逆順デコーダを用いた係り受け構造に基づく Transformer ニューラル機械翻訳

佐々木 拓馬 [1], 田村 晃裕 [1], 出口 祥之 [2], 二宮 崇 [2], 加藤 恒夫 [1]
[1] 同志社大学 理工学部, [2] 愛媛大学 大学院理工学研究科
[1] {cgub0052@mail4, aktamura@mail, tsukato@mail}.doshisha.ac.jp
[2] {deguchi@ai., ninomiya@}cs.ehime-u.ac.jp

1 はじめに

自然言語処理の分野において機械翻訳は古くから盛んに研究されており，近年では，ニューラルネットワークを用いた機械翻訳（Neural Machine Translation; NMT）が主流となっている．NMT の中でも，特に，文内の単語間の関連の強さを捉える self-attention を備えた Transformer NMT [7] が，NMT 初期に広く使われていた RNN や CNN ベースの NMT の翻訳精度を上回り，現在注目を集めている．

これまで，統計的機械翻訳や NMT では，文構造（句構造や係り受け構造など）を考慮することで翻訳精度が改善されている．Transformer NMT においても例外ではなく，原言語文や目的言語文の文構造を活用することで，その翻訳精度が向上することが報告されている [1, 3, 6, 9, 10, 11, 13]．出口ら [13] は，self-attention の一部に対して，係り先の単語に注意が向くような制約を与えて学習することで，文の係り受け構造を捉える self-attention（以降，dependency-based self-attention（DBSA）と呼ぶ）を提案し，Transformer NMT のエンコーダとデコーダで DBSA を使うことで，翻訳時に原言語文と目的言語文の係り受け構造を考慮して翻訳を行うモデルを提案した．そして，Asian Scientific Paper Excerpt Corpus (ASPEC) [5] の日英，英日翻訳において翻訳精度の改善を確認している．

通常の Transformer NMT では，デコーダの self-attention を学習する際，推論時には生成していない単語に対する関連を算出しないように自身より後方の単語をマスクする．出口らのモデルにおいても，デコーダの DBSA が未生成の単語に対して注意を向けないように，自身より後方の単語に係る係り受け関係はマスクして学習される．そのため，デコーダの DBSA は後方に向かう係り受け関係は活用できない．一方，言語によっては，後方に向かう係り受け関係が多く生じる．例えば日本語では，通常，係り先は自分の文節より後方の文節になる．そのような言語が目的言語の場合，デコーダの DBSA の効果が十分に得られないと考えられる．文献 [13] において，日英翻訳では提案モデルにより BLEU が 1.04 ポイント向上したが，英日翻訳では 0.30 ポイントの向上にとどまっていることもこれを裏付けていると考えられる．

そこで本研究では，ほとんどの係り受け関係において係り先が自身より後方になる日本語が目的言語となる翻訳を想定し，目的言語文（日本語文）の文構造を self-attention で効果的に捉えるための Transformer NMT モデルを提案する．具体的には，DBSA に基づくモデル [13] のデコーダにおいて，目的言語文を文頭単語から順に生成するのではなく，文末単語から順に通常の逆順で生成する．この逆順デコーダにより，自身より後方の単語に係る係り受け関係に対して，係り先の単語は係り元の単語より先に生成されるため，マスクされることなくデコーダの DBSA で捉えることが可能となる．

ASPEC の英日翻訳実験において提案モデルと従来の出口らのモデル [13] を比較した結果，目的言語文を逆順で生成することにより，DBSA に基づく Transformer NMT の性能が，学習データが 10 万文対の場合は BLEU で 0.74 ポイント，100 万文対の場合は 0.11 ポイント改善することを確認した．

2 従来モデル：DBSA に基づく Transformer NMT

本節では，提案モデルの基礎となる DBSA に基づく Transformer NMT [13] を概説する．概要図を図 1 に示す．DBSA に基づくモデルは，Transformer NMT [7] のエンコーダとデコーダに，係り受け関係

