マルチソーストランスフォーマと専門用語辞書を用いた訳語の制御方法

石川雄太郎　江原暉将
(一財) 日本特許情報機構 (Japio)　知財 AI 研究センター
yutaro_ishikawa@japio.or.jp, eharate@gmail.com

1 はじめに

ニューラル機械翻訳において、専門用語などの訳語の語彙指定は、実用上、重要なタスクとなっている。特に、特許翻訳に代表される、技術を扱う実務翻訳には重要である。現在、訳語の語彙指定には、大きく分けて２つのアプローチがある。１つ目は、プレースホルダを使用して後処理的に置換を行う方法[1][2]であり、２つ目は、語彙制約翻訳(Lexically Constrained Decoding) である[3][4]。

プレースホルダを使用する方法は、原文側の単語を『<PLS1>』等の特殊な記号～プレースホルダ～に置き換えて翻訳した後、訳文側に出力されたプレースホルダを指定単語で置き換えるものである。しかしながら、この方法は置換前の単語の意味が保持されないことから、翻訳の適切さや流暢さの低下が懸念される[5]。

語彙制約翻訳は、翻訳時のビームサーチを拡張し、指定単語を含むようにビームサーチを行うことで訳文側に指定単語を出力させる方法である。しかしながら、この方法は、多くの翻訳時間が必要な点が指摘されている[6]ほか、訳文側の指定単語と原文側の単語の間に対応関係がないため、原文側の単語が複数回翻訳され得る点が課題として挙げられる[4]。

これらの課題は、訳語指定を、前処理や後処理を用いて行うために生じると推察される。そこで、本研究では、課題に対する１つのアプローチとして、原文・訳文・訳文側の指定単語・原文側の単語を、エンドツーエンド (end-to-end) で学習する手法を提案する。

これによって、翻訳時における原文側単語の意味の保持や、訳文側指定単語と原文側単語の間の対応関係の明示が可能になるだけでなく、複雑なビームサーチが不要となり、翻訳時間が短縮されることが期待される。本稿では、提案手法と、その有用性について議論する。

2 提案手法

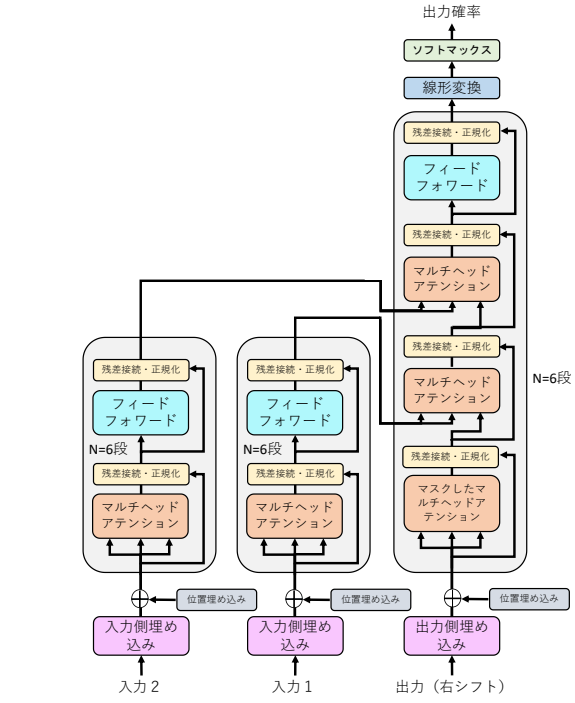
提案手法においては、マルチソーストランスフォーマアーキテクチャを用いる。マルチソーストランスフォーマアーキテクチャでは、複数のエンコーダに異なる種類の入力を行うことが出来る。例えば、原文と画像の両方を入力し、画像情報を考慮した翻訳を行うマルチモーダル翻訳[7]や、原文と機械翻訳結果を入力し、後編集結果を出力する自動後編集[8]等に用いられている。

