英日翻訳のための Positional Encoding を用いた正則化手法

岡佑依　田中貴秋　永田昌明
日本電信電話株式会社　コミュニケーション科学基礎研究所
{yui.oka.vf,takaaki.tanaka.tb,masaaki.nagata.et}@hco.ntt.co.jp

概要

日常会話などで使用される話し言葉，すなわち自由発話における翻訳では，データセット量が少ない，語順の自由度が高いといった問題があり高品質な翻訳のためにはまだ多くの課題が残っている．本研究では，句単位の摂動を用いた正則化手法によってモデル内で様々な書き言葉文を表現し学習することで翻訳精度の改善を試みる．この句単位の摂動はトークンではなく Positional Encoding に付与することで欠陥文をサンプリングする．

1　はじめに

近年，機械翻訳はニューラルネットワークによって大きく発展した．特に，Transformer[1] は Self-Attention，Multi-Head Attention，Positional Encoding という独自の機構を利用して，高い精度の翻訳結果を残した．これは英日翻訳でも同様で，高い精度の翻訳が可能である．しかしながら，日常会話などで使用される話し言葉話における翻訳では，データセットが少ない，間投詞や言い淀みなどを含む，などといった問題があり書き言葉の翻訳と比べると翻訳精度はかなり低い．さらに，日本語など語順の自由度が高い言語では，話し言葉は書き言葉と比べ語順の自由度が大きく異なる傾向にある．

一方，摂動を用いた正則化手法 [2, 3, 4] は，モデルの頑健性の向上に有効であることが示されており，複数の正則化手法を組み合わせることで，推論時の精度をさらに向上させることが知られている．例えば，Word Dropout や Word Replacement などの手法では，トークン単位の摂動を入力文または出力文に加えることで欠陥文をサンプリングして学習する．これらのトークン単位の摂動は学習中に欠陥文をモデル内で表現しており，欠陥文から目的言語文を予測するよう学習することと同じであると考えられる．

本研究では，英日翻訳における書き言葉の翻訳精度を向上させることを目的とし，Positional Encoding 部分に句単位の摂動を付与することで学習中に欠陥のある文を表現し，頑健性を向上させることを試みた．図 1 に提案手法の直感的な例を示す．整数範囲の離散的な摂動を付与することで句単位での摂動を表現し，IBM Models 3 や 4[5] で作成される欠陥文同等の位置情報を表現する．この時，トークン単位ではなく，句単位の摂動を付与することで，意味的に不完全な欠陥文を生成するのではなく，語順が入れ替えられた文を生成する．実験結果から，話し言葉・書き言葉の英日翻訳において若干の BLEU の改善が見られた．

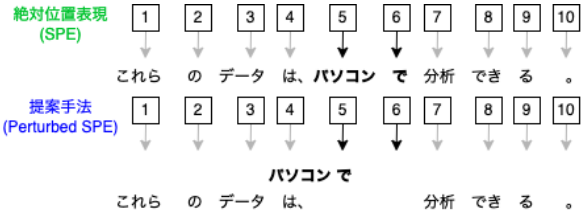

図 1　提案手法における位置表現の直感的な例．通常の位置表現とは異なり，**’パソコンで’** の出現位置がずれた欠陥文同等の位置表現をする．

2　関連研究

2.1　摂動を用いた正則化手法

正則化手法は，機械学習で過学習を防ぐために使われる．NMT においても同様であり，モデルのパラメータの過学習を防ぐために，損失関数に特定の項を足し合わせる手法や，入力にノイズ，すなわち摂動を付与し正確な文を予測するよう学習する手法などが用いられている．本稿では摂動を用いた正則化手法について述べる．

Gal ら [3] は，確率的にトークンの埋め込みをゼロベクトルに確率的に置き換える Word Dropout 法を提案した．Bengio ら [2] は，入力トークンをエンコーダ側またはデコーダ側のスケジュールドサンプリングしたトークンに置き換える Word Replacement 法を提案した．しかし，このようなサンプリングした別

