文内コンテキストを利用した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

長文の翻訳は、ニューラル機械翻訳（NMT）における顕著な課題の一つである。我々は、**分割統治**アプローチに基づいてこの課題に取り組む。提案手法では、(1) 入力文を等位接続詞を基準に節単位に分割し、(2) 分割された各節を、文内コンテキストを考慮できる形で、節翻訳モデルを使用して翻訳し、(3) 翻訳された節を集約して、別の Seq2Seq モデルを用いて最終的な翻訳を得る。ASPEC を用いた英日翻訳の実験結果から、41 単語以上の長い入力文において、mBART をファインチューニングしたベースラインよりも優れた BLEU を実現することが示された。

1　はじめに

近年、NMT の翻訳品質は飛躍的に向上し、より自然で文脈に沿った翻訳が可能となった。しかし、NMT にはいくつかの課題がある [1]。その課題の一つが長文の翻訳である。長文の翻訳精度低下の課題に対して、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、並べ替えて結合する分割統治的手法が提案されている [2]。NMT においては、Pouget ら [3] が、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する手法を提案している。しかし、[3] の手法は分割したセグメントを翻訳した後にそれぞれ前から単純に連結するため、英語と日本語のような語順が大きく異なる言語対を翻訳する際に、出力文が不自然な文になってしまう懸念がある。この語順が不自然になるケースに対応するため Kano ら [4] は NMT における長文のための分割統治的手法を提案している。[4] の手法による翻訳プロセスは大きく 2 つの段階に分かれており、各プロセスはそれぞれ別の Seq2Seq モデルで実施される。1 段階目は文を節単位のセグメントに分割し、分割後の各節の翻訳を行う。2 段階目は翻訳後の各節の連結部分を自然な表現に書き換えたり、必要に応じて節の順番を並び替えることで、最終的な出力文が自然な訳出になるように調整を行う。Kano ら [4] の手法において、二つの課題が挙げられている。一つは節分割の際のセグメンテーション単位の課題、もう一つは節分割後の節の翻訳精度の課題である。

本研究では、上記二つの課題に注目し、英日翻訳における長文の翻訳精度の向上を試みる。具体的に、等位接続詞を基準とした新しい英文の節の分割単位及び、分割後の節翻訳において文内コンテキストを参照して翻訳を行うモデルの学習手法を提案する。

2　分割統治型 NMT とその課題

分割統治型ニューラル機械翻訳の概要とその課題を説明する。

2.1　分割統治型 NMT の概要

Kano ら [4] の手法は主に 4 つの Step からなる。Pouget ら [3] はセグメント翻訳のための最適な分割の境界を見つけるため、文のセグメンテーションに翻訳モデルの出力確率を利用していた。しかし、彼らの手法は文法的に適切な境界で文を分割する保証がない。それに対し、[4] の手法は節という、統語構造上自然な単位に基づくセグメンテーションを採用している。Step1 で入力文を構文解析し、Step2 で文を節単位でセグメントに分割する。節の中に節を含む場合も、すべての節境界で分割が実施される。その後、Step3 で分割された各節は、文単位の対訳コーパスで学習された翻訳モデルによって翻訳される。最後の Step4 で翻訳モデルとは別の Seq2Seq モデルを用いて、最終的な訳出を生成する。Step4 における入力はソース側の各節と翻訳された各節を特殊トークンを使用して結合し、その後、各節同士を別の特殊トークンで結合したものである。分割されたことにより、翻訳された各セグメント同士が、互いに文脈的にどのような関係を持つのかという情報

が失われているため、このようにソース側の情報を付与する必要があるとされている。

2.2 分割統治型 NMT の課題

彼らの手法では、Step3 における節単位の翻訳精度が問題とされている。この節の翻訳精度は、翻訳の最終的な品質に大きな影響を与えるとされる。節翻訳の精度における課題には、次の 3 つの要因が考えられる。

セグメンテーションの単位 すべての節境界で分割を行うため、分割後の節が短くなり過ぎたり、節と節の間を繋ぐ短いセグメントが生じることがある。

文レベルの NMT モデルの利用 節翻訳の後にソース側の情報を翻訳後の各節に付与しているが、節の翻訳は文単位のコーパスで学習されたモデルで行なっているため、トレーニングと推論の間の入力文の不整合により、望ましくない追加や省略を引き起こす可能性がある。