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

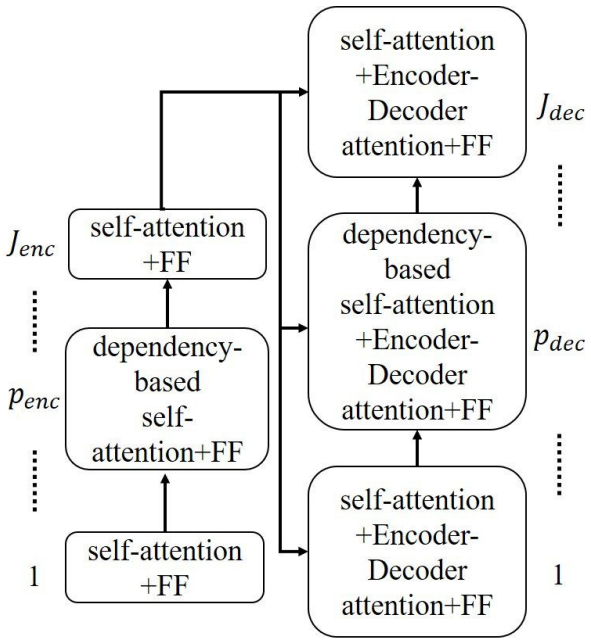

図 1 DBSA に基づく Transformer NMT [13] の概要図

を一部のヘッドで捉える multi-head self-attention である DBSA を導入し，原言語文と目的言語文の係り受け構造を考慮した翻訳を行う．図 1 では，DBSA が p_{enc} 層目のエンコーダと p_{dec} 層目のデコーダに組み込まれている．

DBSA に基づくモデルの学習では，次式の目的関数 $\mathcal{L}$ を最小化することで，翻訳と係り受け解析を同時に学習する．

$$\mathcal{L} = \mathcal{L}_{trans} + \lambda_{enc}\mathcal{L}_{encdep} + \lambda_{dec}\mathcal{L}_{decdep} \quad (1)$$

ここで，$\lambda_{enc} > 0$ と $\lambda_{dec} > 0$ はハイパーパラメータである．$\mathcal{L}_{trans}$ は翻訳に関する誤差であり，ラベル平滑化交差エントロピーにより算出する．また，$\mathcal{L}_{encdep}$ と $\mathcal{L}_{decdep}$ はそれぞれエンコーダ側とデコーダ側の DBSA で捉える係り受け解析に関する誤差であり，交差エントロピーによって算出する．ただし，$\mathcal{L}_{decdep}$ を算出する際は，推論時には予測していない単語に注意を向けないようにするため，自身より後方の単語への係り受け関係はマスクされる．

2.1 Transformer NMT

Transformer NMT は，入力文を中間表現に変換するエンコーダと中間表現から出力文を生成する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．エンコーダやデコーダの入力は，入力単語の埋め込み表現に，入力単語の文における位置情報をエンコードした位置エンコーディングを加えたものである．エンコーダとデコーダでは，それぞれエンコーダレイヤとデコーダレイヤが複数層積み重ねられている．エンコーダレイヤは，入力側から順に，self-attention，位置毎のフィードフォーワードネットワークの 2 つのサブレイヤで構成されている．デコーダレイヤはエンコーダレイヤのサブレイヤに，原言語文と目的言語文間の encoder-decoder attention を加えた 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，残差接続を行った後，層正規化が適用される．

self-attention 及び encoder-decoder attention は multi-head attention により実現される．multi-head attention は, 単語間の関連の強さを捉える機構であり, 単語の埋め込み次元（d 次元）を n_{head} 個の $d_{head}(= d/n_{head})$ 次元の部分空間に射影し，部分空間毎に attention の計算を行う．各部分空間をヘッドと呼ぶ．h 番目のヘッドでは，h 番目の部分空間に射影された multi-head attention の入力 Q_h, K_h, V_h に基づき，次式の演算により，単語間の関連の強さを重みとする重み付き和表現 M_h を得る．

