2020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
第6回特許情報シンポジウム開催報告

2021 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	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 / 東京大学大学院 名誉教授
副 委員長	宇津呂武仁	筑波大学システム情報系 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	日本放送協会 放送技術研究所 主任研究員
	綱川　隆司	静岡大学学術院 情報学領域 講師
	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
オブザーバー	岡　　俊行	(有)アジア産業　研究開発部長
	高　　京徹	(株)高電社 経営企画部 部長
	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 メディア処理研究所 主管研究員
	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）：

大塩　只明	特許情報研究所 調査研究部 研究企画課
長部　喜幸	特許情報研究所 調査研究部 部長
木下　　聡	特許情報研究所 調査研究部 統括研究主幹
久々宇篤志	特許情報研究所 調査研究部 研究企画課 課長
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
前原　義明	特許情報研究所 調査研究部 研究企画課 課長代理
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

2020 年度 AAMT/Japio 特許翻訳研究会・活動履歴

2020 年 6 月 19 日
第 1 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2020 年 8 月 21 日
第 2 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2020 年 10 月 30 日
第 3 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2020 年 12 月 17 日
第 4 回 AAMT/Japio 特許翻訳研究会、第 2 回拡大評価部会
（於オンライン開催）

2021 年 2 月 24 日
第 5 回 AAMT/Japio 特許翻訳研究会
（於オンライン開催）

2021 年 2 月 26 日
第 6 回特許情報シンポジウム
（於オンライン開催）

2021 年 3 月 31 日
『2020 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第 6 回特許情報シンポジウム開催報告』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

2020 年は、コロナ禍によって社会の在り方が大きく変化した年であった。この変化には、コロナ禍の終焉と共に以前の状態に戻る一時的なものもあろうが、この変化には、我々の生活をより豊かなものにする、次の時代にも引き継がれていくべき変化、肯定的な変化でもある。そういう意味では、コロナ禍は社会の変化を加速し、十年の期間をかけて起こる変化を 1 年間に圧縮したものでもある。

このような今後も継続していく変化の一つに、人間の行動の多くが物理的な空間からサイバー空間に移動したことであろう。働き方も、リモートワークの常態化により、オフィスへの出勤がなくなり、職場での同僚とのコミュニケーション、商取引での顧客との会議の多くがサイバー空間へと移行した。これまで日常的に行われていた同僚との会話、打ち合わせのための会議がサイバー空間に移行したことで、本委員会が取り扱う自然言語（日本語や英語、フランス語）の技術も、今後、大きく変化、飛躍することになろう。

この大きな変化は、時間と場所とを共有しないテキストによる情報伝達と、これらを共有した音声による情報伝達という、これまでの区分ではとらえられないコミュニケーションの形態が広く社会に広がったことであろう。

時間と場所を共有した対面での会話や会議では、話し手の身振り、表情、声の調子、あるいは、聞き手の身振り、表情による反応といったパラ言語が、言語による情報伝達を補完していた。逆に、その場で話された内容は記録されず、後日、それを参照したり、他の人に伝えたりすることは不可能であった。この時間と場所を共有した情報伝達が、サイバー空間での会議、あるいは、ライン・スラック・メッセンジャーを使った短いメッセージの交換に置き換えられつつある。

サイバー空間での会議では、話し手の表情や声の調子などのパラ言語的な情報伝達が極めて限定的になる。また、聞き手の反応を話し手が把握することも困難である。このために、話し手と聞き手の頻繁で自然な交代ができず、一方的な情報伝達になりがちとなる。また、場所を共有した会議では、図面や参考資料など言語以外のメディアを適宜使いながらの議論ができるが、サイバーでの会議では、一つの資料の投影が限度であり、頻繁なメディアの使用は困難である。また、ショートメッセージの交換は、発言者の頻繁な切り替えによる双方向的な情報伝達はできるが、パラ言語的な情報伝達はスタンプや絵文字によるものとなり、きわめて限定的となる。

このように、サイバー空間での情報伝達は、時間と空間を共有する情報伝達に比べて、不自由で欠陥のあるものと考えられている。ただ、一方で、時間と空間の制約が弱いサイバー空間での情報伝達が、大きな利点も持っていることもたしかである。時間・空間を共

有するための参加者の物理的な移動コストがなくなったことは大きく、地域を超えた会議参加が可能になった。パラ言語による情報伝達が少なくなったことで、場の雰囲気に左右されない、より客観的で論理的な議論ができるようになったという感想を聞くこともある。

計算機による言語処理の技術開発をしている立場からは、会議やショートメッセージのやり取りが記録として残ることで、後日、それを見直したり、参加者以外と共有したりことができるようになったこと、さらには、サイバー空間での記録を計算機処理の対象にできることになったことが、大きな変化であろう。テキスト処理を中心にした言語処理技術の多くが、サイバー空間での会議や会話の処理に適用可能になった。実際、サイバー会議の内容を音声認識や自動翻訳することで、サブタイトルの付いた会議録を多数の人に配信する、あるいは、会議内容やショートメッセージの交信内容を自動抄録する、会議録を自動作成するといった技術が急速に進歩することが期待できる。

コミュニケーションの大きな部分がサイバー空間に移動することは、この委員会が対象とする言語処理や自動翻訳の技術的な進展にも大きな影響を与える。コロナ禍による社会変動は、もちろん、否定的な側面が大きいことも確かであるが、言語処理をはじめとするデジタル技術の進展に大きな寄与をする肯定的な側面も大きい。今後の技術の進展に期待している。

2. 年間報告書

2.1 依存構造の同期制約に基づくニューラル機械翻訳

愛媛大学　出口　祥之
同志社大学　田村　晃裕
愛媛大学　二宮　崇

2.1.1　はじめに

近年，機械翻訳の分野において，Transformer ニューラル機械翻訳（Neural Machine Translation; NMT）モデル（以下，Transformer NMT）(Vaswani et al. 2017) が従来の回帰型ニューラルネットワークや畳み込みニューラルネットワークベースモデルの翻訳性能を上回り，注目を浴びている．特に，Transformer NMT の特徴の一つである自己注意機構は注意重みと呼ばれる確率分布行列を計算することで文内における単語間の関連の強さを捉えることができ，ここに依存構造を組み込むことで翻訳性能をさらに改善するモデルが提案されている (Deguchi et al. 2019; Wang et al. 2019; Bugliarello and Okazaki 2020)．特許翻訳のように長文が多く出現する文章を翻訳の対象とする場合には，長文に対する翻訳精度の劣化が問題となっているが，依存構造を組み込むことによって長距離の関係が適切に捉えられ，翻訳精度の劣化が抑えられることが期待される．

本稿で提案する同期注意制約は同期文法（Synchronous Grammar）や同期依存文法（Synchronous Dependency Grammar）から着想を得た手法であり，Transformer モデルに対してこの制約を与えて訓練することで翻訳性能の改善を狙う．同期文法及び同期依存文法は，2 言語で定義される文法であり，2 言語の文構造を同時に生成する文法である．これまで，統計的機械翻訳では原言語文と目的言語文間の同期文法や同期依存文法を考慮することで翻訳性能を改善してきた(Jiang et al. 2009; Ding and Palmer 2005)．提案モデルは，原言語側の自己注意機構と目的言語側の自己注意機構にそれぞれの文の依存構造を組み込んだ Transformer + DBSA (Deguchi et al. 2019) に対して，原言語側と目的言語側の自己注意機構間で両側の注意機構の整合性を保つ同期注意制約を与えて訓練する．同期注意制約を与えて訓練することにより，Transformer +DBSA において言語間で依存構造の対応が捉えられることが期待される．図 1 は対訳文における単語アライメント及び各文の依存構造の例を表している．提案モデルでは，図 1 において，“you”の親が“speak”であり，“you”が“du”に，“speak”が“Kannst”に対応しているとき，“du”の親が“Kannst”となるように学習する．自己注意機構の言語間における対応関係は，言語間注意機構によって求める．言語間注意機構は，図 1 の上下を結ぶ線のような単語のアライメント関係を捉えるという実験結果も報告されている(Garg et al. 2019).

提案モデルと従来モデルの翻訳性能を，Asian Scientic Paper Excerpt Corpus（ASPEC）日英翻訳タスク，WMT14 英独翻訳タスク，WMT16 英羅（ルーマニア語）翻訳タスクを用いてそれぞれ BLEU (Papineni et al. 2002) により比較したところ，最大 +0.38 (%) の性能改善を確認した．

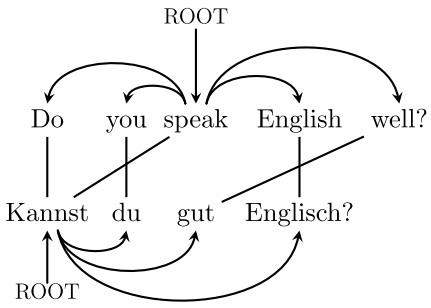

図 1 依存構造とアライメント関係の例

2.1.2 Transformer NMT

本節では，NMT モデルの Transformer NMT (Vaswani et al. 2017) について述べる．Transformer NMT のモデル図を図 2 に示す．

Transformer NMT (Vaswani et al. 2017) は，入力された原言語文 $X = (x_1, x_2, \dots, x_I)$ を符号化する Transformer エンコーダ（以下，エンコーダ）と，エンコーダの出力を受け取り目的言語文 $Y = (y_1, y_2, \dots, y_J)$, $y_j \in \mathcal{V}$ $(\forall j)$ に復号する Transformer デコーダ（以下，デコーダ）を組み合わせたエンコーダ・デコーダモデルである．ただし，$\mathcal{V}$ は出力語彙集合であり，予め訓練データから作成される．目的言語文の生成確率は式 (1) で表される．

$$Pr(Y|X;\theta) = \prod_{j=1}^{J} p\left(y_j \in \mathcal{V} | y_{<j}, X; \theta\right). \tag{1}$$

ここで，θ はモデルパラメータ，$y_{<j}$ は生成済みの単語 $(y_1, y_2, \dots, y_{j-1})$ を示す．訓練時は，訓練データ $\mathcal{D}$ に対して式 (2) に示す目的関数 $\mathcal{L}(\theta)$ を最大化する θ を式 (3) により探索する．

$$\mathcal{L}(\theta) = \frac{1}{J}\sum_{j=1}^{J} \log p\left(y_j | y_{<j}, X; \theta\right), \tag{2}$$

$$\hat{\theta} = arg\max_{\theta} \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathcal{L}(\theta)], \tag{3}$$

$$\text{where } \mathcal{D} = \left\{\left(X^{(s)}, Y^{(s)}\right)\right\}_{s=1}^{|\mathcal{D}|}. \tag{4}$$

ただし，$X^{(s)}$ 及び $Y^{(s)}$ はそれぞれ訓練データ中の s 番目の原言語文と目的言語文を示す．また，翻訳時は生成確率を最大化する目的言語文 Y^* を探索する．

$$Y^* = \arg\max_{Y} Pr(Y|X;\theta). \tag{5}$$

エンコーダ及びデコーダへの各入力単語は埋め込み層によって d_{emb} 次元の単語埋め込み空間に埋め込まれる．ただし，Transformer モデルは各単語を並列に処理するため，そのままでは文における語順の情報を扱うことができない．そのため，位置符号化（positional encoding）によって符号化した各単語の位置情報を単語埋め込み行列に加算し，エンコーダ及びデコーダに入力す

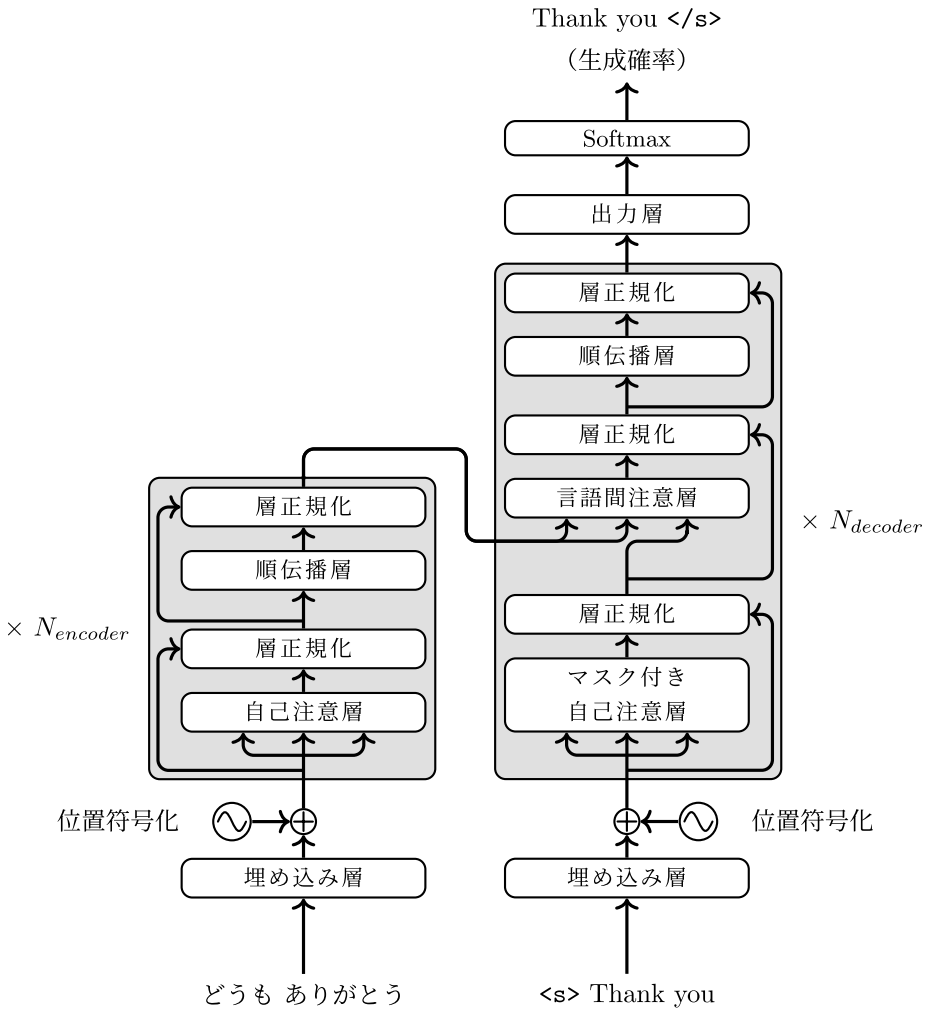

図 2 依存構造とアライメント関係の例

る．位置符号化のための行列 PE は

$$PE_{(i,2d)} = \sin\left(\frac{i}{10000^{\frac{2d}{d_{\text{emb}}}}}\right), \tag{6}$$

$$PE_{(i,2d+1)} = \cos\left(\frac{i}{10000^{\frac{2d}{d_{\text{emb}}}}}\right). \tag{7}$$

