Non-Autoregressive Translation モデルにおける 事前並び替え適用手法の検討

瓦 祐希
大阪大学大学院情報科学研究科
kawara.yuki@ist.osaka-u.ac.jp

Chenhui Chu
京都大学大学院情報学研究科
chu@i.kyoto-u.ac.jp

荒瀬 由紀
大阪大学大学院情報科学研究科
arase@ist.osaka-u.ac.jp

1 はじめに

統計的機械翻訳 (SMT) において、原言語と目的言語間の語順の相違は翻訳精度に大きな影響を与えることが知られている [8]。この問題を解決する手法として、翻訳器に入力する前に原言語文を目的言語文の語順に近づくように並び替える事前並び替え手法が提案されてきた [8, 4]。特に英語・日本語間のように語順の大きく異なる言語対において、事前並び替えを行うことで SMT での翻訳精度は大きく改善した [8, 4]。しかし、Autoregressive Translation (AT) では、事前並び替えを行なった文をそのまま入力として使用すると、事前並び替えを行わなかった文を入力として使用した場合と比較して翻訳精度が低下することが報告されている [1, 4, 5]。

近年では、推論時に翻訳文の全単語を一度に出力する Non-Autoregressive Translation (NAT) が提案され、盛んに研究されている [2, 12, 10]。NAT モデルによって翻訳にかかる速度は上昇した。しかし各単語を独立に出力するため、前に出力した単語を考慮した翻訳が出来ず、 AT モデルと比較して翻訳精度が低下するという問題点がある。また NAT モデルで翻訳を行う際は、デコーダの入力としてエンコーダの出力をそのままの順序で使用するため、語順の相違を考慮することが出来ない。

本研究では、NAT モデルにおける事前並び替えの効果的な適用手法の検討を行う。事前並び替えを行なった原言語文をエンコーダに入力する手法、エンコーダの出力を並び替える手法、エンコーダの入力で事前並び替えの情報を使用する方法で検証を行なった。ASPEC コーパス [9] を使用した英日翻訳実験の結果、NAT モデルでは AT モデルと異なり，事前並び替えを行なった文をそのままエンコーダに入力することで翻訳精度が向上することが明らかとなった。

2 関連研究

代表的な AT モデルである Transformer [13] は、self-attention を使用して翻訳を行うモデルである。このモデルは推論の際に以前に出力した単語に基づいて次の単語の予測を行うため、デコード処理の並列化ができず、翻訳文の長さに比例した時間がかかる。そこで Gu ら [2] は翻訳文を一度に出力する NAT モデルを提案した。NAT モデルは AT モデルと比較すると翻訳速度が大幅に向上したが、各単語を独立に出力するため周辺の単語に基づいた翻訳が出来ず、翻訳精度が低下してしまうといった問題がある。特に英語・日本語のように語順の大きく異なる言語対の翻訳では、語順の考慮が困難である。

Shu ら [12] は潜在変数モデルによる NAT モデル (LaNMT) を提案した。目的言語文の不確かさを低次元の潜在変数によってことでモデリングし、その潜在変数に基づいて翻訳文の出力を行うことでより精度の高い翻訳を可能にしており、最高性能を達成した翻訳手法の 一つである。しかし、エンコーダの出力をそのままの順序で使用してデコーダの入力として使用しているため、並び替えは考慮していない。本研究では LaNMT に対し、効果的に事前並び替えを適用する手法を検討する。

Ran ら [10] は、Transformer により原言語文の語順を並び替えた隠れ変数を NAT のデコーダに入力することで翻訳精度が向上すると報告している。しかし本手法を LaNMT に直接適用することはできず、また Transformer による事前並び替えは、AT モデル

