分割統治的ニューラル機械翻訳

加納保昌　須藤克仁　中村哲
奈良先端科学技術大学院大学
{kano.yasumasa.kw4,sudoh,s-nakamura}@is.naist.jp

1 はじめに

ニューラル機械翻訳によって、統計的機械翻訳よりも流暢な訳文を生成できるようになった。しかし、長い文を翻訳することは未だ難しく、重複訳や、訳抜けなどが発生する。アテンション機構を用いたニューラル機械翻訳モデルである、Transformer [1] でも、その問題は解決できていない [2]。

統計的機械翻訳では、長い文を短く区切って翻訳し、並べ替えて繋げるという分割統治的な手法でこれらの問題を軽減する試みがあった [3]。ニューラル機械翻訳においては、Pouget-Abadie ら [4] は、長い文を短く区切って、各セグメントを RNN を用いて翻訳した。その後、翻訳されたセグメントを、前から順につなぎ合わせたものを翻訳結果としている。しかし、この手法では、語順が大きく異なる言語対を翻訳する際に、出力文が不自然なものになってしまう。

そこで、本研究では、翻訳されたセグメントを、セグメントどうしの関係性を表すトークンを用いて繋げ、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実現し、自然な訳文を生成する。

提案手法はベースラインに比べて、BLEU では顕著な差が見られなかったが、出力文の長さがより参照文へと近くなることがわかった。よって、ニューラル機械翻訳における分割統治的手法が、訳抜けを防ぐために、有効である可能性が示された。今後、並べ替えやセグメント翻訳の精度を上げることにより、さらなる発展が期待される。

2 関連研究

Sudoh ら [3] は、長い文を翻訳するため、統計的機械翻訳に分割統治的手法を用いた。入力文を節の単位に区切って翻訳し、その結果を節の階層構造に基づいて並べ替えることで、精度を向上させていた。

ニューラル機械翻訳においては、Pouget-Abadie ら [4] の研究が Sudoh ら [3] と似た手法を用いている。この手法では、対訳コーパスを用いて学習された、RNN を用いて最適な文の区切れ目を見つけ、その各セグメントを翻訳する。しかし、文法的に適切な境界で分割される保証がないため，特に語順の異なる言語対において結合した翻訳結果が不自然な文になる懸念がある。

機械翻訳以外の分野でも、長い文を短く区切って言い換えるという研究がなされている。Aharoni ら [5] は長い文を、copy 機構を用いた seq2seq モデルに入力し、複数の短い文で言い換えた。しかし、必要な情報が抜けたり、不必要な単語が出力されてしまう問題は解決されていない。

短い単位に区切るものとは異なるアプローチで、長い文の機械翻訳を改善する研究もある。Neishi ら [2] は、Transformer の長い文の翻訳精度を上げるため、RNN を使った相対的位置情報を用いた。

本研究では、Pouget-Abadie ら [4] の手法で問題であった並べ替えを、セグメント翻訳とは別のニューラルネットワークモデルで実行する。それによって、Sudoh ら [3] のような分割統治的な手法を、ニューラル機械翻訳において実現する。本研究は、Sudoh [3] らの手法とは異なり、節単位の対訳対応付けを必要としない。節の翻訳は、文単位の対訳コーパスから学習されたニューラルネットワークモデルにより取得する。

3 提案手法

3.1 提案手法の処理概略

提案手法での翻訳は、大きく分けて、4 つのステップで行われる。図 1 に、英語から日本語への翻訳の概略図を示す。この図では、簡略化のため、サブワード分割はされていない。