と表される．ここで，$i \geq 1$ は単語の位置，$d \geq 0$ は埋め込み空間の各次元である．

エンコーダ及びデコーダは，それぞれ N_{encoder} 層のエンコーダ層，N_{decoder} 層のデコーダ層から成る．各エンコーダ層及びデコーダ層 $Layer()$ は複数のサブレイヤ $SubLayer()$ から成り，全てのサブレイヤの出力には以下の式のように残差接続（He et al. 2016）と層正規化 $LayerNorm()$（Ba et al. 2016）が適用される．

$$Layer(Z) = LayerNorm\big(Z + SubLayer(Z)\big) \tag{8}$$

エンコーダ層は下層より順に自己注意層，順伝播層から構成され，デコーダ層は下層より順に自

己注意層，言語間注意層，順伝播層から構成される．

自己注意層内のネットワーク（以下，自己注意機構）や言語間注意層内のネットワーク（以下，言語間注意機構）には，複数ヘッド注意機構 $Attn(Q,K,V)$ が用いられる．複数ヘッド注意機構では，d_{emb} 次元の単語埋め込み空間から H 個の $d_k\left(=\frac{d_{\mathrm{emb}}}{H}\right)$ 次元のヘッド（部分空間）に射影し，各ヘッド毎に次式のように注意を計算する．

$$Attn(Q,K,V) = [M_1; M_2; \cdots; M_H]W^{\mathrm{M}}, \tag{9}$$

$$M_h = A_h V_h, \tag{10}$$

$$A_h = \mathrm{softmax}\left(\frac{Q_h K_h^{\mathrm{T}}}{\sqrt{d_k}}\right), \tag{10}$$

$$Q_h = QW_h^{\mathrm{Q}}, K_h = KW_h^{\mathrm{K}}, V_h = VW_h^{\mathrm{V}}. \tag{12}$$

なお，本稿では簡単のため，行列に対しての softmax() は各行のベクトルに対して独立に softmax 関数を適用する．複数ヘッド注意機構の入力 Q,K,V は，それぞれパラメータ行列 $W_h^{\mathrm{Q}}, W_h^{\mathrm{K}}, W_h^{\mathrm{V}} \in \mathbb{R}^{d_{\mathrm{emb}} \times d_k}$ によって，$Q_h, K_h, V_h\ (1 \le h \le H)$ に射影される．ここで，注意重み行列 A_h の各要素の値は Q の各単語と K の各単語の間の関連の強さを示している．続いて，注意重み行列 A_h と V_h の行列積を計算することで，Q_h に対応する表現を V_h から荷重和の形で得られる．その後，全ヘッドの M_h（すなわち，$M_1, M_2, \ldots, M_H$）を結合し，パラメータ行列 $W^{\mathrm{M}} \in \mathbb{R}^{Hd_k \times d_{\mathrm{emb}}}$ によって d_{emb} 次元の単語埋め込み空間に射影する．なお，自己注意機構では，入力の Q,K,V はいずれも直前のサブレイヤの出力であり，原言語文，目的言語文それぞれにおいて文中の全ての単語との関連の強さを計算する．ただし，翻訳時に目的言語文の各単語は自己回帰的に生成されるため，デコーダ側の自己注意機構は各単語の後方の単語を指さないようマスクをかけて訓練する．デコーダ側のみで用いられる言語間注意機構では，入力 Q は直前のサブレイヤの出力，K，V はどちらもエンコーダの最終層の出力であり，目的言語文の各単語と原言語文の各単語との関連の強さを計算する．

順伝播層には，次式で表される 2 層の順伝播ネットワーク（$FFN()$）が用いられる．

$$FFN(Z) = max(0, ZW^{\mathrm{in}})\, W^{\mathrm{out}}. \tag{13}$$

ただし，$W^{\mathrm{in}} \in \mathbb{R}^{d_{\mathrm{emb}} \times d_{\mathrm{inner}}}$，$W^{\mathrm{out}} \in \mathbb{R}^{d_{\mathrm{inner}} \times d_{\mathrm{emb}}}$ はパラメータ行列である．

最後に，デコーダの最終層の出力を $|\mathcal{V}|$ 次元空間に線形変換し，各単語毎に softmax 関数を適用することで，単語出力確率を得る．

$$p(y_j | y_{<j}, X; \theta) = \mathrm{softmax}(z_j W^{\mathcal{V}}). \tag{14}$$

ただし，z_j はデコーダ最終層出力の j 単語目のベクトルであり，$W^{\mathcal{V}} \in \mathbb{R}^{d_{\mathrm{emb}} \times |\mathcal{V}|}$ はパラメータ行列である．

2.1.3 提案モデル

本稿で提案するモデルは，原言語側と目的言語側の自己注意機構でそれぞれの文の依存構造を捉え，それらを同期注意制約によって対応付ける．

2.1.3.1 依存構造に基づいた自己注意機構

はじめに，提案モデルの基礎となる依存構造に基づいた自己注意機構（dependency-based self-attention; DBSA）（Deguchi et al. 2019）について説明する．DBSA を用いた Transformer NMT（以下, Transformer+DBSA）は, エンコーダ及びデコーダの l_{dep} 層目の自己注意機構を拡張し，原言語文及び目的言語文の依存構造を捉える．Transformer＋DBSA の訓練は，従来の NMT の目的関数の代わりに次の目的関数 $\mathcal{L}(\theta)$ を最大化し，翻訳と依存構造解析を同時に学習する．

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{TM}}(\theta) + \lambda_{\mathrm{dep}}\mathcal{L}_{\mathrm{dep}}(\theta), \tag{15}$$

$$L_{\mathrm{dep}}(\theta) = \frac{1}{I}\sum_{i=1}^{I} \log p(head(x_i)|X;\theta) + \frac{1}{J}\sum_{j=1}^{J} \log p\left(head\left(y_j\right)\middle|y_{<j},X;\theta\right). \tag{16}$$

ここで, $\mathcal{L}_{\mathrm{TM}}(\theta)$ は式 (2) のような従来の翻訳モデルの目的関数, $\mathcal{L}_{\mathrm{dep}}(\theta)$ は原言語文と目的言語文それぞれの各単語の親を予測する目的関数, $head(t)$ は単語 t の親を返す関数, $\lambda_{\mathrm{dep}} > 0$ は $\mathcal{L}$ において $\mathcal{L}_{\mathrm{dep}}$ が影響する度合いを調整するハイパーパラメータである．

$\mathcal{L}_{\mathrm{dep}}$ は具体的に以下の手順に従って算出する．まず，依存構造を捉えるヘッドを h_{dep} とすると，注意重み行列 $A_{h_{\mathrm{dep}}}$ は次のような双線形変換（Dozat and Manning 2017）によって求められる．

$$A_{h_{\mathrm{dep}}} = \mathrm{softmax}\left(\frac{Q_{h_{\mathrm{dep}}} U K_{h_{\mathrm{dep}}}^{\top}}{\sqrt{d_k}}\right). \tag{17}$$

ここで, $U \in \mathbb{R}^{d_k \times d_k}$ はパラメータ行列である．また，$A_{h_{\mathrm{dep}}}$ の要素 $A_{h_{\mathrm{dep}}}[t,q]$ は，以下のように，単語 q が単語 t の親である確率を示す．

$$p(q = head(t)|\mathcal{X}) = A_{h_{\mathrm{dep}}}[t,q]. \tag{18}$$

ここで，$\mathcal{X}$ はモデルへの入力文である．ただし，親がROOTとなる単語 t_{ROOT} は自分自身を指すよう，以下のように定義する．

$$t_{\mathrm{ROOT}} = head(t_{\mathrm{ROOT}}). \tag{19}$$

次に，以下のように $A_{h_{\mathrm{dep}}}$ と $V_{h_{\mathrm{dep}}}$ の行列積を計算することで単語間の依存関係を重みとする荷重和による表現を得る．

$$M_{h_{\mathrm{dep}}} = A_{h_{\mathrm{dep}}} V_{h_{\mathrm{dep}}}. \tag{20}$$

最後に，通常の自己注意機構と同様，他のヘッドと結合し，単語埋め込み空間に線形変換する．

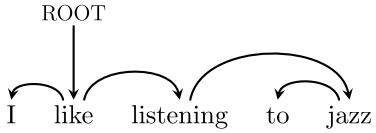

図 3 依存構造の例

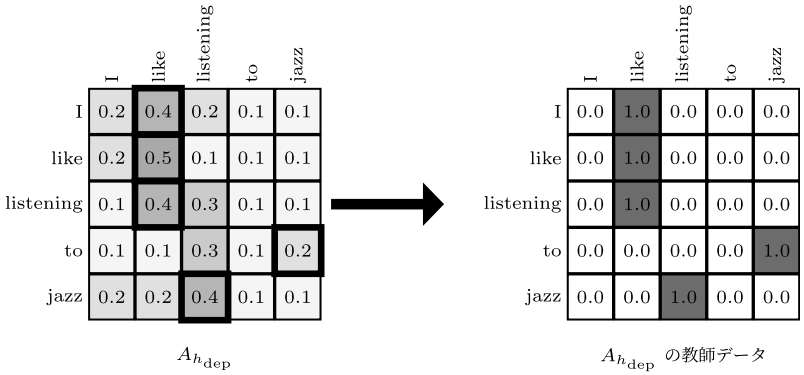

図 4 DBSA の概要図

$$M = \left[M_{h_{\mathrm{dep}}}; M_1; \ldots; M_{H-1}\right] W^{\mathrm{M}}, \tag{21}$$

以上のとおり，DBSA は文中の単語間の依存関係を第 l_{dep} 層の自己注意機構内のヘッド h_{dep} で捉える．

なお，翻訳時は通常の Transformer NMT モデルと同様，目的言語文の各単語を自己回帰的に生成するため，デコーダ側の自己注意機構は生成済みの単語（自身より前方の単語）に対する依存関係しか捉えられない．そのため，親が後方にある単語については学習しないようにマスクをかけて訓練する．

図 3，図 4 にそれぞれ依存構造の例とそれに対応する注意重み行列と教師データの例を示す．図 4 の行列の各要素は行毎に正規化された確率分布となっており，色が濃いほど値が高いことを示す．図 4 の $A_{h_{\mathrm{dep}}}$ は図 3 の依存構造に基づいた教師データに近づくように学習される．

DBSA は，現在 NMT で広く用いられているサブワード単位（Sennrich et al. 2016; Kudo and Richardson 2018）にも拡張されている．ある単語が複数のサブワードに分割された場合，その単語を構成する各サブワードの親は隣接する後方のサブワードとし，単語の終端となるサブワードの親はその単語の親に設定する．また，親となる単語が複数のサブワードに分割されている場合，親は先頭のサブワードとする．

2.1.3.2 提案手法：同期注意制約

提案モデルの目的関数は，式 (15) で示した目的関数の代わりに，次の目的関数 $\mathcal{L}(\theta)$ を最大化し，翻訳と依存構造解析に加え，言語間の依存構造対応を学習する．

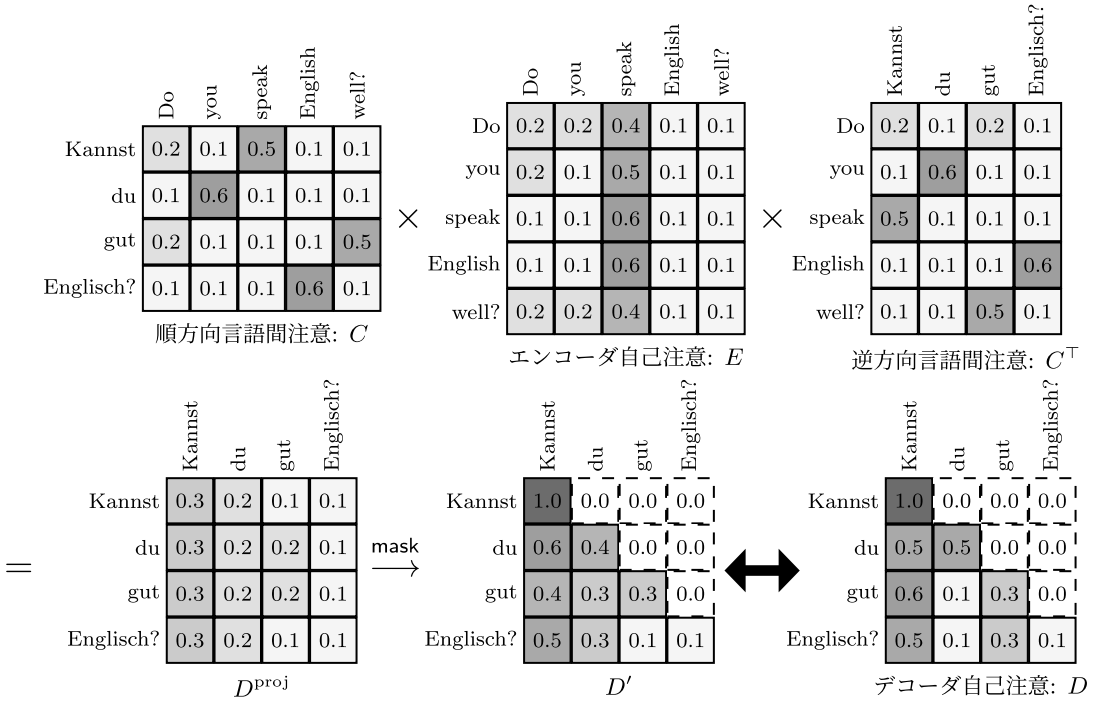

図 5 同期注意制約の計算例

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{TM}}(\theta) + \lambda_{\text{dep}}\mathcal{L}_{\text{dep}}(\theta) + \lambda_{\text{sync}}\mathcal{L}_{\text{sync}}(\theta). \tag{22}$$

なお，λ_{sync} は目的関数において $\mathcal{L}_{\text{sync}}$ が影響する度合いを調整するハイパーパラメータである．

同期注意制約 $\mathcal{L}_{\text{sync}}$ は，言語間注意機構によって，原言語側のエンコーダ自己注意機構を目的言語側の確率空間に変換し，デコーダ自己注意機構との差を計算することで得られる．目的言語側の空間に変換されたエンコーダ自己注意機構を D^{proj} とすると，以下の式によって D^{proj} が求められる．

$$D^{\text{proj}} = CEC^{\top}. \tag{23}$$

ただし，E と D はそれぞれエンコーダ及びデコーダの l_{dep} 層目の自己注意機構の注意重み行列 $A_{h_{\text{dep}}}$ であり，E に対して左から掛ける C は順方向（目的言語側から原言語側へ）の注意，右から掛ける $C^{\top}$ は逆方向（原言語側から目的言語側へ）の注意となる．なお，Q_h と K_h はデコーダの第 l_{sync} 層の言語間注意の入力である．次に，デコーダ自己注意機構は各単語における後方の単語をマスクして訓練するため，同様に D^{proj} に対しても後方の単語を指さないようにマスクをかけて softmax 関数を適用する．

$$D' = \text{softmax}\Big(\text{mask}(D^{\text{proj}})\Big). \tag{24}$$

ただし，mask() は各単語における後方の単語をマスクする関数である．

最後に，D' と D の自乗誤差を計算し，それらの分布の近さ $\mathcal{L}_{\text{sync}}$ を得る．

表 1 提案手法の同期注意制約に関するハイパーパラメータ $\boldsymbol{\lambda_{\text{sync}}}$

	λ_{sync}
WMT14 英独	0.5
WMT16 英羅	0.1
ASPEC 日英	10.0

表 2 Transformer モデルのハイパーパラメータ

ラベル平滑化	$\epsilon = 0.1$
ドロップアウト	$p = 0.1$
最適化器	Adam ($\beta_1 = 0.9, \beta_2 = 0.98$)
学習率ウォームアップ	4000 回更新まで線形的に増加
学習率調整	更新回数の逆平方根に比例して減衰
最大学習率	7e-4
文長正規化	$\alpha = 0.6$ (Wu et al. 2016)

$$\mathcal{L}_{\text{sync}}(\theta) = -\frac{1}{J}\sum_{j=1}^{J}\sum_{t}\left(D'_{j,t} - D_{j,t}\right)^2. \tag{25}$$

図 5 に同期注意制約の例を示す. 図 5 に示すように, エンコーダ自己注意 E は言語間注意 C, $C^{\top}$ によって目的言語側の確率空間に変換される．制約はデコーダ自己注意機構の注意重み行列 D と求めた D' との差から計算される.