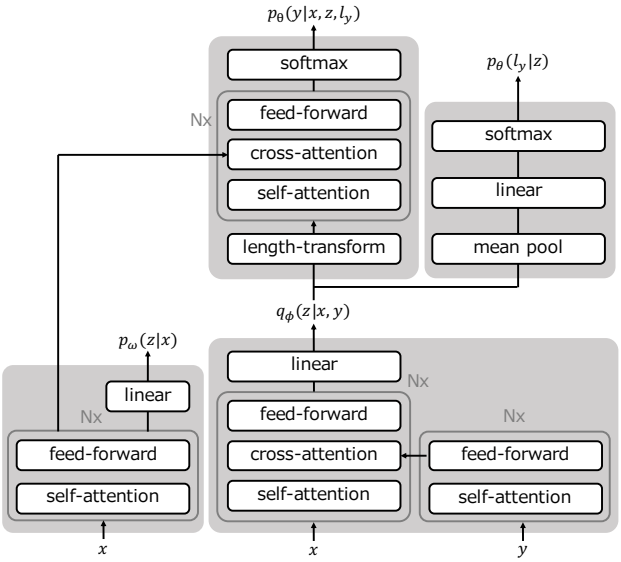

図 1　LaNMT モデルの全体図。

と同様、文の長さに比例した時間がかかるという課題がある。

3　LaNMT における事前並び替えの適用

3.1　前提知識：LaNMT

LaNMT モデルの全体図を図 1 に載せる。LaNMT は大きく四つのサブモデルからなり、原言語文である x から潜在変数 z を予測するモデル、x と目的言語文 y から z を予測するモデル、z から目的言語文の長さ l_y を予測するモデル、x、z、l_y から y を予測するモデルである。学習は、以下の式 (1) で示される変分下限を最大化することを目的に学習を行う。

$$\begin{aligned}&\mathcal{L}(\omega,\phi,\theta)\\&=\mathbb{E}_{z\sim q_\phi}[\sum_{i=1}^{|y|}\log p_\theta(y_i|x,z,l_y)+\log p_\theta(l_y|z)]\\&\quad-\sum_{k=1}^{|x|}\mathrm{KL}[q_\phi(z_k|x,y)||p_\omega(z_k|x)]\end{aligned}\tag{1}$$

ω はエンコーダ側の、x が分かっている時の事前分布をモデル化するパラメータ、ϕ はエンコーダ側の、x と y が分かっている時の事後分布をモデル化するパラメータ、θ はデコーダ側のパラメータを表す。事前分布、事後分布は予測された潜在変数 z から linear 層によって計算される。推論時は y は存在しないため、まず原言語文 x から z の計算を行い、それをデコーダの入力として使用して翻訳文 y' の予測を行う。それから出力した y' を使用して z の再計算を行い、新たな翻訳文の出力を行う。これを決められた回数行い、最後の出力を最終的な翻訳文とする。

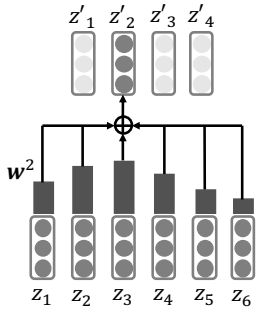

図 2　Length-transform の例。この例では、エンコーダの出力は 6 単語、デコーダへの入力は 4 単語としている。

原言語文の長さと目的言語文の長さは異なることがあるため、デコーダの入力はエンコーダの出力から計算を行う必要がある。LaNMT では length-transform という機構によってエンコーダの出力からデコーダの入力を計算している。図 2 に例を示す。エンコーダの出力 z からデコーダの入力 z' は、以下の式に従って各位置を基準に計算された重みを使用し z の重み付き和によって計算される。

$$z'_j=\sum_{k=1}^{|x|}w_k^j z_k\tag{2}$$

$$w_k^j=\frac{\exp(a_k^j)}{\sum_{k'=1}^{|x|}\exp(a_{k'}^j)}$$

$$a_k^j=-\frac{1}{2\sigma^2}(k-\frac{|x|}{l_y}j)^2\tag{3}$$

j はデコーダの入力のインデックス、k はエンコーダの出力のインデックス、a_k^j はデコーダの入力に対するエンコーダの出力のアテンション、w_k^j はデコーダの入力に対するエンコーダの出力の重みを表す。各 z' に対する重みは式 (3) によってそれぞれの位置を中心とした正規分布によって計算される。

3.2　事前並び替えの適用

本研究では、(1) 事前並び替えを行なった原言語文をエンコーダに入力するシンプルな手法、(2) 潜在変数 z を並び替える手法、(3) 事前並び替えのインデックスの position encoding を足し合わせる手法によって事前並び替えの利用の検証を行う。以降では (2) および (3) の手法について説明する。

(2) 潜在変数 z を並び替える手法　LaNMT では原言語文の潜在変数 z の順番をそのまま使用している

		BLEU	RIBES
AT モデル	ベースライン	35.53	83.68
	BTG	31.20$^{\downarrow}$	81.15$^{\downarrow}$
	Gold-standard	44.84$^{\uparrow}$	89.85$^{\uparrow}$
NAT モデル	ベースライン	22.14	79.42
	BTG	24.01$^{\uparrow}$	77.92$^{\downarrow}$
	Gold-standard	32.39$^{\uparrow}$	86.26$^{\uparrow}$
	z の並び替え (BTG)	13.49$^{\downarrow}$	78.21$^{\downarrow}$
	z の並び替え (Gold-standard)	15.71$^{\downarrow}$	81.55$^{\uparrow}$
	absolute position encoding (BTG)	15.49$^{\downarrow}$	78.18$^{\downarrow}$
	absolute position encoding (Gold-standard)	21.05$^{\downarrow}$	83.75$^{\uparrow}$

