事前学習モデルによる分割統治ニューラル機械翻訳

石川 隆太　加納 保昌　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{ishikawa.ryuta.il7,kano.yasumasa.kw4,sudoh,s-nakamura}@is.naist.jp

概要

ニューラル機械翻訳におけるハルシネーションや繰り返し、訳抜けなどの課題に対し、文をセグメントに分割して翻訳し、並べ替えて繋げるニューラル機械翻訳における分割統治的手法 [1] が提案されている。本研究では、[1] で課題とされていた、セグメント翻訳と、セグメントの並べ替えと編集の精度の２つに対し、事前学習モデルの mBART を用いることに加え、セグメント翻訳のモデルを節単位の擬似対訳データでファインチューニングすることで、セグメント翻訳と、並べ替えと編集の精度の改善を試みた。実験では提案手法による BLEU の改善は認められなかったが、過剰な長さの訳出が大きく減少し、ハルシネーションや繰り返しを抑制できている傾向が確認できた。

1　はじめに

グローバル化に伴い翻訳の需要が高まっており、ニューラル機械翻訳 (NMT) が注目されている。NMT は質の高い訳文の生成が期待できる一方で、入力される文が長くなるとハルシネーションや訳抜け [2] などの問題が発生することがある。この課題に対し、統計的機械翻訳では、長文を短いセグメントに分割して翻訳し、並べ替えて繋げるという分割統治的手法 [3] が提案されている。また、[3] を参考にした、NMT における長文のための分割統治的手法 [1] も提案されている。本研究では、[1] において課題とされていたセグメント翻訳の精度と、翻訳されたセグメントの並べ替えと編集の精度改善に向け、セグメント翻訳と並べ替えと編集のモデルに事前学習モデルの mBART を用い、さらにセグメント翻訳のモデルを節単位の擬似対訳データでファインチューニングする手法の検討を行った。英日翻訳の実験において提案手法による BLEU の改善は認められなかったが、過剰な長さの訳出についてはベースラインから大きく削減できており、ハルシネーションや繰り返しを提案手法により抑制できている傾向が確認できた。

2　関連研究

NMT において、長文を分割し、分割された各セグメントを翻訳した後に前から順に結合し直す手法として Pouget-Abadie ら [4] の自動分割の手法がある。この手法は RNN を用いて、長文を分割して翻訳する際の最適な位置を予測し、モデルが翻訳を行いやすいように入力文を分割する手法である。しかし、この手法は英語からフランス語への翻訳のように、ソース言語とターゲット言語がほとんど同じ語順を持たない限り、長文分割後のセグメントの並べ替えの問題に対応することはできない。特に、英語と日本語のような語順が大きく異なる言語対において、結合後の翻訳結果は不自然な文になりやすい。この課題に対し、加納ら [1] は NMT における分割統治的手法を提案している。

2.1　分割統治型ニューラル機械翻訳

入力文を節単位でセグメントに分割し、分割されたセグメントを文単位の対訳コーパスで学習された翻訳モデルを用いて翻訳し、翻訳された各セグメントを、セグメント同士の関係性を表すトークンを用いて繋げ、翻訳モデルとは別のニューラルネットワークモデルに入力する。それによって、セグメントの並べ替えと編集を行い、自然な訳文の生成を実現している。ニューラル機械翻訳の分割統治的手法において、文分割後のセグメント翻訳の精度と翻訳されたセグメントの並べ替えと編集の精度は課題とされており、これら２つの精度は最終的な翻訳結果に大きく影響すると考えられる。そこで、本研究ではセグメント翻訳と翻訳後のセグメントの並べ替えと編集の精度を向上させるため、加納らの手法を基に次の２つを実施した。

2.2 mBART を用いた分割統治型ニューラル機械翻訳

加納らの手法においてはセグメント翻訳と翻訳されたセグメントの並べ替えと編集のモデルに Transformer[5] を利用しているのに対し、本研究では事前学習モデルの BART[6] を多言語の大規模なモノリンガルコーパスで事前学習した mBART[7] を利用した。mBART のファインチューニングによる機械翻訳 [8] は、10 万から 1000 万文対規模の比較的少量の対訳コーパスで学習する場合に最も効果を発揮し、mBART と同じ Transformer のエンコーダ・デコーダモデルを同じ対訳コーパスでゼロから学習するよりも、顕著に高い性能を発揮することが知られている。そのため、事前学習モデルの mBART を利用することで少量のコーパスで比較的性能の良いベースラインの翻訳モデルを作成することができ、セグメント翻訳の精度の向上が期待できる。さらに mBART は入力文のトークンのマスクや削除、文の順序をシャッフルするなど人工的なノイズを加えた上で、元の文を復元する事前学習を行なっている。そのため、本研究における翻訳したセグメントの並べ替えと編集を行うタスクと mBART の事前学習方法は似ているため、相性が良く、並べ替えと編集のモデルにおいても mBART を利用することで、並べ替えと編集の精度の向上が期待できる。