2.1.4 実験

2.1.4.1 実験設定

本実験では，提案手法の有効性を確認するため，提案手法を適用したモデルと従来モデルとの翻訳性能を比較する．Transformer モデルには *base* (Vaswani et al. 2017) モデルを使用した．翻訳性能の評価実験は，WMT14 英独翻訳タスク，WMT16 英羅（羅: ルーマニア）翻訳タスク，Asian Scientic Paper Excerpt Corpus（ASPEC）(Nakazawa et al. 2016) 日英翻訳タスクを用いた．ASPEC 日英において訓練データは上位 150 万文対を抽出して用いた．単語分割器は，英語文，独語文，羅語文には Moses トークナイザを，日本語文には KyTea (Neubig and Mori 2010) を用いた．単語分割した後，訓練データから学習した BPE (Sennrich et al. 2016) によりサブワード分割を行った．ただし，ASPEC 日英においては原言語側と目的言語側で独立して BPE を学習した．BPE の語彙量は，WMT14 英独で 37,000，WMT16 英羅で 40,000，ASPEC 日英で原言語側と目的言語側で独立してそれぞれ 16,000 に設定した．DBSA の教師データは依存構造解析器の出力結果を用い，日本語には EDA (Flannery et al. 2011)，それ以外の言語には Stanza

表 3 実験結果（BLEU (%)）

モデル	WMT14 En→De	WMT16 En→Ro	ASPEC Ja→En
Transformer	27.23	23.83	28.94
+DBSA	27.31	24.13	29.57
++SyncAttn	**27.69**	**24.33**	**29.84**

表 4 ASPEC 日英翻訳における提案手法の切除実験

+DBSA	+SyncAttn	BLEU (%)
—	—	28.94
○	—	29.57
—	○	29.48
○	○	**29.84**

(Qi et al. 2020) を使用した．ただし，解析器は訓練時の教師データ作成のためのみに使用し，翻訳時には使用していない．

目的関数の $\mathcal{L}_{\mathrm{TM}}$ にはラベル平滑化を用いた．目的関数における項の影響する度合いを調整するハイパーパラメータは依存構造に関する項においては $\lambda_{\mathrm{dep}} = 0.5$，提案手法の同期注意制約に関する項においては表 1 に示す．同期注意制約を与える層 l_{dep} は Transformer *base* モデルの第 1 層目とし，D' を求める際に必要な言語間注意の層 l_{sync} は第 5 層目（Garg et al. 2019）とした．モデルのパラメータ更新回数は 100,000 回とした．ミニバッチの大きさは WMT14 英独で約 25,000 トークン, WMT16 英羅で約 6,000 トークン, ASPEC 日英で約 12,000 トークンになるよう設定した．翻訳文の生成にはビーム探索を用い，ビーム幅は 4 とした．モデルの詳細なハイパーパラメータは表 2 に示す．

2.1.4.2 実験結果

実験結果を表 3 に示す．表中の SyncAttn は同期注意制約を表す．表 3 より，提案モデルの ++SyncAttn の性能がベースラインモデルの Transformer や +DBSA と比べて全実験において改善されていることがわかる．具体的には，++SyncAttn によって，+DBSA と比較して，WMT14 英独，WMT16 英羅，ASPEC 日英においてそれぞれ 0.38, 0.20, 0.27 BLEU (%) の性能改善を確認した．

2.1.4.3 切除実験

本節では，依存構造を組み込まない従来の Transformer モデルに対して同期注意制約を組み込んだモデルの翻訳性能を評価した．翻訳実験には ASPEC 日英翻訳タスクを用いた．

実験結果を表 4 に示す. 表より, DBSA と SyncAttn を組み合わせることで, 従来の DBSA,

SyncAttn 単体で用いたモデルよりそれぞれ +0.27，+0.36 BLEU (%) の性能改善を確認した．

2.1.5 おわりに

本稿では，Transformer NMT において，自己注意機構によって原言語文の依存構造と目的言語文の依存構造を捉え，言語間注意機構によって言語間の依存構造対応を学習するモデルを提案した．提案モデルを用いることで，WMT14 英独翻訳，WMT16 英羅翻訳，ASPEC 日英翻訳タスクにおいて，従来モデルよりも最大 +0.38 BLEU (%) の性能改善が確認された．今後は，依存構造の対応付けだけでなく，句構造など他の文構造に対しても提案手法の有効性を確認したい．

参考文献

- Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). ``Layer Normalization.''.
- Bugliarello, E. and Okazaki, N. (2020). ``Enhancing Machine Translation with Dependency-Aware Self-Attention.'' In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1618--1627, Online.
- Deguchi, H., Tamura, A., and Ninomiya, T. (2019). ``Dependency-Based Self-Attention for Transformer NMT.'' In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 239--246, Varna, Bulgaria.
- Ding, Y. and Palmer, M. (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, Ann Arbor, Michigan.
- Dozat, T. and Manning, C. D. (2017). ``Deep Biaffine Attention for Neural Dependency Parsing.'' In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.
- Flannery, D., Miayo, Y., Neubig, G., and Mori, S. (2011). ``Training Dependency Parsers from Partially Annotated Corpora.'' In Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 776--784, Chiang Mai, Thailand.
- Garg, S., Peitz, S., Nallasamy, U., and Paulik, M. (2019b). ``Jointly Learning to Align and Trans-late with Transformer Models.'' In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4453--4462, Hong Kong, China.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). ``Deep Residual Learning for Image Recognition.'' In Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770--778.
- Jiang, H., Yang, M., Zhao, T., Li, S., and Wang, B. (2009). ``A Statistical Machine Translation Model Based on a Synthetic Synchronous Grammar.'' In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 125--128, Suntec, Singapore.
- Kudo, T. and Richardson, J. (2018). ``SentencePiece: A simple and language independent

subword tokenizer and detokenizer for Neural Text Processing." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66--71, Brussels, Belgium.

- Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ``ASPEC: Asian Scientific Paper Excerpt Corpus." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2204--2208, Portoroz, Slovenia.
- Neubig, G. and Mori, S. (2010). ``Word-based Partial Annotation for Efficient Corpus Construction." In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). ``Bleu: a Method for Automatic Evaluation of Machine Translation." In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311--318, Philadelphia, Pennsylvania, USA.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). ``Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 101--108, Online.
- Sennrich, R., Haddow, B., and Birch, A. (2016). ``Neural Machine Translation of Rare Words with Subword Units." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715--1725, Berlin, Germany.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). ``Attention is All you Need." In Advances in Neural Information Processing Systems 30, pp. 5998--6008.
- Wang, X., Tu, Z., Wang, L., and Shi, S. (2019). ``Self-Attention with Structural Position Representations." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1403--1409, Hong Kong, China.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Lukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). ``Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.".

2.2 特許翻訳における Multi-Source ニューラル機械翻訳の有効性

磯部 僚也,
(筑波大学大学院システム情報工学研究科知能機能システム専攻)
宇津呂 武仁,
(筑波大学システム情報系知能機能工学域)
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

多言語 NMT において，複数原言語の対訳文を参照して目的言語の一文に翻訳を行う(n-to-1) Multi-Source 翻訳手法[7]においては，モデルの訓練および評価においては全言語の対訳文が揃うことが前提とされるが，実世界のコーパスにおいてこの条件が満たされる状況は極めて限られる．一方，複数翻訳方向の 1-to-1 翻訳が混在した Google の多言語翻訳[1]に用いられるモデルにおいても，三言語以上の対訳関係を無視し独立な二言語対とみなすため，評価時に複数の原言語の文が存在したとしても，その利点をいかすことができない点が弱点である．そこで本論文では，三言語対訳コーパスが利用できる場合を対象として，これを有効活用する Multi-Source Many-to-One 手法を提案する．

本論文では，複数の原言語文を文区切り記号連結したものを入力とする文連結型 Multi-Source 翻訳と通常の Single-Source 翻訳(1-to-1 翻訳)を，入力に言語タグを付与することにより区別して一つのモデルとして訓練する方法を Multi-Source Many-to-One 翻訳と定義する．具体的には，英中日の三言語対訳，英日対訳，中日対訳があるとき，英語文と日本語文を連結した入力には“<ENZH2JA>”タグを，英語文の入力には“<EN2JA>”タグを，中国文の入力には“<ZH2JA>”タグを，それぞれ付与して，日本語文を出力する一つのモデルを訓練する．

また，本論文の比較対象として，Nishimura ら[3]の方法では，そもそも複数エンコーダで複数言語入力を実現しているので，単一言語入力の翻訳モデルとパラメータを共有することは不可能である．しかし，単一言語入力の場合には，Nishimura ら[3]の方法において三言語対訳の欠落を表現するのに使われている“<NULL>”トークンを使用することにより，文連結型 Multi-Source 翻訳において Nishimura ら[3]の方法を近似したものを，本研究の一つの比較対象とする．一方，Google 多言語翻訳のうちの Many-to-One 翻訳は，出力言語が同一である複数の言語対に対して，一つのモデルとして訓練する方法である．しかし，この方法では，英中日の三言語対訳データのように，三つ以上の言語が互いに対訳になっている場合に，これらが互いに対訳になっていることを利用することができない．そこで，本論文では，この Google の Many-to-One 翻訳を参考にした提案法の非 Multi-Source 版(すなわち，Many-to-One 翻訳)をもう一つの比較対象とする．以上の二つの比較実験を行うために，特許の三言語対訳データから表 1 のデータを作成し，図 1 に示す形でモデルの比較を行う．

文番号	英	中	日
1-200,000	20 万	20 万	20 万
200,001-600,000	40 万	X	40 万
600,001-1,000,000	X	40 万	40 万

表 1:特許三言語対訳コーパスの欠落設定 (各数字は欠落のない文数を示す．“X”は欠落した部分を示す)

2.2.2 多言語 NMT の関連研究

2.2.2.1 Multi-Source 翻訳

Zoph ら[7]はマルチエンコーダの再帰型ニューラルネットワーク(RNN)モデルを用いて Multi-Source 翻訳を実現した．各エンコーダは各言語の原言語文を処理する．1-to-1 翻訳モデルよりも翻訳精度が改善しているが，全言語の対訳訓練文を用意する必要がある．実世界のコーパスにおいてこの条件が満たされる状況は，非常に限定される．この問題に対して，Nishimura らは，欠落した言語の部分を，トークン[3]や自動生成した疑似対訳文[2]で置き換える手法を提案した．

2.2.2.2 Google 多言語翻訳

Johnson ら[1]は，英語を含む三言語コーパスを中心に，異なる翻訳方向の 1-to-1 文対を混在させた翻訳手法を提案した．この手法では，三言語以上の文対応関係が含まれる場合もこれを利用せず，全ての言語対を 1-to-1 の対応に変換して翻訳モデルの訓練を行う．原言語を英語とし，目的言語を他二言語とする One-to-Many 翻訳，原言語を他二言語とし，目的言語を英語とする Many-to-One 翻訳，原言語と目的言語両方を三言語とする Many-to-Many 翻訳の三種類の設定に分けられている．出力時の目的言語を指定するため，入力文の先頭に，該当する言語タグを付与する．本論文では，上述の三種類の設定の中でも，特に Many-to-One 翻訳を参考にして提案法の非 Multi-Source 版(すなわち，Many-to-One 翻訳)を作成し，比較評価を行う．

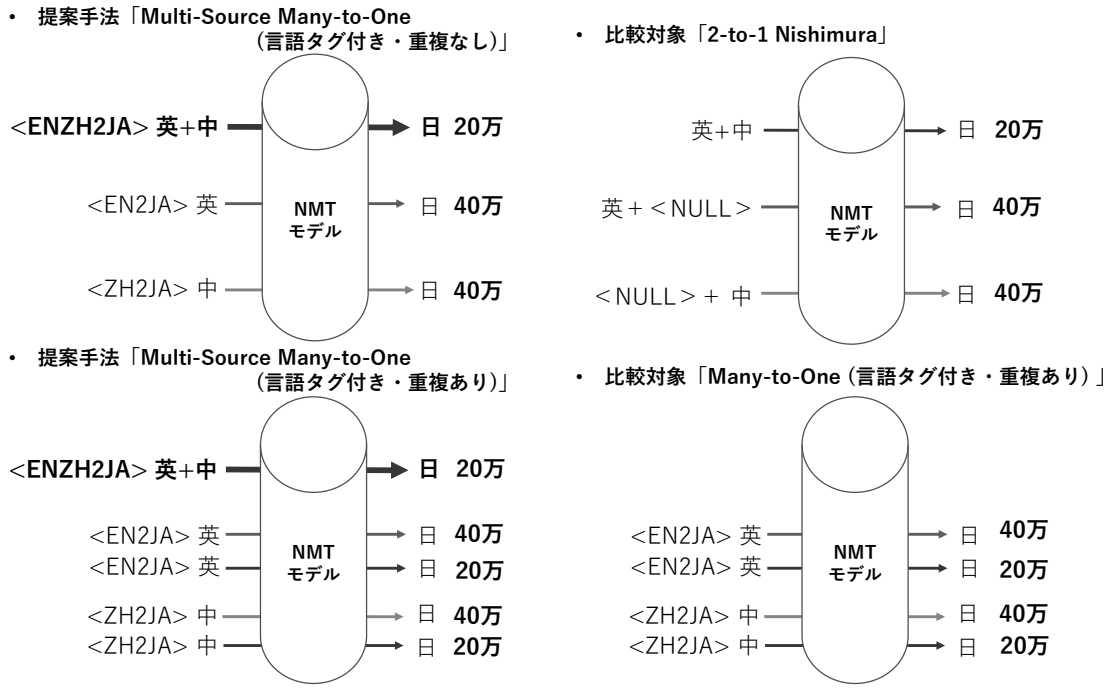

図 1: 提案手法と先行研究のモデルの比較

SOURCE	*TARGET*
The present invention relates to a cover glass . <CONCAT> 本发明涉及覆盖玻璃。	本発明は、カバーガラスに関する。
The sample S has a known shape . <CONCAT> 样品S为已知的形状。	サンプルSは既知の形状である。
The process may terminate thereafter . <CONCAT> 该过程此后可以结束。	その後、プロセスは終了してもよい。