図１に、提案手法で用いるマルチソーストランスフォーマアーキテクチャの概要を示す。通常のトランスフォーマと異なる点は、複数のエンコーダを備えた点と、ソースターゲットマルチヘッドアテンション、残差接続・正規化からなるコンポーネントが各エンコーダに対応して、デコーダレイヤ内に繰り返しスタックされている点である。

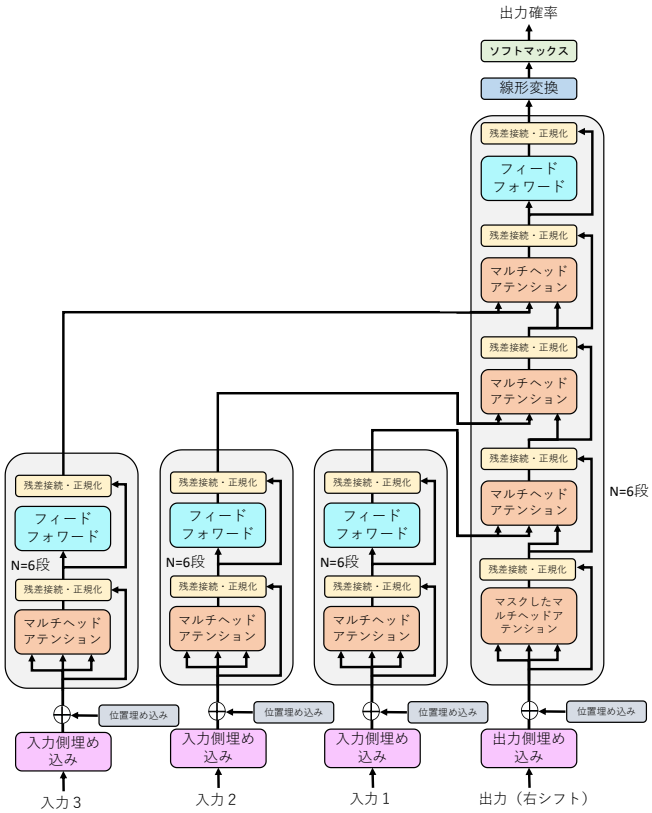
なお、本研究では、訳文側の指定単語（複合語を含む）・原文側の単語（複合語を含む）を１つのエンコーダに交互に入力するか（図１の（ａ）ダブルエンコーダ)、または、２つのエンコーダに分けて入力するか（図１の（ｂ）トリプルエンコーダ）の２パターンを試した。提案手法の入出力については、以下の通り。

（ａ）ダブルエンコーダ
・入力１：原文
・入力２：訳文側の指定単語と原文側の単語とを交互に並べたもの（複数）
・出力：訳文

（ｂ）トリプルエンコーダ
・入力１：原文
・入力２：訳文側の指定単語（複数）
・入力３：原文側の単語（複数）
・出力：訳文



（a）ダブルエンコーダ



（b）トリプルエンコーダ

図１　提案手法で用いるマルチソーストランスフォーマアーキテクチャ

3 実験

本実験では英日特許機械翻訳を対象にして、提案手法に加えて、手法の有効性の確認のため、通常のトランスフォーマ（バニラ）、プレースホルダを用いた手法について、指定単語出力割合と翻訳品質を比較する。

3.1 データの作成

対訳データは、当財団で独自に作成したクリーニング済みの英日特許対訳コーパスの中から、学習：100 万文対、開発とテスト：各 2000 文対をランダムに抽出した。また、専門用語辞書としては、特許公開公報をベースに当財団で独自に抽出した人手チェック済みの英日専門用語辞書から１つの英単語・フレーズに対して複数の日本語訳語が対応するものを、約 117 万エントリー抽出した。そして、各英単語・フレーズについて、より長い英単語・フレーズを優先するように、対訳データと辞書の文字列マッチングを、英日両側で行い、対訳データと辞書が英日両側ともにマッチングする単語のみを抽出し、これを訳文側の指定単語と原文側の単語とした。