2.3 節単位の擬似対訳データによる節翻訳モデルの作成

加納らの手法では文分割後の節の翻訳を文単位の対訳データで学習した翻訳モデルで実施していたが、節の翻訳において文単位のコーパスで学習された翻訳モデルが適しているとは限らない。そこで、本研究では節単位の擬似対訳データを作成し、mBART をセグメント翻訳用にファインチューニングすることで節翻訳モデルを作成した。これにより、セグメント翻訳の精度の向上を試みた。

3 提案手法

提案手法の概略を図 1 の (a) ベースラインと節翻訳モデルの作成、(b) 節翻訳モデルを用いた分割統治的を用いて示す。提案手法の手順の太字の部分と図 1 の太枠部分が [1] における分割統治的手法を基に新たに実施、または変更した部分である。

1. JParaCrawl コーパスからサンプリングした 300 万文対を学習データとして **mBART をファインチューニング**しベースラインとなる翻訳モデルを作成する。
2. コーパスの英語の文を構文解析し、接続詞で分割する。
3. 英文を分割してできた英語節をベースラインの翻訳モデルで翻訳し、擬似日本語節とし、**節単位の英日の擬似コーパスを作成**する。
4. 3で作成した節単位の擬似対訳データで mBART をファインチューニングし**節翻訳モデルを作成**する。
5. 1 で作成した学習データの英文を構文解析し接続詞で分割し、4 で作成した**節翻訳モデルで分割した節を翻訳**する。
6. 5 で翻訳した英語と翻訳結果の日本語のペアを”@”を 3 つ繋げた対訳関係を示す記号で結合し、分割した節同士を別の”|”を 3 つ繋げた文の区切りを表す記号で結合する。
7. 6 で作成した英語と日本語を特殊トークンで結合した文をソース、分割前の英語に対応する日本語の文をターゲットとして **mBART をファインチューニングし、並べ替えと編集のモデルを作成**する。

表 1: トークン長ごとの評価データの文数

トークン長	全て	1-20	21-40	41-60	61-
文数	30000	15122	11641	2517	720

4 実験

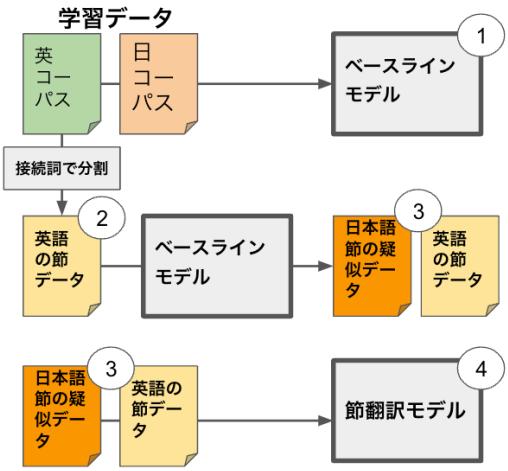
提案手法の有効性を検証するため、以下の実験を行った。

4.1 実験設定

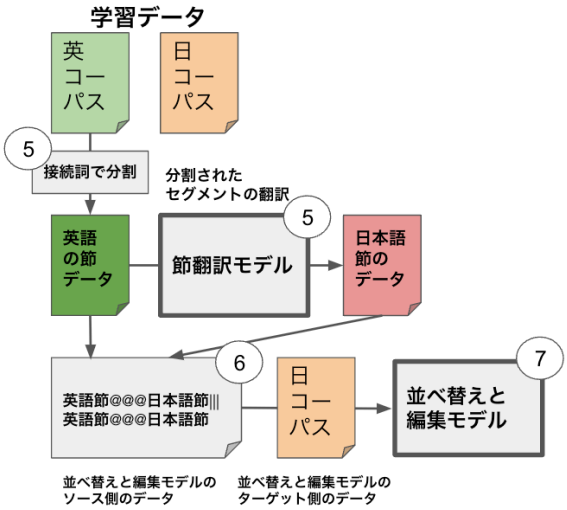
今回実施した実験の詳細設定を以下に示す。

モデルの学習と評価データの対訳コーパスとして、JParaCrawl.ver2 [9] を用いた。具体的に JParaCrawl.ver2 から、評価データを 3 万文サンプリングし、残りのデータから学習データを 300 万文サンプリングすることで、それぞれ評価データと学習データとした。評価データは sentencepiece でサブワードに分割した後に、英語のサブワードトークンを基準とし、トークン長ごとに 4 分割した評価データを作成した。分割したトークン長と対応する英日の文対の数は表 1 のようになった。評価データの BLEU は、MeCab [10] によって単語分割し、sacrebleu [11] を用いて算出した。