図 2: 二文連結による{英, 中}-to-日の訓練データ例
(実際の訓練過程では，訓練前に 4 節の前処理を適用する)

2.2.3 特許三言語対訳コーパス

本論文では，日本，中国，米国の 2005~2017 年特許データからパテントファミリーに基づいて対訳文書対を抽出し，文献[5]の手法により文対応を付与し，英語と中国語を原言語，日本語を目的言語として[1]，日中英三言語対訳コーパスを作成した．その結果，100 万訓練文対，2 万開発文対，および，2 万評価文対の三言語対訳コーパスが作成された．

[1] 英語の Tokenization は Moses Tokenizer (https://github.com/moses-smt/mosesdecoder/)，中国語の形態素解析は Jieba (https://github.com/fxsjy/jieba)，日本語は文字単位に分割した．英語文と中国語文の合計文長が 1,000 トークン以上となる対訳文対は除外した．

2.2.4 提案手法

2.2.4.1 文連結型 Multi-Source 翻訳

本論文では，文連結手法を用いて三言語対訳コーパスを有効利用する手法を提案する．本論文では，“<CONCAT>”トークンを用いて複数の言語の原言語文を連結する．連結後の対訳文対の例を図 2 に示す．このように，連結された二文を一文として扱い，翻訳モデルの訓練・評価を行う．このアプローチによれば，単エンコーダモデルによって 2-to-1 翻訳が実現可能となるため，マルチエンコーダの実装が不要となる．

さらに，本論文では，Nishimura ら[3,2]が提案した手法において，同様の文連結手法を導入する方式の翻訳精度の評価も行った．文献[2]に従う場合には，No.1~No.200,000 の文を用いて{英，日}-to-中と{中，日}-to-英の翻訳モデルを訓練した後，表 1 の欠落部分“X”に相当する翻訳文を自動生成し全体を疑似対訳文として扱い，この手法を“2to1 Nishimura (疑似対訳) ”と呼ぶ．一方，文献[3]に従う場合には，表 1 の欠落部分“X”を“<NULL>”トークンに置き換え，この手法を“2to1 Nishimura (NULL)”と呼ぶ．

2.2.4.2 文連結型 Multi-Source 翻訳

本論文では，欠落を含む三言語対訳コーパス(表 1)において，日本語を目的言語とした Many-to-One 翻訳に着目し，Multi-Source Many-to-One 翻訳の方式を提案する．本論文の各提案手法，および，比較対象の手法との間の関係を図 1 に示す．

この手法においては，まず，4.1 節の文連結 2-to-1 翻訳手法を No.1~No.200,000 の三言語対訳文に適用し，{英，中}-to-日翻訳訓練用対訳文対を生成し，連結後の原言語文対の先頭に“<ENZH2JA>”タグを付与する．次に，No.200,001~No.600,000 の英日対訳文から，1-to-1 翻訳用に英-to-日翻訳訓練用対訳文対を作成し，各原言語文の先頭に“<EN2JA>”タグを付与する．さらに， No.600,001~No.1,000,000 の中日対訳文から，1-to-1 翻訳用に中-to-日翻訳訓練用対訳文対を作成し，各原言語文の先頭に“<ZH2JA>”タグを付与する．以上の訓練用対訳文対を用いて訓練した翻訳モデルを“Multi-Source Many-to-One 翻訳(言語タグ付き・重複なし)”モデルと呼ぶ．このモデルの直接の比較対象は，Nishimura ら[3]が提案した“2to1 Nishimura (NULL)”モデルである．

さらに，No.1~No.200,000 の三言語対訳文から，1-to-1 翻訳用の英-to-日翻訳訓練用対訳文対と中-to-日翻訳訓練用対訳文対をそれぞれ生成し，“Multi-Source Many-to-One 翻訳(言語タグ付き・重複なし)”モデルの訓練用対訳文対に追加して訓練を行った翻訳モデルを，“Multi-Source Many-to-One 翻訳(言語タグ付き・重複あり)”モデルと呼ぶ．このモデルの直接の比較対象は，Google の Many-to-One 翻訳を参考にした提案法の非 Multi-Source 版(すなわち，Many-to-One 翻訳)である．このモデルの訓練用対訳文対は，“Multi-Source Many-to-One 翻訳(言語タグ付き・重複あり)”モデルの訓練用対訳文対から{英，中}-to-日翻訳訓練用対訳文対を除外したもの(図 1 参照)であり，また，Johnson ら[1]のモデルと比べて原言語を識別するタグを付与している点が異なる[2]

[2] 図 1 中の赤色の 1-to-1 翻訳用の英-to-日翻訳訓練用対訳文対 20 万文対と中-to-日翻訳訓練用対訳文対 20 万文対における日本語文が重複している点に注意する．

2.2.5 評価

評価結果を表 2 に示す[3].

欠落なしの訓練文規模 100 万文対において，2-to-1 翻訳の BLEU スコアが 1-to-1 翻訳より 4.19 ポイント有意に改善した．次に，欠落を考慮した特許コーパスにおける BLEU 評価結果では，欠落なしの場合のベースライン翻訳モデルとして，(i) 2-to-1 翻訳モデルでは，表 1 の No.1~No.200,000 の文を用いて訓練した「欠落なし (20 万対訳文対)」のモデル，(ii)1-to-1 翻訳モデルでは，英-to-日翻訳において No.1~No.600,000 の文を用いて訓練した「欠落なし (60 万対訳文対)」モデル，および，中-to-日翻訳において No.1~No.200,000 の文と No.600,001~No.1,000,000 の文を用いて訓練した「欠落なし (60 万対訳文対)」モデル，をそれぞれ評価した．2-to-1 翻訳においては，本論文で提案した文連結型 2-to-1 翻訳モデルは，Nishimura らの手法[2,3]に対しても適用可能であることが示された．さらに，本論文の提案手法である“Multi-Source Many-to-One 翻訳(言語タグ付き)”の「重複あり・なし」両モデルとも，ベースラインモデルの BLEU を有意に改善した[4]．一方，1-to-1 翻訳においては，“Many-to-One 翻訳(言語タグ付き・重複あり)”モデル，および，提案手法の“Multi-Source Many-to-One 翻訳(言語タグ付き・重複あり)”モデルは，ベースラインの BLEU を有意に改善した．“2to1 Nishimura (NULL)”モデルとの比較においては，“Multi-Source Many-to-One 翻訳(言語タグ付き・重複あり)”モデルによる 2-to-1 翻訳の BLEU 70.63 は“2to1 Nishimura (NULL)”モデルの BLEU70.04 を有意($p<0.01$)に改善し，1-to-1 翻訳の BLEU 66.29 および 66.71 は“2to1 Nishimura (NULL)”モデルの BLEU64.60 および 62.68 を有意($p<0.01$)に改善した．一方，“Many-to-One 翻訳(言語タグ付き・重複あり)”モデルの 1-to-1 翻訳との比較においては，“Multi-Source Many-to-One 翻訳(言語タグ付き・重複あり)”モデルにより同等の BLEU を達成しており，それに加えて，単一のモデルにより 2-to-1 翻訳も行うことができている．

[3] 本論文では，fairseq ツールキット[4]を用いて実装された Transformer モデル[6]を用いた．Head の数を 4，エンコーダとデコーダを各 6 層，単語分散表現を 512 次元，隠れ層を 1,024 次元，学習率を 0.0003 とし，Adam optimizer を使用した．ドロップアウトを 0.3 として，50 エポックの訓練を行う．ハードウェアとして，NVIDIA Tesla P100 16GB GPU 1 枚を使用した．評価時，モデルの日本語出力文は MeCab の IPADic 辞書の形態素単位に分割して評価を行った．BLEU スコアの評価においては，Moses デコーダー(https://github.com/moses-smt/mosesdecoder)のスクリプト(multi-bleu.perl)を使用した．BLEU スコアの有意差検定においては，mteval Toolkit (https://github.com/odashi/mteval)を使用した．

[4]「重複あり・なし」両モデル間の BLEU の差においても有意差がある($p<0.01$)．“Multi-Source Many-to-One 翻訳(言語タグ付き・重複あり)”モデルにおいては，擬似対訳を用いなくても，疑似対訳を用いる“2to1 Nishimura (疑似対訳)”モデルと同等の翻訳精度を達成できており，訓練効率の点からも優位性が大きいと言える．

表 2: 特許コーパスのBLEU評価における欠落の有無の比較評価(単一モデルにおいて複数の翻訳方向を評価する場合を＊で示す．1-to-1翻訳ベースラインに対して有意差がある(p<0.01)場合を†で示す．2-to-1翻訳ベースラインに対して有意差がある(p<0.01)場合を‡で示す．)

(a) 欠落なし

設定	英-to-日	中-to-日	{英，中}-to-日
欠落なし(100 万対訳文対)	66.42	66.99	71.18
欠落なし(20 万対訳文対)	62.59	62.36	66.69 (2-to-1 翻訳ベースライン)
欠落なし(60 万対訳文対) (1-to-1 翻訳ベースライン)	65.89	64.99	N/A

(b) 欠落あり(訓練事例：100 万対訳文対)

設定		英-to-日	中-to-日	{英，中}-to-日
2to1 Nishimura (疑似対訳) [2]		N/A	N/A	70.63‡
*2to1 Nishimura (NULL) [3]		64.60	62.68	70.04‡
* Many-to-One 翻訳 (言語タグ付き・重複あり)		66.31†	66.03†	N/A
*Multi- Source Many-to-One 翻訳 (言語タグ付き)	重複なし	64.32	62.70	69.84‡
	重複あり	66.29†	66.71†	70.63‡

2.2.6 実例分析

{英，中}-to-日翻訳モデルによる Multi-Source 翻訳の例を表 3 に示す．英単語“breaker”の日本語訳は「器」と「装置」の可能性があるが，中国語では日本語と同じ漢字「装置」が使われているので，中日方向の翻訳においては訳語の曖昧性の問題がない．このため，“breaker”は，中-to-日翻訳モデル，および，{英，中}-to-日翻訳モデルにおいては，正しく日本語訳された．

Multi-Source 翻訳過程における自己注意ヒートマップを図 3 に示す．この図において，注意の値は 4 つの Head の平均値となる．“Normally”-「通常」，および，“breaker”-「切断装置」等，英語と中国語の原言語文の間での注意の対応が適切にとれていることが分かった．提案手法である文連結型 2-to-1 翻訳手法においては，この例のように，連結された Multi-Source 文において，このように自己注意が適切に機能することによって，高い翻訳精度を達成できていると考えられる．

表 3:特許データにおける翻訳例 ({ 英，中 }-to-日翻訳)

原言語文・参照文・モデル	文	BLEU
原言語文(英)	Normally , the IGBT 14 of the current breaker 10 is off .	N/A
原言語文(中)	在通常时，电流切断 装置 10 的 IGBT14 断开 。	N/A
参照文	通常時は、電流遮断 装置 10 の IGBT14 はオフしている。	N/A
英-to-日	通常、電流遮断 器 10 の IGBT14 はオフである。	37.79
中-to-日	通常時には、電流遮断 装置 10 の IGBT14 がオフする。	54.41
{英，中}-to-日	通常時には、電流遮断 装置 10 の IGBT14 はオフしている。	85.79

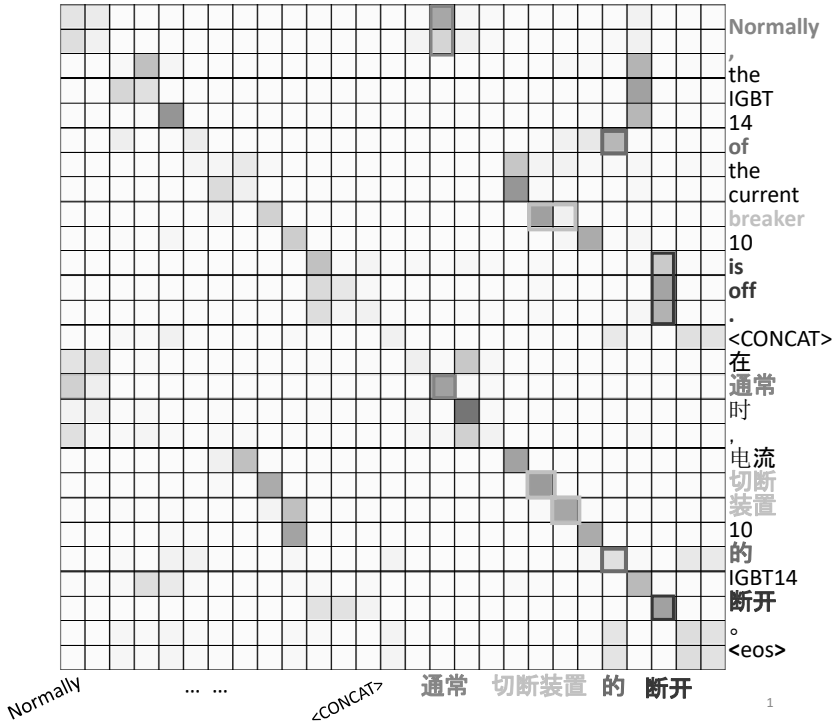

図 3:エンコーダ 6 層目の自注意機構におけるヒートマップの例

2.2.7 おわりに

本論文では，出力言語が同じである複数の言語対を一つのモデルにする Many-to-One 翻訳の考え方を拡張した Multi-Source Many-to-One 翻訳モデルを提案した．評価実験においては，Nishimura ら[3]のマルチエンコーダを用いた Multi-Source 翻訳モデルを文連結により実現したモデルと比較して，2-to-1 翻訳・1-to-1 翻訳とも BLEU を有意に改善することを示した．この方法は見方を変えると，Many-to-One 翻訳において，三言語以上の対訳が存在している場合には，三言語対訳データであることを活用する拡張となっている．さらに，評価実験においては，1-to-1 翻訳においては Google の Many-to-One モデルを参考にした”Many-to-One 翻訳(言語タグ付き・重複あり)”モデルと同等の BLEU を達成し，かつ，単一のモデルにより 2-to-1 翻訳も行うことができることを示した．

参考文献

[1] M. Johnson, M. Schuster, Q. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viegas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of ACL, Vol. 5, pp. 339–351, 2017.