次に、対訳データと訳文側の指定単語と原文側の単語を各モデルの入力に対応させるべく、表１のような加工を行った。プレースホルダを用いた手法では、プレースホルダにインデックスを付与する標準的な手法を用いた。提案手法のトリプルエンコーダ、ダブルエンコーダでは、訳文側の指定単語と原文側の単語をそれぞれ制御文字[SEP]で区切っている。

その後、英日各々について、語彙数 8000 のセンテンスピース[9]で、制御文字やプレースホルダを区切らないように留意しつつ、トークナイズを行った。

3.2 学習と翻訳

学習はオープンソースソフトウェアである Marian NMT[10]を利用して行った。ダブルエンコーダについては、type オプション “multi-transformer” を指定し、トリプルエンコーダについては、“multi-transformer” のエンコーダ数を 3 つにしたものを用いた。モデルのパラメータの最適化は Adam を用いて行った。

表１　学習時・翻訳時における各モデルへの入力と出力

バニラ （既存手法）	入力：next, the lens surface shape of the free-form surface lens 10 for correcting scanning distortion will be described. 出力：次に、走査歪み補正用の自由曲面レンズ10のレンズ面形状について説明する。
プレースホルダ （既存手法）	入力：next, the <PLS1> of the <PLS2> lens 10 for correcting <PLS3> will be described. 出力：次に、<PLS3>補正用の<PLS2>レンズ10の<PLS1>について説明する。
ダブルエンコーダ （提案手法）	入力１：next, the lens surface shape of the free-form surface lens 10 for correcting scanning distortion will be described. 入力２：[SEP] lens surface shape [SEP] レンズ面形状 [SEP] free-form surface [SEP] 自由曲面 [SEP] scanning distortion [SEP] 走査歪み 出力：次に、走査歪み補正用の自由曲面レンズ10のレンズ面形状について説明する。
トリプルエンコーダ （提案手法）	入力１：next, the lens surface shape of the free-form surface lens 10 for correcting scanning distortion will be described. 入力２：[SEP] lens surface shape [SEP] free-form surface [SEP] scanning distortion 入力３：[SEP] レンズ面形状 [SEP] 自由曲面 [SEP] 走査歪み 出力：次に、走査歪み補正用の自由曲面レンズ10のレンズ面形状について説明する。

表２　各モデルの BLEU 値と指定語の出力割合（2000 文）

	BLEU値	指定語出力割合
バニラ（既存手法）	46.56	5722/7464
プレースホルダ（既存手法）	49.22	7293/7464
ダブルエンコーダ（提案手法）	50.93	7197/7464
トリプルエンコーダ（提案手法）	50.86	7184/7464

表３　人手によるエラーカウント結果（300 文、各項目に重複有り）

	湧き出し	訳抜け	その他誤訳	エラー文総計
バニラ（既存手法）	8	54	78	122
プレースホルダ（既存手法）	13	62	80	127
ダブルエンコーダ（提案手法）	6	62	71	119
トリプルエンコーダ（提案手法）	4	55	70	109

デフォルトからパラメータを変更した箇所は、エンコーダ・デコーダ層数（6 層）、最大文長（200 トークン）、label-smoothing の指定、exponential-smoothing の指定、warmup と decay ステップ数の指定、ネットワーク各部のドロップアウト率の設定である。各々１万ステップごとに開発セットで評価を行い、クロスエントロピーが最も低いモデルを最良のモデルとして用いた。また、翻訳後にはデトークナイズを行い、プレースホルダのみ翻訳後に訳文側の指定単語に戻す処理を行った。

3.3 結果と考察

表２が各モデルの BLEU 値と指定語の出力割合である。テストセットは、2000 文を用いた。BLEU 値は mecab でトークナイズした後、Moses の multi-bleu.perl で算出した。指定語出力割合は全ての指定語（2000 文中の 7464 単語）のうち、実際に訳文に出力された語の割合を示す。なお、バニラには、語を指定する機能は無いが、何も指定しない場合、周辺文脈だけでどの程度希望の語が出力されるかの比較のために掲載した。