まず、原言語文を構文解析し、節の境界を探す。

次に、節の始めと終わりを境界としてセグメントに区切る。節の中に節を含む場合も、すべての節境

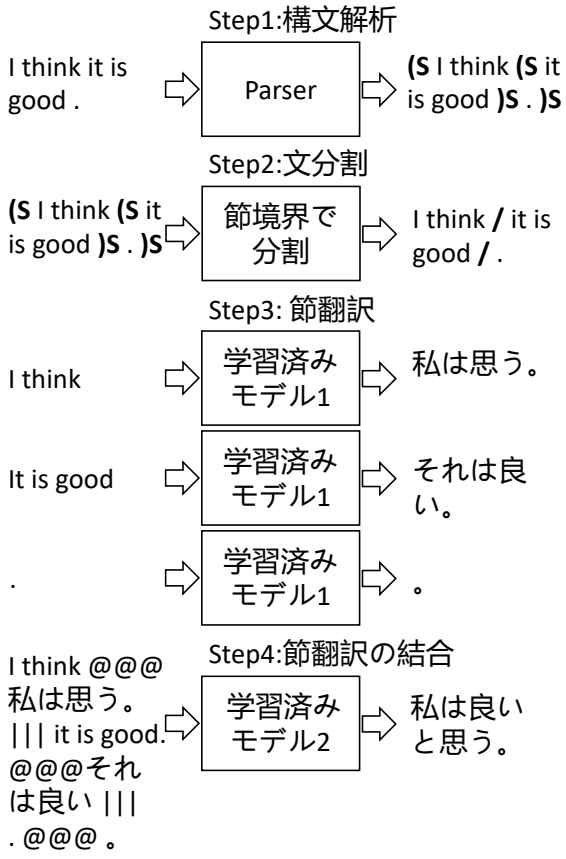

図 1　提案手法の処理概略

界で分割を行う。Pouget-Abadie ら [4] は、2 つの学習済み RNN モデルを用いて最適な境界を探索していたが、必ずしも、文法的に適切な境界で区切られるとは限らない。よって、本研究では節単位で区切った。各セグメントの翻訳は、周囲のセグメントの影響を受けるが、節という、比較的大きな意味のある単位で区切ることによって、その影響を抑えることが期待できる。

その次に、各セグメントをモデル 1 で翻訳する。このモデル 1 として、Transformer を用いる。推論時には、モデル 1 のビーム探索のスコアが最高となる節の翻訳を出力する。

最後に、モデル 2 を用いてこの結果を並べ替え、編集することで、最終的な訳文を出力する。このモデル 2 にも Transformer を用いるが、モデル 1 とは別に学習されたものである。モデル 2 の入力は、原言語の節とその翻訳結果を繋げたものである。具体的には、原言語の各節とその翻訳結果を “@@@” で繋げた。そして、それらの節のペアを “|||” で繋げた。“@@@” や “|||” は元の訓練データに含まれておらず、特殊な一つのトークンとして語彙に含めた。例えば、「I think it is good.」を日本語に翻訳する時には、「I think @@@ 私は思う。||| it is good @@@ それは良い。 ||| . @@@ 。」のような形で、モデル 2 に入力されることが想定されている。翻訳された各セグメントどうしが、どのような関係性をもつのかという情報を持っていないため、このように、原言語の情報を付与する必要がある。Xu ら [6] は、原言語の文と似た目的言語の文を繋げたものを、モデルに入力する手法を提案した。本研究では、似ている文ではなく、翻訳された複数の節を繋げることによって、並べ替えと編集をモデル 2 によって実現した。

3.2　学習

学習時には、まず、元の対訳コーパスを用いて、英日翻訳をモデル 1 で学習する。その後、図 1 のステップ 1 からステップ 3 までを、全ての学習データに対して行う。最後に、ステップ 3 の節の翻訳を図 1 のステップ 4 と同様の形式で繋げたものを入力とし、目的言語文を出力とすることで、モデル 2 を学習させる。

モデル 2 の学習時には、モデル 1 が出力する、複数の翻訳候補を利用して学習を行う。本研究ではビームサイズを 4 としているので、4 つの候補が出力される。よって、節の翻訳の候補でスコアが最も高いもののみを繋げた文、２番目に高いもののみを繋げた文、というように、四種類の形式をモデル 2 の入力として与えることができる。このように、英語の節の部分は同じだが日本語の部分のみ違うという入力文を用いることで、最終的に、モデル 2 のトレーニングデータは 4 倍になった。対応する目的言語の文は、４種類に対して、全て同じものを使う。

