)のドイツ語では全く同様に、“umfassend” (含む) という、同じく現在分詞で規定され、(c)の日本語では、「特徴とする」という句で締めくくられている。日本語名詞句の係り部分に機能語がない場合の、請求項のある種の典型的なパターンの一つである。英語やドイツ語特許でも、詳しくは延べないが、出現する動詞は異なるものの、典型的なパターンの一つであると言える。

内部の係りの句を見ると、この英語の例では、不定冠詞 “a” を改行やセミコロンまで入れて羅列し、それに係る名詞句には、定冠詞 “the” を用いている。ドイツ語特許では、改行やセミコロンこそないが、不定冠詞や定冠詞を英語と同様に用いている。日本語では、定冠詞の部分に「前記」を用いて、それぞれの名詞を並列助詞「と」で羅列する形を取っている。

英語の最後の部分では、関係副詞 “wherein” を用いて前の名詞句を修飾する形を、同様に、ドイツ語では、関係代名詞句 “in dem” や、関係副詞 “wobei” を用いて前の名詞句を規定している。日本語の場合には、この部分は比較的長めの修飾句となることが多い。

このような各国別の特許の特徴を比較し、自動翻訳結果と比較するのも興味あることであるが、本稿ではこれ以上立ち入らない。

今日、自然言語処理や機械翻訳の分野では、大量のデータをニューラルネットワークを用いて処理する手法が主流であり、Transformer や、そのある種の応用ともいえる BERT を用いた研究が盛んに行われている。研究のペースが加速しており、次々に新しい手法が現れて、全体像がなかなか見えにくいのも事実であるが、ある程度全体の動向をとらえた論文も見られる (たとえば [2])。

本稿では、請求項に見られるパターンを引き続き人手で調査し、その結果について報告する。ニューラルネットワーク全盛の時代に、このようなやり方はなじまないかもしれないが、たとえば、学習済みのネットワークを別の少数のラベル付きデータに適用して精度を高めようという、ファインチューニングの手法なども提案されており [3]、本稿のような研究データをこのような手

法に用いる可能性も考えられる。

本稿は、主としてこれまでの研究[4]に基づくものであるが、データを増強し、例文を代え、内容を改めたものである。

2.3.2 資料・調査方法

例年と同様に、2003 年の特許公開情報（特開 2003-180995～181100）の中から、請求項 1 の記載が 120 文字以上のものを 100 文選び出した。表 1 にその内訳を示す。表には、合わせて、前回まで[4]の調査 550 文（特開 2003-180203～180994）の内訳も示す。

表 1 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	前回まで	232	185	83	50
	今回	21	34	23	22
	合計	253	219	106	72

今回調査した 100 文は、前回同様、遊技機に関するものが非常に多く、以前のものに比べて長い文が多い印象がある。最も短い請求項で 131 文字、長いものでは、一文で 886,992 文字といったものがあり、平均でも 300 文字を超える長さであった。

以下で示す調査項目は以前と同じである。調査項目別に内容を示す。

2.3.3 調査結果

調査した特許は、公開からかなりの年数が経過しているが、以下で見るように、最近の特許請求項（たとえば図 1(c)）の書き方とほとんど変わっていないので、調査項目や調査の対象を変える必要はないと考えられる。以下に、項目ごとの調査結果を示す。

(a) 機能語「であって、」、「において、」による句の分割

請求項に含まれる文の典型は、図 1(c)のように、最後に「特徴とする」(b)で述べる）が来る長い名詞句になるものであるが、その多くが、すでに述べたように、機能語「であって、」や「において、」を伴って記述される。しかもその場合に句の前後で名詞句が分割される場合が多いことは、すでに述べてきた。つまり、

- ・[名詞句 A] であって、～する（した）[名詞句 A]
- ・[名詞句 A] において、～する（した）[名詞句 A] (1)

の形になっているときに、読点で名詞句を二つに分割できるケースが非常に多いということを繰り返して述べてきた。

図 2 に、「であって、」を含む典型例、図 3、図 4 に、「において、」を含む典型例、非典型例をそれぞれ示す。

複数の識別情報画像が変動表示又は停止表示される表示部を備え、前記表示部において、前記複数の識別情報画像のうち少なくとも一つ以上の識別情報画像が常時変動表示され、遊技者の発射するパチンコ球が所定の位置を通過又は入球したことを契機に、前記常時変動表示されている識別情報画像が停止表示されるパチンコ遊技機であって、前記常時変動表示されている識別情報画像が停止表示されたとき又は停止表示された後に、新たに常時変動表示されている識別情報画像を表示させることを特徴とするパチンコ遊技機。

図 2 「であって、」を含む典型例（特開 2003-181014）

遊技板を含む遊技機本体を、枠体に着脱可能に組み付けてなり、その遊技機本体を、遊技内容が異なる他の遊技機本体に組み替え可能とした弾球遊技機において、前記枠体は、遊技内容が異なる複数種類の遊技機本体に、共通使用される共通部材と、前記枠体に組み付けた遊技機本体の種類に応じて固有使用されて、前記共通部材に着脱可能に取り付けられる非共通部材とから構成されたことを特徴とする弾球遊技機。

図 3 「において、」を含む典型例（特開 2003-181076）

図柄表示装置に複数の識別図柄を変動状態および変動停止状態で順に表示し、複数の識別図柄の変動停止時の態様に基づいて大当りの抽選結果を報知する構成のものにおいて、前記図柄表示装置の画面の今後の態様を予告する予告パターンを表示する表示制御手段を備え、前記表示制御手段は、複数の識別図柄が変動開始してから所定の期間内に同一系統の予告パターンを複数のタイミングで表示することを特徴とする遊技機。

図 4 「において、」を含む非典型例（特開 2003-181002）

典型例とは、図 2, 3 に示すように、上の形の機能語と「特徴とする」から成る例、非典型例（図 4）とは、機能語の前の名詞句と、「特徴とする」の後の締めくくりの名詞句とが異なるものを示す。図 2 で示すように、機能語「であって、」と「において、」がともに、あるいは複数出現する場合があるが、その場合でも句の分割には一つしか寄与しない。

表 2 には、機能語の統計を示す。表で、「前回まで」とあるのは、以前の 550 文であり、今回の 100 文と合わせて示してある。今回、図 2 に示すように、句の分割に直接寄与しない機能語は、分割に寄与する機能語とともに使われた例がすべてであり、請求項全体が、この機能語を含むのに分離できない例はなかったので、この表からは、前回までにのせてあった、句の分割にかかわらない機能語の統計は省いてある。

表 2 機能語別の統計

		典 型	非典型
であって、	前回まで	141	21
	今回	26	4
	計	167	25
において、	前回まで	163	21
	今回	51	4
	計	214	25

(b) 「特徴とする」の調査

「特徴」という語は、今回調査した 100 文中に 89 例出現し、そのすべてが最後の名詞句の前に置かれたものであった。そのうち、「特徴とする」は 87 例（そのうちで、「特徴とする、」という形で、最後の名詞句の前に読点があるものが 2 例）、「特徴とした」が 2 例であった。多くは(a)で述べた「であって、」、「において、」といった機能語を伴っているが、すべてがこれらを伴うものではないことは、従来の調査と矛盾しない。機能語とのかかわりについては、4 節で述べる。

(c) 助詞「と」による並列構造

背景図と共に図柄の変動表示が可能な表示手段と、
始動条件の成立に基づいて前記表示手段による前記図柄の変動表示の開始を指示する変動開始指示手段と、
通常時には複数の異なる状況を連続させてなる背景図内を所定の条件に基づいて移動する移動ポインタと該移動ポインタを含む前記背景図の少なくとも一部とが前記表示手段に表示されるよう該表示手段を制御し、前記変動開始指示手段により変動表示の開始の指示がなされたときには前記移動ポインタがポインティングしている前記背景図における状況に応じた変動態様で前記図柄が前記表示手段に変動表示されるよう該表示手段を制御する表示制御手段と、
を備える遊技機。

図 5 助詞「と」による並列構造の例（特開 2003-181040）

図 1(c)にも示したように、名詞句の並列構造を助詞「と」で担わせる請求項は多い。今回調査した 100 文でも、約半数の 52 文にこのような例が見られた。その一つの例を図 5 に示す。この例は、機能語も「特徴」も含まない比較的珍しい例であるが、助詞「と」で、並列構造が比較的確に表されている。「と、」の後で改行している例も多い。

(d) 動詞による並列構造

遊技盤が装着される本体部と、その本体部に対し一側が開閉機構によって開閉可能に取り付けられ、他側が閉止機構によって閉じ状態に保持される開閉扉と、を備え、前記開閉扉には、前記遊技盤の遊技領域を透視可能な開口窓が形成されるとともに、その開口窓を塞ぐ透明板が設けられている遊技機であって、
前記遊技盤前面と前記透明板後面との間の距離を検知する距離検知手段を備えている遊技機。

図 6 動詞による並列構造の例（特開 2003-181085）

並列構造に、「備える」、「有する」、サ変動詞などの動詞が関わっている場合も非常に多い。図 6 に、動詞による並列構造の例を示す。今回の調査では、100 文中の約 7 割が、動詞が並列構造の修飾句に関わっている。その多くは、助詞「と」とも併用している場合が多い。

(e) 改行による係り受けの明確化

すでに示した多くの例では、修飾句の区切り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にしている場合も多い。今回も図 2～6 を含む 85 例がそれに相当する。中には、改行を適当に用いることによって、かえって係り受けを不明確にしている場合もあるが、多くはこの情報を用いることによって、正しい意味関係の解析に寄与することが可能である。