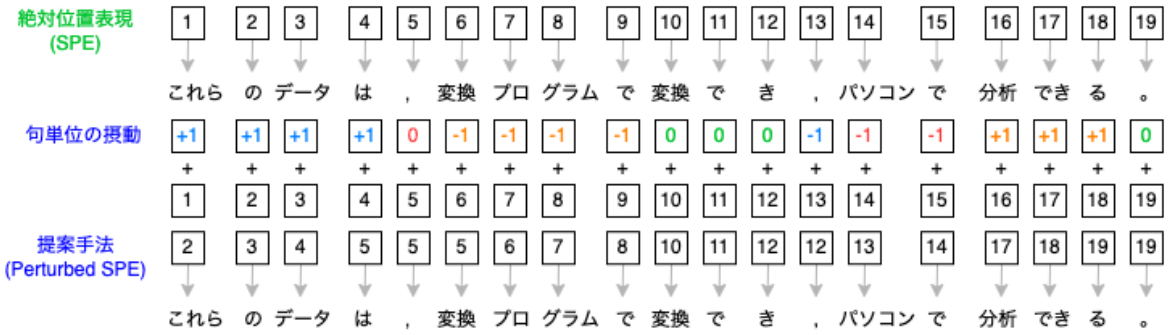

図 2 提案手法による位置表現の例．通常はトークンごとに絶対的位置表現 pos を決定するが（1 行目），提案手法では摂動範囲 [-1,+1] 内で確率的に句単位の摂動 per を決定し（2 行目）通常の絶対位置表現 pos に足し合わせる．（最終行目）．

トークンを摂動として使用する手法では，摂動を計算する計算量がかかり学習時間が長くなる傾向がある．高瀬ら [4] は，単純な正則化手法を様々に組み合わせて，学習時間を比較した．その結果，比較的単純な手法である Word Dropout や Word Replacement およびそれらを組み合わせた手法が，計算速度だけでなく翻訳精度においても従来の Transformer を上回ることがわかった．

2.2 Positional Encoding

Sinusoidal Positional Encoding（以下，SPE）は，Transformer のエンコーダ・デコーダ両者において各埋め込み表現に対し，文中における絶対位置を足し合わせることで位置情報を与える役割を持つ．その時足し合わせる値は正弦関数と余弦関数の式で表される．トークンの位置を pos，埋め込み表現の次元数を d とすると i 番目の次元の埋め込み表現に足し合わせる SPE は以下の式 (1,2) で表される．このとき，偶数次元は正弦関数，奇数次元は余弦関数で定義される．

$$SPE_{(pos,2i)} = sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{1}$$

$$SPE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{2}$$

また，上記のような式ではなく，完全に学習可能な Embedding で絶対位置を表現し，同じように各単語埋め込み表現に足し合わせる Position Embedding という位置表現手法も存在する．

Relative Position Representation（以下，RPE）は，文中における各単語の相対的な位置を表現する [6]．SPE が各単語埋め込みに足し合わせるのに対し，Self-Attention 内部で表現する．

3 提案手法

従来の手法ではトークン単位で摂動を付与していた．しかし，自然な文をサンプリングするにはトークン単位ではなく句単位で摂動を付与する方が適切であると考えられる．一方で従来の手法はトークン単位の摂動でありながら翻訳精度の改善が期待できるため従来の手法と組合せな可能な手法を提案する．本研究では，SPE 内のトークン位置 pos にトークン単位ではなく句単位の整数摂動を付与することで日本語の書き言葉文に近い欠陥文をモデル内で表現する手法を提案する．図 2 に提案手法の位置表現例を示す．SPE に摂動を組み込むことで，他の正則化手法と同様に NMT の頑健性を向上させることができ，また，モデル構造中の位置表現を操作するため他の正則化手法と補完することが期待される．さらに，トークン単位ではなく句単位で摂動を付与することで書き言葉文に近い文を表現し学習することで書き言葉における翻訳での精度向上が期待される．提案手法は SPE を拡張した式 (3,4) で表される．

$$perSPE_{(pos,2i)} = sin\left(\frac{pos+per}{10000^{\frac{2i}{d}}}\right) \tag{3}$$

$$perSPE_{(pos,2i+1)} = cos\left(\frac{pos+per}{10000^{\frac{2i}{d}}}\right), \tag{4}$$

ここで，per は一様分布からの整数摂動である．また，開始トークンの per は常に 0 とする．例えば，摂動範囲が [-1,+1] の場合，開始トークンを除いて-1,0,+1 の整数がランダムに決定する．[7] で提案されている on-the-fly subword sampling のように，摂動は文・エポック毎にサンプリングされ，文・エポック毎に異なる SPE が各トークンに付加される．摂動は学習中にのみ付与され，生成時は通常の SPE を使用する．

4 実験

実験は大きく次の 3 ステップに分けて行った．(1) まず事前実験として PE の Ablation study を行った．これはエンコーダ・デコーダどちらに摂動を付与するべきか，それとも従来のまま絶対位置を付与するべきか検討するために行った．(2) 次に話し言葉における実験を行った．(3) 最後に書き言葉においても同様に効果があるのかを検証するために書き言葉における実験を行った．