4　実験

提案手法の有効性を検証するために以下の実験を行った。

4.1　実験設定

対訳コーパスとして、ASPEC [7] の英日 100 万文 (train-1) を用いた。Transformer [1] は、fairseq [8] で実装されたものを用いた。基本的には、Vaswani ら [1] の結果を再現するためのウェブページ[1)] に従ってハイパラメーターを設定した。dropout は 0.3 とし、length penalty は 1 とした。checkpoint はパラメータを 1000 回更新する毎にごとに保存し、開発データの loss が最小となる checkpoint を選択した。これ

1）　https://sgithub.com/pytorch/fairseq/issues/1352

表 1　入力トークン長ごとの BLEU

モデル	all	1-20	21-40	41-60	61-
Transformer	**41.7**	40.8	41.5	**43.5**	39.1
Linearized tree	41.6	41.0	**41.8**	42.7	39.1
Not reordering	28.8	32.4	28.2	28.8	26.4
Proposed	**41.7**	**41.3**	41.4	**43.5**	**39.9**

表 2　入力トークン長ごとの ratio

モデル	all	1-20	21-40	41-60	61-
Transformer	0.960	**0.944**	0.956	0.971	0.975
Linearized tree	0.958	0.939	0.956	0.970	0.974
Not reordering	1.272	1.136	1.276	1.343	1.301
Proposed	**0.962**	0.943	**0.958**	**0.977**	**0.979**

以上の詳細は付録に添付する。サブワード分割には、SentencePiece [9] を用い、語彙の大きさを 16000 とし、日英で語彙を共有した。英語の構文解析には、stanza [10] を用いた。モデル 1 の入力には、構文解析で出力された単語を空白で繋ぎ、それを SentencePiece でサブワードに分割したものを利用した。このモデル 1 は、ベースラインの Transformer であり、モデル 2 を学習する際に必要となるモデルでもある。

実験のベースラインとして，モデル 1 として用いた通常の Transformer の他に以下の二つの手法を比較した．一つ目は、線形化された構文木 [11]（構文木をカッコとラベル、単語の系列で表したもの）を Transformer に入力したものである。二つ目は、Pouget-Abadie らの手法と同様に、モデル 1 から出力された節の翻訳を、前から順に繋げたものを最終的な訳文とするモデルである。

テストデータの BLEU [12] スコアは、MeCab [13] によって単語分割し、sacrebleu [14] を用いて算出した。

4.2　実験結果

表 1 と表 2 が英日翻訳の結果である。これらは test データの文を、モデル 1 の入力文のサブワードトークン長さ別に分け、それぞれに対して BLEU と ratio を算出した結果を示している。

ratio は出力文と参照文の長さの比を表し、1 より小さいということは、訳抜けが発生している可能性を示している。全てのテストデータに対する ratio は、線形化された構文木のモデルや Transformer に比べ、提案手法の方が大きかった。その一方で、翻訳精度を表す BLEU は Transformer と提案手法では同等である。この結果から、提案手法は訳抜けを防ぐために長い訳文を出力できるが、参照文に含まれていない情報も出力している可能性がある。また、モデル 2 を使った並べ替えを行わず、モデル 1 の出力を前から順に繋げたモデルは、ratio は 1 を大きく上回っており，また BLEU も比較モデルの中で最も低かった。この結果は、不必要な情報が翻訳結果に多く含まれて訳質が低下していることを示している。また、提案手法ではモデル 2 の処理によってそうした不必要な情報を取り除いて訳文を構成できていることも分かる。

次に、長さ別のスコアを見ると、原言語のサブワードトークン数が 60 よりも大きいものに対して、提案手法の BLEU は比較対象の中で最も高い。2 番目に高い Transformer の BLEU を 0.8 上回っており、ratio も提案手法の方がわずかながら高いことから、提案手法が長文の翻訳において有効であることが示唆される。