表 1 英日対における翻訳結果。それぞれのモデルのベースラインと $p < 0.05$ で有意差があり、精度が向上したものを↑で、精度が低下したものを↓で示す。

			BLEU	RIBES
NAT モデル	ベースライン	w/o knowledge distillation	22.14	79.42
		w/ knowledge distillation	21.93	79.64
	BTG	w/o knowledge distillation	24.01$^{\uparrow}$	77.92$^{\downarrow}$
		w/ knowledge distillation	21.87	78.01$^{\downarrow}$
	Gold-standard	w/o knowledge distillation	32.39$^{\uparrow}$	86.26$^{\uparrow}$
		w/ knowledge distillation	28.17$^{\uparrow}$	86.25$^{\uparrow}$

表 2 英日対における翻訳結果。それぞれのモデルのベースラインと $p < 0.05$ で有意差があり、精度が向上したものを↑で、精度が低下したものを↓で示す。“w/o knowledge distillation” と “w/ knowledge distillation” の行はそれぞれ knowledge distillation をせず訓練した結果と knowledge distillation して訓練した結果を表す。

が、そのままでは語順の相違を考慮することが出来ない。そこで潜在変数 z の変換の際に以下の式 (4) に示すように並び替えた後のインデックスを使用することで語順の相違を考慮した翻訳が可能になると期待される。

$$z'_j = \sum_{k=1}^{|x|} w_k^j z_{r_k} \tag{4}$$

r_k は原言語文の k 番目の単語の、並び替えた後のインデックスである。z_k を並び替えてから重み付き和を計算することで、対応した目的言語文の位置に近い z_{r_k} の重みが大きくなり、並び替えた後の文における周辺単語をより考慮した z' が計算できる。

(3) 事前並び替えのインデックスの position encoding を足し合わせる手法 ここでは事前並び替えを利用するため、事前並び替えのインデックスの position encoding を足し合わせる手法を設計する。Kawara ら [5] は AT モデルにおいて事前並び替えのインデックスによる position encoding を足し合わせることで翻訳精度が向上したことを報告しているが、NAT モデルにおける効果は定かではない。NAT モデルにおいて事前並び替えのインデックスによる position encoding を足し合わせることで、エンコーダで語順を考慮した潜在変数の計算を行うことが可能になる。

4 翻訳実験

4.1 実験設定

ASPEC コーパス [9] を使用して英日翻訳実験を行なった。ASPEC コーパスに含まれている訓練データは 300 万分、開発データは 1,790 文、テストデータは 1,812 文である。翻訳器の学習には上位 200 万文のうち、原言語、目的言語がともに 3 単語以上 50 単語以下で、かつ単語数の比が 9 未満である約 180 万文を使用した。

英語文の単語分割および品詞タグ付けには Stanza[1] を使用した。日本語の形態素解析は Juman[2] で行なった。

事前並び替え手法として BTG [8] を使用した。事前並び替えモデルは、訓練データのうち上位 50 万

1） https://stanfordnlp.github.io/stanza/
2） http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

文からランダムにサンプリングした 10 万文を使用して 20 イテレーションの訓練を行なった。単語クラスタ数は 256 に設定し、単語アライメントは MGIZA[3] を使用して計算した。

事前並び替えの適用手法 (2) および (3) は Shu らによる LaNMT の公開モデル[4]上に実装した。訓練は 50000 ステップ行なった。分散表現の次元数および学習率は Shu らの設定に従った。また Shu らと同様、翻訳は 1 度 refinement した。翻訳結果において同じ単語が 2 回以上繰り返されている場合は 2 回目以降を除去した。