(f) 「前記」などによる階層構造

すでに上のいくつかの図で示したように、「前記」を用いる階層構造も多い。今回調査した 100 文中、図 2～6 を含む 9 割弱が「前記」を複数用いている。今回は、「該」や「上記」といった表現はあまり見られなかった。ただ、すでに述べたように、「前記」という表現は、句の中のさまざまなレベルで使われるので、たとえば図 1(c)のように、英語の“the”との明確な対照を詳細に行わない限り、系統的な解析が困難な実情がある。

2.3.4 機能語と「特徴」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

	であって		において		特徴 含まず
	典	非	典	非	
前回まで	59	9	93	8	42
今回	20	4	50	4	11
計	79	13	143	12	53

機能語と「特徴」の組み合わせパターンを表3に示す。表で、「前回まで」とは、これまでに調査した300文についての結果である。今回は、新たに100文について調査した結果を加えてある。

2節の(1)で示したパターンとの違いは次のようである。すなわち、

- ・[名詞句 A] であって、～を特徴とする [名詞句 A]
- ・[名詞句 A] において、～を特徴とする [名詞句 A] (2)

で、後ろの名詞句の直前が、「特徴とする」になっているものを、表3の「典」の列にカウントしてある。ここでは、「特徴とした」はカウントせず、後ろに読点のついた「特徴とする、」はカウントに入れてある。すでに示した図2、図3が「典」に相当する。表2と表3を比較すると、今回対象とした文の中で、典型例だが「特徴とする」がないものは、「であって」で6、「において」で1ある。また、そもそも機能語も「特徴とする」もない例(図5,6)が4ある。これらを合わせた11例が表3の「特徴含まず」にカウントされている。機能語がなく、「特徴とする」が含まれる例(調査の対象ではないが、図1(c)はこの例に相当する)は、10例あるが、それはこの表にはカウントされていない。

これまで述べてきたのと同様に、請求項の過半数が(2)のようなパターンを持っている。

2.3.5 終わりに

人手で継続的に調査を行うことによって、大量データの統計処理では見落とされがちな細かい点がある程度解明できると考えて研究を行ってきた。さらにデータを精査して、ファインチューニングなどの知見に貢献できればよいと考えている。

また、1節に述べたように、パテントファミリーと機械翻訳との比較対照も機会があれば行ってみたい。

参考文献

[1] 横山晶一：特許文請求項からの定型パターンの抽出と調査、2018年度 AAMT/Japio 特許翻訳研究会報告書 (2019) pp.25-32,

http://aamtjapio.com/kenkyu/files/kenkyu05/AAMT_Japio_20190422.pdf

[2] 須藤克仁：ニューラル機械翻訳の進展—系列変換モデルの進化とその応用—、人工知能 Vol. 34, No.4 (2019) pp.437-445

[3] Alberto Poncelas, Gideon Maillette de Buy Wenniger and Andy Way: Transductive Data-Selection Algorithms for Fine-Tuning Neural Machine Translation, PSLT2019 (The 8th Workshop on Patent and Scientific Literature Translation) (2019)

[4] 横山晶一：特許請求項定型文の分類と解析、Japio Year Book 2019, pp.288-291 (2019)

2.4 BERT によるテキスト分類における折り返し翻訳による

データ拡張の効果の検討

静岡大学 福田 治輝 綱川 隆司 大島 純
大島 律子 西田 昌史 西村 雅史

2.4.1 はじめに

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) は自己注意機構のみを使用した Transformer をベースとする汎用言語表現モデルであり、適用する自然言語処理タスクにファインチューニングすることで多くのタスクにおいて従来モデルより高い性能を示すことが知られている。しかしファインチューニングに用いるデータ量が十分でないときはその限りではない。本稿では協調学習場面における学習者の調整能力評価手法 CSSER におけるテキスト評価を BERT による分類器を用いて推定する際に、既存の評価済みテキストを折り返し翻訳によってデータ拡張したときの効果について検討した。

2.4.2 研究背景と目的

近年、教育現場にて主体的な学習の重要性が認知されてきている。これを達成するための一つの方法としてグループワークが取り入れられており、グループワークにおいては協調学習すなわち「複数の学習者が各自の理解や考え方を交換し合い、他人から提供されたアイディアを加味しつつ一般的で抽象度の高い知識を自分自身で作り上げていく学習」(三宅, 2008) の形態がとられる。協調学習においては複数のメンバーで協力してうまく学習を進めるためには、課題解決能力や知識の共有といった社会認識的調整と、グループメンバー間の人間関係に関する調整を行う社会感情的調整が必要となる。CSSER (Collaboration Scenario-based Scale for Emotion Regulation) (Oshima and Oshima, 2015) はこれらの調整能力を評価するために開発された手法である。

CSSER では協調学習で起こるあるシナリオをマンガ形式で評価対象者に見せ、シナリオの特定の状況において学習者がどのような思考や発言をするかをテキストで記入させ、これをルーブリックに基づいて人手で評定することにより社会認識的調整能力および社会感情的調整能力の 2 つの側面から評価を行う。図 1 に CSSER のシナリオの例と記入すべき吹き出しを示す。



図 1 CSSER のシナリオ例

CSSER による学習者の評価をグループワークの事前と事後に行うことで、あるグループワークにおける調整能力向上効果の測定や授業の改善に寄与することなどが期待されるが、一方で評価結果を直接フィードバックするのは評価を人手で行うための時間とコストから困難である。

本稿では、BERT によるクラス分類により学習者の回答テキストの自動評定を行うことで、評定結果の即時のフィードバックを可能にすることをねらう。ここで、学習に用いるデータはそれまでに人手で評定を行った学習者の回答テキストであり、BERT のファインチューニングに用いるデータとして必ずしも十分な量のデータが存在しない。

そこで、本稿では折り返し翻訳による言い換え表現の拡張 (Sun and Zhou, 2012; Mallinson et al., 2017) によりトレーニングに用いるデータの拡張を試みた。この方法の利点は機械翻訳の構築に用いた対訳コーパスに別言語を介して現れる言い換え表現を得られるところにある。ただし、機械翻訳を 2 回適用することにより、言い換えとはいえない文や文法的に正しくない文も相当数得られることになる。しかし本タスクにおけるデータ拡張の目的は、既存のデータに含まれる回答の表現に対して BERT のベースモデルのみでは得られない言い換え表現をファインチューニングのためのデータに加えることであり、必ずしも意味的または文法的に正確な言い換え表現を得る必要はない。このため、本稿では折り返し翻訳によって得られたテキストを言語モデルから得られるパープレキシティによりフィルタリングしてデータ拡張に用いることとした。

2.4.3 BERT によるテキストのクラス分類

日本語 Wikipedia 記事 (約 1800 万文) によってプレトレーニングされたモデル (柴田ら, 2019) を用いて、BERT への入力を CSSER の「心の声」または「発言」のテキスト、出力を社会認知的側面または社会感情的側面の 1~5 点の評定とし、それぞれクラス分類を行うようにファインチューニングを行う。

日本語 Wikipedia 記事のテキストは、半角文字を全角文字に正規化し JUMAN++ で分かち書きを行ったあとバイトペアエンコーディング (Sennrich et al., 2016) を適用しサブワードに分割され、プレトレーニングが行われている。ファインチューニングの際の入力も同様に、全角文字への正規化後に JUMAN++ (Morita et al., 2015) で分かち書きを行い、プレトレーニングの際に生成された語彙リストを用いてサブワードに分割され、BERT モデルに入力される。

BERT のクラス分類タスクにおいてテストデータの予測結果はそれぞれのクラスに対して、どれほどそのクラスらしいかという尤度が出力される。本タスクでは、1 から 5 の各評定について求めた尤度が最大である評定を、自動評定の結果として採用する。

2.4.4 折り返し翻訳による言い換えテキストの生成

Google 翻訳 API の v3 NMT モデル¹を利用し、CSSER の各回答テキストを日本語から韓国語へ翻訳を行い、再度それぞれ日本語へ翻訳を行うことで元の回答テキストの言い換えテキストを生成する。なお、予備実験として英語および中国語を介した折り返し翻訳も行ったが、韓国語を用いた場合にデータ拡張の効果がみられたため本稿では韓国語を採用した。

ここで、日本語として自然な折り返し翻訳テキストを選別するため、BERT as Language Model²を利用し折り返し翻訳された個々のテキストに対してパープレキシティを算出し、パープレキシティが相対的に小さいテキストを言い換えテキストとして用いる。BERT as Language Model においても、同様に日本語 Wikipedia 記事（約 1800 万文）によってプレトレーニングされたモデル（柴田ら, 2019）を用いる。

2.4.5 実験

2.4.5.1 実験設定

グループワークが主となる静岡大学の演習系講義の 2015 年度から 2017 年度の受講生に対して、初回講義の冒頭および最終回講義の冒頭の 2 回 CSSER を実施し、「心の声」と「発言」合わせて 10293 件の回答が得られた。これらのシナリオ別の回答データおよび人手での評定を実験対象データとする。

実験対象となるデータを 3:1:1 の割合でランダムにトレーニングセット、バリデーションセット、テストセットに分割し、5 分割交差検証を行う。ベースラインとして、「心の声」と「発言」のテキストを Doc2Vec (Le and Mikolov, 2014) でベクトル化し、コサイン類似度が最も大きくなるものの評定を自動評定の値として採用する手法と比較する。言い換えテキストをトレーニングデータに追加した場合の検証に関しては、トレーニングセットから生成した言い換えテキストのうちパープレキシティの値の小さいものから順に、25%, 50%, 75%, 100%の割合でトレーニングデータに含める場合と、全く含めない場合で評価実験を行った。

評価指標として、Cohen の一致係数（以下 κ 係数と呼ぶ）を用いる。 κ 係数は、2 人以上の評定者の評価の一致率を表す係数であり、その評定基準の信頼性を求めることができる指標である

3. BERT の主なパラメーターとして、語彙サイズ (サブワード含む) 32,000, バッチサイズ 32, エポック数 3 を用いた。

2.4.5.2 実験結果と考察

¹ <https://cloud.google.com/translate/docs?hl=ja>

² <https://github.com/xu-song/bert-as-language-model>

³ Fleiss (1981) によれば、0.75 を超える κ 値で概ね優れた一致を示すという目安が示されている。

BERT および Doc2Vec を用いたクラス分類による各評定の κ 係数を比較した結果を表 1 に示す.