長さが 20 以内の短い文に対しても、提案手法は Transformer の BLEU を、0.5 上回った。短い文では、長い文に比べて、節の数が少ない。その結果、分割が発生しないこともあり、モデル 2 の入力が、原言語文に、その文の翻訳を繋げただけのものもある。これは、Xu ら [6] の用いた似た訳文の代わりに、モデル 1 による翻訳を用いたことで、同様の効果があったと推定される。

表 3 は、各モデルの実際の出力例である。Transformer では、「伊那谷と木曽谷を結ぶ」という部分の地名が抜け、代わりに「インターチェンジ・ゾーンエリアを結ぶ」となっている。「交流圏域」と類似した意味をもつ「インターチェンジ・ゾーンエリア」という表現が入っているが、これは、本来ならば、参照訳にあるような「交流圏域の拡大」という文脈のみで使われる単語である。また、線形化された構文木のモデルでは、「振興・創出 について検討した。」となっているが、「振興・創出をめざす地域整備の方策を検討した。」となるべきである。

これに比べて、提案手法は、より訳抜けを改善し、統語構造も参照訳に近いものとなっている。「インダニと木曽谷を結ぶ」の部分では、「木曽谷」という地名を出力し、「振興と創出を目指した地域整備の方策を検討した。」という、参照文に近い構造を持っている。

実際のモデル 2 の入力は付録に記載するが、表 3 の、モデル 1 による節の翻訳例を見ると、“National

表 3 モデルごとの出力例

モデル	出力例
英語の節分割	The measures for regional improvement aiming at expansion of the interchanging zone area **/**, and **/** promotion and creation of the industry in the area were examined , **/** taking advantage of the time that **/** National Route 361 connecting Inadani and Kisodani was approved in the business as a high standard road **/** .
参照訳	伊那谷と木曽谷を結ぶ国道 361 号線が高規格道路として事業認可されたのを機に, 交流圏域の拡大と域内産業の振興・創出をめざす地域整備の方策を検討した。
節の翻訳結果	交流圏域の拡大を目指した地域整備の方策。 **/** 又, 及び, また, そのためのものである。 **/** 地域における産業の振興と創出について検討した。 **/** 時間の利点を活かした。 **/** インダニと木曽谷を結ぶ国道 361 号が高規格道路として事業で認可された。 **/** その他。
Transformer	インターチェンジ・ゾーンエリアの拡大を目指した地域整備の方策と, インターチェンジ・ゾーンエリアを結ぶ国道 361 号を高規格道路として, ビジネスで認可された時期を背景に, エリアの産業の振興・創造について検討した。
Linearized tree	高規格道路として Inadani と Kisodani を結ぶ国道 361 号が事業で承認された時期をふまえて, 交流圏域の拡大を目指した地域整備に向けた施策と, 同地域の産業の振興・創出について検討した。
Not reordering	交流圏域の拡大を目指した地域整備の方策。又, 及び, また, そのためのものである。地域における産業の振興と創出について検討した。時間の利点を活かした。インダニと木曽谷を結ぶ国道 361 号が高規格道路として事業で認可された。その他。
Proposed	インダニと木曽谷を結ぶ国道 361 号が高規格道路として事業に認可された時の利点を活かし, 交流圏域の拡大, 産業の振興と創出を目指した地域整備の方策を検討した。

Route 361 connecting Inadani and Kisodani was approved in the business as a high standard road" という英語の節が、「インダニと木曽谷を結ぶ国道 361 号が高規格道路として事業で認可された。」とモデル 1 によって翻訳されている。この節の翻訳がモデル 2 に入力されていることで、「木曽谷」を正確に翻訳できていると言える。しかし、節の翻訳は、「地域における産業の振興と創出について検討した。」のように、構文木を使ったモデルと同じような統語構造を持っている。これに対し、その節と一つ前の節の「交流圏域の拡大を目指した地域整備の方策」がモデル 2 によって並べ替え、編集されることによって正しい統語構造が出力されていることが分かる。