[2] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura. Multi-source neural machine translation with data augmentation. In Proc. 15th IWSLT, pp. 48–53, 2018.

[3] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura. Multi-source neural machine translation with missing data. In Proc. 2nd WNGT, pp. 92–99, 2018.

[4] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. Fairseq: A fast, extensible toolkit for sequence modeling. In Proc. NAACL-HLT: Demonstrations, pp. 48–53, 2019.

[5] M. Uchiyama and H. Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In Proc. 41th ACL, pp. 72–79, 2003.

[6] 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.

[7] B. Zoph and K. Knight. Multi-source neural translation. In Proc. NAACL-HLT, pp. 30–34, 2016.

2.3 複数の単語分散表現モデルに基づく

単語ベクトルベースの自動評価法における性能評価

北海学園大学　越前谷　博

2.3.1　はじめに

自動評価はディープラーニング技術の進展により、初期の n-gram や編集距離などの表層情報に基づく単語ベースの自動評価法[1][2][3][4][5]に加え、ニューラルネットに基づく自動評価法の研究が盛んに行われている。表層情報に基づく自動評価法がモデルフリーの自動評価法と呼ばれるのに対して、ニューラルネットのモデルを用いる自動評価法はモデルベースの自動評価法と呼ばれる。また、モデルベースの自動評価はトレーニングベースの自動評価法[6][7][8][9][10]とノントレーニングベースの自動評価法[11][12][13][14]に大きく分類される。このモデルベースの自動評価法は近年より大きな注目を集めており、盛んに研究が行われている。

このような自動評価法の現状において、本報告では複数の事前学習済み単語分散表現モデルを組み合わせた単語ベクトルベースの自動評価法を提案する。提案手法では訳文と参照訳を構成する単語を単語分散表現モデルによりベクトル化する。そして、得られた単語のベクトル表現を訳文と参照訳それぞれの単語の特徴量とみなすことで、2 つの分布間の距離を求めるための Earth Mover's Distance (EMD)[15]に適用し、評価スコアを計算する。本報告ではこの単語ベクトルを用いた EMD に基づく自動評価法[16]における複数の単語分散表現モデルの利用とその有効性について述べる。その際、単語分散表現モデルとして fastText[17][18]と BERT[19][20]を使用する。さらに性能評価実験は WMT18 と WMT19 の Metrics Shared Task データ[21][22]を用いてセグメントレベルの相関係数を求めることで行う。性能評価実験の結果、fastText 及び BERT を単独で使用するよりも、この 2 つのモデルを組み合わせることにより人手評価との間でより高い相関係数が得られることを確認した。

2.3.2 提案手法

2.3.2.1 Earth Mover's Distance（EMD）に基づく自動評価法

単語分散表現モデルを用いた EMD に基づく自動評価法[16]では、EMD を利用するにあたり特徴量には訳文と参照訳の構成単語の単語ベクトル、その単語ベクトルに付与する重みには文レベルの *tf・idf* を用いる。ただし、訳文と参照訳それぞれにおいて *tf・idf* の合計が 1.0 となるように正規化する。また、単語ベクトル間の距離計算にはコサイン類似度を用いる。

ただし、EMD を自動評価法に適用することの問題点は、語順の違いが反映されないため訳文と参照訳間で語順が大きく異なっていても同じ単語で構成されていると高いスコアになってしまうことである。そこで提案手法では語順の違いをスコアに反映させるために、事前に単語ベクトルに基づく単語アライメントを行い、相対的な位置のずれをコサイン類似度に対する負の重みとして使用する。以下の式（1）に提案手法における単語の位置情報を付与した距離計算式を示す。

$$d_{ij} = \begin{cases} 1.0 - \cos\_sim(t_i, r_j) \times e^{-pos\_inf(T_i, R_j)} \\ \qquad \text{if } T_i \text{ corresponds to } R_j \\ 1.0 \qquad \text{if } T_i \text{ does not correspond to } R_j \end{cases} \quad (1)$$

式（1）の $cos_sim(t_i, r_j)$は訳文の単語のベクトル t_iと参照訳の単語のベクトル r_jの間のコサイン類似度を示す。単語アライメントの結果、訳文の単語 T_iと参照訳の単語 R_jが対応関係にあると判断された場合には、T_iと R_j間の位置の相対的なずれである $pos_inf(T_i, R_j)$に基づく重みを $cos_sim(t_i, r_j)$に付与する。また、単語アライメントの結果、対応関係にないと判断された場合には最大値である 1.0 を距離として使用する。式（1）中の T_iと R_j 間の位置の相対的なずれである $pos_inf(T_i, R_j)$は以下の式（2）より計算される。

$$pos\_inf(T_i, R_j) = \left| \frac{pos(T_i)}{m} - \frac{pos(R_j)}{n} \right| \quad (2)$$

式（2）の m は訳文の長さ、n は参照訳の長さを示している。また、$pos(T_i)$は訳文における単語 T_iの出現位置、$pos(R_j)$は参照訳における単語 R_jの出現位置を示している。

2.3.2.2 複数の単語の分散表現モデルの組み合わせ

これまで式（1）の $cos_sim(t_i, r_j)$は 1 つの単語分散表現モデルを用いて求められていた。それに対して本報告では複数の単語分散表現モデルを組み合わせることで $cos_sim(t_i, r_j)$を得る、新たな手法を提案する。具体的には 2 つの単語分散表現モデルそれぞれから得られたコサイン類似度の平均及び積を式（1）の $cos_sim(t_i, r_j)$として用いる。式（3）に 2 つのモデルによるコサイン類似度の平均を用いた場合の計算式を示す。

$$cos\_sim(t_i, r_j) = \frac{\alpha \times cos\_sim(t_i\_A, r_j\_A) + \beta \times cos\_sim(t_i\_B, r_j\_B)}{\alpha + \beta} \quad (3)$$

式（3）の t_i_A は単語分散表現モデル A より得られる訳文中の単語 T_iの単語ベクトル、r_j_A は単語分散表現モデル A より得られる参照訳中の単語 R_jの単語ベクトルを示している。したがって、$cos_sim(t_i_A, r_j_A)$は単語分散表現モデル A より得られるコサイン類似度である。また、t_i_B は単語分散表現モデル B より得られる単語 T_iの単語ベクトル、r_j_B は単語分散表現モデル B より得られる単語 R_jの単語ベクトルを示している。したがって、$cos_sim(t_i_B, r_j_B)$は単語分散表現モデル B より得られるコサイン類似度である。また、α と β はモデル A、B の比重を制御するためのパラメータである。式（4）には 2 つのモデルによるコサイン類似度の積を用いた場合の計算式を示す。

$$cos\_sim(t_i, r_j) = cos\_sim(t_i\_A, r_j\_A)^{\alpha} \times cos\_sim(t_i\_B, r_j\_B)^{\beta} \quad (4)$$

式（4）の α と β はモデル A、B のコサイン類似度に対する重みを制御するためのパラメータである。コサイン類似度は 1.0 以下であるためパラメータの値に 1.0 未満を使用するとそのモデルのコサイン類似度は大きくなり、1.0 以上を用いるとそのモデルのコサイン類似度は小さくなる。

2.3.3 性能評価実験

2.3.3.1 実験手順

単語分散表現モデルを用いた EMD に基づく自動評価法[16]において、複数の単語分散表現モデルの組み合わせが評価精度の向上に有効であるかどうかを確認するために性能評価実験を行った。その際、式（3）と式（4）において $cos_sim(t_i_A, r_j_A)$と $cos_sim(t_i_B, r_j_B)$を得るための単語分散表現モデルには fastText と BERT を用いた。すなわち、式（3）と式（4）におけるモデル A として fastText、モデル B として BERT を用いた。また、実験は WMT18 と WMT19 の Metrics Shared Task データを用い、セグメントレベルの相関係数を求めることにより行った。

2.3.3.2 実験結果

表 1 に WMT18 における“to-English”（多言語から英語への翻訳）の実験結果、表 2 に WMT18 における“out-of-English”（英語から多言語への翻訳）の実験結果、表 3 に WMT19 における“to-English”の実験結果、表 4 に WMT19 における“out-of-English”の実験結果、そして、表 5 に WMT19 における“not involving English”（英語を含まない言語間の翻訳）の実験結果を示す。表 1 と表 2 の WMT18 においては 14 の言語ペアを用いている。その内訳は中国語(zh)、チェコ語(cs)、ドイツ語(de)、エストニア語(et)、フィンランド語(fi)、ロシア語(ru)、トルコ語(tr)と英語(en)間の双方向である。また、表 3 と表 4 の WMT19 においては 15 の言語ペアを使用している。その内訳はチェコ語(cs)、ドイツ語(de)、フィンランド語(fi)、グジャラート語(gu)、カザフ語(kk)、リトアニア語(lt)、ロシア語(ru)、中国語(zh)と英語(en)間の双方向から cs-en を除いたものである。表 5 で使用された言語ペアは英語を含まない言語ペアであり、de-cs、de-fr、そして、fr-de の 3 つである。fr はフランス語を意味する。また、表中の“Avg.”は全言語ペアの相関係数の平均である。

提案手法においては大きく fastText 及び BERT のみを単独で用いた場合と式（3）及び式（4）を用いて 2 つのモデルを組み合わせた場合に分類される。“提案手法(fastText)”は単語の分散表現モデルに fastText のみ、“提案手法(BERT)”は BERT のみを用いたことを示している。“提案手法(平均：α=2、β=1)”はコサイン類似度として式（3）の平均を用いたことを示し、パラメータ α と β の値にはそれぞれ 2 と 1 を用いたことを示している。したがって、fastText から得られたコサイン類似度の重みと BERT から得られたコサイン類似度の重みを 2 対 1 の比で用いたことを示している。また、“提案手法(積：α=0.5、β=2.0)”はコサイン類似度を式（4）の積より求めたことを示している。その際のパラメータ α と β の値にはそれぞれ 0.5 と 2.0 を用いたことを示している。この場合、fastText より得られるコサイン類似度 $cos_sim(t_i_A, r_j_A)$の値はパラメータ α が 0.5 と 1.0 未満であるため本来の値よりも大きくなる。それに対して、パラメータ β は 2.0 であるため、BERT より得られるコサイン類似度 $cos_sim(t_i_B, r_j_B)$の値は本来の値よりも小さくなる。すなわち、“提案手法(積：α=0.5、β=2.0)”においては BERT よりも fastText により得られるコサイン類似度の方を重要視していることを意味する。

表 1　WMT18 における“to-English”の実験結果

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Avg.
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR	
n	5110	77811	56721	15648	10404	8525	33357	
Correlation	τ	τ	τ	τ	τ	τ	τ	
提案手法（fastText）	0.298	0.462	0.329	0.210	0.270	0.230	0.205	0.286
提案手法（BERT）	0.286	0.448	0.323	0.207	0.254	0.213	0.204	0.276
提案手法 (平均： α =1, β =1)	0.295	0.455	0.329	0.211	0.266	0.228	0.211	0.285
提案手法 (平均： α =2, β =1)	0.288	0.453	0.331	0.206	0.262	0.223	0.203	0.281
提案手法 (積: α =0.5, β =2.0)	0.325	0.478	0.347	0.229	0.286	0.223	0.229	0.302
提案手法 (積: α =0.5, β =3.0)	0.321	0.480	0.348	0.230	0.290	0.228	0.232	0.304
提案手法 (積: α =0.5, β =4.0)	0.328	0.481	0.349	0.233	0.294	0.233	0.233	**0.3073**
提案手法 (積: α =0.5, β =5.0)	0.327	0.483	0.349	0.234	0.295	0.231	0.230	0.3070
提案手法 (積: α =1.0, β =1.0)	0.305	0.471	0.343	0.224	0.280	0.228	0.225	0.297
BEER[6]	0.295	0.481	0.341	0.232	0.288	0.229	0.214	0.297
BLEND[7]	0.322	0.492	0.354	0.226	0.290	0.232	0.217	0.305
CHARACTER	0.256	0.450	0.286	0.185	0.244	0.172	0.202	0.256
CHRF[13]	0.288	0.479	0.328	0.229	0.269	0.210	0.208	0.287
CHRF+[13]	0.288	0.479	0.332	0.234	0.279	0.218	0.207	0.291
ITER	0.198	0.396	0.235	0.128	0.139	-0.029	0.144	0.173
METEOR++	0.270	0.457	0.329	0.207	0.253	0.204	0.179	0.271
RUSE[8]	0.347	0.498	0.368	0.273	0.311	0.259	0.218	<u>0.325</u>
SENTBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178	0.234
UHH_TSKM	0.274	0.436	0.300	0.168	0.235	0.154	0.151	0.245
YiSi-0[11]	0.301	0.474	0.330	0.225	0.294	0.215	0.205	0.292
YiSi-1[11]	0.319	0.488	0.351	0.231	0.300	0.234	0.211	0.305
YiSi-1_SRL[11]	0.317	0.483	0.345	0.237	0.306	0.233	0.209	0.304
newstest2018								

表 2　WMT18 における“out-of-English”の実験結果

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh	Avg.
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR	
n	5413	19711	32202	9809	22181	1358	28602	
Correlation	τ	τ	τ	τ	τ	τ	τ	
提案手法（fastText）	0.472	0.644	0.531	0.488	0.382	0.387	0.331	0.462
提案手法（BERT）	0.468	0.633	0.471	0.452	0.349	0.330	0.347	0.436
提案手法 (平均： α =1, β =1)	0.458	0.636	0.508	0.488	0.368	0.351	0.347	0.451
提案手法 (平均： α =2, β =1)	0.458	0.641	0.517	0.491	0.370	0.368	0.344	0.456
提案手法 (積: α =0.5, β =2.0)	0.502	0.673	0.534	0.520	0.391	0.371	0.353	**0.478**
提案手法 (積: α =0.5, β =3.0)	0.496	0.675	0.535	0.524	0.396	0.368	0.350	**0.478**
提案手法 (積: α =0.5, β =4.0)	0.492	0.674	0.535	0.517	0.396	0.364	0.346	0.475
提案手法 (積: α =0.5, β =5.0)	0.488	0.671	0.531	0.512	0.392	0.370	0.347	0.473
提案手法 (積: α =1.0, β =1.0)	0.488	0.662	0.539	0.515	0.389	0.380	0.351	0.475
BEER	0.518	0.686	0.558	0.511	0.403	0.374	0.302	0.479
BLEND	-	-	-	-	0.394	-	-	-
CHARACTER	0.414	0.604	0.464	0.403	0.352	0.404	0.313	0.422
CHRF	0.516	0.677	0.572	0.520	0.383	0.409	0.328	0.486
CHRF+	0.513	0.680	0.573	0.525	0.392	0.405	0.328	<u>0.488</u>
ITER	0.333	0.610	0.392	0.311	0.291	0.236	-	-
SENTBLEU	0.389	0.620	0.414	0.355	0.330	0.261	0.311	0.383
YiSi-0	0.471	0.661	0.531	0.464	0.394	0.376	0.318	0.459
YiSi-1	0.496	0.691	0.546	0.504	0.407	0.418	0.323	0.484
YiSi-1_SRL	-	0.696	-	-	-	-	0.310	-
newstest2018								