表 1 BERT と Doc2Vec による手法の評定性能 (κ 係数)

評価対象		手法	シナリオ 1	シナリオ 2	シナリオ 3	シナリオ 4	平均
社会 認識的側面	発言	BERT	0.409	0.592	0.408	0.408	0.454
		Doc2Vec	0.091	0.289	0.132	0.137	0.173
	心の声	BERT	0.332	0.346	0.333	0.386	0.349
		Doc2Vec	0.099	0.115	0.150	0.088	0.111
社会 感情的側面	発言	BERT	0.457	0.350	0.475	0.376	0.415
		Doc2Vec	0.143	0.239	0.173	0.135	0.162
	心の声	BERT	0.409	0.392	0.334	0.377	0.378
		Doc2Vec	0.102	0.069	0.145	0.127	0.113

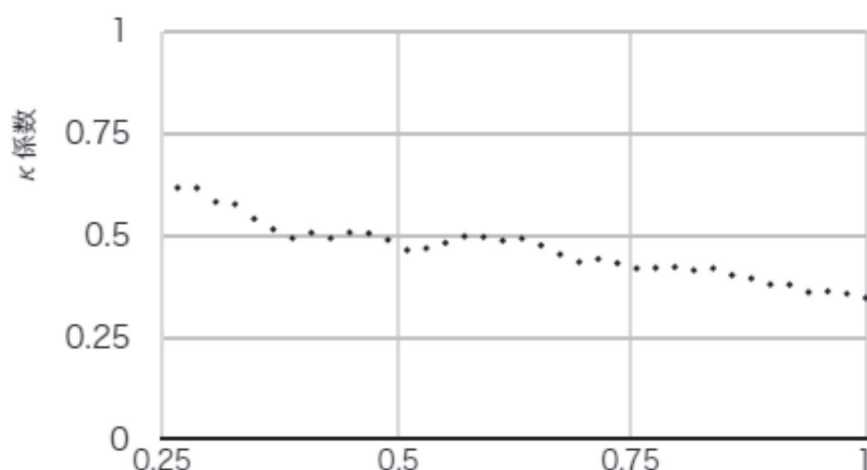


図 2 評価対象に用いたデータ量 (高尤度順) と評定性能 (κ 係数)

全ての結果において, BERT による評定は Doc2Vec による評定よりも高い性能となった. 全体的な傾向として, 心の声より発言の方が高い性能で評定できている.

BERT によって出力される 1 から 5 の各評定について個別で評定される尤度は回答データ毎にばらつきがあり, 一つの評定に対して非常に高い尤度がつくものから, 上位の複数の評定が近い値の尤度になる場合も見られる. そこで, テストセットの中で出力した評定の尤度で降順にソートを行い, 上位の回答データのみで κ 係数の計算を行った. 縦軸が κ 係数, 横軸が κ 係数を算出するのに利用したデータの割合として, シナリオ 2 の発言データの社会感情的側面のデータを用いて計算を行った結果を図 2 に示す. 尤度が高いテキストに絞るほど, 高い評定性能が得られるといえる.

表 2 トレーニングデータに追加するデータの割合と評定性能 (κ 係数)

評価 対象	データ 追加割合	発言				心の声			
		シナリオ 1	シナリオ 2	シナリオ 3	シナリオ 4	シナリオ 1	シナリオ 2	シナリオ 3	シナリオ 4
社会 認識的 側面	0%	0.457	0.350	0.475	0.376	0.409	0.392	0.334	0.377
	25%	0.446	0.389	0.460	0.349	0.428	0.382	0.336	0.342
	50%	0.492	0.433	0.496	0.352	0.459	0.339	0.331	0.345
	75%	0.509	0.382	0.504	0.378	0.429	0.392	0.348	0.380
	100%	0.444	0.425	0.549	0.351	0.414	0.329	0.342	0.419
社会 感情的 側面	0%	0.409	0.592	0.408	0.408	0.332	0.346	0.333	0.386
	25%	0.379	0.592	0.421	0.383	0.355	0.360	0.339	0.357
	50%	0.437	0.646	0.435	0.382	0.342	0.333	0.372	0.380
	75%	0.383	0.638	0.426	0.368	0.346	0.373	0.340	0.341
	100%	0.364	0.611	0.429	0.404	0.346	0.367	0.327	0.357

次に、言い換えテキストをトレーニングデータに追加した場合の結果を表 2 に示す。社会認識的側面、社会感情的側面のいずれの側面においても、多くの設定において言い換えテキストをトレーニングデータとして含めた場合の方が高い κ 係数が得られた。また、全体的には折り返し翻訳のうち Perplexity の大きいテキストを取り除くことは有効であると考えられる。

2.4.6 まとめと今後の課題

本稿では、協調学習内で行われる調整活動プロセスにおける調整能力の評価手法 CSSER におけるテキストの自動評定タスクにおいて、BERT による分類器を用いた性能評価を行い、また折り返し翻訳を用いた言い換えテキストをトレーニングデータに追加した場合の効果を検討した。実験の結果、BERT が出力する尤度が高いテキストについては人手による評定と同程度の κ 係数を得ることができた。また、折り返し翻訳を用いた言い換えテキストの追加により κ 係数が改善することが確認できた。

今後はモデルの改善等により評定性能のさらなる向上を目指し、人手での評定と併用する場合、類似テキストの提示等の評定支援の機能に関しても今後検討していく。

参考文献

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.
- Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley.
- Le, Q. and Mikolov T. (2014). Distributed Representations of Sentences and Documents. In *Proc. of ICML 2014*, pp. 1188-1196.
- Morita, H., Kawahara, D., and Kurohashi, S. (2015). Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model. In *Proc. of EMNLP 2015*,

- pp. 2292-2297.
- Oshima, R. and Oshima, J. (2015). Collaboration Scenario-based Scale for Emotion Regulation: Measuring Learners' Agency to Regulate Own, Others' and Group Emotions. In *Proc. of EdMedia + Innovate Learning 2015*, pp. 796-801.
- Sennrich, R., Haddow, B., and Birch, A.. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proc. of ACL 2016*, pp. 1715-1725.
- Sun, H. and Zhou M. (2012). Joint Learning of a Dual SMT System for Paraphrase Generation. In *Proc. of ACL 2012, short papers*, pp. 38-42.
- 柴田知秀, 河原大輔, 黒橋禎夫. (2019). BERT による日本語構文解析の精度向上. 言語処理学会 第 25 回年次大会発表論文集, pp. 205-208.
- 三宅なほみ. (2008). 協調的な学習と AI. 人工知能学会誌, Vol.23, No.2, pp. 174-183.

3. 機械翻訳評価手法

3.1 WMT2017 データにおける単語の位置情報と分散表現に基づく

自動評価法の有効性について

北海学園大学 越前谷 博

3.1.1 はじめに

近年、ニューラルネット翻訳の急速な進展に伴い、自動評価法においても従来の表層的な情報に基づくアプローチから単語及び文ベクトルに基づく意味情報を利用したアプローチへと変わりつつある。その傾向は国際会議 Second Conference on Machine Translation^[1]（以降、WMT17 と記す）の自動評価タスク（Metrics Task）^[2]に参加している自動評価法からも見てとれる。このような状況において、本報告では WMT17 の自動評価タスクのデータを用いて、従来より提案している単語の位置情報と分散表現に基づく自動評価法^[3]を含めた様々な自動評価法によるメタ評価を行う。そして、表層情報に基づく自動評価法と他の意味情報を利用した自動評価法との比較に基づき、提案手法の有効性について述べる。

WMT では従来よりベースラインとして 7 つの表層情報に基づく自動評価法（BLEU、SentBLEU、NIST、WER、PER、TER、そして、CDER）が使用されている。また、WMT17 においては、7 つのベースラインに対して 9 つのチームが提案する自動評価法を加えて、メタ評価が行われた。本報告ではこれらの自動評価法に加えて、これまでに提案している 2 つの自動評価法を含めたメタ評価を行う。1 つは単語の位置情報と分散表現に基づく自動評価法^[3]である。提案手法では、訳文と参照訳中の単語を分散表現に変換する。そして、変換後の訳文と参照訳を 2 つの分布とみなし、2 つの分布間の距離を求める Earth Mover's Distance（EMD）を用いることで訳文と参照訳間の類似度を求める。また、EMD を用いる際には単語の位置情報を反映させることで言語文に適合させる。もう一つは表層情報に基づく自動評価法である IMPACT（Recursive Acquisition of Intuitive comMon Parts ConTinuum）^[4]である。IMPACT は訳文と参照訳間の表層的な共通部分をチャンクとし、チャンクの長さに重みづけを行う。さらに、チャンクの出現位置の情報をパラメータとすることで語順の違いをスコアに反映させる。本報告ではこれらの自動評価法を含めた様々な自動評価法によるメタ評価を WMT17 の自動評価タスクデータを用いて行い、その結果より提案手法の有効性について述べる。