節翻訳モデルと並べ替えと編集のモデルには



(a) ベースラインと節翻訳モデルの作成



(b) 節翻訳モデルを用いた分割統治的手法

図 1: ニューラル機械翻訳における分割統治的手法の概略

表 2: 入力トークン長ごとの BLEU

モデル	全て	1-20	21-40	41-60	61-
ベースライン	**28.6**	**29.2**	**28.8**	**26.8**	**28.8**
提案手法	27.8	28.4	28.1	25.8	27.6

表 3: 入力トークン長ごとの BERTScore (F1)

モデル	全て	1-20	21-40	41-60	61-
ベースライン	**0.822**	**0.828**	**0.819**	**0.806**	**0.800**
提案手法	0.821	0.827	0.818	0.805	0.798

mBART を使用し、fairseq [12] で実装されたものを利用した。ベースラインと節翻訳モデル、並べ替えと編集のモデルの学習時のパラメータの設定は Github で公開されている英語とルーマニア語における mBART のファインチューニングのデフォルトの設定 [13] と同じ設定を用いた。

節翻訳モデルを作成する際の節単位の擬似対訳データは JParaCrawl.ver2 からサンプリングした英文 200 万文を構文解析し、接続詞で分割し、200 万文の内、分割された英語節のみを英語側の節単位のデータとし、それらを全てベースラインで日本語節に翻訳することで作成した。分割された英語節と翻訳によって得られた日本語節の対訳節の数は約は 190 万となった。構文解析には spaCy [14] を利用した。接続詞で分割される英文は次の例のようになる。

I don't like studying math, but / I like studying English.

提案手法の手順 3 と 5 における節対訳データの日本語節と並べ替えと編集のモデルの入力データの日本語節を生成する際のベースラインと節翻訳モデル

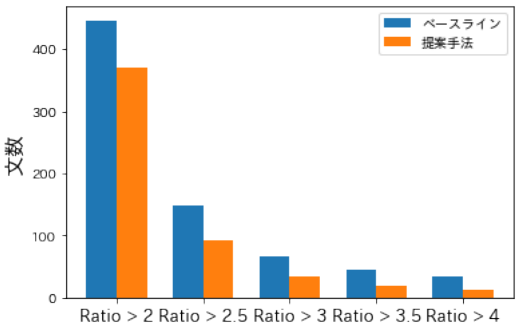


図 2: 参照文に対するモデルの出力文の長さ比が 2 以上となった数

のビームサーチのサイズはどちらも 4 とし、評価の際のビームサイズも 4 とした。

4.2 実験結果

表 2 はベースライン、提案手法の入力トークン長ごとの BLEU を示している。表 2 から分かるように、提案手法はベースラインに比べて全体としての BLEU が 0.8 ポイント低かった。表 3 は BERTScore[15] による評価結果で、表 2BLEU に比べて差が小さいと言えるが提案手法はベースラインを上回れなかった。

一方、ハルシネーションや繰り返しの影響を明らかにするため、参照文のモデルの出力文の長さ比が大きい事例数を調べた結果を図 2 に示す。特に長さ比が 3 以上となるような極端に長い翻訳出力がベースラインでは 67 文、提案手法では 33 文と半減し、ratio の平均±標準偏差はベースラインで 1.035 ± 0.360、提案手法で 1.017 ± 0.339 になる等、提

表 4: 提案手法によるハルシネーションの改善例

入力文
Original Evaluation of the Stratum Corneum Exfoliated by Tape Stripping Method (Part2)—Seasonal and age-related differences of exfoliated pattern ...
参照文
原著角層剥離パターンによる角層評価 (第 2 報)—季節変化と年代差について—…
ベースライン
テープストライプ法によるストラトムコネム浸潤のオリジナル評価 (第 2 部)—浸潤パターンの季節と年齢関係の差 ... テープストライプ法によるストラトムコネム浸潤の Original Evaluation of Stratum Corneum Exfoliated by Tape Stripping Method (Part2)—浸潤パターンの季節と年齢関係の差 ...
提案手法
テープストリップ法によるストラトム角膜剥離のオリジナル評価 (パート 2)—季節と年齢に関わる剥離パターンの違い

