階層型 Transformer デコーダを用いた文字レベル機械翻訳

髙﨑環
東京大学工学部電子情報工学科
takasaki@logos.t.u-tokyo.ac.jp

鶴岡 慶雅
東京大学大学院情報理工学系研究科
tsuruoka@logos.t.u-tokyo.ac.jp

1 はじめに

文字レベル機械翻訳では、単語レベルや部分語レベルでの機械翻訳に比べて、一文あたりのトークン数が増加する．それに伴い、予測の際に長距離間の文脈を把握することがより難しくなることが考えられる．そこで本研究では、翻訳文を出力するデコーダを、単語レベルのデコーダと文字レベルのデコーダを階層型に組み合わせたアーキテクチャとすることで、長距離間の文脈を把握する性能を向上させることを目指す．本研究では、文字レベル機械翻訳の先行研究では未だ行われていない Transformer デコーダの階層化を行い、通常の Transformer モデルとの性能の比較を行った．提案するアーキテクチャでは、階層型モデルを使用する際に、単語レベルデコーダへの入力のために、文字列を単語レベルに分割する必要が生じる．そこで本研究では、単語境界での単語分割に対し、文字数での単語分割を行った場合と比較することで、適切な単語分割手法を検討した．その結果、階層化した Transformer デコーダ [1] を用いる際に、文字数による分割を行った場合では精度が低く、単語境界での分割の方が適していることが明らかになった．また、単語境界によって分割した階層型 Transformer は、従来の Transformer モデルと比較して BLEU スコアの精度は下回ったものの、階層型 Transformer では文章が途切れずに翻訳できるケースが複数あるなど、長距離間での文脈保持に一定の効果があることが観察できた．

2 関連研究

2.1 文字レベル機械翻訳

ニューラル機械翻訳などの自然言語処理では、文を単語レベル、部分語レベル、文字レベルで分割する処理が行われる．従来は単語レベルでの分割が主流であったが、計算コストやメモリ容量の関係で、語彙サイズが大きい際には低頻度語を辞書に登録せず、未知語として扱う必要がある．そのため、高頻度語と字面が近い活用語などの単語でも、未知語として処理されてしまうという問題が生じる．この問題を防ぐために、語彙サイズを抑えた文の分割手法が提案されてきた．文字レベル機械翻訳は、Costa-jussa ら [2] によって、単語の形を文字レベルから直接生成するモデルが提案されて以降、研究が進められている．また、Lee ら [3] は、単語の境界状態などのセグメンテーションを全く用いない文字レベル機械翻訳アーキテクチャを考案した．また、最近では、ニューラル機械翻訳で Transformer ベースのモデルが広く用いられるようになったことから、文字レベル機械翻訳に最適に改良された Transformer モデル [1] が、Banar ら [4] や Gao ら [5] によって提案されている．

2.2 階層型デコーダ

エンコーダ・デコーダから構成される機械翻訳モデルにおいて、デコーダを階層化する工夫をした先行研究がいくつか存在する．Luong ら [6] は、単語レベルの分割手法で語彙サイズを一定にすると定頻度語が未知語として扱われる問題に対して、階層型デコーダを用いることを提案した．通常は単語レベルで自己回帰的に復号し、未知語が出力された場合に限り文字レベルで自己回帰的に復号し、単語レベルと文字レベルの損失関数の和を最小化するように学習することで、翻訳品質を向上させた．また、Ataman ら [7] は、RNN[8] を用いた文字レベル機械翻訳において、単語レベルと文字レベルデコーダからなる階層型デコーダを用いることで、部分語に分割した場合に比べてより少ないパラメータで同等の翻訳精度が得られることを示すとともに、従来の文字レベル機械翻訳に比べて長距離の文脈・文法依存性をより学習できることを主張した．階層型デコーダは、文章生成タスクでも研究が進められている．Serban ら [9] は、文章の時系列を予測する文単位でのデコーダと、文ごとに単語の時系列を予測する単語単位でのデコーダを用いた RNN モデルを用いて、文脈を考慮した文章生成の研究を行った．また、渥美ら [10] は、このモデルを階層型