データセット　英日翻訳の実験のためのデータセットには，書き言葉の実験には対訳コーパス ASPEC[8]，話し言葉の実験には IWSLT2017[9]，加えて JParaCrawl[10] からランダムに選択した 400 万文を用いた．ASPEC は 1,783,817 文対の学習データ，1,790 文対の開発データ，1,812 文対のテストデータからなり，今回学習には 100 万文対の学習データである train-1.txt のみを使用した．IWSLT2017 はオリジナルの 232,423 文対の学習データから Moses[11] の `clean-corpus-n.perl`[1)] を使いクリーニングを行った 9,250 文対を学習データとして用い，開発には dev2010 の 871 文対，テストには tst2015 の 1,194 文対を使用した．JParaCrawl はオリジナルの 1,000 万分からランダムに 400 万文抽出し学習データとして使用した．英語及び日本語の入出力はサブワードとし，Sentencepiece[12] を使いトークナイズを行った．このとき，語彙サイズは 16,000 とし，言語間で共有した．

実験設定　実装には fairseq[13] を用いた．バッチサイズは実験 4.1 において 2048，実験 4.2 において 8192 にした．それ以外のハイパーパラメータは全てにおいて [1] と同じにした．

摂動　英日翻訳の目的言語側，すなわち日本語側で句構造の解析を行い，解析結果から句単位の摂動を付与する．MeCab[14] を使って形態素解析を行い，MeCab が出力した品詞列に対して正規表現で句単位に区切るルールを定義し，nltk[15] の `RegexpParser` でチャンキングを行った．このチャンキングは学習時に行うと学習時間が膨大となるため，学習データ作成時に行いチャンキング結果を目的言語文と共にモデルに渡すことでベースラインと変わらない速度での学習を行った．摂動範囲は全てにおいて [-1,+1] に統一し，一様分布に基づいて選択される．

1）https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl

表 1　ASPEC 英日翻訳における BLEU

エンコーダ	デコーダ	BLEU
SPE	SPE	40.8
なし	SPE	22.8
SPE	なし	41.0
RPE	RPE	39.1
なし	RPE	19.3
RPE	なし	40.3
なし	なし	22.8

評価　翻訳文の評価手法には機械翻訳の自動評価として一般的な BLEU[16] を用い，sacreBLEU[17] で計算した．

4.1 事前実験

摂動をエンコーダ・デコーダどちら側に付与するべきなのか検証するため，エンコーダ側，デコーダ側それぞれにおいて通常の SPE(絶対位置)，RPE(相対位置) を付加しない場合の実験を行った．本研究では [18] と同様，Query と Key のみに適用する簡略化された RPE を用いた．実験結果を表 1 に示す．

エンコーダに SPE を付加しない場合，デコーダにおける SPE の有無関係なしに大きく BLEU が下がった．一方で，デコーダ側の SPE の有無は翻訳精度に大きく影響を与えることはなかった．RPE の場合でも同様の傾向が見られた．したがってエンコーダ側の絶対的な位置表現が翻訳精度に大きく影響を与えていることがわかる．

4.2 話し言葉・書き言葉における実験

表 2　IWSLT17 英日翻訳における BLEU

エンコーダ	デコーダ	$En \rightarrow Ja$
SPE	SPE	8.5
SPE	なし	8.4
SPE	perSPE	9.3

表 3　ASPEC 英日翻訳における BLEU

エンコーダ	デコーダ	$En \rightarrow Ja$
SPE	SPE	41.0
SPE	なし	40.9
SPE	perSPE	41.4

話し言葉における実験　事前実験から，エンコーダ側には SPE による絶対的な位置情報を与えた方が良く，デコーダ側の SPE の位置情報の与え方には改善の余地があると考えられた．そこでデコーダ側に提案手法を適用し実験を行った．IWSLT2017 データセットと JParaCrawl データセットを使った英日翻訳

の結果を表 2 に示す．通常，話し言葉の翻訳では，話し言葉のデータセットのみでの fine-tuning を行うが，この実験結果では fine-tuning は行っていない．結果から，提案手法がベースラインと比べて 0.8 ポイント BLEU が向上したことがわかる．また，デコーダ側に SPE を与えなかった場合，ベースラインと比べて 0.1 ポイント BLEU が下がった．