3.1.2 単語の位置情報と分散表現に基づく自動評価法

3.1.2.1 Earth Mover's Distance（EMD）の概要

提案手法では訳文と参照訳の構成単語を単語の分散表現モデルを用いてベクトルに変換する。そして、変換後の文を 2 つの分布とみなし、EMD に基づいて分布間の距離を求める。EMD は輸送問題における仕事量 $WORK$ を最小化するフロー f_{ij} を求めることで得られる。仕事量 $WORK$ は以下の式（1）より得られる。 d_{ij} は特徴量間の距離を示す。

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

また、フロー f_{ij} に対しては4つの制約が設けられている。それぞれを式(2)から式(5)に示す。

$$f_{ij} \geq 0 \quad 1 \leq i \leq m, \quad 1 \leq j \leq n \quad (2)$$

$$\sum_{j=1}^n f_{ij} \leq w_{p_i} \quad 1 \leq i \leq m \quad (3)$$

$$\sum_{i=1}^m f_{ij} \leq w_{q_j} \quad 1 \leq j \leq n \quad (4)$$

$$\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min \left(\sum_{i=1}^m w_{p_i}, \sum_{j=1}^n w_{q_j} \right) \quad (5)$$

式(2)は必ず P から Q に輸送することを示している。式(3)は P にある荷物以上は輸送できないことを示し、式(4)は Q の容量以上は受け付けることができないことを示している。また、式(5)は輸送量の上限は P と Q の荷物の総容量の小さい方となることを示している。そして、最終的な P と Q の距離は以下の式(6)より得る。式(6)はEMDが仕事量 $WORK$ を最小化するフローを求めることで得られることを意味している。

$$EMD(P, Q) = \frac{\min(WORK(P, Q, F))}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (6)$$

次いで、EMDの計算例について述べる。ここでは説明を簡略化するために、2つの分布 P と Q を構成する特徴量を2次元空間の座標 (x, y) とする。分布 P と Q はそれぞれ表1と表2に示すように2次元座標を特徴量に持ち、それぞれの特徴量には重み w が付与されている。

表1 2次元空間における特徴量 P の例

P_1		P_2		P_3		P_4	
p_1	w_{p1}	p_2	w_{p2}	p_3	w_{p3}	p_4	w_{p4}
(1, 5)	0.6	(5, 5)	0.6	(1, 1)	0.6	(5, 1)	0.6

表2 2次元空間における特徴量 Q の例

Q_1		Q_2		Q_3	
q_1	w_{q1}	q_2	w_{q2}	q_3	w_{q3}
(2, 3)	0.8	(4, 3)	0.8	(3, 2)	0.8

図1は表1と表2の特徴量を x 軸と y 軸の2次元座標に丸印で示し、重みを丸印の中の数値で示している。白丸が分布 P の特徴量、グレーの丸が分布 Q の特徴量を示している。

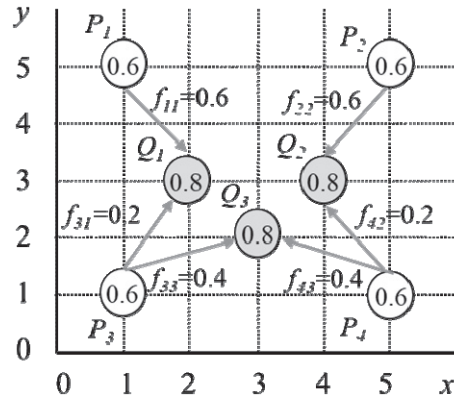


図 1 特徴量 P と Q の 2 次元空間におけるフローの例

例えば、 P_4 は 0.6 の重みを持っており、 Q_2 と Q_3 にそれぞれ $f_{42}=0.2$ 、 $f_{43}=0.4$ の重みを配分している。つまり、条件式 (3) を満たしている。また、 Q_2 は 0.8 の容量を受け付け可能であり、 P_2 と P_4 からそれぞれ $f_{22}=0.6$ 、 $f_{42}=0.2$ の合計 0.8 の重みを受け付けている。したがって、条件式 (4) を満たしている。また、距離についてはユークリッド距離を用いた場合について述べる。例えば、 P_4 から Q_2 へのフロー f_{42} に対応する距離 d_{42} は x 軸は 1、 y 軸は 2 離れた位置にあるため、ユークリッド距離 d_{42} は 2.236 ($=\sqrt{1^2 + 2^2}$) となる。同様に、 d_{11} 、 d_{22} 、 d_{31} 、 d_{33} 、そして、 d_{42} もまた 2.236 となる。したがって、仕事量 $WORK$ は 5.366 ($=2.236 \times (0.6+0.6+0.2+0.4+0.2+0.4)$) となり、式 (5) の値は 2.4 となるため EMD の値としては 2.236 が得られることになる。

3.1.2.2 位置情報に基づく単語アライメント

提案手法では、始めに訳文と参照訳間で単語の意味情報と位置情報に基づき単語アライメントを行う。単語間の対応関係は単語の分散表現を用いたコサイン類似度に基づくスコアにより決定する。スコアはコサイン類似度に対して、単語の位置情報を重みとして付与することで得られる。式 (7) に単語間のスコアの計算式を示す。

$$align_score = \cos_sim(t_i, r_i) \times (1.0 - pos_inf(T_i, R_j)) \quad (7)$$

式 (7) の $\cos_sim$ は訳文の単語 T と参照訳の単語 R のそれぞれの分散表現 t と r の間のコサイン類似度を示す。また、 pos_inf は訳文の単語 T と参照訳の単語 R の出現位置の相対的なずれを示している。 pos_inf の計算式は以下の式 (8) となる。

$$pos_inf(T_i, R_j) = \left| \frac{pos(T_i)}{m} - \frac{pos(R_j)}{n} \right| \quad (8)$$

式 (8) の pos_inf の値は出現位置の相対的なずれが大きいくほど 1.0 に近い大きな値となる。図 2 に訳文と参照訳間のアライメントの具体例を示す。

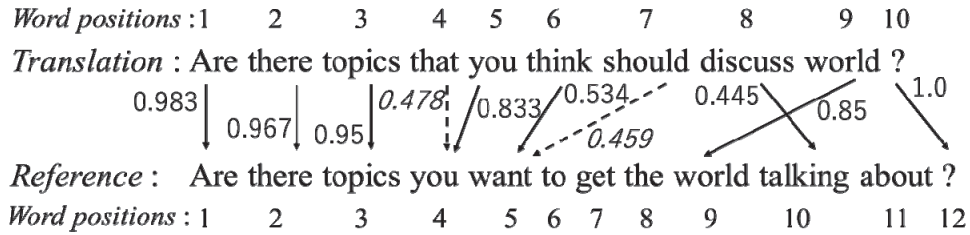


図2 単語アライメントの例

図2の実線の矢印は対応関係が成立していることを示している。破線の矢印は対応関係が不成立であることを示している。また、実数値は式(7)より得られるスコアである。例えば、訳文の単語 *discuss* は参照訳の単語 *talking* と対応している。その際のスコアは 0.445 である。単語 *discuss* と単語 *talking* との間のコサイン類似度は 0.460 であり、この値は訳文の単語 *discuss* と参照訳の単語 *topics* との間のコサイン類似度 0.477 より小さな値である。したがって、コサイン類似度のみを用いた場合には単語 *topics* が単語 *discuss* と対応する単語になってしまう。しかし、提案手法ではさらに式(8)の位置情報を重みとして付与することにより単語 *discuss* と単語 *talking* の間のスコアの方が単語 *discuss* と単語 *topics* の間のスコアより高くなり、より高い精度での単語アライメントが可能になる。また、訳文の単語 *that* においては参照訳の全単語の中で *you* とのスコアが最も高かったが、訳文の単語 *you* もまた参照訳中の単語 *you* とのスコアが最も高かった。このように複数の単語が競合した場合にはスコアの高い単語が優先されるため、参照訳の単語 *you* とは訳文の単語 *you* と対応付けが行われ、訳文の単語 *that* は対応する単語が存在しないことになる。

3.1.2.3 文レベルの *tfidf* による重み

提案手法では EMD を適用するにあたり、特徴量には単語の分散表現、重みには文レベルの *tfidf* を用いる。単語の分散表現は既存のモデルから取得する。重みは以下の式(9)を適用することにより得る。

$$w = tf \times \left(\log \frac{N}{df} + 1.0 \right) \quad (9)$$

式(9)の N は文の数、 tf はある任意の単語が文中に出現する回数、そして、 df はある任意の単語が出現する文数を示す。ある任意の単語が一つの文に何度も出現する可能性は低いため、基本的には tf の値は小さく、単語ごとに大きな違いは生じない。しかし、 df は機能語と内容語とでは値は異なると考えられる。すなわち、式(9)における w の値は機能語は小さく、内容語は大きくなると考えられる。