分割された節をそれぞれ独立に翻訳 Step3 の節翻訳の際に各節はそれぞれ独立に翻訳されるため、文中の文脈情報が活用できず、不適切で一貫性のない単語選択を引き起こす可能性がある。

3 提案手法

本研究では上記の 3 つの課題を考慮して Kano ら [4] の手法を基盤とし、新しい分割統治ニューラル機械翻訳手法を提案する。

3.1 等位接続詞を基準とした節のセグメンテーション

[4] の手法ではすべての節境界で分割を行うため、分割の際に [2] の手法のように階層構造を考慮していない。本研究では、節と節を連結する等位接続詞を基準として、その前後に S 構造を持つ文にのみ分割の単位を限定することで、この問題に対処する。この節分割単位に限定することで節の中に節を含む場合の分割は行われない。したがって、節の階層構造を考慮せずに、分割された節が短くなり過ぎることが防げる。節分割のルールの詳細を実例を用いて示す。次の 2 つの文があるとき、

1. I like football and baseball.
2. She wants to travel the world, but her job keeps her busy, and her savings are not enough for the trip.

これらに構文解析を適用して、次の表現を得る。

1. (S (S She wants to travel the world),
2. (CC but) (S her job keeps her busy), (CC and) (S her savings are not enough for the trip).)

最初の文には、2 つの名詞を接続する and が含まれているが、これはセグメンテーションの条件を満たしていない。2 番目の文を接続詞を基準にセグメント化すると、3 つの節と 2 つの接続詞が導き出される。

- (S She wants to travel the world,)
- (CC but)
- (S her job keeps her busy,)
- (CC and)
- (S her savings are not enough for the trip.)

我々のセグメンテーションルールは、入力文全体が S を形成する場合、結果として生じるセグメントが S または CC であることを保証する。

3.2 節単位の対訳データのアライメント

節翻訳に文レベルの NMT モデルを利用することによって生じる可能性のある問題を軽減するために、節の対訳データを使用して節翻訳のための NMT モデルが作成できることが望ましい。本研究では、節レベルの対訳データを作成し、このデータに基づいて mBART を節翻訳用にファインチューニングした。我々は、Kano ら [5] の手法を参考にした、文レベルの対訳データから節レベルのアライメントを得る手法を提案する。この手法は主に以下の 3 つのプロセスから構成されている。

対訳コーパスのソース側の構文解析 ソース側の文を構文解析し、接続詞の前後に節 S を持つ文を特定し、接続詞の位置に基づいて節に分割する。

ソース側の各節の翻訳 分割された各ソース節をベースラインの翻訳モデルを使用して翻訳し、ターゲット側の節としてペアにすることで、節レベルで擬似的な対訳データを構築する。

節アライメントの獲得 疑似的な節単位の対訳データのターゲット節はベースラインによって翻訳されたものである。しかし、ソース側の各節は独立して翻訳されるため、翻訳後のターゲット側の各節は分割前の元のソース文から一部の文脈情報を失う。したがって、翻訳されたターゲット側の各節を並べ替えて繋げ直してたとしても、元のターゲット文を再現できる保証はない。したがって、節のアライメントは元の対訳データから獲得する。このプロ

<<< En-sentence1 >>> + En-sentence2 --- Ja-sentence1
En-sentence1 + <<< En-sentence2 >>> --- Ja-sentence2
<<< En-sentence2 >>> + En-sentence3 --- Ja-sentence2
En-sentence2 + <<< En-sentence3 >>> --- Ja-sentence3

<<<En-clause1>>> + En-clause2 + En-clause3 --- Ja-clause1
<<<En-clause2>>>+ En-clause3 --- Ja-clause2
En-clause1 + En-clause2 + <<<En-clause3>>> --- Ja-clause3

図 1: 文レベルのコンテキストを考慮した対訳データ (左) と節レベルのコンテキストを考慮した対訳データ (右) の形式（タグで囲われた部分のみを翻訳するようにモデルを微調整する）