案手法により過剰な長さの訳出が抑制できていることが分かり、ハルシネーションや繰り返しが減少していることが示唆される (出力単語数の散布図を付録の図 3 に示す)。実際にベースラインと提案手法の翻訳結果を確認すると、ベースラインにおいてハルシネーションや繰り返しが起きていた文を提案手法では適切に翻訳できている例が確認できた（表 4 に例を示す)。

4.3 考察

提案手法の BLEU がベースラインより低下してしまった原因として、提案手法における節翻訳モデルの性能が低いことが考えられる。文単位のベースライン翻訳モデルと節翻訳モデルにより文単位の翻訳を行ったときのトークン長ごとの BLEU を表 5 に示す。節翻訳モデルは分割された節を適切に翻訳し、節翻訳の精度を向上させる目的で作成した。しかし、ベースラインと比較して BLEU が 3.9 ポイント低いことから、節翻訳の段階で質の低い訳出文を生成してしまい、並べ替えと編集のモデルの最終的な出力文の質が低下した可能性がある。

さらなる比較のため、提案手法における節翻訳モデルをベースラインのモデルに差し替え分割統治型翻訳を行った。その結果を表 6 に示す。表 5 から分かるように、提案手法における節翻訳モデルをベースラインのモデルに差し替えた場合、全体として BLEU が 2.6 ポイント低下した。これは、提案手法の節翻訳での利用において本稿の節翻訳モデルがベースラインよりも有効であったことを示唆している。表 5 で示した通り BLEU が 3.9 ポイント低いにもかかわらずこのような結果になった原因として、ベースラインは文単位のコーパスで学習されているため、分割された節を翻訳する際に分割によって失われた節の前後の文を勝手に補おうとしたことが考えられる。その具体例を表 7 に示す。こうした事象により、節翻訳にベースラインモデルを利用した場合に重複やハルシネーションが生じ、最終的な出力の質が低下することが考えられる。

表 5: 節翻訳モデルの入力トークン長ごとの BLEU

モデル	全て	1-20	21-40	41-60	61-
ベースライン	**28.6**	**29.2**	**28.8**	**26.8**	**28.8**
節翻訳モデル	24.7	25.6	24.7	23.3	24.5

表 6: 節翻訳モデルをベースラインに差し替えた際の入力トークン長ごとの BLEU

差し替え	全て	1-20	21-40	41-60	61-
あり	25.2	26.0	25.5	23.2	24.6
なし	**27.8**	**28.4**	**28.1**	**25.8**	**27.6**

表 7: ベースラインが節翻訳の際にハルシネーションを起こした例

入力文
The US is likewise bankrupt but
出力文
米国は同様に破産していますが、米国は同様に破産しています。

5 おわりに

本稿では、ニューラル機械翻訳による分割統治的手法において、節単位疑似対訳と、後段の並べ替えと編集のモデルに事前学習モデル mBART を利用する手法を提案した。提案手法は BLEU ではベースラインを下回る結果となったが、ハルシネーションやを起こしていた文を提案手法では適切に翻訳できていた例が確認できた。

本稿の手法では、節翻訳後の節の並べ替えと編集の工程において、[1] で利用されていた Transformer を事前学習モデルの mBART に置き換えることで並べ替えと編集の精度の改善を試みたが、精度面での効果は確認できなかった。文全体のコンテキストの情報やセグメント同士の関係性を考慮した節翻訳結果の並べ替えと編集が改善に向けた課題である。また、今回作成した節翻訳モデルは疑似的な節単位の対訳データで学習されたモデルであるため、節翻訳モデルの精度は改善の余地があると考えられる。

謝辞

本研究の一部は科研費 21H03500 と 21H05054 の助成を受けたものです。

参考文献

[1] 加納保昌, 須藤克仁, 中村哲. 分割統治的ニューラル機械翻訳. 言語処理学会 第 27 回年次 大会 発表論文集., 2021.

[2] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. *Multilingual Denoising Pre-training for Neural Machine Translation.* **arXiv:2202.03629**, 2022.