表 3　WMT19 における“to-English”の実験結果

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	Avg.
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR	
n	85365	38307	31139	27094	21862	46172	31070	
提案手法（fastText）	0.158	0.284	0.251	0.416	0.297	0.168	0.354	0.275
提案手法（BERT）	0.154	0.275	0.248	0.401	0.310	0.152	0.357	0.271
提案手法 (積: α =0.5, β =2.0)	0.172	0.298	0.275	0.417	0.324	0.170	0.372	0.290
提案手法 (積: α =0.5, β =3.0)	0.173	0.298	0.277	0.420	0.328	0.170	0.375	0.292
提案手法 (積: α =0.5, β =4.0)	0.174	0.299	0.278	0.423	0.333	0.173	0.376	**0.294**
提案手法 (積: α =0.5, β =5.0)	0.173	0.299	0.276	0.421	0.335	0.173	0.375	0.293
提案手法 (積: α =1.0, β =1.0)	0.165	0.295	0.268	0.416	0.315	0.167	0.366	0.285
BEER	0.128	0.283	0.260	0.421	0.315	0.189	0.371	0.281
BERTR[10]	0.142	0.331	0.291	0.421	0.353	0.195	0.399	0.305
CHARACTER	0.101	0.253	0.190	0.340	0.254	0.155	0.337	0.233
CHRF	0.122	0.286	0.256	0.389	0.301	0.180	0.371	0.272
CHRF+	0.125	0.289	0.257	0.394	0.303	0.182	0.374	0.275
EED[14]	0.120	0.281	0.264	0.392	0.298	0.176	0.376	0.272
ESIM[10]	0.167	0.337	0.303	0.435	0.359	0.201	0.396	0.314
HLEPORA _BASELINE	-	-	-	0.372	-	-	0.339	-
METEOR++_2.0 (SYNTAX)	0.084	0.274	0.237	0.395	0.291	0.156	0.370	0.258
METEOR++_2.0 (SYNTAX+COPY)	0.094	0.273	0.244	0.402	0.287	0.163	0.367	0.261
PREP	0.030	0.197	0.192	0.386	0.193	0.124	0.267	0.198
SENTBLEU	0.056	0.233	0.188	0.377	0.262	0.125	0.323	0.223
WMDO	0.096	0.281	0.260	0.420	0.300	0.162	0.362	0.269
YiSi-0	0.117	0.271	0.263	0.402	0.289	0.178	0.335	0.265
YiSi-1	0.164	0.347	0.312	0.440	0.376	0.217	0.426	0.326
YiSi-1_SRL	0.199	0.346	0.306	0.442	0.380	0.222	0.431	<u>0.332</u>
newstest2019								

表 4　WMT19 における“out-of-English”の実験結果

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	Avg.
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR	DARR	
n	27178	99840	31820	11355	18172	17401	24334	18658	
提案手法 (fastText)	0.437	0.319	0.465	0.469	0.496	0.403	0.521	0.291	0.425
提案手法(BERT)	0.421	0.301	0.413	0.446	0.469	0.353	0.470	0.299	0.397
提案手法 (積: α =0.5, β =2.0)	0.465	0.326	0.469	0.482	0.504	0.412	0.529	0.324	0.4389
提案手法 (積: α =0.5, β =3.0)	0.469	0.328	0.474	0.482	0.503	0.412	0.531	0.325	**0.441**
提案手法 (積: α =0.5, β =4.0)	0.467	0.328	0.472	0.477	0.501	0.408	0.531	0.328	0.4390
提案手法 (積: α =0.5, β =5.0)	0.467	0.328	0.470	0.477	0.500	0.407	0.531	0.328	0.4385
提案手法 (積: α =1.0, β =1.0)	0.464	0.322	0.473	0.476	0.498	0.418	0.519	0.312	0.435
BEER	0.443	0.316	0.514	0.537	0.516	0.441	0.542	0.232	0.443
CHARACTER	0.349	0.264	0.404	0.500	0.351	0.311	0.432	0.094	0.338
CHRF	0.455	0.326	0.514	0.534	0.479	0.446	0.539	0.301	0.449
CHRF+	0.458	0.327	0.514	0.538	0.491	0.448	0.543	0.296	0.452
EED	0.431	0.315	0.508	0.568	0.518	0.425	0.546	0.257	0.446
ESIM	-	0.329	0.511	-	0.510	0.428	0.572	0.339	-
HLEPORA _BASELINE	-	-	-	0.463	0.390	-	-	-	-
SENTBLEU	0.367	0.248	0.396	0.465	0.392	0.334	0.469	0.270	0.368
YiSi-0	0.406	0.304	0.483	0.539	0.494	0.402	0.535	0.266	0.429
YiSi-1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355	<u>0.484</u>
YiSi-1_SRL	-	0.368	-	-	-	-	-	0.361	-
newstest2019									

表 5　WMT19 における“not involving English”の実験結果

	de-cs	de-fr	fr-de	Avg.
Human Evaluation	DARR	DARR	DARR	
n	35793	4862	1369	
提案手法（fastText）	0.324	0.288	0.271	**0.294**
提案手法（BERT）	0.264	0.234	0.221	0.240
提案手法(積: α =0.5, β =2.0)	0.314	0.268	0.264	0.282
提案手法(積: α =0.5, β =3.0)	0.311	0.270	0.265	0.282
提案手法(積: α =0.5, β =4.0)	0.308	0.271	0.274	0.284
提案手法(積: α =0.5, β =5.0)	0.306	0.267	0.277	0.283
提案手法(積: α =1.0, β =1.0)	0.312	0.260	0.265	0.279
BEER	0.337	0.293	0.265	0.298
CHARACTER	0.232	0.251	0.224	0.236
CHRF	0.326	0.284	0.275	0.295
CHRF+	0.326	0.284	0.278	0.296
EED	0.345	0.301	0.267	0.304
ESIM	0.331	0.290	0.289	0.303
HLEPORA_BASELINE	0.207	0.239	-	-
SENTBLEU	0.203	0.235	0.179	0.206
YiSi-0	0.331	0.296	0.277	0.301
YiSi-1	0.376	0.349	0.310	<u>0.345</u>
YiSi-1_SRL	-	-	0.229	-
newstest2019				

2.3.3.3 考察

表 1 から表 5 の実験結果に基づいて複数の単語分散表現モデルを用いた提案手法の有効性について述べる。表中の“Avg.”に着目した場合、fastText と BERT の比較においては全ての実験結果で“提案手法（BERT)”よりも“提案手法（fastText)”の方が高い相関係数を示した。また、fastText と BERT のいずれかを単独で用いた場合と 2 つのモデルを組み合わせた場合の比較においては、表 5 を除き、組み合わせた場合の方が高い相関係数を示した。したがって、複数の単語分散表現モデルを利用することの有効性を確認した。2 つのモデルの組み合わせにおいても式（3）による平均と式（4）による積との比較を行った結果、式（4）による積を用いた方が高い相関係数が得られた。そのため表 3 から表 5 までの WMT19 の実験結果においては式（3）の平均による提案手法は求めていない。また、上述した通り、単独の利用においては BERT よりも fastText を用いた方が高い評価精度を示したため、組み合わせを用いる場合には fastText より得られるコサイン類似度を重視した。したがって、式（4）の積においては α の値には 1.0 以下、β の値に

は 1.0 以上を用いた。

式（4）の積を用いた場合におけるパラメータ α と β の最適な値の組み合わせに関しては α に 0.5、β に 3.0 を用いた際の組み合わせと α に 0.5、β に 4.0 を用いた際の組み合わせが高い相関係数を示した。表 1 においては“提案手法（積：α=0.5、β=4.0)”、表 2 においては“提案手法（積：α=0.5、β=2.0)”と“提案手法（積：α=0.5、β=3.0)”、表 3 においては“提案手法（積：α=0.5、β=4.0)”、表 4 においては“提案手法（積：α=0.5、β=3.0)”、そして、表 5 においては“提案手法（積：α=0.5、β=4.0)”が“Avg.”において最も高い相関係数を示した。したがって、全体として“α=0.5、β=3.0”または“α=0.5、β=4.0”が最適なパラメータの組み合わせであるといえる。

さらにメタ評価の観点より提案手法と他の自動評価法との“Avg.”の比較について述べる。その結果、表 1 では“提案手法（積：α=0.5、β=4.0)”の相関係数は他手法の中で最も相関係数が高かった RUSE より 0.018 低く、表 2 では“提案手法（積：α=0.5、β=2.0)”と“提案手法（積：α=0.5、β=3.0)”の相関係数は他手法で最も相関係数が高かった CHRF++より 0.01 低くい値を示した。したがって、WMT18 において提案手法は他の自動評価法に比べ大きく下回ることはなく同等のレベルであるといえる。一方、WMT19 においては他の自動評価法に対して不十分であった。表 3 においては“提案手法（積：α=0.5、β=4.0)”は全体的には“Avg.”は低くないが、最も高い相関係数を示した YiSi-1_SRL に対しては 0.038 低かった。表 4 と表 5 においては提案手法の“Avg.”は他の自動評価法と比較して低い値であった。言語ペアに着目した場合には、提案手法は zh-en と en-zh において他の自動評価法に対して比較的高い相関係数を示した。各言語ペアと相関係数の関係性については今後精査する必要がある。

2.3.4 まとめ

本報告では EMD に基づく自動評価法において複数の単語分散表現モデルを用いることの有効性について述べた。性能評価実験の結果、英語を含む言語ペアについては単語分散表現モデルとして fastText と BERT の両方を組み合わせたコサイン類似度を用いた方がそれぞれを単独で用いた場合よりも高い相関係数が得られた。しかし、メタ評価においては他の自動評価法と比較して WMT18 では同等レベルであったが、WMT19 では下回る評価精度となり十分とはいえない。

今後は他の自動評価法との比較においても有効性が得られるように提案手法の改良を行う予定である。

参考文献

[1] 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 (ACL02), pp. 311-318.

[2] George Doddington (2002) "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, " Proceedings of the Second International Conference on Human Language Technology Research (HLT02), pp.138-145.

[3] 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 Machine Translation and/or Summarization, pp.65-72.
[4] Abhaya Agarwal and Alon Lavie (2008) "Meteor, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output," Proceedings of the Third Workshop on Statistical Machine Translation (WMT08), pp. 115–118.
[5] Gregor Leusch, Nicola Ueffing and Hermann Ney (2006) "CDER: Efficient MT Evaluation Using Block Movements," Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL06), pp.241-248.
[6] Miloš Stanojević and Khalil Sima'an (2014) "BEER: BEtter Evaluation as Ranking," Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT14), pages 414–419.
[7] Qingsong Ma, Yvette Graham, Shugen Wang and Qun Liu (2017) "Blend: a Novel Combined MT Metric Based on Direct Assessment — CASICT-DCU submission to WMT17 Metrics Task," Proceedings of the Conference on Machine Translation (WMT17), pages 598–603.
[8] Hiroki Shimanaka, Tomoyuki Kajiwara and Mamoru Komachi (2018) "RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation," Proceedings of the Third Conference on Machine Translation (WMT18), pp.751-758.
[9] Thibault Sellam, Dipanjan Das and Ankur P. Parikh (2020) "BLEURT: Learning Robust Metrics for Text Generation," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL20), pages 7881–7892.
[10] Nitika Mathur, Timothy Baldwin and Trevor Cohn (2019) "Putting Evaluation in Context: Contextual Embeddings improve Machine Translation Evaluation," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL19), pp. 2799–2808.
[11] Chi-kiu Lo (2019) "YiSi - A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources," Proceedings of the Fourth Conference on Machine Translation (WMT19), pp. 507–513.
[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger and Yoav Artzi (2020) "BERTSCORE: Evaluating Text Generation with BERT," https://arxiv.org/pdf/1904.09675.pdf
[13] Maja Popović (2015) "CHRF: character n-gram F-score for automatic MT evaluation," Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT15), pp.392-395.
[14] Peter Stanchev, Weiyue Wang and Hermann Ney (2019) "EED: Extended Edit Distance Measure for Machine Translation," Proceedings of the Fourth Conference on Machine Translation (WMT19), pp.514-520.

[15] 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.
[16] Hiroshi Echizen'ya, Kenji Araki and 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-HLT19), pp.1874-1883.
[17] Tomas 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.
[18] "Word vectors for 157 languages," https://fasttext.cc/docs/en/crawl-vectors.html
[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," https://arxiv.org/abs/1810.04805
[20] "BERT-Base, Multilingual Cased," https://github.com/google-research/bert
[21] Qingsong Ma, Ondřej Bojar and Yvette Graham (2018) "Results of the WMT18 Metrics Shared Task," Proceedings of the Third Conference on Machine Translation (WMT18), pp.671-688.
[22] Qingsong Ma, Johnny Tian-Zheng Wei, Ondřej Bojar and Yvette Graham (2019) "Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges," Proceedings of the Fourth Conference on Machine Translation (WMT19), pp.62-90.