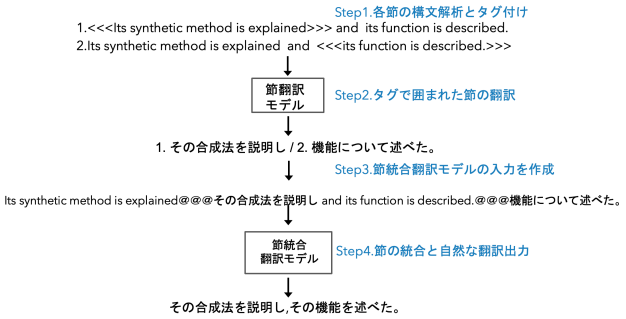

図 2: 提案手法による翻訳処理の概要

セスは以下の 4 つの Step で行われる。（付録の図 3 に節アライメント獲得のプロセスを示す。）

1. mBART のトークナイザーを使用して、分割されたソース文に対応するターゲット文をトークン化する。このとき、最後のトークンから始め、最初のトークンに向かって、一つずつトークンを連結しながら、分割されたソース節とのアライメント候補を作成する。
2. 翻訳された文の最後の節から最初の節に向かって順序を追いながら、Step1 で作成したアライメント候補（最後から順に連結されたトークン化されたターゲット文）に対して BERTscore（F1）を測定する。そして、最も高いスコアを持つアライメント候補をソース側の節とアライメントする。
3. ターゲット側のアライメントされた文字列を元のターゲット文から削除する。
4. ソース側のアライメントされていない節が残り 1 つのみの場合は、それを残りのターゲット文とアライメントし、操作を終了する。2 つ以上のソース節が残っている場合は、削除された文字列を除いた文を新しいターゲット文として扱い、Step1 に戻る。

この手法を使用することで、ソース側のセグメント化された各節とターゲット側の文との間の対応が確立され、擬似的な対訳データを使用せずに節レベルでのアライメントが獲得される。その結果、節単位の対訳データの品質が向上することが期待できる。Step1 において、[5] の手法のように、ターゲット文の最初からトークンを連結し、アライメント候補を作成することも試みたが、後方からの連結と比較してアライメントの精度が低かったため、後者のアプローチを採用した。

3.3 文内コンテキストを利用した節単位翻訳モデル

節翻訳モデルのデータ形式 文内コンテキストを利用した節翻訳モデルは複数の節が組み合わされた入力から、タグ <<< >>> で囲まれた節のみを翻訳するようにファインチューニングされる。図 1 は、文レベルおよび節レベルの文脈を考慮できる形の対訳データの形式を示している。このデータ形式を採用することにより、[6] の手法のように入力の文脈情報が拡張される。その結果、タグに囲まれた部分の外側の文脈情報を参照しながら、タグで指定された部分のみを翻訳するモデルを作成することが可能になる。

節翻訳モデルのファインチューニング 節翻訳モデルは次の二段階でファインチューニングされる。

- 文レベルのコンテキストを考慮した対訳データを用いて、mBART の初期のファインチューニングを行う。
- 節レベルでのコンテキストを考慮した対訳データを使用して、節翻訳に特化するための二段階目のファインチューニングを行う。

節レベルのコンテキストだけでなく、文レベルのコンテキストを考慮した対訳データを使用して段階的にファインチューニングを行う理由は、節レベルのコンテキストを考慮した対訳データの量が比較的少ない (300K) ためである。これを補うために、今回の実験では、文レベルのコンテキストを考慮した対訳データも使用して、節翻訳モデルのファインチューニングを行った。

通常、対訳コーパスに含まれる文が互いに文脈的なつながりを持つ保証はない。しかし、文書 ID と文 ID の情報があれば、元の文書を再構築できる可能性がある。本研究では、そのようなコーパスが利

単語数	1-20	21-40	41-60	61-	All
全文	932	784	89	7	1812
節分割された文	54	179	24	2	259

表 1: 単語長ごとのテストデータの文数

単語数	1-20	21-40	41-60	61-	全て
ベースライン	**40.3**	**42.1**	40.0	46.5	41.3
提案手法	40.2	**42.1**	**40.7**	**47.8**	**41.4**
文数	932	784	89	7	1812