評価指標として BLEU [6] と RIBES [3] を使用した。有意差検定はブートストラップ法 [7] に従って行なった。

4.2 実験結果

実験結果を表 1 に示す。ここで Gold-standard は単語アライメントの計算をした後、アライメントの交差がなくなるように並び替えたものを指す。AT モデルでは先行研究同様、事前並び替えをそのまま使用するとベースラインと比較して BLEU および RIBES が低下した (それぞれ −4.33 ポイント、−2.53 ポイント)。一方、NAT モデルでは事前並び替えをそのまま使用するとベースラインと比較して BLEU が向上した (+1.87 ポイント) 。Gold-standard を使用して学習を行なった翻訳実験では、AT モデルにおいても NAT モデルにおいても大幅に翻訳精度が向上した (AT モデルで BLEU が +9.31 ポイント、RIBES が +6.17 ポイント、NAT モデルで BLEU が +10.25 ポイント、RIBES が +6.84 ポイント)。これは、翻訳の学習において並び替えの情報が有用であることを示唆している。

z を並び替える方法では、Gold-standard の順に並び替える方法においても BTG の順に並び替える方法においてもベースラインより BLEU が低下した (Gold-standard で −6.43 ポイント、BTG で −8.65 ポイント)。また、position encoding に並び替えのインデックスを使用する方法でも BLEU は低下した (Gold-standard で −1.09 ポイント、BTG で −6.65 ポイント)。翻訳精度が低下した原因として、z を並び替えることで潜在変数同士の依存関係が崩れてしまったためであると考えられる。また、position encoding を足し合わせる手法では、z の変換では原言語文の語順をそのまま使用しているため、並び替え後のインデックスが離れている潜在変数の重みが大きくなる場合があるためであると考えられる。

4.3 Knowledge Distillation に対する影響

NAT の訓練では、訓練データの参照訳をそのまま使用するのではなく、AT モデルで翻訳した出力を使用した訓練を行う Knowledge Distillation により、翻訳性能が大幅に向上することが知られている [2, 12]。表 2 に、事前並び替えを適用した各モデルにおいて、Knowledge Distillation をし訓練を行なった結果を示す。ここで、BTG で AT モデルの訓練を行うとベースラインよりも精度が低下し結果として NAT モデルの訓練結果も低下すると考えられるため、Knowledge Distillation は全て Gold-standard な並び替えを行なった文によって AT モデルを訓練し翻訳したものを使用している。先行研究と異なり、Gold-standard による Knowledge Distillation で訓練を行うと、 全てのモデルで Knowledge Distillation しなかったものよりも精度が低下した。今回の実験では Byte Pair Encoding (BPE) [11] を適用しておらず未知語が出力されるため、その影響によって Knowledge Distillation による翻訳結果が低下してしまったと考えられる。Knowledge Distillation における BPE の有無の影響調査は今後の課題である。

5 結論

本研究では LaNMT モデルにおける事前並び替えの適用手法について調査を行なった。NAT モデルでは AT モデルの場合と異なり、事前並び替えを行なった文をそのまま入力することで翻訳精度が向上することが明らかとなった。一方、潜在変数を並び替える手法や事前並び替えのインデックスの position encoding を足し合わせる手法では翻訳精度が低下した。今後の課題として、よりうまく語順の情報を活用した NAT モデルの検討が挙げられる。

謝辞 本研究は、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および科研費#19K20343 の助成を受けたものである。

参考文献

[1] Jinhua Du and Andy Way. Pre-reordering for neural machine translation: Helpful or harmful? *The Prague Bulletin of Mathematical Linguistics (PBML)*, Vol. 108, pp. 171–182, June 2017.

3） https://github.com/moses-smt/mgiza
4） https://github.com/zomux/lanmt

[2] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *Proc. of the ICLR*, Vancouver, Canada, Apr-May 2018.

[3] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. of the EMNLP*, pp. 944–952, Cambridge, USA, October 2010.

[4] Yuki Kawara, Chenhui Chu, and Yuki Arase. Recursive neural network based preordering for english-to-japanese machine translation. In *Proc. of the ACL-SRW*, pp. 21–27, July 2018.

[5] Yuki Kawara, Chenhui Chu, and Yuki Arase. Preordering encoding on transformer for translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[6] Papineni Kishore, Roukos Salim, Ward Todd, and Zhu Wei-Jing. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the ACL*, pp. 311–318, Philadelphia, USA, July 2002.

[7] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proc. of the EMNLP*, pp. 388–395, Barcelona, Spain, July 2004.

[8] Tetsuji Nakagawa. Efficient Top-Down BTG Parsing for Machine Translation Preordering. In *Proc. of ACL-IJCNLP*, pp. 208–218, Beijing, China, July 2015.

[9] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proc. of the LREC*, pp. 2204–2208, Portorož, Slovenia, May 2016.

[10] Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. Guiding non-autoregressive neural machine translation decoding with reordering information. *AAAI*, February 2021.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proc. of the ACL*, pp. 1715–1725, Berlin, Germany, 2016.

[12] Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior. *AAAI*, February 2020.

[13] 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 *NeurIPS*, pp. 5998–6008. Long Beach, USA, December 2017.