$$A_h = softmax(d_{head}^{-0.5} Q_h K_h^T) \quad (2)$$

$$M_h = A_h V_h \quad (3)$$

ここで，A_h が単語間の関連の強さを表す行列である．ただし，デコーダの self-attention を学習する際は，推論時には予測していない単語との関連を求めないように，Q_h の各単語に対する後方の単語を表す K_h の各成分をマスクしてから A_h を求める．全ヘッドの計算が終わったら，それらの出力 $M_{1,\ldots,n_{head}}$ を結合し，単語の埋め込み次元に線形変換した以下の M が，multi-head attention の出力となる．

$$M = W^M [M_1; M_2; \ldots; M_{n_{head}}] \quad (4)$$

ここで，$W^M \in \mathbb{R}^{d \times d}$ はパラメータ行列である．

目的言語文は，デコーダの最終層の出力を単語の語彙数次元に線形変換した後，ソフトマックス関数をかけることで得られる単語出力確率分布に基づき生成する．

2.2 DBSA

DBSA は，multi-head attention の一部のヘッドにおいて，Deep Biaffine parser [2] の要領で係り受け関係を捉える．具体的には，係り受け関係を捉えるヘッドの入力を Q_{dep}，K_{dep}，V_{dep} とすると，まず次式の通り，bi-affine 変換によって単語間の係り受け関

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

係を示す行列 A_{dep} を求める．

$$A_{dep} = softmax(Q_{dep}U^{(1)}K_{dep}^{\top} + Q_{dep}U^{(2)}) \quad (5)$$

ここで，$U^{(1)} \in \mathbb{R}^{d_{head} \times d_{head}}$，$U^{(2)} = \overbrace{(\mathbf{u} \dots \mathbf{u})}^{n}$，$\mathbf{u} \in \mathbb{R}^{d_{head}}$ はパラメータ行列である．また，A_{dep} の要素 $A_{dep}[v, w]$ は，単語 w が単語 v の係り先である確率を示している．その後，A_{dep} と V_{dep} をかけ合わせることで単語間の係り受け関係の強さを重みとする重み付き和表現 M_{dep} を得る．

$$M_{dep} = A_{dep}V_{dep} \quad (6)$$

multi-head self-attention において，ある 1 つのヘッドを M_{dep}，残りを通常のヘッドとしたものが DBSA である．つまり，M_{dep} と $n_{head} - 1$ 個の通常のヘッド（$M_{2,\dots,n_{head}}$）を結合し，単語の埋め込み次元に線形変換した以下の $M^{(p)}$ が，DBSA の出力である．

$$M^{(p)} = W^{M^{(p)}}[M_{dep}; M_2; \dots; M_{n_{head}}] \quad (7)$$

ここで，$W^{M^{(p)}} \in \mathbb{R}^{d \times d}$ はパラメータ行列である．

DBSA はサブワード列に適用できるように拡張されている．サブワード単位の DBSA は，まず，単語単位の係り受け構造をサブワード単位の係り受け構造に変換する．具体的には，1 つの単語が複数のサブワードを含む場合，右端以外の各サブワードの係り先は隣接する右側のサブワードとし，右端のサブワードの係り先は元の単語の係り先[1)]とする．その後，変換したサブワード単位の係り受け構造に対して DBSA を適用することで，サブワード列に対して DBSA に基づくモデルが適用できる．

3　提案モデル：逆順デコーダを用いた DBSA に基づく Transformer NMT

2 節で述べた通り，従来の DBSA に基づくモデルは，デコーダの DBSA を学習する際，自身より後方の単語に係る係り受け関係はマスクされるため，目的言語文内の後方への係り受け関係を捉えることができない．一方で，日本語は自分より後方の文節に係るという特徴があるため，日本語が目的言語である場合，従来モデルでは日本語文の係り受け関係をほとんど活用できない．