[3] K. Sudoh, K. Duh, H. Tsukada, T. Hirao, and M. Nagata. *Divide and translate: improving long distance reordering in statistical machine translation.* In Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Met- ricsMATR(SMT ‘10), p. 418–427, 2010.

[4] Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merrien-boer, Kyunghyun Cho, and Yoshua Bengio. *Overcoming the curse of sentence length for neural machine translation using automatic segmentation.* **In Proceedings of SSST-8 Elighth Workhop on Syntax Semantics and Structure in Statistical Translation**, p. 78–85, 2014.

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

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. *BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension.* **In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, p. 7871–788, 2020.

[7] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. *Multilingual Denoising Pre-training for Neural Machine Translation.* **arXiv:2001.08210**, 2020.

[8] 岡崎直観, 荒瀬由紀, 鈴木潤, 宮尾裕介, 鶴岡慶雅. 自然言語処理の基礎. オーム社, 2022.

[9] 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.

[10] Taku Kudo. Mecab : Yet another part-of-speech and morpho- logical analyzer., 2006. https://taku910.github.io/mecab/.

[11] Matt Post. *A call for clarity in reporting bleu scores.* **arXiv:1804.08771**, 2018.

[12] 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. 2020.

[14] Matthew Honnibal and Ines Montani. *spaCy 2*: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[15] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

A　付録

表 8: ベースラインと提案手法の出力例

ハルシネーション

入力文
Djiboutian Honorary Consulate in Montreal, Canada usacan i have a djiboutienne visa plsssssssssssssssssssssssssss, i will like to see my country. Post a comment on this page We invite you to share your experiences with the Djiboutian Honorary Consulate — obtaining visas and other services, locating the building, and so on.

参照文
この手法はカナダのオイルサンド産出地域の観測データに適用し、成功しています。

ベースライン
モントリオール, カナダのドジブティアン名誉領事館 このウェブサイトは名誉領事館により運営されているものではなく、あなたの意見や質問が必ずしもそちらの職員により確認されるというわけではありません。こちらはドジブティアンの外交政策について幅広い討論をするフォーラムではないことにご注意ください。そのような話題は削除されます。

提案手法
Montreal, カナダの Djiboutian 名誉領事館 usacan 私は djiboutienne ビザ plsssssssssssssssssssssss を持っています, 私は私の国を訪問したい. このページにコメントを投稿 私たちは、Djiboutian 名誉領事館であなたの経験を共有することをお勧めします — ビザやその他のサービス、建物の検索など.

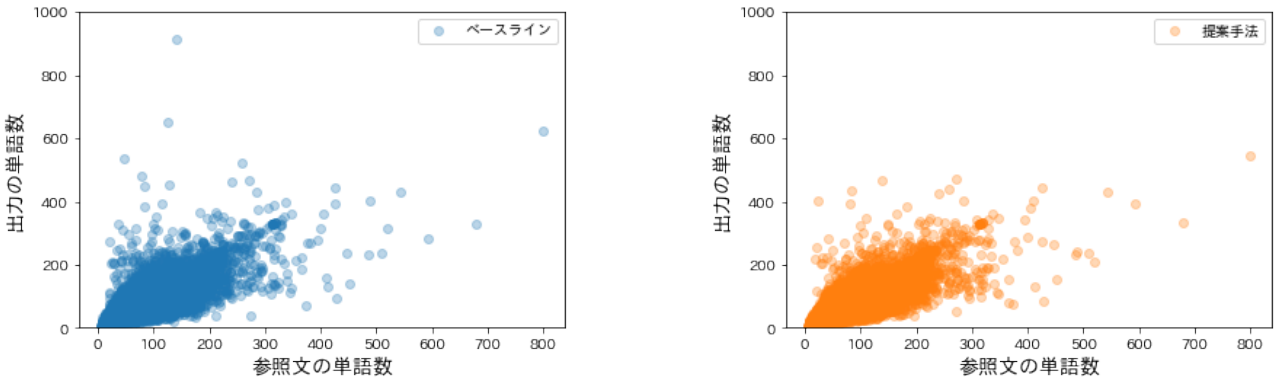
繰り返し

入力文
So this is Kinomoto Sakura?

参照文
「で、こちらが木之本桜さん?」「はい、そうです、国王陛下。

ベースライン
「さあ、さくらさくら?」「そう、さくらさくらは?」「さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらは、さくらさくらさくらは、さくらさくらさくらは、さくらさくらさく

提案手法
「これは木本さくらですか?」



(a) ベースライン　　(b) 提案手法

図 3: 参照文の単語数に対する出力文の単語数