書き言葉における実験　話し言葉と同様に，デコーダ側に提案手法を適用し実験を行った．表 3 に ASPEC データセットを使った書き言葉における実験結果を示す．結果より，ベースラインと比べると 0.4 ポイントの BLEU の向上が見られた．また，デコーダ側に SPE を与えなかった場合，ベースラインと比べて 0.1 ポイント BLEU が下がった．

5　考察

事前実験　事前実験の結果から，デコーダの位置表現が翻訳精度に大きく影響を与えるとは考えにくい．生成時，Transformer は自己回帰的に生成を行い，生成した前の単語 (時刻 t の単語) から次の単語 (時刻 t+1 の単語) の予測を行うようデコーダの Self-Attention ではマスクを行う．このように時系列に沿った生成を行うため，位置情報は必要ではないと考えられる．また，絶対位置表現，相対位置表現両方において同じ傾向であったことから，位置表現の違いやどこで位置情報を足し合わせるかは精度に大きく関係しないと考えられる．

提案手法　話し言葉，書き言葉における実験では翻訳精度の改善が見られた．さらに，デコーダ側で SPE を付与しない時と比べても翻訳精度の改善が見られた．デコーダ側の位置表現が必須でないならば，位置表現において正則化を行うことで若干の改善が可能であることを示唆している．若干の改善しか見られなかったのは，デコーダ側の位置情報が翻訳精度にそれほど大きな影響を与えないことが原因であると考えられる．このように単語埋め込みに直接作用する摂動を用いた正則化手法 [2, 3, 4] と比べると大きく改善するとは言い難いが，提案手法はこれらの手法と組み合わせが容易に可能であるため，組み合わせることで更なる精度の改善が期待できる．

トークン単位の摂動　また，先行研究のように句単位ではなくトークン単位で摂動を付与した場合，翻訳精度が改善するのか実験を行った．表 1 と同じパラメータで実験を行った結果を表 4 に示す．トークン単位の摂動を付与した場合でも若干の翻訳精度の改善 (+0.4 の BLEU の改善) が見られた．また，SPE をデコーダに付与しない場合と比較しても若干の改善が見られた，これはデコーダ側では絶対位置を付与するのではなく，正則化を行う提案手法を適用することで翻訳の改善が見込めることを示すと考えられる．一方で，エンコーダに摂動を付与したとき翻訳精度がベースラインより下がったことから，エンコーダ側ではやはり絶対的な位置表現をしたほうがよいと考えられる．

表 4　ASPEC 英日翻訳においてトークン単位の摂動を付与した場合の実験結果．perSPE の摂動範囲はすべて [-2, +2] である．

Enc	*Dec*	*En → Ja*
SPE	SPE	40.8
perSPE	SPE	37.4
SPE	perSPE	41.2
perSPE	perSPE	23.7
SPE	なし	41.0

6　まとめ

本稿では，書き言葉における英日翻訳のため，句単位の摂動を Positional Encoding に付与する正則化手法を提案した．実験結果から，書き言葉・話し言葉における英日翻訳それぞれにおいて若干の翻訳精度の改善が見られた．一方で，他の正則化手法と比べると精度の改善が小さいという問題点もある．今後の課題としては，他正則化手法との組み合わせの検証，他ランダムシードでの検証などが挙げられる．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **CoRR**, Vol. abs/1706.03762, , 2017.

[2] S. Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam M. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In **NIPS**, 2015.

[3] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, **Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain**, pp. 1019–1027, 2016.

[4] Sho Takase and Shun Kiyono. Rethinking perturbations in encoder-decoders for fast training. In **Proceedings of the 2021 Conference of the North American Chapter of**

the Association for Computational Linguistics: Human Language Technologies, pp. 5767–5780, Online, June 2021. Association for Computational Linguistics.

[5] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. **Computational Linguistics**, Vol. 19, No. 2, pp. 263–311, 1993.

[6] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[7] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)**, pp. 2204–2208, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[9] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. Overview of the iwslt 2017 evaluation campaign. 2017.

[10] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of The 12th Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[11] Philipp Koehn, Hieu T. Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **ACL**, 2007.

[12] Taku Kudo and John Richardson. 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, November 2018. Association for Computational Linguistics.

[13] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of NAACL-HLT 2019: Demonstrations**, 2019.

[14] Taku Kudo. Mecab : Yet another part-of-speech and morphological analyzer. **http://mecab.sourceforge.net/**, 2005.

[15] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In **Proceedings of the ACL Interactive Poster and Demonstration Sessions**, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics.

[17] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.