2.4 特許文献の語彙翻訳評価のための
中日テストセット試験文の収集と分析

（株）富士通研究所　長瀬 友樹
元・山梨英和大学　江原 暉将

2.4.1　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 。

・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

昨年度までの取組みの中で、下記のようなテストセットの課題が明らかになっている。

・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
・訳文の表記ゆれへの対応

今年度は、これらの課題の中で、「技術用語の翻訳」に着目したテストセットの拡充を行い、あわせていくつかの中日機械翻訳システムを評価した。

2.4.2　試験文の採集方法

昨年度から使用している 100 万文対の中日特許文コーパスから試験文を採集した。今年度のポイントは訳語の正しさを計測する試験文を採集する点にある。NMT においては、特に低頻度語で訳語を誤ることがある。SMT においては訓練データに含まれる原語と訳語のペアのみが翻訳に用いられるが、NMT ではその保証がない。そこで訳語の正しさを計測することは、構文の正確さを計測することと並んで重要な評価項目である。

100 万文対のデータの中国語部分から名詞(複合名詞を含む)を抽出した。特許文では技術用語の誤訳が特に問題となり、その多くは名詞である。その中から低頻度語として頻度 100 未満 5 以上の名詞を対象にした。一方、100 万文対の中日データから GIZA++を用いて中国語の名詞に対応する日本語訳語を求めた。同一の中国語名詞に複数の日本語訳語が対応する場合が特に問題である。例えば、「上梁」に対応する日本語として「アッパーメンバー」が 7 回「上レール」が 7 回出現しており、曖昧でありかつ頻度も拮抗している。このような名詞の翻訳を周辺の文脈を用いて正しく選択する必要がある。

上記のような戦略を中心として今年度は 130 の試験文を選択し、日本語の訳語パターンを設定した。130 の試験文には、同一の中国語名詞に複数の日本語訳語が対応する場合に訳語を正しく選択できるかどうかを評価するための文が含まれている。例えば、「上梁」に対応する日本語訳の候補が「アッパーメンバー」、「上レール」の 2 通りの場合、訳語の選択が正しくできるかどうかを評価するために、それぞれの訳語を含む試験文がテストセットに含まれている。同様に、日本語訳の候補が 3 通り、4 通りある場合は、それぞれ 3 文、4 文の試験文がテストセットに存在する。そのため、今回設定した試験文の数は 130 であるが、評価対象としている技術用語の種類は 81 語となっている。内訳を表 1 に示す。

表 1　語彙評価のために追加した試験文の内訳

中国語単語に対応する日本語訳の種類	試験文の数	評価する技術用語の語数
単一の日本語訳のみが対応	46 文	46 語
２つの日本語訳が対応	46 文	23 語
３つの日本語訳が対応	30 文	10 語
４つの日本語訳が対応	8 文	2 語
計	130 文	81 語

2.4.3　中日語彙翻訳評価テストセットを用いた機械翻訳の評価

今年度整備した語彙翻訳評価のためのテストセットを用いて、実際に機械翻訳の評価を行い、翻訳エンジンの能力に照らして評価対象にしている技術用語の翻訳難易度は妥当かどうか、翻訳エンジンの違いによってどのくらい評価の幅が出てくるかという観点で確認を行った。評価に用いた機械翻訳エンジンは、インターネット上で無料サービスとして公開されている代表的な 3 つの翻訳エンジンである。これらの 3 つの翻訳エンジンを、以下では便宜的に「翻訳エンジンＡ」、「翻訳エンジンＢ」、「翻訳エンジンＣ」と呼ぶことにする。

語彙評価テストセットを用いて３つの翻訳エンジンを評価した結果を表 2 に示す。

表２　語彙評価テストセットを用いた 3 種の翻訳エンジンの評価結果

中国語単語に対応する日本語訳の種類	試験文の数	正解数：翻訳ｴﾝｼﾞﾝA	正解数：翻訳ｴﾝｼﾞﾝB	正解数：翻訳ｴﾝｼﾞﾝC
単一の日本語訳のみが対応	46 文	35	33	35
２つの日本語訳が対応	46 文	28	26	24
３つの日本語訳が対応	30 文	9	9	12
４つの日本語訳が対応	8 文	5	6	8
計	130 文	77（59.2%）	74（56.9%）	79（60.8%）

正解率が最も高かったのは翻訳エンジンＣの 60.8%で、最も低かった翻訳エンジンＢとの差は 3.9 ポイントであった。内訳を見ると、翻訳エンジンＣは、他のエンジンと比較して 3 つ以上の日本語訳が対応している単語の選択がうまくいっているという傾向が見て取れる。

次に、各試験文に対して 3 つの翻訳エンジンのうち正解を訳出できたエンジンの数で試験文を分類した結果を表３に示す。

表３　正解した翻訳エンジンの数別の試験文の文数

	試験文の数
３つのエンジンで不正解	26 文（20%）
１つのエンジンで正解	28 文（22%）
２つのエンジンで正解	28 文（22%）
３つのエンジンで正解	48 文（37%）
計	130 文

3 つのエンジンがすべて同じ評価となる試験文は 74 文で、全体に占める割合 57%であった。1 つのエンジンのみで正解の文と２つのエンジンで正解の文はどちらも 28 文（22%）であった。

語彙のみに着目した評価と文全体の評価との関係を見るために、３つの翻訳エンジンの BLEU スコアを求めた（表４）。

試験４　各エンジンの BLEU スコア

	翻訳エンジンＡ	翻訳エンジンＢ	翻訳エンジンＣ
BLEU スコア	0.400	0.413	0.295

BLEU スコアは、翻訳エンジンＢのスコアが最も高くて 0.413、最もスコアが低い翻訳エンジン

Cは0.295であり、差が大きいことを確認した。BLEUスコアは高い方から、B　＞　A　＞　Cの順序だが、語彙評価テストセットの結果は、まったく逆の　C　＞　A　＞　Bの順であった。

2.4.4　中日語彙評価テストセットの課題

ここでは、今年度整備した語彙評価のための試験文を個々に見たときの気づきや課題について論じる。

(1)用語末尾の揺れ

テストセット評価では、正規表現で記載されたチェック項目と訳文とのマッチングによって正誤を判断する方式である。この方式では末尾の長音記号などの有無にかかわらず正解とみなしてしまうという問題がある。

［例］　条件: 熱可塑性エラストマ
　　訳文：熱可塑性エラストマー　（評価：○）

(2)記号で接続された連語

特許文では連語を構成する単語の間にハイフンやコロンを入れることがあるが、区切り記号の有無、区切り記号の種類の違いが条件判定で考慮されない。(テストセットの検討以前に、区切り記号のみが異なる場合に正解とみなすべきかどうかの方針を明確にする必要がある。)

［例］　条件：リチウムニッケルコバルトマンガン
　　訳文：リチウム-ニッケル-コバルト-マンガン　（評価：×）

(3)一般語（非専門用語）の同義語

特許文献の語彙評価テストセットでは、評価対象を専門用語に絞るべきという考え方があるが、機械的に専門用語と一般語を区別することは難しい。テストセットに一般語（非専門用語）の翻訳チェックを含む場合、正解の条件に訳語として許容される同義語が列挙されている必要があるが、現在は特定の単語のみが条件となっているため、本来正解と評価されるべきところが不正解にされる例が見られる。

［例］　条件：メディア　　（記録媒体の意味で使われる）
　　訳文：媒体　（評価：×）

2.4.5　まとめ

昨年までに整備してきたテストセットは、主に特許文献に特有の表現・言い回しを評価するためのものだったのに対し、今年度は特許文献中の技術用語の正しさを計測する試験文の採集を行った。また、採取できた試験文を用いて3つの翻訳エンジンの評価を実施した。

語彙翻訳評価テストセットは、技術用語の訳語の正しさ、文脈による訳語の選択の正しさを個

別に評価するツールとして利用できることを確認できた。BLUE スコアのような総合的な評価指標と語彙評価テストセットを組み合わせて用いることで、翻訳エンジンの多角的な評価を提供できる可能性があることも確認できた。

語彙翻訳評価の観点として、あらゆる入力文に対して安定的に正しい翻訳ができるかどうかという点も重要である。この観点の評価を行うためには、同一の訳語が正解となる複数の試験文をテストセットに含めて評価を行う必要がある。語彙翻訳評価テストセットの大規模化については今後の課題である。

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. サーベイ論文

3.1 訳抜け・過剰訳対策

日本放送協会　後藤　功雄
愛媛大学　二宮　崇

訳抜け・過剰訳に関連する対策としてこれまでにいくつかのアプローチが提案されている。それらを分類すると以下のように分けられる。

(1)　NMT 内のネットワーク構造でのモデル化
(2)　過不足に関するスコアとの組み合わせ
(3)　低頻度語・新語対策
(4)　出力長制御
(5)　訓練データのノイズ対策
(6)　NMT の動作分析

(1) は、NMT のネットワーク構造を SMT のカバレッジモデルを参考にして一部変更することで訳抜け・過剰訳を減らす試みである。(2) は、アテンションの確率や逆翻訳スコアで出力のスコアを補正するものである。(3) は、訳抜けしやすいのは低頻度語・新語であるため、サブワードに分割して語彙の頻度を向上させたり、変数に置き換えて翻訳し、出力中の変数を辞書などから取得した訳に置換したり、低頻度語と類似する高頻度語を低頻度語に置き換えてデータ増強し学習データでの低頻度語の頻度を増やすなどがある。(4) は、出力長を制御することで訳抜け・過剰訳の抑制を目指すものである。(5) は、訓練データでの訳抜けを学習してしまう問題への対策である。(6) は、NMT の動作を分析する手法で、原因の調査に役に立つと考えられる。

3.2 機械翻訳における文章中での訳語の統一

静岡大学　綱川　隆司
元・山梨英和大学　江原　暉将
東京大学　中澤　敏明

自然言語では、同一の物事を複数の異なる表現で表すことができる。一方で、論文などの説明的文章では、同一の意味を表すのに統一した表現を用いることで無用な誤解を防ぐことが一般的である。また、ある概念に対して決まった専門用語を用いるような場合には、用語集を用いて出力文中の語彙の統制が必要となる。さらに、テキスト検索時にも異なる表現が混在すると、それらの表記揺れに対応しない限り網羅的な検索が困難になってしまうなど、テキストの機械的処理に支障をきたすことがある。

基本的な機械翻訳モデルは入力テキストを一文ごとに訳出しており、入力テキストにおいて同一の意味に対して統一した表現が用いられていたとしても同じ訳出がされる保証はない。例えば、英日翻訳において“apple”は“りんご”、“リンゴ”、あるいは“林檎”のいずれにも訳される可能性があり、特別な意図なく同じ文書で混在させるのは望ましくないが、文単位で訳出する機械翻訳システムは両方の訳を混在させる可能性がある。

機械翻訳において訳語を統一するための方法としては、用語集・対訳辞書を用いる方法と、翻訳履歴や文脈を考慮する方法の二つが挙げられる。以下それぞれについて概要を述べる。

訳語を統一する必要のある語彙に対して既定の訳語を定めた用語集あるいは対訳辞書を予め用意する方法では、訳出された文を後編集したり機械翻訳の過程で対訳辞書を適用したりすることにより訳語を統一する。後編集は用いる機械翻訳システムに依らず適用ができるため一般的な戦略であるが、語のもつ曖昧性により誤った置換が生じるおそれや、そもそも訳出すべき語が抜けたり誤って訳された場合に対応できないことなどの課題に対処する必要が生じる。一方で、並列コーパスにより翻訳器の学習を行うデータ駆動型の機械翻訳モデルでは対訳辞書を明示的に用いないため、ニューラル機械翻訳においては対訳辞書を考慮した翻訳モデルや、訳文生成時に語彙制約を課した探索を行う方法、翻訳器の学習に用いる並列コーパスを対訳辞書により事前に編集する方法が提案されている。

また、入力テキストの各文を独立に翻訳するのではなく、ある文の翻訳時に前後の文あるいは入力テキスト全体の文脈を考慮に入れることで、入力テキスト全体を考慮して統一した訳語が出力されることが期待できる。統計的機械翻訳では訳語の統一により精度向上の可能性が示唆されており、訳語を統一する手法も提案された。ニューラル機械翻訳においても、翻訳履歴を保持することで以前訳した語の情報を考慮に入れるモデルや、文脈情報を考慮するモデルが提案されており、近年活発な研究が行われている。

3.3 長文対策

愛媛大学 二宮 崇
静岡大学 綱川 隆司

Koehn ら（Koehn & Knowles, 2017）によって現在のニューラル機械翻訳における 6 つの課題が指摘されているが、その一つに長文問題がある。機械翻訳における長文問題とは、1 文あたりの語数が大きい文(長文)を翻訳する際に生じる問題のことであり、(1) 文長に対し計算量が線形ではない計算量となってしまう問題、および (2) 語数の大きい文に対し翻訳精度が著しく劣化する問題に分けられる。特許翻訳は特に長文からなる文章の翻訳を特徴の一つとしており、実用的な特許翻訳の実現のために長文の効率化および高精度化が必要とされている。

長文対策のための様々な手法が提案されており、大きく分類すると、まず、上記の問題に対応して、機械翻訳の効率化による長文対策と、長文に対する翻訳精度の劣化を防ぐ長文対策に分けられる。それぞれをさらに分類すると、下記のように分けられる。

(A) 機械翻訳の効率化

(A-1) 文長に対する線形計算量を目的とするニューラル機械翻訳

(A-2) 非自己回帰モデルによるデコーダーの並列化

(B) 翻訳精度の改善

(B-1) 文分割

(B-2) ニューラル機械翻訳モデルの構造的特徴(依存構造や位置情報、自己アテンションなど)を利用したニューラル機械翻訳の改良

(A-1)は、文長に対する計算量を線形にすることを目的とする研究であり、基本的には、文長nに対し自己アテンションの計算量を$O(n^2)$から$O(n \log n)$や$O(n)$にすることで計算量の削減を目指している。Reformer や Linformer、Big Bird などが有名である。(A-2)は非自己回帰ニューラル機械翻訳(non-autoregressive NMT)と呼ばれる研究群である。従来の自己回帰的なモデルでは目的言語文の各単語を並列に生成することができないが、文長の予測を行い、単語間の関係をモデルに導入することで、翻訳生成の並列化を実現している。並列に翻訳を生成することでデコーダーの効率化を実現している。

(B-1)は、文構造等の特徴を用いて文全体を分割し、分割された個々の部分文を翻訳することで長文の精度劣化の問題を解決する手法である。(B-2)は、ニューラル機械翻訳のモデルを改良することで、長文に対応する手法である。アテンションそのものが長文に対応する手法となっているが、さらに自己アテンションに依存構造を導入する手法や、位置情報を改良する手法、自己アテンションに被覆率を導入する手法などが提案されており、長文における性能劣化を抑えている。

参考文献

Philipp Koehn, Rebecca Knowles (2017) Six Challenges for Neural Machine Translation, Proceedings of the First Workshop on Neural Machine Translation (WNMT 2017), pp. 28-39.

