Positional Encoding への摂動付与による長さ制御を用いた非自己回帰型機械翻訳のための知識蒸留

岡 佑依　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{oka.yui.ov2, sudoh, s-nakamura}@is.naist.jp

1 はじめに

近年，ニューラルネットを用いた機械翻訳（NMT）の手法が多く考案されている．これらの手法において，注視機構を用いた自己回帰型エンコーダ・デコーダモデルは自然性があり高い精度の翻訳結果を残している．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．我々の研究 [2] では，高瀬ら [3] が提案した長さ制約付き Positional Encoding を機械翻訳に適用し，摂動を加えることによって翻訳精度を改善した．さらに通常の Transformer と比べ長く出力することを可能とした．また，推論時に与えられる長さが参照訳の長さと一致する場合，大きく翻訳精度を改善した．

非自己回帰型エンコーダ・デコーダモデルは，自己回帰型モデルと比べ高速な翻訳を可能とするが，翻訳精度は低く，通常の自己回帰型モデルと比べてさらに短い文を生成する傾向にある．

本研究では，長さ制約つき Positional Encoding への摂動を用いて知識蒸留をすることで非自己回帰型モデルの翻訳精度の改善を試みた．さらに，既存の非自己回帰型モデルの一つである Levenshtein Transformer[4] に長さ制約付き Positional Encoding への摂動を適用することで長い文を生成し，訳抜けを改善することを試みた．

2 関連研究

2.1 Positional Encoding による出力長制御

Positional Encoding（以下，PE）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，その位置に対応した絶対的な値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 PE は以下のようになる．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．

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

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

高瀬ら [3] は デコーダ側の PE の式に所望の出力長の値を組み込んだ．これにより，文生成時に所望の出力長までの残りのトークン数を考慮することが可能である．提案された式は終端までの比率に応じた $LRPE$(length-ratio positional encoding)，終端までの差に応じた $LDPE$(length-difference positional encoding) の 2 種類がある，例えば LDPE は以下のように表される．

$$LDPE_{(pos,len,2i)} = sin\left(\frac{len-pos}{10000^{\frac{2i}{d}}}\right) \tag{3}$$

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

len は所望の出力長を表す．$LRPE$，$LDPE$ の値はベースラインの Transformer の PE と同じように各埋め込み表現に足し合わせる．また，エンコーダ側にはベースラインの Transformer と同様の PE の式が適用される．

2.2 長さ制約付き Positional Encoding への摂動

機械翻訳において，LRPE，LDPE を用いると，出力長制御性と引き換えに翻訳精度が大きく落ちることがわかっている．これを改善するため，長さ制約付き PE への摂動 (Perturbation into Length-aware Positional Encoding) を，我々は [2] で提案した．LDPE に摂動を与える場合，以下の式で定義される．

$$perLDPE_{(pos,len,2i)} = sin\left(\frac{len - pos + perturbation}{10000^{\frac{2i}{d}}}\right) \tag{5}$$

$$perLDPE_{(pos,len,2i+1)} = cos\left(\frac{len - pos + perturbation}{10000^{\frac{2i}{d}}}\right) \tag{6}$$

摂動 *perturbation* は学習時，ある特定の整数の範囲から一様分布に基づいて選択され，文単位ごとに足し合わされる．我々は [2] において，摂動範囲 $[-2, 2]$，$[-4, 4]$ の負の値も含める範囲で実験を行った．さらに，大規模言語モデル BERT を用いて推論時の出力長の予測を行うことで翻訳精度を改善した．また，参照訳と同じ長さを LDPE，LRPE に入力するとき，翻訳精度は大きく改善することが明らかになっている．

2.3 Levenshtein Transformer

非自己回帰型ニューラル機械翻訳モデル (以下，NAT) は，各ステップの出力トークンを次のステップの入力に利用する自己回帰 (autoregression) に基づいて翻訳するモデル (以下，AT) とは異なり，自己回帰を用いず文内のトークンを並列に予測するプロセスを繰り返して翻訳を行う [5]．これによって並列計算が可能になり，非常に高速な翻訳を実現する．Jiatao Gu ら [4] が提案した，NAT モデルの一つである Levenshtein Transformer は，空トークン <PLH>を挿入する (placeholder) 機構，空トークンに単語を挿入する (insert) 機構，不必要なトークンを削除する (deletion) 機構の 3 つのデコーダを持つ．デコーダでは全て Position Embedding が用いられており，各 embedding はエンコーダ，デコーダで共有される．翻訳時に，これら 3 つの処理を繰り返し行うことで高速かつ Transformer に近い翻訳文生成を可能にした．