また、提案手法では式(10)と式(11)を用いて正規化する。具体的には文中の全単語の重みの総和に対する個々の単語の重みの割合を求める。したがって、文中の全単語の重みの総和は 1.0 になる。

$$\widetilde{w}_{t_i} = \frac{w_{t_i}}{\sum_{i=1}^m w_{t_i}} \quad (10)$$

$$\widetilde{w}_{r_j} = \frac{w_{r_j}}{\sum_{j=1}^n w_{r_j}} \quad (11)$$

3.1.2.4 距離計算

EMD を求める際の単語間の距離は以下の式（12）を用いて計算する。

$$d_{ij} = \begin{cases} 1.0 - \cos_sim(t_i, r_j) \times e^{-pos_inf(T_i, R_j)} & \text{if } T_i \text{ corresponds to } R_j \\ 1.0 & \text{if } T_i \text{ does not correspond to } R_j \end{cases} \quad (12)$$

式（12）の $\cos_sim$ は単語 t と単語 r の分散表現間のコサイン類似度を意味する。訳文の単語 T と参照訳の単語 R が対応関係にある場合には、コサイン類似度に式（8）で示した出現位置の相対的なずれに基づく重みを付与することで語順の違いを EMD の値に反映させる。さらに、1.0 から引くことにより対応関係にある単語間ほど距離 d_{ij} の値は小さくなる。また、訳文の単語 T と参照訳の単語 R が単語アライメントの結果より対応関係にないと判断された場合には、距離を一律に 1.0 とする。そして、EMD では式（12）を用いて全単語間の距離を要素とする距離行列を用いる。図2で用いた訳文と参照訳に対する距離行列を図3に示す。

		Reference											
		Are	there	topics	you	want	to	get	the	world	talking	about	?
t r a n s l a t i o n	Are	0.017	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	there	1.0	0.033	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	topics	1.0	1.0	0.049	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	that	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	you	1.0	1.0	1.0	0.154	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	think	1.0	1.0	1.0	1.0	0.456	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	should	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	discuss	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.555	1.0	1.0
	world	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.139	1.0	1.0	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3 距離行列の例

3.1.2.5 提案手法のスコア

提案手法では訳文と参照訳間の EMD を用いて、以下の式（13）より最終的なスコアを得る。

$$\text{スコア} = 1.0 - \frac{\min(\text{WORK}(T, R, F))}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (13)$$

EMD の値は 0.0 から 1.0 の値となり、距離に近いほど値は小さくなる。しかし、一般的な自動評価法では類似しているほどスコアが高くなる。したがって、式 (13) では 1.0 から EMD の値を引くことにより、訳文と参照訳が一致するほどスコアが高くなるように変換し、一般的な自動評価法と一致させている。

3.1.3 性能評価実験

3.1.3.1 実験の手順

提案手法と IMPACT を含めた様々な自動評価法において WMT17 の自動評価タスクデータを用いたメタ評価を行った。WMT ではシステムレベルとセグメントレベルの 2 つの観点よりメタ評価を行っているが、本報告ではシステムレベルよりも評価が困難とされているセグメントレベルについてのみメタ評価を行った。WMT17 のデータセットにはニュース翻訳タスクで使用されたデータセット “newstest2017” と “EU project HimL” により公開された “HUME Test Set Round 2” のサブセットである “himltest2017” の 2 種類がある。“newstest2017” の言語ペアは英語 (en) と中国語 (zh)、チェコ語 (cz)、フィンランド語 (fi)、ドイツ語 (de)、ラトビア語 (lv)、ロシア語 (ru)、そして、トルコ語 (tr) のペアであり、その双方向における翻訳文と参照訳を用いて各自動評価法はスコアを算出する。“himltest2017” ではルーマニア語については 2 つのデータセットが使用されており、1 つは英語からチェコ語 (cs)、ドイツ語 (de)、ポーランド語 (pl)、ルーマニア語 (ro) の 4 つの言語ペアに含まれているデータセット “himltest2017a” である。もう一つは英語からルーマニア語の 1 つの言語ペアのみのデータセット “himltest2017b” である。

また、提案手法では単語の分散表現を得るために学習済み単語分散表現モデルを用いた。今回は全ての言語において fastText^[5]の単語分散表現モデルを用いた。

3.1.3.2 実験結果

提案手法と IMPACT を含めたメタ評価実験の結果を示す。表 3 と表 4 に “newstest2017” における多言語から英語と英語から多言語へのそれぞれのセグメントレベルの相関係数を示す。表 5 には “himltest2017a” におけるセグメントレベルの相関係数を示す。また、表 6 には “himltest2017b” におけるセグメントレベルの相関係数を示す。表中の “Correlation” にある $|r|$ はピアソンの相関係数の絶対値を示し、 τ はケンドール τ を示す。

表 3 “newstest2017” における多言語から英語の訳文を用いた実験結果

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
Human Evaluation	DA	DA	DA	DA	DA	DA	DA
n	560	560	560	560	560	560	560
Correlation	$ r $	$ r $	$ r $	$ r $	$ r $	$ r $	$ r $
Proposed metric	0.547	0.493	0.692	0.550	0.564	0.602	0.589
IMPACT	0.494	0.489	0.630	0.469	0.505	0.581	0.570

AUTODA	0.499	0.543	0.673	0.533	0.584	0.625	0.583
BEER	0.511	0.530	0.681	0.515	0.577	0.600	0.582
BLEND	0.594	0.571	0.733	0.577	0.622	0.671	0.661
BLEU2VEC_SEP	0.439	0.429	0.590	0.386	0.489	0.529	0.526
CHRF	0.514	0.531	0.671	0.525	0.599	0.607	0.591
CHRF++	0.523	0.534	0.678	0.520	0.588	0.614	0.593
MEANT_2.0	0.578	0.565	0.687	0.586	0.607	0.596	0.639
MEANT_2.0-NOSRL	0.566	0.564	0.682	0.573	0.591	0.582	0.630
NGRAM2VEC	0.436	0.435	0.582	0.383	0.490	0.538	0.520
SENTBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512
TREEAGGREG	0.486	0.526	0.638	0.446	0.555	0.571	0.535
UHH_TSKM	0.507	0.479	0.600	0.394	0.465	0.478	0.477
newstest2017							

表 4 “newstest2017” における英語から多言語の訳文を用いた実験結果

	en-cs	en-de	en-fi	en-lv	en-ru	en-tr	en-zh
Human Evalution	DARR	DARR	DARR	DARR	DA	DARR	DA
n	32,810	3,227	3,270	3,456	560	247	560
Correlation	τ	τ	τ	τ	$ r $	τ	$ r $
Proposed metric	0.399	0.322	0.552	0.407	0.524	0.490	0.754
IMPACT	0.292	0.264	0.423	0.304	0.500	0.433	0.719
AUTODA	0.041	0.099	0.204	0.130	0.511	0.409	0.609
AUTODA-TECTO	0.336	-	-	-	-	-	-
BEER	0.398	0.336	0.557	0.420	0.569	0.490	0.622
BLEND	-	-	-	-	0.578	-	-
BLEU2VEC_SEP	0.305	0.313	0.503	0.315	0.472	0.425	-
CHRF	0.367	0.336	0.503	0.420	0.605	0.466	0.608
CHRF+	0.377	0.325	0.514	0.421	0.609	0.474	-
CHRF++	0.368	0.328	0.484	0.417	0.604	0.466	0.602
MEANT_2.0	-	0.350	-	-	-	-	0.727
MEANT_2.0-NOSRL	0.395	0.324	0.565	0.425	0.636	0.482	0.705
NGRAM2VEC	-	-	0.486	0.317	-	-	-
SENTBLEU	0.274	0.269	0.446	0.259	0.468	0.377	0.642
TREEAGGREG	0.361	0.305	0.509	0.383	0.535	0.441	0.566
newstest2017							

表 5 “himltest2017a” における実験結果

	en-cs	en-de	en-pl	en-ro
n	879	891	1,020	354
Correlation	$ r $	$ r $	$ r $	$ r $
Proposed metric	0.476	0.394	0.428	0.453
IMPACT	0.387	0.351	0.378	0.351
AUTODA	0.391	0.445	0.442	0.127
AUTODA-TECTO	0.400	-	-	-
BEER	0.400	0.428	0.442	0.508
CHRF	0.383	0.454	0.445	0.477
CHRF+	0.395	0.451	0.445	0.474
CHRF++	0.400	0.445	0.444	0.477
MEANT_2.0	-	0.479	-	-
MEANT_2.0-NOSRL	0.473	0.463	0.489	0.479
SENTBLEU	0.347	0.338	0.329	0.261
TREEAGGREG	0.323	0.374	0.450	0.481
himltest2017a				

表 6 “himltest2017b” における実験結果

	enro
n	350
Correlation	$ r $
Proposed metric	0.320
IMPACT	0.281
BEER	0.293
CHRF	0.305
CHRF+	0.314
CHRF++	0.310
MEANT_2.0-NOSRL	0.370
SENTBLEU	0.254
TREEAGGREG	0.244
himltest2017b	

3.1.3.3 考察

表 3 の多言語から英語の訳文を用いた実験結果においては全体としては **BLEU** が多くの言語で最も高い相関係数を示しており、提案手法は他手法と比べて特段高い相関係数を示してはいない。

したがって、多言語から英語の訳文に対する評価精度については表層情報に基づく自動評価法に対しては比較的高いが、学習に基づく自動評価法においては不十分と考えられる。

また、表 4 の英語から多言語の訳文を用いた実験結果においては全言語ペアの相関係数が存在しない自動評価法もあるため十分な比較は困難であるが、全言語ペアの相関係数が得られている自動評価法に限定した比較においては提案手法の相関係数は上位に位置する。全言語ペアの相関係数の平均を求めたところ、全言語ペアの相関係数が得られている自動評価法 IMPACT、AUTODA、BEER、CHRF、CHRF++、MEANT_2.0-NOSRL、SENTBLEU、TREEAGGREG、そして、提案手法の中では提案手法の平均は 0.493 であり、MEANT_2.0-NOSRL の平均 0.505 に次いで高い値であった。特に、en-zh の言語ペアにおいては提案手法の相関係数が 0.754 と最も高い値を示した。したがって、提案手法は英語から多言語の訳文の評価においては比較的高い評価精度を有していると考えられる。