提案手法は、プレースホルダを用いた手法と比較して、指定語出力割合が 1%強（100 語程度）低いにも関わらず、BLEU 値は約 1.6～1.7 高くなっており、より参照訳に近い訳文が出力されていることが分かった。

次に、人手によるエラーカウント結果について、表３に示す。テストセット中から 300 文をランダム抽出し、人手で湧き出し（原文にない形態素～文節等が出現）、訳抜け（原文にある形態素～文節等が消失）、その他誤訳（関連語の混同や係り受けの間違い等）が存在するか否かを文単位でカウントした。但し、指定語が出力されていなかった場合、許容可能な単語が出力されていれば、誤訳とはカウントしていない。また、上述のいずれかのエラーが存在した文を、エラー文総計にカウントした。

比較手法のプレースホルダを用いた手法は、湧き出し、訳抜け、その他誤訳の全ての項目において件数が多く、従来から指摘されている通り、適切さに問題が生じている一方で、湧き出しとその他誤訳は共に提案手法の方が改善されており、結果としてエラー文総数もバニラと同等（ダブルエンコーダ）もしくは減少（トリプルエンコーダ）となっている。プレースホルダを用いた手法の方が、指定語出力割合が高いにも関わらず、BLEU 値が低いという今回の結果については、このような原因であると推察される。表４、５に実例を示す。

表４　湧き出し、訳抜け、その他誤訳の実例

	原文	提案手法（ダブルエンコーダ）	プレースホルダ
湧き出し	as shown in FIG. 6, the roof 22 has flanges 23 disposed along both ends, respectively, and a pair of horizontal supporting bars 25 secured to the frame 11 passing through the flanges, whereby the roof 22 is secured to the frame.	ルーフ22は、図6に示すように、両端にそれぞれ形成されたフランジ23と、このフランジを通るフレーム11に固定された一対の水平支持バー25とを有し、フレームに固定される。	図6に示すように、ルーフ22は、両端に沿って配置されたフランジ23と、このフランジを貫通するフレーム11に固定された一対の水平な支持バー25とを有し、フレーム22に固定されている。
訳抜け	a photoresist pattern 3 is formed on the insulating layer 2, and the insulating layer 2 is etched using it, thereby forming an opening 20 in the insulating layer 2 (opening forming step), as illustrated in FIG. 1B.	次に、図1(b)に示すように、絶縁層2上にフォトレジストパターン3を形成し、これをエッチングすることにより、絶縁層2に開口部20を形成する(開口形成工程)。	次に、絶縁層2上にフォトレジストパターン3を形成し、図1(b)に示すように、絶縁層2に開口部20を形成する(開口形成工程)。
誤訳その他	in the second current mirror CM2 that forms another part of the input of the bias circuit Bias_Gen, current proportional to the collector reference current IQref is subtracted from the APC current Iapc proportional to the APC current Iapc0 from the voltage-current converter V/I.	バイアス回路Bias_Genの入力の別の部分を形成する第2のカレントミラーCM2では、電圧電流変換器V/Iからの APC電流Iapc0に比例したAPC電流Iapcからコレクタ基準電流IQrefに比例した電流を減算する。	バイアス回路Bias_Genの入力の一部を構成する第2のカレントミラーCM2では、コレクタ基準電流IQrefに比例した電流が、電圧-電流変換器V/IからAPC電流Iapc0に比例したAPC電流Iapcから減算される。 **（不自然な訳語）**