2.4 知識蒸留

上述した NAT モデルは高速な翻訳を実現するが，翻訳精度は AT と比べ大きく劣る．これを改善するため，NAT モデルを訓練する際，既存の AT モデルを用いて知識蒸留した訓練データを用いる．この時，AT モデルは教師モデル (通常，Transformer が用いられる)，NAT モデルは生徒モデルとなる．これによって，高品質な AT モデルの出力を NAT モデルに模倣させることが可能であり，NAT モデルの翻訳精度は教師モデルである AT モデルに依存する [6]．

3 提案手法

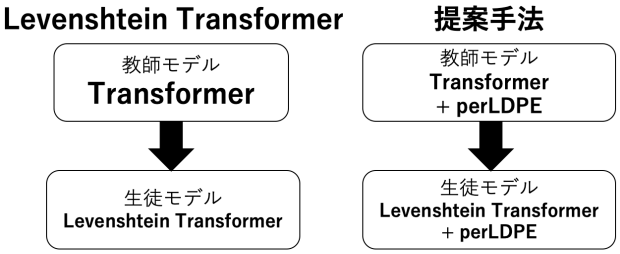

図 1 Levenshtein Transformer と提案手法の比較

通常の知識蒸留を用いた Levenshtein Transformer と提案手法の流れを図 1 に示す．

3.1 長さ制約つき Positional Encoding への摂動を用いた知識蒸留

本研究では，この知識蒸留で教師モデルとして用いる Transformer に長さ制約付き PE への摂動を適用することで，生徒モデルである NAT モデルの翻訳精度を改善する手法を提案する．知識蒸留における訓練データの翻訳時において出力すべき長さが既知 (正解参照訳長が存在する) であるため，翻訳時に長さ制約付き PE に入力する長さは参照訳の長さを入力する．これにより教師モデルである AT モデルが出力する翻訳文の精度を改善することで，生徒モデルである NAT モデルの翻訳精度の改善が期待される．

3.2 長さ制約つき Positional Encoding への摂動を用いた Levenshtein Transformer

さらに，Levenshtein Transformer の空トークン <PLH>を挿入する機構 (placeholder) に長さ制約付き PE への摂動を適用する．トークン長を制御する機構は placeholder のみであることから，挿入機構 (insert)，削除機構 (deletion) には通常の position embedding を適用した．また，エンコーダ側も position embedding を適用した．先行研究では $[-2, 2]$ のように負の値も PE への摂動範囲に加えるが，本研究では $[0, 2]$ のように正の値のみを摂動範囲として適用する．これは，モデルができるだけ長い文を生成するように学習するためである．また，Levenshtein Transformer ではエンコーダ・デコーダ間の embedding を共有して学習するが，提案手法では共有しないとする．そして翻訳時では，[2] と同様，学習済み言語モデルで予測した長さを入力する場合と，[7] と同様に，原言語文の長さを入力する場合を検証する．

4 実験

4.1 実験設定

提案手法による知識蒸留と Levenshtein Transformer の性能を調べることを目的とし，実験を行った．本研究では，英日・英独翻訳をタスクとした．データセットには，英日翻訳には対訳コーパス ASPEC[8]，英独翻訳には WMT14[9] 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．英語及び日本語の入出力はサブワードとし，Sentencepiece[10] 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．WMT14 は，Stanford NLP group[1] で配布された前処理済みのデータセットを用いた．学習データは 440 万文対で構成され，各文は 50 語以内で構成される．開発データには 3,000 文対で構成される newstest2013，テストデータには 2,737 文対で構成される newstest2014 を用いた．トークナイズ方法は ASPEC と同様であり，語彙サイズのみ 32,000 とした．実装には fairseq[11] を用いた．ハイパーパラメータは全てにおいて[2] と同じにした．