表 5 と表 6 の “himltest2017” データにおける実験結果では、MEANT_2.0-NOSRL が最も高い相関係数を示している。提案手法の相関係数が MEANT_2.0-NOSRL の相関係数を上回った言語ペアは en-cs だけであった。特に en-de の言語ペアの相関係数が他手法に比べて低い傾向にある。したがって、この点については今後精査する必要があると考えられる。

3.1.4 まとめ

本報告では、WMT17 の自動評価タスクデータを用いたメタ評価により提案手法の有効性について述べた。実験の結果、“newstest2017” における英語から多言語の訳文に対する評価においては他手法と比べても比較的高い相関係数が得られた。しかし、“newstest2017” における多言語から英語の訳文に対する評価においては、BLEND、MEANT に比べて低い相関係数となっており、より高い精度での評価を実現するための改善が必要である。

今後は WMT18 や WMT19 を用いたさらなるメタ評価を行う。また、より高い評価精度を得るための改良を行う予定である。

参考文献

- [1] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia and Marco Turchi (2017) “Findings of the 2017 Conference on Machine Translation (WMT17),” Proceedings of the Conference on Machine Translation (WMT), pp.169-214.
- [2] Ondřej Bojar, Yvette Graham and Amir Kamran (2017) “Results of the WMT17 Metrics Shared Task,” Proceedings of the Conference on Machine Translation (WMT), pp.489-513.
- [3] Hiroshi Echizen'ya, Kenji Araki, Eduard Hovy (2019) “Word Embedding-Based Automatic MT Evaluation Metric using Word Position Information,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), pp.1874-1883.

- [4] Hiroshi Echizen-ya, Kenji Araki (2007) “Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum,” Proceedings of the Eleventh Machine Translation Summit, pp.151-158.
- [5] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. (2018) “Learning word vectors for 157 languages,” Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), pp.3483- 3487.

3.2 新コーパスを用いた中日テストセット試験文の採集と分析

元・山梨英和大学 江原 暉将
(株) 富士通研究所 長瀬 友樹

3.2.1 はじめに

機械翻訳評価の一手法として、翻訳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翻訳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¹⁾。

- ・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
- ・テストセットの拡充

その結果、昨年度までに 1,151 文の試験文と対応する翻訳パターンを構築できた。しかしながら、これらの試験文は研究の初期に収集した「中日特許文平行コーパス」から採集されたものであり、訓練データとの重複については考慮されていない。試験文と訓練文が重複していると試験の公平性が担保できなくなるため重複は避けるのが望ましい。今年度は、訓練データとは別に収集した中日特許文コーパスが利用可能になったため、新たに試験文を採集しなおした。本文では、その採集方法と採集結果の分析について述べる。

3.2.2 試験文の採集方法

今年度新たに利用可能になった中日特許文コーパスは、化学、電気、機械、物理の各分野から 25 万文対ずつ合計 100 万文対の中日対訳文から構成されている。構成部分別の統計としては、要約(ABS)が 2,509 文対、請求範囲(CLM)が 10,751 文対、詳細説明(DES)が 986,740 文対である。発明の名称(TIT)は本コーパスには含まれていない。

昨年度までに構築した中日テストセット(今後、旧テストセットと呼ぶ)は 1,151 文の中国語試験文と中日翻訳結果に含まれるべき日本語表現(翻訳パターン)から構成されている。例えば

- ・中国語試験文：提供一种具有锗治疗效果的眼镜用鼻托及眼镜。
- ・日本語表現：有する.+提供する。

という試験データがある。これは中国語試験文の中日翻訳結果に「有する.+提供する。」という日本語表現が含まれているかどうかを試験するものである(「.+」は正規表現であり、1 文字以上の任意文字列を意味する)。この旧テストセットにはさらに中国語表現も含まれている。上記の例では「提供一种具有.+的」がそれである。

上記の旧テストセットのデータを利用して、新テストセットに含めるべき試験文を以下の手順で採集した。第一段階として、中日特許文コーパスの中国語部分が中国語表現に適合し、日本語

¹ 平成 25 年度から平成 30 年度の AAMT/Japio 特許翻訳研究会報告書を参照。

部分が日本語表現に適合するデータを集める。ただし、この照合において任意文字列を全ての文字列としたのでは適合範囲が過多となるため、句読点(、。)以外の文字列として照合した。その結果、上記の例では 100 万文対中 7 文が適合した。例えば次の文対がある。

中国語文：在一些实施方案中，用于取代的位点特异性位置提供一种具有优异折叠性能的抗体。

日本語文：特定の実施形態では、置換のための部位特異的位置は、有利なフォールディングを有する抗体を提供する。

次に第二段階として、収集されたデータを標本化した。第一段階で抽出できた文対は総数で 642,938 文対であり、これらをすべて試験文とするには多すぎる。そこで、適合した文数が 2 文以下ならばすべての文を採用し、3 文以上であれば、2 文または 3 文が採集できるように等間隔に間引いた。その結果、2,007 文対が採集できた。同一の試験文に対して異なる翻訳パターンが対応する場合もあるため、試験文と翻訳パターンの対としては 2,123 データとなった。なお、旧テストセットに存在するが新テストセットでは採集できなかった翻訳パターンは 139 パターンあった。例えば、

中国語表現：需注意的是

日本語表現：留意されたい

がある。

採集できた 2,123 データの構成部分と分野の統計は表 1 のとおりである。

表 1：新テストセットの構成部分と分野の統計

構成部分\分野	化学	電気	機械	物理	合計
発明の名称(TIT)	0	0	0	0	0
要約(ABS)	164	3	6	3	176
請求範囲(CLM)	93	13	14	10	130
詳細説明(DES)	616	295	471	435	1817
合計	873	311	491	448	2123

3.2.3 採集結果の分析

採取できた 2,123 データに関して、実際の中日機械翻訳システムによる翻訳結果を用いて分析を行った。手順は以下のとおりである。

1. 2,123 の中国語試験文を中日機械翻訳システムで翻訳
2. 各翻訳結果（日本語）に対して翻訳パターン（正規表現）とのマッチングを試し、マッチするものを「○」、マッチしないものを「×」として評価
3. 同一のパターンに紐づいている複数の試験文の評価において「○」と「×」の評価が混在するパターンについて、評価が分かれる原因を調査

2,123 の中国語試験文における、中日機械翻訳システムによる翻訳結果と翻訳パターンとのマッチングの結果を表 2 に示す。

表 2: 試験文全体の翻訳パターンとのマッチング結果

	文数	割合 (%)
マッチ成功 (○)	1605	76.0
マッチ失敗 (×)	508	24.0

今回、試験文採取に用いた翻訳パターンのユニーク数は 945 であるが、そのうち 139 はマッチする原文が見つからなかったため、採取できた 2123 のデータに含まれるパターンのユニーク数は 806 である。各パターンには 2 ないし 3 文の試験文が対応している。同一のパターンに紐ついた試験文は翻訳結果のパターンマッチ結果も同じであると考えられるが、実際には評価が分かれる例が多数見つかる。その件数と割合は表 3 のとおりである。

表 3: 翻訳パターンごとの複数試験文の評価

	パターン数	割合 (%)
すべての試験文の評価が同じ	536	67.5
試験文によって評価が分かれる	270	33.5

およそ 1/3 のパターンで、対応する複数の試験文の評価が分かれるという結果になった。以下にその原因について述べる。

➤ 翻訳エンジンの不安定性によるもの

同一の中国語表現に対して翻訳エンジンが異なる日本語訳を出力する場合があります、同じ翻訳パターンに対応する複数の試験文の翻訳結果の評価が分かれる原因になっている。翻訳エンジンの不安定性によるものであり、採取できた試験文に問題があるわけではない。

例えば、下記の例では、「**不局限**」という中国語表現に対して、試験文 1 では「限定されない」と訳出してパターンにマッチするが、試験文 2 では「限られない」と訳出されてパターンマッチに失敗している。

翻訳パターン：

中国語パターン	日本語パターン
不局限	限定((する される)(ものでは ことは) され(てい))ない。

試験文 1：当然、本发明并不局限于所述和附图所示的实施方式。

訳文 1：当然ながら、本発明は、上述した実施形態に**限定されない**。【評価：○】

試験文 2 : 进一步的, 输出检查端子 20 的露出于箱体 11 的外部的地方不局限于第二端部 20b。

訳文 2 : さらに、出力検査端子 20 の筐体 11 の外部に露出する箇所は、第 2 端部 20 b に限られない。【評価：×】

また、次の例では、試験文 1 では「使用される」、試験文 2 では「用いる」という異なる表現で訳出されており、試験文 2 の訳文のみ「×」の評価になっている。

翻訳パターン：

中国語パターン	日本語パターン
使用([^、,。]+)溶剤	使用される。 使用する。

試験文 1 : 在一个实施方案中，作为吸收介质，使用溶剂或溶剂混合物，其用作后续萃取蒸馏中的萃取介质。

訳文 1 : 一実施形態では、吸収媒体として溶媒または溶媒混合物を使用し、それは後続の抽出蒸留における抽出媒体として使用される。【評価：○】

試験文 2 : 从多孔膜去除增塑剂的增塑剂去除工序中，为了去除增塑剂而使用萃取溶剂。

訳文 2 : 多孔質膜から可塑剤を除去する可塑剤除去工程では、可塑剤を除去するために抽出溶媒を用いる。【評価：×】

このような訳文中の表記の揺れは多くの機械翻訳システムで普通に起きることであるが、事前に訳出のバリエーションをすべて予見してパターンに反映することは困難である。訳出が正しいにもかかわらずパターンマッチに失敗している事例を集め、継続的にパターンをメンテナンスしていく必要があると考える。