そこで提案モデルでは，目的言語文を文末から文頭に向かって生成する逆順デコーダを導入することで前記問題点を解決する．具体的には，提案モデル

1)　係り先の単語が複数のサブワードを含む場合，左端のサブワードを係り先とする．

図 2　従来モデルと提案モデルのデコーダ側の DBSA の例

を学習する際の目的言語文として，対訳データの目的言語文を逆順に並べ替えた文を用いる．これにより，元の文では自身より後方にあった係り先が並び替え後は前方に位置することになり，その係り先への係り受け関係がマスクされずに学習できる．そして推論時には，学習した逆順デコーダを用いて，目的言語文を文末から文頭に向かって生成する．これにより，元の文では後方に向かう係り受け関係を捉えながら翻訳を行うことが可能となる．そして最後に，逆順デコーダで生成した逆順の目的言語文を逆から並べ変えた文を翻訳文として出力する．こうすることで，自分より後方に係るという特徴を持つ日本語を目的言語とする翻訳において，提案モデルは日本語文の係り受け構造を活用でき，翻訳精度の向上が期待できる．

図 2 に従来モデルと提案モデルのデコーダ側の DBSA の例を示す．例文は ASPEC の日本語文の実例である．図 2 において，斜線のセルはマスクされる要素，黒いセルは係り先を表す．つまり，学習時には，係り受け関係を捉えるヘッドは黒いセルの確率が 1 に近づくように学習される．図より，従来モデルでは，ほとんどの係り受け関係がマスクされて考慮されないことが分かる．これは日本語の係り受け関係が文頭に近い方から文末に近い方に係る特徴があるためである．また，従来モデルをサブワード

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

列に適用する際，単語内の各サブワードは，右端のサブワードを除き，右隣に係るように係り受け構造の変換が行われる（2.2 節参照）ため，単語内のサブワード間の係り受け関係も考慮されない．一方で，提案モデルのデコーダでは，目的言語文を逆順にして扱うため，全ての係り受け関係がマスクされずに考慮できていることがわかる．

4 実験

4.1 実験設定

提案モデルの有効性を確かめるため，ASPEC の日英翻訳タスクにおいて，提案モデルを従来の DBSA に基づくモデル [13] と比較する．それぞれのモデルを「DBSA（逆順）」,「DBSA（正順）」と表記する．また，係り受け構造を活用しない Transformer NMT に対しても，逆順デコーダの有効性を確認するため，Transformer NMT モデル [7] とその Transformer NMT において目的言語文を逆順で扱うモデルの性能も評価する．それぞれのモデルを「Transformer（正順）」,「Transformer（逆順）」と表記する．データセットは ASPEC[5] の日英対訳コーパスを使用した．学習データは train-1.txt から抽出した上位 10 万文対と，train-1.txt と train-2.txt から抽出した上位 150 万文対の 2 種類を用いた．英語の単語分割には Moses Tokenizer，日本語文の単語分割には KyTea を用いた．また，各文は BPE によりサブワードに分割した．原言語側と目的言語側で独立に BPE モデルを学習し，それぞれの語彙数は 16,000 トークンとした．英語文の係り受け解析には Stanza，日本語文の係り受け解析には EDA を用いた．従来モデル及び提案モデルの設定は，基本的には文献 [13] に従った．文献 [13] の設定からの変更点は，DBSA を組み込む層を p_{enc}=3，p_{dec}=3 とし，ミニバッチの大きさを 100 文，エポック数を学習データが 10 万文対の時は 50 エポック，150 万文対の時は 20 エポックとした点である．また，ハイパーパラメータ λ_{enc}, λ_{dec} は，0.05, 0.1, 0.5, 1.0 を試し，開発データで最も良い性能であった値を評価時に採用した．