表 2: テストセット全体の BLEU

用可能であると仮定する。まず、対訳コーパスのソース側の文書 ID と文 ID を使用して、ソース側コーパス内の元の文書を再構築する。次に、各文書内で、文脈を維持しながら文を 2 つずつ組み合わせる。この際に組み合わされた文の一方に、単一の文単位を示すタグ <<< >>> が追加し、タグ付きの文に対応するターゲット側の文がペアになる。

2 文の文脈を考慮する対訳データと同様に、アライメントされたソースとターゲットの節を基に、節の文脈を考慮した対訳データを作成する。これにより、各セグメントされた節を、その節の前後の文脈を参照して翻訳することが可能になる。その結果、Kano ら [4] の手法と比較して、より効果的な節レベルの翻訳が期待される。図 2 は、提案された手法による翻訳プロセスの概要を示すもので、文脈を考慮できる形の節翻訳の実装を例示している。

4 実験

提案方法の有効性を調査するため、我々は長文の翻訳に焦点を当てた以下の実験を実施した。

4.1 実験設定

我々は、文書と文の順序情報が利用可能で、我々の提案手法が適応可能なバイリンガルコーパスである ASPEC [7] を基に、英日翻訳の実験を行った。

評価指標は、BLEU、BLEURT、COMET を用いた。日本語のトークン化には `ja-mecab` [8] を使用し、`sacrebleu` [9] を用いて計算した。文の長さの影響を調べるために、テストセットを英文の単語数に基づいて 4 つビン (1-20、21-40、41-60、61 以上) に分割した。提案手法は、入力英語文が等位接続詞を持ち、その前後に S 構造を持つ文にのみ働き、それ以外の場合にはベースラインが使用される。そのためテストセット全ての文に提案手法が適応されるわけではない。表 1 は、テストセット全体と提案手法が適応される文の数をそれぞれ示している。

単語数	1-20	21-40	41-60	61-	全て
ベースライン	**49.7**	44.8	40.1	31.4	44.7
提案手法	48.8	**44.9**	**42.5**	**32.5**	**45.0**
文数	54	179	24	2	259

表 3: 入力単語数ごとの BLEU(分割が行われた文のみ)

4.2 結果

表 2 はテストセット全体のベースラインと提案手法の入力単語数ごとの BLEU スコアを比較している。全体的な結果から、提案手法はベースラインに対し、0.1 ポイント高い結果となった。単語長 41-60、61 以上の範囲においてはそれぞれベースラインと比較して、0.7 ポイント、1.3 ポイント BLEU の改善が確認できた。

表 3 はベースラインと提案手法における構文解析の結果、分割が行われた文のみを対象とした場合の入力単語数ごとの BLEU を比較している。これは、ベースラインと提案手法の性能差を明確に示すための結果である。提案手法はベースラインに対し、1-20 単語長以外の全ての範囲において、高い BLEU を示している。特筆すべきは単語長 41-60、61 以上の区分において、提案手法はベースラインよりもそれぞれ 2.4 ポイント、1.1 ポイント高い BLEU を達成していることである。これは提案手法が長文の翻訳において特に効果的であることを示唆している。より詳細な分析については付録で述べる。

5 おわりに

本研究では、長文の翻訳品質の向上を目指し、NMT において分割統治手法を基盤とした、新たな文のセグメンテーション単位及び、文内コンテキストを利用した節翻訳モデルの学習方法を提案した。実施した実験により、提案された手法は単語数 41 以上の長文の翻訳において BLEU スコアを向上させることが確認された。

提案手法が適応される文の割合を増やすことによるテストセット全体としての翻訳品質のさらなる向上が期待できる。そのため、等位接続詞以外での新しい節分割パターンの追加などを行い、接続詞を持たない長文にも対応できるように提案手法を拡張することを今後の課題とする。

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]Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In **Proceedings of the First Workshop on Neural Machine Translation**, pp. 28–39, Vancouver, August 2017. Association for Computational Linguistics.

[2]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.

[3]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.

[4]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.