➤ パターン設定において文脈の考慮が必要なもの

今回、試験文抽出に用いた翻訳パターンには、文脈によって日本語訳が日本語パターンに合致しない場合がある。つまり、翻訳パターンの条件が弱すぎるために想定外の文脈で使われる試験文を抽出し、このことが原因で同一の翻訳パターンに対応する試験文の間で評価の相違を起こすことがある。

下記は、同じ翻訳パターンから抽出した二つの試験文とその訳文である。「有」の主語に注目すると、試験文 1 は抽象物（「インパクト」）であり、試験文 2 では具体物（「センサ」）であることがわかる。翻訳システムは、主語が具体物の場合には「備える」を出すように訳し分けているが（訳

文 1、訳文 2)、翻訳パターン（訳語を「有する」に限定）」で訳し分けの考慮がされておらず、マッチングに失敗している。

翻訳パターン：

中国語パターン	日本語パターン
提供一种具有〔^、。〕+的	有する〔^、。〕+提供する。

試験文 1：本发明有利地提供一种具有独特视觉冲击的容器。

訳文 1：本発明は、有利には、独特の視覚的インパクトを**有する容器を提供する。**

【評価：○】

試験文 2：本发明还提供一种具有上述传感器的充电桩。

訳文 2：本発明はまた、上記のセンサを**備えた充電カラムを提供する。**【評価：×】

次も同様に、翻訳パターンから抽出した 2 つの試験文とその訳文の例である。試験文 1 では、キーワード「示意图」の前に「的剖面」が付いたことで、パターンに定義されている「概略図」という訳語は適切ではなくなっている。つまり、この翻訳パターンに対応する試験文は、「示意图」が単独で用いられている例文を採用するのが安全であろう。

翻訳パターン：

中国語パターン	日本語パターン
图\d+〔^、。〕*示意图	図¥d+は(、)?〔^、。〕*概略図である。

試験文 1：图 4 及图 5 分别是图 3 中剖面 A-A' 及剖面 B-B' 的剖面示意图。

訳文 1：図 4 及び**図 5 は**、それぞれ図 3 における線 A-A' 及び線 B-B' の断面模式図である。

試験文 2：图 1 是人眼的示意图。

訳文 2：**図 1 は**、人間の眼の**概略図である。**

これらの例のように、翻訳パターンの中には文脈の考慮が必要なものが含まれており、文脈を

踏まえて翻訳パターンを適切に修正することにより、抽出される試験文もより適切なものが得られるようになると考えられる。

3.2.4 中日テストセットの分析

2.1.3 では、今回、新たに抽出した中国語試験文について分析を行った。本節では、これまで構築してきた中日テストセット全体に関して、機械翻訳評価用テストセットとしての観点から分析を行う。なお、本分析は中国語ネイティブ（中日バイリンガル）の協力を得て行った。

➤ 中日パターン間の要素の過不足

中国語と日本語で情報の過不足のあるパターン対が散見された。
例えば、下の例では、日本語のパターンの中に、中国語の「本発明」に対応する部分が見つからない。

翻訳パターン

中国語パターン	日本語パターン
本发明提供([^、,。]+)优异的	優れた([^、,。]+)提供する。

一方、この例とは反対の中国語パターンの中に欠落要素が含まれる例はほとんど見つからなかった。翻訳パターンの用途を機械翻訳結果の評価に限れば、中国語と日本語のパターンの要素の不一致（中日翻訳用のパターンでは日本語側の情報の欠落）があっても特に問題はないと考える。

➤ 表記の揺れ

パターン対で表記の揺れに対応しきれずに、本来「○」評価になるべき例文が「×」にされるケースが見られた。表記ゆれには、以下に示す例のように多くの種類がある。

・数字

翻訳パターン：

中国語パターン	日本語パターン
中任一项所述的([^、,。]+), 其特征在于	であることを特徴とする請求項 ([^、,。]+)いずれか1項に記載の ([^、,。]+)。

訳文：14. 前記導体基部がメタライズ層であることを特徴とする請求項8～13のいずれか一項に記載の配線基板。 【評価：×】

パターンは算用数字を条件としているが、訳文は漢数字を出力しており、「×」評価とされた。
(特許文において漢数字は不適切であるが、このパターンで評価するものではない)

・漢字・かな

翻訳パターン：

中国語パターン	日本語パターン
根据权利要求¥d+或¥d+所述的	請求項¥d+又は¥d+((に) 記載の)に記載された)

訳文：3. 前記制御装置（１８）は、対応する呼吸適用において前記呼吸時定数を決定するために使用される、請求項１または２に記載の呼吸器（１０）。【評価：×】

パターンでは漢字「又は」での訳出を必須としているが、訳文ではひらがな「または」を出力しており、「×」と評価された

・能動・受動

翻訳パターン：

中国語パターン	日本語パターン
([^、,、。]+)中所示的([^、,、。]+)	([^、,、。]+)に示した([^、,、。]+)

訳文：図４および図５に示されるプランジャ５０は、同じアクチュエータによって作動され、それらの各チャンバ８０内で往復運動するように摺動する。【評価：×】

パターンの「示した」と訳文の「示される」の不一致のため「×」評価となった。

・同義語

翻訳パターン：

中国語パターン	日本語パターン
已知	知られている。 わかっている。

訳文：この点に関して、本開示によれば、使用可能な無菌水性媒体は、当業者に公知である。

【評価：×】

日本語パターンの同義語定義に「公知である。」が抜けていたため、「×」と評価された。

以上のように、個々のパターンを詳細に見ていくと、種々の改善すべき点が含まれていることが判った。なお、テストセット全体のなかで書換えが必要なデータの割合は、感触としては一割程度である。

3.2.5 まとめ

昨年度までに収集できた中日表現パターンを利用して新しく利用可能になった中日対訳コーパスから新テストセットを構築した。ここで利用した中日対訳コーパスは訓練データとは別に収集されたものであり、試験としての公平性が担保できるものである。

今後の課題として以下のものがある。

- ・旧テストセットに含まれるが、新テストセットでは採集できなかったパターンへの対応。
- ・技術用語や数量表現など用語の翻訳に着目したテストセットの構築。
- ・文脈を考慮した翻訳パターンの精緻化。
- ・より広い表記ゆれへの対応。

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

3.3 独自テストセットによる中日機械翻訳の評価

奈良先端科学技術大学院大学 須藤 克仁
株式会社ディープランゲージ 王 向莉

3.3.1 はじめに

2018 年度までに AAMT/Japio 特許翻訳研究会拡大評価部会において検討を進めてきた翻訳評価に関する種々の検討事項を活用するための試みとして、2019 年度はテストセットグループで構築された評価用テストセットを利用した中日翻訳評価を行った。実施にあたってはワーキンググループとして須藤が全体統括と評価集計を、王がテストセットの整備を担当し、両名で評価結果の確認・分析を行った。

本件評価にあたっては、研究会内で協力を依頼し 7 システムの翻訳結果をご提示いただいた。本件評価と合わせ、江原委員には表現パターンに基づく中日翻訳評価セットに基づく評価についてご協力をいただいた。

3.3.2 評価対象システム

ワーキンググループにご提示いただいた 7 システムの翻訳結果はランダムに採番しシステム 1 からシステム 7 と名付け、両テストセットにおける評価を行った。ただし、システム 3、4、7 は同一の翻訳結果を出力していたことから、最終的な評価対象は 5 システムとなった。以下システム 4、システム 7 の結果は省略する。

また、別途オンラインで試用可能な機械翻訳エンジンを利用し、システム X として評価に利用した。

3.3.3 文法能力評価用テストセット（科学技術文書）における評価

3.3.3.1 テストセットについて

今回使用したテストセットは 108 文で構成され、5 つの述語（放置, 収集, 提供, 降低, 添加）を中心とする中国語の文法項目が日本語に正しく訳出されているかどうかをパターンマッチングで検証するためのものである。詳細は平成 30 年度報告書 3.2 節を参照のこと。

以下に一例を示す。パターンは正規表現で記載し、プログラムによるマッチングを簡便に行えるようにした。

原文：将其放置在了三维微工作台上的样品架上。

参照訳：それを三次元マイクロステージ上のサンプルホルダに載せた。

格要素パターン：“それを”, “三次元マイクロステージ上のサンプルホルダに”, “載せた”

述語・格パターン：“.*を.*に(載せ|置い)た”, “.*に.*を(載せ|置い)た”

3.3.3.2 評価方法

評価はテストセット中で定義された格要素パターン（それぞれ独立にマッチング）、述語・格パターン（候補のいずれかにマッチング）について Python スクリプトでマッチングを行い、そのマッチ数を求めた。

3.3.3.3 結果（格要素パターン）

格要素パターンマッチの結果（正解率(%)）を次表に示す。なお、項目横の括弧付きの数字は対応するパターン数を表す。網掛け太字のセルは各要素毎の最高正答率を示したものを表す。