4.2 実験結果

表 1 に実験結果を示す．翻訳性能は BLEU により評価した．表 1 より，学習データが 10 万文対と 150 万文対の両方の場合において，DBSA（正順）より提案モデルである DBSA（逆順）の方が翻訳精度が高いことが分かる．具体的には，DBSA（逆順）は DBSA（正順）と比べて，学習データが 10 万文対の場合は BLEU で 0.74 ポイント，学習データが 150 万文対の場合は 0.14 ポイント翻訳性能が高い．これより，逆順デコーダを用いることで，DBSA に基づく Transformer NMT の英日翻訳性能が改善でき，提案手法が英日翻訳において有効であることを確認した．

表 1 実験結果

学習データ	モデル	BLEU (%)
10 万文対	Transformer（正順）	29.76
	Transformer（逆順）	30.29
	DBSA（正順）	31.13
	提案：DBSA（逆順）	31.87
150 万文対	Transformer（正順）	44.05
	Transformer（逆順）	44.29
	DBSA（正順）	44.21
	提案：DBSA（逆順）	44.35

また，係り受け構造を活用しない Transformer NMT においても，逆順デコーダを用いることで，学習データが 10 万文対の場合は BLEU が 0.53 上がり，150 万文対の場合は BLEU が 0.24 上がっていることが分かる．これは，文献 [8] などで報告されている通り，Transformer の self-attention の一部では係り受け関係などの文構造を捉えている可能性があり，係り受け構造を教師信号として与えなくても，逆順デコーダを用いることで潜在的に係り受け関係に近い構造を self-attention で捉えることが可能になり，DBSA に基づく Transformer モデルと同様の効果が表れたのではないかと思われる．

5 おわりに

本研究では，係り先が自身より後方になる特徴を持つ日本語が目的言語の場合の翻訳精度を改善するため，DBSA に基づく Transformer NMT において目的言語文を逆順で生成する逆順デコーダを活用する手法を提案した．ASPEC の英日翻訳タスクの評価実験を通じて，提案手法により，学習データが 10 万文対の場合は BLEU が 0.74 ポイント，100 万文対の場合は 0.14 ポイント改善することを確認した．今後は中日翻訳などの英日翻訳以外の翻訳対に対する有効性を確認したい．また，文献 [4] や [12] などのように正順方向と逆順方向のデコーダを組み合わせた方法への拡張も行いたい．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Emanuele Bugliarello and Naoaki Okazaki. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1618–1627, 2020.

[2] Timothy Dozat and Christopher D Manning. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*, 2016.

[3] Chunpeng Ma, Akihiro Tamura, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. Improving neural machine translation with neural syntactic distance. In *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)*, pp. 2032–2037, 2019.

[4] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. NTT neural machine translation systems at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pp. 99–105, 2019.

[5] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pp. 2204–2208, 2016.

[6] Yutaro Omote, Akihiro Tamura, and Takashi Ninomiya. Dependency-based relative positional encoding for transformer nmt. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2019*, pp. 854–861, 2019.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. 2017.

[8] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 2019.

[9] Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 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*, pp. 1403–1409, 2019.

[10] Shuangzhi Wu, Dongdong Zhang, Zhirui Zhang, Nan Yang, Mu Li, and Ming Zhou. Dependency-to-dependency neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 11, pp. 2132–2141, 2018.

[11] Meishan Zhang, Zhenghua Li, Guohong Fu, and Min Zhang. Syntax-enhanced neural machine translation with syntax-aware word representations. In *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)*, pp. 1151–1161, 2019.

[12] Long Zhou, Jiajun Zhang, and Chengqing Zong. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 91–105, 2019.

[13] 出口祥之, 田村晃裕, 二宮崇. 係り受け構造に基づく attention の制約を用いた transformer ニューラル機械翻訳. 自然言語処理, Vol. 27, No. 3, pp. 553–571, 2020.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).