[5]Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. Simultaneous neural machine translation with prefix alignment. In **Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)**, pp. 22–31, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[6]Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors, **Proceedings of the Third Workshop on Discourse in Machine Translation**, pp. 82–92, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

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

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

A 付録

単語数	1-20	21-40	41-60	61-	全て
ベースライン	0.792	0.729	0.726	**0.653**	0.742
提案手法	**0.797**	**0.738**	**0.732**	0.618	**0.749**
文数	54	179	24	2	259

表 4: 入力単語数ごとの BLEURT(分割が行われた文のみ)

単語数	1-20	21-40	41-60	61-	全て
ベースライン	0.924	0.896	0.888	0.908	0.901
提案手法	**0.929**	**0.897**	0.888	**0.909**	**0.903**
文数	54	179	24	2	259

表 5: 入力単語数ごとの COMET(分割が行われた文のみ)

テストデータ	1-20	21-40	41-60	61-
節の平均単語数	8.6	13.4	21.6	24.4
文の平均単語数	14.8	27.7	46.8	68.1

表 6: テストデータの節と文の平均単語数

表 4、表 5 から BLEURT、COMET での評価では、全体として BLEU と同様に提案手法がベースラインを上回っていることが確認できた。しかし、BLEURT においては単語長 61 以上の範囲ではベースラインを下回る結果となった。今回の実験において全テストデータ 1812 文の内提案手法が適応されている文は 259 文で、割合としては約 14%と非常に少ないため、テストセット全体としての BLEU の向上は 0.1 ポイントに留まったと考えられる。単語数 41-60、61-の範囲で提案手法の BLEU が改善した原因として、節単位のセグメンテーションによって長文の入力と出力の長さが翻訳モデルにとって扱いやすい単位になったことが考えられる。図 4 は単語数ビンごとの文数を示している。学習データにおける文の長さの平均値、中央値はそれぞれ 22.8、21 であり、単語数 10-30 の範囲に学習データが集中していることが分かる。このことから、単語数 10-30 の範囲が翻訳モデルにとって扱いやすい文長であることが考えられる。表 6 はテストデータのそれぞれの単語長の範囲における節と文の平均単語数を示している。単語数 41-60、61 以上の範囲の節の平均単語数に注目すると、どちらも単語数 10-30 の範囲に収まっている。このことから長文の翻訳において入力文を翻訳モデルが扱いやすい単位に文を分割して扱うことでより効率的に入力文の情報を処理することができ、結果的に翻訳の精度が向上したことが考えられる。

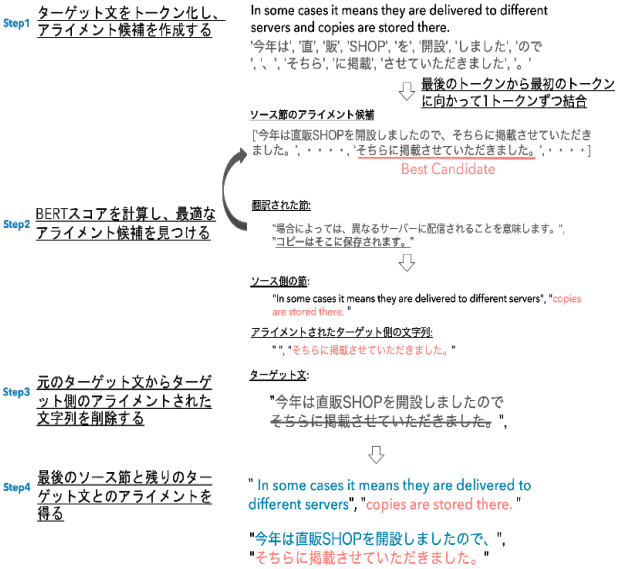

図 3: 提案手法による分割された英語文節と対応する日本語文節のアライメントを得る処理の概要

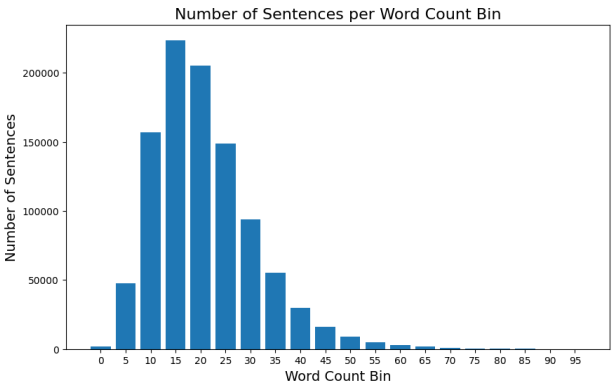

図 4: ASPEC コーパスの単語数ビンごとの文数