	Total (319)	放置 (71)	収集 (61)	提供 (60)	降低 (64)	添加 (63)
システム 1	26	18	20	32	28	33
システム 2	24	20	16	27	25	35
システム 3	22	17	20	22	23	29
システム 5	13	14	6.6	13	20	13
システム 6	28	21	21	32	28	37
システム X	34	69	39	23	19	14

3.3.3.4 結果

述語・格パターンマッチの結果（正解率(%)）を次表に示す。

	Total (108)	放置 (24)	収集 (22)	提供 (20)	降低 (21)	添加 (21)
システム 1	16	25	14	20	0	19
システム 2	21	21	18	35	9.5	24
システム 3	19	21	23	30	4.8	19
システム 5	12	21	9.7	10	0	19
システム 6	25	25	27	40	9.5	24
システム X	33	79	32	30	9.5	9.5

3.3.3.5 議論

評価システム間でのシステムの差異があまり大きくなかったためか、各システムのパターン正解率の差は比較的小さい範囲に留まったと言える。ただ、ベースラインシステムであるシステム 3 と比較して多少の差分が出ているため、翻訳技術の工夫が若干の性能向上に寄与した可能性が高い。

一方、システム X は「放置」「収集」では他より高いパターン正解率であったが、「降低」「添加」では正解率が低い傾向にあり、科学技術文献特有の訳語選択ができていなかった（「下げる」「追加する」等）ことによる影響が考えられる。こうした訳語選択や訳し分けの問題は実際の機械翻訳においても重要な課題であり、文法能力の評価という観点のみならず今後考慮すべき点であると言えよう。

また、パターンマッチングが表層に依拠していることにより、特に述語において異なるが許容

できる訳語を過度に棄却する傾向にあり、単語分散表現の利用等の拡充が今後重要となるであろう。

3.3.4 表現パターンテストセット（特許文書）における評価

3.3.4.1 テストセットについて

本テストセットは 1,151 個の中国語表現パターンを含む中国語特許文に対して日本語の翻訳パターン設問を付した設問型のテストセットであり、Web サイトに翻訳結果をアップロードして評価結果を得ることができるようになっているものである。詳細は平成 30 年度報告書 3.3 節および平成 29 年度報告書 3.4 節を参照のこと。

3.3.4.2 結果

次表に表現パターンテストセットによる評価結果を示す。各評価項目毎に 100 点満点の評価値が付き、さらに評価項目数に応じた加重平均値を全体の評価値として示している。

	名称	要約	請求項	詳細説明	加重平均
システム 1	95	64	61	71	71
システム 2	59	29	21	44	42
システム 3	54	30	18	45	43
システム 5	40	16	14	26	25
システム 6	63	23	20	43	41

3.3.4.3 議論

前節と異なりシステム 1 が他システムを大きく上回る結果となった。これはシステム 1 が科学技術論文でなく特許に注力して構築されていることが主因と考えられる。その他のシステムについては前節と似たような差となっている（システム 5 が若干劣る）。

3.4 おわりに

本報告では昨年度までに構築されたテストセットを実際に利用した機械翻訳評価について、ごく簡単な形ではあるが評価を試みた結果について述べた。

各テストセットがそれぞれ目指していた方向性での評価がある程度実現できたと考えられる一方で、翻訳における別表現の許容度についての問題がある。パターンマッチングによる簡便な評価と単語分散表現等の外部リソースを連携させた実用的機械翻訳評価の実現が次のステップの課題と言えるだろう。

4. 国際ワークショップ開催報告：

PSLT 2019

4 国際ワークショップ開催報告：PSLT 2019

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、2005 年の第 1 回から数えて 8 回目となるワークショップを開催した。開催日は機械翻訳サミットの本会議前日にあたる 8 月 20 日（火）であった。オーガナイザは須藤が代表となり、綱川、宇津呂の両委員とともに務めた。プログラム委員として研究会委員及び 6 名の国内外の研究者にご協力いただき、技術講演論文の査読を依頼した。

招待講演者の選定は昨年度末から研究会での議論を通じて行い、特許庁、機械翻訳研究者、技術文書翻訳に関連して医療言語処理研究者に依頼した。

技術講演に関しては、数日投稿期限を延長した結果 4 件の投稿があり、プログラム委員による査読の結果そのすべてを採択した。

従来本ワークショップは 1 日開催としていたが、本年は技術論文の投稿件数が伸び悩んでいることを考慮して半日の開催とした。しかしながら、結果的に 4 件の技術論文講演を実施することとなり、短い時間ながらも密度の高いプログラムとなった。ワークショップ当日の参加者は 20 名強（事前登録 17 名）であり、およそ半数が日本からの参加者であった。

4.2 ワークショップ報告

まず、招待講演について簡単に報告する。

1 件目は、University of Amsterdam の Christof Monz 准教授による、ニューラル機械翻訳におけるドメイン適応についての講演であった。ニューラル機械翻訳が広く使われるようになるにつれて、実用において対象分野の対訳コーパスの不足という問題が重要となっており、現在は世界的に様々な手法が検討されている状況である。講演では、大量の分野外コーパスによる事前学習とごく少量の対象分野コーパスによる fine-tuning（微調整）を行う方式において、繰り返し学習の中で徐々に分野外コーパスの重みを下げていくための手法について紹介された。

2 件目は、特許庁の松谷洋平氏による、特許庁における機械翻訳の導入についての講演であった。特許庁では NICT の技術を活用し、日本語と英語／中国語／韓国語の間の機械翻訳技術を導入しており、J-platpat のような情報公開サイトや審査業務への利用が進んでいることが示された。

3 件目は、LIMSI-CNRS の Aurélie Névél 氏による、医療言語処理と医療分野における機械翻訳についての講演であった。医療分野における言語処理の研究は英語を中心に行われており、実際の医療現場における様々な言語への対応が難しいという課題があり、フランス語等欧州言語について様々なコーパスの整備を進めていること、また日本でも日本語の医療言語処理データが整備され利用可能であることが示された。

招待講演はいずれも機械翻訳の実応用に関係する内容で、本研究会の活動との関連が深いものであった。特にドメイン適応に関しては大量の特許対訳の別ドメインでの適応、また特定分野の対訳の特許翻訳への適応、の両側面からその応用に期待が持てる。

続いて、技術講演 4 件について報告する。

技術講演 1 件目は NICT の小野氏らによる、多層 RNN に基づくニューラル機械翻訳における計算の効率化についての発表であった。複数の GPU を利用可能な場合に、各層の RNN を異なる GPU に担当させる（モデル並列）とともに、注視層や出力層のデータも分散させる（データ並列）というアイデアに基づく技術である。

技術講演 2 件目は Dublin City University（アイルランド）の Poncelas 氏らによる、ニューラル機械翻訳の適応についての発表であった。通常分野適応の問題はモデル学習時に大量の分野外データと少量の対象分野データを利用して対処するが、提案手法では翻訳対象であるテストセットの原言語側の情報に基づき学習データ中の類似する対訳文を選択的に利用してモデル適応を行った。

技術講演 3 件目は筑波大の飯田氏らによる（発表は共著者の Hung 氏が行った）、RNN に基づくニューラル機械翻訳における多段注視機構についての発表であった。複数ヘッドによる注視機構が Transformer をはじめよく使われているが、提案手法ではさらに異なるヘッドによる注視を混合して注視機構を多段化することで、特に長文における翻訳において精度改善を果たした。

技術講演 4 件目は Tampere University（フィンランド）の Nurminen 氏による、企業等の知財・特許担当者の視点から機械翻訳が業務における意思決定にどう貢献するかを、実際のユーザへのヒアリングによって調査した研究であった。本ワークショップではユーザサイドの話題がこれまで非常に少なかったことから、実際に機械翻訳がどう使われるかという意味で興味深いものであった。

4.3 所感

前回と比較してニューラル機械翻訳が広く浸透し、基礎研究に加えて、実用的な研究・検討が進んできている印象を受けた。その一方で特許や技術文書の翻訳に関する多言語のリソースは医療関係の EMEA、論文抄録の ASPEC 等存在はするものの、機械翻訳研究のメインストリームとなっている WMT の News タスクのように広く用いられているとは言い難いのが実情である。特許等技術文書の翻訳は比較的大規模な対訳コーパスが入手しやすく、文脈に依存する代名詞や省略等の問題も比較的小さく、また長文の翻訳等の問題もあって機械翻訳研究の対象として有益であると考えられる。しかしながら、実際に共通タスクとして利用できるリソースが限定されていること、それと関連し世界的な認知度がまだ低いのが実情と言えよう。他方、日本の特許庁における NICT 技術の利用や、欧州特許庁と Google との連携等実応用も進みつつあり、もはや基礎研究の対象ではないという見方もあるのかもしれない。ただ、（筆者の私見ではあるが）専門用語や長文の問題のように統計的機械翻訳の時代からその困難さが取り沙汰されてきた問題についての顕著な改善はまだ得られておらず、今後のさらなる発展が期待される場所である。

本ワークショップについても、技術的な内容のみならず、ユーザ視点での分析・提言や研究開発のためのリソースの提案等、これまで以上に幅広い展開を検討してもよいのではないかと考えられる。AAMT/Japio 特許翻訳研究会としてこの分野の貢献にどのような役割を果たすことができるか、継続して議論をする必要があろう。

————— 禁 無 断 転 載 —————

2019年度AAMT/Japio特許翻訳研究会報告書

発 行 日 2020 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