4.3 考察

提案手法は Transformer に比べ、テストデータ全体に対して、ratio が若干改善したが、BLEU の改善には結びつかなかった。その理由の一つとして、提案手法の推論時に、ビーム探索によるスコア最大の節翻訳結果のみを考慮していることが挙げられる。節の翻訳は、隣の節にも依存するため、ビーム探索の結果の 1-best ではなく、k-best を考慮したモデルを考えることで、より精度の高い節翻訳結果の選択と結合が行える可能性がある。また、本研究では節の始めと終わりで区切っているため、節と節の間を繋ぐ、短いセグメントが生じ得る。そのような短い単位の翻訳では、重複訳が発生する傾向があった。例えば、", and" は、「又, 及び, また, そのためのものである。」のようにモデル 1 によって翻訳されている。そのため、節の分割の方法に対する更なる検討が必要と言える。

また、線形化された構文木のモデルの結果から、構文情報を入力文に付与することが必ずしも長い文の精度向上に役立つわけではないということが示された。

5 おわりに

本研究では、長い文を分割して翻訳し、その結果を並べ替えて繋げるという分割統治的なニューラル機械翻訳の手法を提案した。それによって、訳抜けを防止できる可能性を示した。コーパス単位の BLEU で顕著な差はなかったが、長文においては一定の効果が見られ、出力長を参照文に近づけることができた。今後はセグメント分割の手法、セグメント単位の翻訳の結合の方法についてさらに検討を進める。

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

andIllia Polosukhin. Attention is all you need. *CoRR*, p. Vol.abs/1706.03762, 2017.

[2]Masato Neishi and Naoki Yoshinaga. On the relation between position information and sentence length in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 328–338.*, 2019.

[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 MetricsMATR(SMT'10). pages 418–427.*, 2010.

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

[5]Roee Aharoni and Yoav Goldberg. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 719–724. Association for Computational Linguistics.*, 2018.

[6]Jitao Xu, Josep Crego, and Jean Senellart. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1580–1590, Online, July. Association for Computational Linguistics.*, 2020.

[7]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).

[8]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.

[9]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.*, 2018.

[10]Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101–108, Online.*, 2020.

[11]Roee Aharoni and Yoav Goldberg. Towards string-to-tree neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 132–140. Association for Computational Linguistics.*, 2017.

[12]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.*, 2002.

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

[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.*, 2018.

A 付録

A.1 実験設定の詳細

開発データの loss が 5 回改善が見られない時に学習を止めた。提案手法のモデル 2 では、入力の長さがベースラインの約 2 倍となる。1 バッチあたりに含まれる文の数をベースラインに合わせるため、モデル 2 の学習では、update-freq をベースラインの 2 倍にした。同様の理由により、線形化された構文木を入力とするモデルの update-freq をベースラインの 4 倍にした。

A.2 実際のモデル 2 の入力例

The measures for regional improvement aiming at expansion of the inter ch ang ing zone area @@@ 交流 圏 域 の拡大 を目指した 地域 整備 の 方策 。 ||| , and @@@ 又 , 及び , また , その ための ものである 。 ||| promotion and creation of the industry in the area were examined , @@@ 地域 における 産業 の 振興 と 創 出 について検討した 。 ||| taking advantage of the time that @@@ 時間 の利点 を 活 か した 。 ||| National Ro ute 36 1 connecting In ad ani and K is od ani was approv ed in the business as a high standard road @@@ イ ン ダ ニ と 木 曽 谷 を 結 ぶ 国 道 36 1 号 が 高 規格 道路 として 事業 で 認 可 された 。 ||| . @@@ その他 。

A.3 モデル 2 の入力例を SentencePiece でデコードしたもの)

The measures for regional improvement aiming at expansion of the interchanging zone area@@@ 交流圏域の拡大を目指した地域整備の方策。||| , and@@@ 又, 及び, また, そのためのものである。||| promotion and creation of the industry in the area were examined ,@@@ 地域における産業の振興と創出について検討した。||| taking advantage of the time that@@@ 時間の利点を活かした。||| National Route 361 connecting Inadani and Kisodani was approved in the business as a high standard road@@@ インダニと木曽谷を結ぶ国道 361 号が高規格道路として事業で認可された。||| .@@@ その他。