本研究では *len* に出力文のトークン長を与える．このトークン長は Sentencepiece でトークナイズされたときのものである．教師モデルのベースラインには，sinusoidal Positional Encoding を用いた base Trasnformer を，生徒モデルのベースラインには，Position Embedding を全デコーダに適用した Levenshtein Transformer を用いた．提案手法では，知識蒸留に用いる教師モデル Transformer に適用する長さ制約付き PE は LDPE を用い，学習時に与える摂動範囲は英日翻訳の時 $[-4, 4]$，英独翻訳の時 $[-6, 6]$ とした．また，生徒モデル Levenshtein Transformer に適用する長さ制約付き PE は LDPE を用い，学習時に与える摂動範囲は $[0, 2]$ のみとした．

翻訳時，英日翻訳では [2] と同様に，BERT[12] のエンコーダーの最後の層にある [CLS] ベクトルの出力を回帰問題として出力長を予測した．英独翻訳では [7] と同様に原言語文長の長さを LDPE の入力とした．さらに，両方の翻訳において参照訳の長さを入れた場合も比較した．

1) https://nlp.stanford.edu/projects/nmt/
2) https://github.com/pytorch/fairseq/tree/master/examples/nonautoregressive_translation

表 1 知識蒸留の教師モデルとして用いた AT モデルの比較 (Train データ)

Model	ASPEC	WMT14
	BLEU	
Transformer (baseline)	31.7	30.1
提案手法	**32.4**	**31.2**

表 2 日英対訳コーパス ASPEC における NAT モデルの翻訳精度

Model	入力長	BLEU	LR
Transformer		37.1	0.948
Transformer による知識蒸留			
LevT (baseline)		34.0	0.912
LevT + $[0, 2]$	予測長	**34.1**	0.920
LevT + $[0, 2]$	参照訳長	34.6	0.975
提案手法による知識蒸留			
LevT (baseline)		**34.3**	0.900
LevT + $[0, 2]$	予測長	**34.2**	0.922
LevT + $[0, 2]$	参照訳長	34.3	0.989

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[13] とサブワード単位の Length ratio を用い，sacreBLEU[14] で計算した．また，知識蒸留の教師モデルの評価にも [6] と同様，BLEU を用いる．

4.2 実験結果

表 1 に知識蒸留の教師モデルとして用いた AT モデルの学習データの翻訳精度を示す．提案手法の知識蒸留法が既存手法よりも文単位の知識蒸留において優れていることがわかる．

表 2, 表 3 に各コーパスにおける生徒モデルの翻訳精度を示す．太文字はベースラインである Levenshtein Transformer より BLEU が向上したもの，下線部は参照訳長が既知の時，すなわち正解長を入力した時 Levenshtein Transformer より BLEU が向上したものを示す．

英日翻訳では，提案手法である知識蒸留を用いた場合，全ての生徒モデルにおいて翻訳精度が改善した．通常の Levenshtein Transformer で比べると，0.3 ポイントの BLEU 値の改善が見られた．さらに，提案する知識蒸留法を用いない場合でも，提案手法である Levenshtein Transformer に LDPE への摂動を適用した時，翻訳精度は改善することがわかった．参照訳長を用いた場合，0.6 ポイントの BLEU 値の改善 (34.6) が見られ，入力長によっては最大 0.6 ポイントの BLEU 値の改善が見込めることがわかった．さらに，学習済みモデルで長さを予測し入力することで LR が上がったことも BLEU 値の改善に繋がっ

表 3 英独対訳コーパス WMT14 における NAT モデルの翻訳精度

Model	入力長	BLEU	LR
Transformer		30.1	0.960
Transformer による知識蒸留			
LevT (baseline)		28.7	0.905
LevT + [0, 2]	原言語文長	26.9	0.955
LevT + [0, 2]	参照訳長	31.0	0.962
提案手法による知識蒸留			
LevT (baseline)		27.2	0.878
LevT + [0, 2]	原言語文長	25.9	0.933
LevT + [0, 2]	参照訳長	29.7	0.940

たことがわかる．しかしながら，全ての生徒モデルにおいて，BLEU 値は教師モデルとして用いた AT モデルである通常の Transformer に劣る結果となった．また，BERT による長さ予測の精度は，参照訳長との平均トークン誤差が 3.0，トークン誤差分散が 19.92 であった．原言語文長と参照訳長との平均トークン誤差は 6.54，トークン誤差分散は 72.45 であった．これは，英日翻訳では，原言語文長ではなく，BERT によって予測された長さの方が参照訳長に近いことを示している．