3.4 低リソース言語対対策

元・山梨英和大学　江原　暉将
情報通信研究機構　今村　賢治

3.4　低リソース言語対対策

低リソース言語対では NMT の翻訳精度が著しく低下することが指摘されている。そのための対策として、1) 間に第 3 の言語を挟んで高リソースとする、2) 単言語コーパスから疑似的な 2 言語コーパスを作成する、3) 高リソース言語対で事前訓練を行い低リソース言語対で追加訓練を行う、4) 多対多の翻訳システムの一部として低リソース言語対を含める、などが提案されている。

1) 言語 X から言語 Y への翻訳システム(本節ではこのような翻訳システムを X→Y と略記する)を構築したい場合、X と Y が低リソース言語対である場合でも他の言語 Z を間にはさむことで X と Z および Z と Y は高リソース言語対とできる場合がある。例えば、英語や中国語を言語 Z とすることが考えられる。このような方式をピボット方式と呼ぶ。ピボット方式の最も単純な方法として X→Z の後に Z→Y を直列につなぐカスケード方式がある。カスケード方式では X→Z と Z→Y は独立に訓練されるが、交差項を用いて非独立に X→Y を構築する方式が提案されている。

2) X または Y の単言語コーパスとして大規模なものが得られる場合がある。特に Y の単言語コーパスを何らかの Y→X を用いて翻訳し疑似的な二言語コーパスを得て訓練データに加える方式は逆翻訳方式と呼ばれ有効性が確認されている。逆翻訳方式の場合 Y→X として何を用いるかが問題となる。Y と X が言語的に近い場合は Y の語を単語辞書を用いて X に置き換える逐語訳が可能である。ピボット方式と組み合わせると次が可能となる。Z と Y は高リソースだが、X と Z は低リソースである場合でも、X と Z が言語的に近いと、Z の単言語コーパスから逐語訳による Z→X によって逆翻訳し、X と Z も疑似的に高リソースにでき、結果、ピボット方式が可能となる。

3) X と Y が低リソースでも Z と Y が高リソースの場合、まず Z→Y を訓練し、次に X→Y を転移学習(追加訓練)する方式が提案されている。本方式でも X と Z が言語的に近い方が有利である。また X と Z を表記する文字が異なる場合はラテン文字転写をして文字を揃えることが考えられる。転移学習方式には、原言語 X が共通の場合もある。つまり X と Y が低リソースでも X と Z が高リソースの場合、まず X→Z を訓練し、次に X→Y を転移学習する。

4) 転移学習方式を発展させて多対多の NMT とする方式が提案されている。多対多の中の一つとして低リソース言語対 X と Y を含ませる。多対多となるために原言語の文の先頭に目標言語名を示す特殊な記号を追加する。多くの言語対を一つのシステムでモデル化することでお互いに影響を受け、低リソース状態で X→Y を単独で訓練したベースラインを上回る性能が示されている。

低リソース言語対としては TED の多言語コーパスや聖書のコーパスから採られる研究が多く、特許文書を対象としたものは知られていない。特許分野では日本語と対になる低リソース言語として東南アジアの言語などがある。WAT2020 においてヒンディ/タイ/マレー/インドネシアの各言語と英語との翻訳について研究されている。今後、日本語を含むアジア諸言語についての低リソース言語対対策の研究が進むことが望まれる。

3.5 機械翻訳技術に基づく自動評価法の変遷

北海学園大学　越前谷　博
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

3.5.1　統計的機械翻訳とモデルフリーの自動評価法

自動評価法は機械翻訳技術と共に発展してきた。1990 年前後に提案された統計的機械翻訳はオープンソースの機械翻訳システムであり、多くの研究者にとって容易に入手可能であったことから盛んに研究が行われた。それに伴い開発サイクルを円滑に進めるために自動的に訳文を評価する技術が求められるようになる。そうした中で自動評価法 BLEU[1]が提案された。BLEU は訳文に対して人手による参照訳と比較することにより評価スコアを算出する。そして、この BLEU 以降に自動評価法の研究は大きな注目を集め、様々な自動評価法が提案されるようになる。

BLEU を始め単語及び文字の表層情報に基づく自動評価法はモデルフリーの自動評価法と呼ばれる。モデルフリーの自動評価法は高速に訳文を評価できることが最大のメリットである。その反面、表層レベルの処理に留まっていることから単語や文の意味といった深いレベルの処理に基づく評価が困難であることが問題となる。なお、自動評価法における評価は人手評価との相関係数を求めることで定量化するのが一般的である。

3.5.2　ニューラル機械翻訳とモデルベースの自動評価法

2010 年代に入るとニューラル機械翻訳が翻訳技術の主流となる。そこでは単語や文をベクトルに変換することで表層処理から意味処理へとより深いレベルでの言語処理が可能となった。このような機械翻訳技術の進展に伴い自動評価法においても単語の分散表現及び文ベクトルに基づく手法が数多く研究されるようになった。これらの自動評価法は上記のモデルフリーの自動評価法に対して、学習したモデルの使用を前提とするためモデルベースの自動評価法と呼ばれる。

また、モデルベースの自動評価法は事前学習済みモデルを使用する手法と自動評価用に新規に構築したモデルを用いる手法に大別される。前者のモデルベースの自動評価法は多言語対応や分野適応のための学習を必要としないため利便性が高い。しかし、評価対象によっては固定されたモデルを使用することでバイアスが取り込まれている危険性があることが問題となる。後者のモデルベースの自動評価法では学習のための多大なコストと言語資源を必要とするが、言語と分野に柔軟に適応した評価が可能であり、より高い評価精度を得られることが知られている。このように自動評価法は機械翻訳技術の変遷と共に進展しており、今後もさらなる発展が期待される。

参考文献

[1] 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 (ACL02), pp. 311-318.

3.6 デプロイ対策

国立研究開発法人　情報通信研究機構　今村　賢治
東芝デジタルソリューションズ株式会社　　園尾　聡

3.6.1 概要

ニューラル機械翻訳（NMT）を特許翻訳で実際に使用するときの問題点について、現状を調査している。特許翻訳は、対象文書が膨大であるため、翻訳速度が問題となる。また、複数のコンピューターで並列処理しようとすると、GPU が搭載されていないような小型コンピューターでも処理を行いたい。これらの問題に対して、われわれはシステムの翻訳速度、および使用メモリー量の削減に焦点を当てて調査を行った。

3.6.2　実施内容

まず、Workshop on Generation and Machine Translation[1, 2, 3]で実施された Efficient MT 共有タスクの調査を行った。本タスクは、参加者のシステムを同じ仮想環境で動作させ、翻訳品質のほかに、翻訳速度と使用メモリー量を比較するというものである。この共有タスクに参加したシステムで使われている、高速化や省メモリーのための技術の調査を行った。

- モデル自体の高速化：モデル縮小、サブワード、自己注視機構の置き換え、非自己回帰型デコーダー
- 訓練・テストで高速化：知識蒸留、ビーム幅、動的語彙選択、ミニバッチ構成法
- 実装による高速化：半精度浮動小数点、8 ビット整数量子化、注視機構のキャッシュ、C++実装

また、知識蒸留を行ったときのモデル縮小方法（レイヤー削減、次元数削減、語彙削減）の効果について、実験的に調査した。

参考文献

[1] A. Birch, A. Finch, M. Luong, G. Neubig, and Y. Oda. 2018. Findings of the second workshop on neural machine translation and generation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 1-10, Melbourne, Australia, July.
[2] H. Hayashi, Y. Oda, A. Birch, I. Konstas, A. Finch, M. Luong, G. Neubig, and K. Sudoh. 2019. Findings of the third workshop on neural generation and translation. In Proceedings of the 3rd Workshop on Neural Generation and Translation, pages 1-14, Hong Kong, November.
[3] K. Heafield, H. Hayashi, Y. Oda, I. Konstas, A. Finch, G. Neubig, X. Li, and A. Birch. 2020. Findings of the fourth workshop on neural generation and translation. In Proceedings of the Fourth Workshop on Neural Generation and Translation, pages 1-9, Online, July.

4. シンポジウム開催報告

4 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
（シンポジウム実行委員長）

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「第 6 回特許情報シンポジウム」を、令和 3 年 2 月 26 日（金）午後にオンラインで開催した。

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 6 回目を数えた。折からの社会情勢により開催体制を見直しオンラインでの開催となったこと、またシンポジウムの企画の遅れや日程確保の都合により開催時期が例年より遅い 2 月末となったこと、そして一般論文の募集・選考期間が十分確保できないために招待講演とパネル討論による構成となったこと、と従前とは異なる形式での実施となった。シンポジウムの企画は研究会委員を中心とする実行委員会で行い、当日は須藤が実行委員長として司会およびパネル討論モデレータの役目を務め運営を行った。参加者は最大で 100 名程度あり、イベントとしては十分な規模となった。

4.2 講演の概要

4.2.1 招待講演（1）

中西 聡 様（特許庁 総務部 総務課 情報技術統括室 室長補佐（特許情報利用推進班長））
講演題目「特許庁における機械翻訳の活用」

近年の日本国内からの海外出願率の増加や全世界的な特許・実用新案の出願の急増を背景として、特許の審査情報を各国で活用するための機械翻訳技術はさらにその重要性を増している。日本国特許庁においても日本語と外国語双方向の機械翻訳を活用しており、現在はパブリッククラウドを用いた機械翻訳プラットフォームとして様々な翻訳エンジンを活用するに至っている。また機械翻訳の持続的改善に向けた各種調査活動も継続的に実施している。

4.2.2 招待講演（2）

成瀬 博之 様（特許庁 審査第四部 インターフェイス（検索・記憶管理）先任上席審査官）
講演題目「令和２年度特許出願技術動向調査-機械翻訳-」

特許庁では毎年いくつかの技術分野を対象とした出願動向調査を行い報告しているが、令和 2 年度の調査対象の一つとして機械翻訳技術が選ばれた。本講演ではその途中経過が報告された。詳細は令和 3 年 4 月頃に報告書が公開される予定であり、本報告では詳細を述べないが、米国が市場・技術を主導する中で、中国籍の出願が急伸しており注目に値する。また、日本で特徴的な技術領域や、総務省のグローバルコミュニケーション計画 2025 のような戦略プロジェクトが存

在することを踏まえ、今後の機械翻訳関連技術の研究開発への提言が行われた。

4.2.3 招待講演（3）

清野 舜 様（国立研究開発法人理化学研究所 革新知能統合研究センター）

講演題目「機械翻訳コンペティション参加報告」

機械翻訳研究では長年に渡り様々なコンペティション（共通のデータを利用して性能を競うイベント）が行われており、近年最も注目を集めている国際会議 WMT の 2020 年新聞記事翻訳タスクにおいて、東北大・理研・NTT の合同チームがトップの成績を収めた。技術としてはニューラル機械翻訳（NMT）に基づくものであるが、教科書的な単純な構成による翻訳システムではなく、様々な工夫を凝らした「泥臭い」技術の集合体とも言える大規模かつ複雑なものとなった。それ故に必要となる計算資源も膨大となっているが、一方で機械翻訳に限らず様々な自然言語処理が同様のアプローチを用いて実現されるようになり、技術の垣根が下がり新たなチャンスが生まれてきているとも言える。

4.2.4 招待講演（4）

谷川 英和 様（IRD 国際特許事務所 所長・弁理士）

講演題目「特許文書の品質評価の標準化について　～特許文書品質特性 モデル～」

産業日本語研究会 特許文書分科会では特許文書を対象とした文書品質モデルの検討や評価を行っている。特許文書は発明を解説する技術文書という側面と知財権を確保するための権利文書という側面があり、一口に品質といっても多角的に捉える必要がある。着目すべき品質特性は権利化に向けたフェーズ毎、また評価する立場や対象によって異なるため、的確な特許文書の作成・評価のためには品質特性を把握し活用できるようにならなければならない。特許文書分科会ではそのための助けとなる教育用テキストを作成しており、弁理士をはじめとした知財関連人材の研修目的等での展開を予定している。

4.2.5 招待講演（5）

桐山 勉 様（日本特許情報機構 特許情報研究所 IP リサーチフェロー）

講演題目「COVID-19 に絡む特許情報の分析 ～ COVID-19(敵)を知り、己を守る～」

COVID-19 に関係する特許の分析を通じた提言。社会的影響が甚大であったため COVID-19 に関連する研究開発は世界各国で驚異的なスピードで進んでおり、関連する早期出願が急増する等熾烈なワクチン開発競争が行われていることが窺える。COVID-19 に関しては膨大な情報が存在するが、そうした中から重要な情報を特許情報で確認・検証することが重要である。その際に機械翻訳は有用なツールであり、世界各国のオリジナル特許の早期のアクセスを可能にする。

4.2.6 招待講演（6）

内山 将夫 様（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 上席研究員）

講演題目「自動翻訳技術の概要：なにができるか／できるようになってきているか」

情報通信研究機構（NICT）では「みんなの自動翻訳@TexTra」という機械翻訳サイトを運営しており、現在は第 2 世代の NMT と言える Transformer が主軸を担っている。実用に資するためには対象に合わせたチューニング（アダプテーション）を重視し、また実務では同様の文を繰り返し翻訳する事例が多いことから用例ベース翻訳（EBMT）も合わせて活用している。対訳データの収集にあたっては、「翻訳バンク」と呼ばれる対訳データの提供を機械翻訳技術のライセンス料負担軽減に繋げる仕組みによってデータの蓄積が進んでいる。こうした大規模データに加え分野アダプテーションや用語のカスタマイズ等を活用することで機械翻訳精度は大きく向上した。一方で依然として正確性が保証されるとはなっておらず、利用には引き続き注意を要する。

4.3 パネル討論

シンポジウムの最後に招待講演者（中西様・清野様・谷川様・桐山様・内山様）をパネリストとしたパネル討論を行った。特許情報処理の課題や今後について議論すべく、須藤が次の 3 つのトピックを事前にパネリストに提示し、討論の時間に聴講参加者からの質問を受けつつ議論を行った。事前に提示したトピックは以下の 3 点である。

1. 特許のような専門的な業務において人と機械の得手不得手は何で、何を人が担い、何を機械が担うべきと考えられるか？
2. 西欧言語・中韓以外の言語、特にアジア諸語をどう扱っていくか。仮に主要言語ほど精度が出なくても用途はあるか？また非主要言語におけるチャレンジとは何か？
3. 特許という大規模な公開技術文書資源は広く諸用途に活用できないか？

トピックを機械翻訳に寄せすぎたこともあり特許実務の先生方のご意見を十分くみ取れなかった点は反省点と言えるが、一口に特許と言っても様々な側面があり、一概に特許だからこうであるというほど単純なものではないということが（企図したものではないが）一つの結論と言えよう。ただ、近年の機械学習技術の進展は言語の違いや分野の違い、さらにはメディアの違いすらも超えて様々なタスクに「適応」が可能な汎用性を持つと謳うモデルを生むに至っており、特に資源の多い領域の知見を十分に集めることの重要性は変わらないと言えよう。

4.4 所感

オンライン開催のシンポジウムであったために参加者の反応が分からなかったのが残念であるが、様々な観点からの講演で充実したものであったと言って良いだろう。ただ、一つ開催側として心残りだった点は機械翻訳以外に特許実務と先端研究のギャップを埋めるような議論が今回提供できなかったことである。自然言語処理業界はこの 3-4 年で急速に変化を遂げており、国内外問わず技術開発競争が非常に激しくなっている。そうした観点からは特許実務における自然言語処理をはじめとした最新技術の適用可能性についてより踏み込んだ議論ができる場にしたい。

2020年度AAMT/Japio特許翻訳研究会報告書

発 行 日　2021 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　一般社団法人　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