表５　訳語指定の実例（下線部は指定単語　〇：指定通り　×：指定外）

	原文	参照訳	提案手法（ダブルエンコーダ）	プレースホルダ
プレースホルダが優れていた例	the diagnosis result creation processing unit 10 creates a diagnosis result of the module battery 2 to be diagnosed based on a processing result of the operation result monitoring processing unit 8 and the manufacturing/usage environment factor classification processing unit 9.	診断結果作成処理部10は、稼動実績監視処理部8及び製造・使用環境の要因分類処理部9の処理結果に基づいて診断対象となるモジュール電池2の診断結果を作成する。	診断結果作成処理部（〇）10は、**処理結果監視処理部（×）**8および稼動**実績（×）要因（×）分類処理部（〇）**9の処理結果に基づいて、診断対象のモジュール電池（〇）2の診断結果（〇）を作成する。	診断結果作成処理部（〇）10は、**稼動実績監視処理部（〇）**8及び製造**環境（〇）要因（〇）分類処理部（〇）**9の処理結果に基づいて、診断対象のモジュール電池（〇）2の診断結果（〇）を作成する。
提案手法（ダブルエンコーダ）が優れていた例	the increase in the surface roughness of Ru derived from formation of the second intermediate layer 15, which is thin and is composed of a metal or an alloy having the fcc lattice structure, on the first intermediate layer composed of Ti having the hcp structure has presumably relation to the melting point.	hcp構造を有するTi第一中間層上に薄いfcc構造を有する金属もしくは合金からなる第二中間層15を形成した際の粗さの増加は、融点と関係があると考えられる。	hcp構造を有するTiからなる第1中間層上に、薄い、fcc格子構造を有する**金属（〇）**または**合金（〇）**からなる第2中間層15の形成に由来するRuの面粗さの増大は、融点に関連していると考えられる。	hcp構造を有するTiからなる第1の中間層において、薄い（×）（×）第2の中間層15の形成に由来するRuの表面粗さの増加は、融点に関連していると考えられる。

まとめると、提案手法は従来のプレースホルダを用いた訳語の指定方法に比して、指定性能は少し下がるものの、より適切な文を出力可能であり、既存手法と同程度以上の翻訳品質を保ったまま、訳語指定を実現できることが分かった。

4 おわりに

本稿では、原文・訳文・訳文側の指定単語・原文側の単語を、エンドツーエンドで学習することで、モデルアーキテクチャそのものに語彙指定翻訳機能を付与したシステムを提案し、その有効性について検討した。その結果、翻訳の適切さを保ちつつ、全体の 96%程度の語彙指定を行うことが出来ることを確認した。

今後については、語彙制約翻訳の１つである Post らの DBA[11]や、同じくエンドツーエンドの訳語指定モデルとしての Dinu らのモデル[12]との性能比較を行いたい。また、一文献内での訳語統一や用語集を用いた人手翻訳支援に提案手法が適用可能か検討し、その有効性も検証していきたい。

参考文献

[1] Crego, Josep, et al. "Systran's Pure Neural Machine Translation Systems." arXiv preprint arXiv:1610.05540 (2016).
[2] Wang, Yuguang, et al. "Sogou Neural Machine Translation Systems for WMT17." Proceedings of the Second Conference on Machine Translation. 2017.
[3] Hokamp, Chris, and Qun Liu. "Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.
[4] Hasler, Eva, et al. "Neural Machine Translation Decoding with Terminology Constraints." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018.
[5] Song, Kai, et al. "Code-Switching for Enhancing NMT with Pre-Specified Translation." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.
[6] Post, Matt, et al. "An Exploration of Placeholding in Neural Machine Translation." Proceedings of Machine Translation Summit XVII Volume 1: Research Track. 2019.
[7] Takushima, Hiroki, et al. "Multimodal Neural Machine Translation Using CNN and Transformer Encoder." No. 873. EasyChair, 2019.
[8] Junczys-Dowmunt, Marcin, and Roman Grundkiewicz. "MS-UEdin Submission to the WMT2018 APE Shared Task: Dual-Source Transformer for Automatic Post-Editing." Proceedings of the Third Conference on Machine Translation: Shared Task Papers. 2018.
[9] Kudo, Taku, and John Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018.
[10] Marian NMT <https://marian-nmt.github.io/>
[11] Post, Matt, and David Vilar. "Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018.
[12] Dinu, Georgiana, et al. "Training Neural Machine Translation To Apply Terminology Constraints." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.