英独翻訳では，英日翻訳のような結果は見られなかった．提案手法である知識蒸留を用いると，同じ Levenshtein Transformer においても BLEU の向上は見られなかった．さらに，提案する NAT モデルにおいて，原言語文長を用いた場合，BLEU は下がった．しかしながら，提案する NAT モデルに参照訳長を用いた時，通常の Transformer より 0.9 ポイント BLEU 値の改善する (31.0) ことがわかった．

4.3 摂動範囲による翻訳精度の推移

提案した NAT モデルの摂動範囲を変えることで，翻訳精度が改善するのかを英日・英独翻訳それぞれで検証した．知識蒸留に用いた教師モデルは表 1 のベースライン Transformer，実験設定は 4.1 と同じである．検証した摂動範囲は [0,2] に加え，[0,4]，[0,6] である．表 4 に，実験結果を示す．英日・英独翻訳両方において，摂動範囲が大きくなるにつれて，Length ratio が下がっていることがわかった．英日翻訳において，摂動範囲 [0,6] の時，摂動範囲 [0,2] を用いた時と比べて 0.1 の BLEU の改善が見られたが，それ以外で改善は見られなかった．また，英独翻訳では摂動範囲を大きくしてもベースラインの Levenshtein Transformer と比べ翻訳精度の改善は見られなかった．

表 4 摂動範囲ごとの NAT モデルの翻訳精度

Model	入力長	BLEU	LR
ASPEC 英日翻訳			
LevT + [0, 2]	予測長	**34.1**	0.920
LevT + [0, 2]	参照訳長	34.6	0.975
LevT + [0, 4]	予測長	33.2	0.900
LevT + [0, 4]	参照訳長	33.9	0.940
LevT + [0, 6]	予測長	**34.2**	0.919
LevT + [0, 6]	参照訳長	34.5	0.957
WMT14 英独翻訳			
LevT + [0, 2]	原言語文長	26.9	0.955
LevT + [0, 2]	参照訳長	31.0	0.962
LevT + [0, 4]	原言語文長	25.1	0.955
LevT + [0, 4]	参照訳長	28.8	0.956
LevT + [0, 6]	原言語文長	26.0	0.935
LevT + [0, 6]	参照訳長	30.0	0.938

4.4 考察

実験結果より，提案した知識蒸留は英日翻訳における生徒モデルの翻訳結果の向上に有効であると考えられる．また，生徒モデルである NAT モデルに長さ制約付き PE への摂動を用いた提案手法において，参照訳長を用いた場合より BERT による予測長を用いた場合において翻訳精度が下がった原因は長さ予測の精度にあると考えられる．これは BP の値を見比べたとき，参照訳長を用いた場合 BP の値が大きく変化していることからもわかる．英独翻訳では提案した知識蒸留と NAT モデル両方において有効性は見られなかった．しかしながら，提案した NAT モデルに参照訳長を入力した時，翻訳精度は大きく改善することから，入力する長さを改善することで翻訳精度が改善すると考えられる．

5 おわりに

本稿では，長さ制約付き PE への摂動を知識蒸留に用いる教師モデル，そして生徒モデルである Levenshtein Transformer へ適用することを提案した．結果として，最大 0.3 ポイントの BLEU 値の向上が見られた．英日翻訳において．提案した知識蒸留と NAT モデルは有効であることがわかった．英独翻訳では出力長予測精度の影響で改善が見られず，より高精度な出力長予測が今後の検討課題である．

謝辞

本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものである．

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]Yui Oka, Katsuki Chousa, Katsuhito Sudoh, and Satoshi Nakamura. Incorporating noisy length constraints into transformer with length-aware positional encodings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3580–3585, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[3]Sho Takase and Naoaki Okazaki. Positional encoding to control output sequence length. 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. 3999–4004, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4]Jiatao Gu, Changhan Wang, and Junbo Zhao. Levenshtein transformer. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 32, pp. 11181–11191. Curran Associates, Inc., 2019.

[5]Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*, 2018.

[6]Chunting Zhou, Jiatao Gu, and Graham Neubig. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*, 2020.

[7]Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. Controlling the Output Length of Neural Machine Translation. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT 2019)*, October 2019.

[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]Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[10]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.

[11]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.

[12]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, Vol. abs/1810.04805, , 2018.

[13]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.

[14]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.