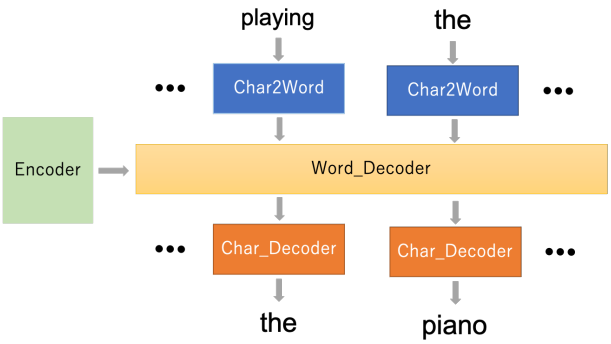

図 1　階層型デコーダ

Transformer に拡張したモデルを提案し、生成精度が向上されることを主張した．

3　提案手法

3.1　階層型 Transformer デコーダ

本実験では、長距離間の文脈把握能力の向上に期待し、階層型 Transformer デコーダを実装した．これは、Ataman ら [7] の研究で用いられている階層型 RNN デコーダを、現在機械翻訳タスクで用いられることの多い Transformer に拡張したものである．階層型デコーダの大まかな構成を図 1 に示す．詳細な構造を Ataman らの実装した階層型 RNN デコーダと比較しながら説明する．

3.1.1　文字表現から単語表現への変換

単語ごとの文字レベルトークン列を、Character Embedding 層によってそれぞれの文字ベクトルを獲得した後に、単語表現に変換する必要がある．Ataman ら [7] の階層型 RNN デコーダでは、単語内の文字ベクトルを双方向 RNN に入力し、最終出力を単語表現としていた．本実験では、実行速度の観点から、双方向 RNN ではなく、畳み込み層と Highway 層からなるアーキテクチャを構成し、単語表現を獲得している．

3.1.2　単語レベル Transformer デコーダ

変換された単語表現を元に、次の単語を予測する必要がある．Ataman ら [7] の階層型 RNN デコーダでは、エンコーダに対しての注意機構付き RNN を用いて次の単語を予測する表現を獲得していた．本実験では、エンコーダの出力を Source-Target Attention とした Transformer を用いて実装を行った．

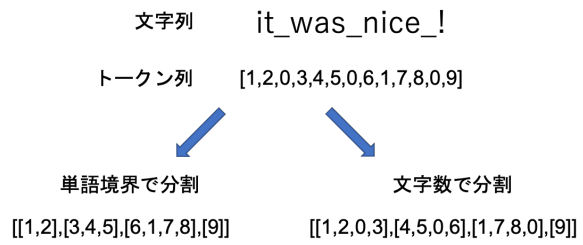

図 2　単語への分割方法

3.1.3　文字レベル Transformer デコーダ

単語レベル Transformer によって得られた次の単語の予測表現を元に、単語を構成する文字を復号する必要がある．Ataman ら [7] の階層型 RNN デコーダでは、文字レベルデコーダも同様に RNN を用いている．そして、単語レベルの RNN デコーダの出力を、文字レベルデコーダの状態として入力し、単語頭を表すトークンを入力することで文字を逐次的に復号する方法が提案されていた．本実験では、文字レベルデコーダも Transformer で実装した．文字レベル Transformer デコーダの Source-Target Attention の Source として、過去に生成した単語の予測表現を利用することで、現在生成する単語についての情報を文字レベルデコーダに与えるようにした．

3.2　単語への分割方法

階層的にデコーディングを行うに当たって、文字レベルでのトークン列を、単語などの大きな単位で分割し、次元を一つ増やす整形をする必要がある．図 2 に、トークン列の整形を行う方法を示す．本実験では、二つの方法でトークン列を整形した．一つ目は、単語レベルでの分割手法である．これは、単語が空白によって分割できる言語において、空白を境界として単語ごとに整形し直す手法である．単語の境界を活用する方法は、Ataman ら [7] の手法でも用いられており、単語の境界の情報を与えることによって、文字レベル機械翻訳の精度が向上する場合があると考察されている．しかし、この研究では、同じ階層型デコーダを用いる際に単語の境界情報を与えない場合については検証されていない．そこで、本研究では、単語の境界情報を与えない整形方法として、文字数を基準とした分割を採用し、比較することにした．

表 1　データ数の詳細

訓練用データ	検証用データ	評価用データ
160,240 文	7,284 文	6,751 文

4　実験

4.1　データと前処理

今回の実験では、IWSLT’14 独英翻訳コーパス[1)]を利用した．全ての実験において一字ごとにトークン化し、文頭トークン・文末トークンを必要に応じて付加した．データ数の詳細を表 1 に示す．

階層型 Transformer を用いたモデルについては、変換したトークン列を次のいずれかの方法で分割し、整形した．

- スペースによる分割（単語の境界による分割）
- 文字数による分割（文字数は 4 とした）

分割したトークンの前後には単語頭トークン・単語末トークンを必要に応じて付与した．

4.2　モデル

翻訳モデルのエンコーダには、Gao ら [5] が提案している、Convolutional Transformer を利用した．デコーダには、次の二つのいずれかを利用した．

- 通常の Transformer デコーダ
- 階層型 Transformer デコーダ

通常の Transformer デコーダでは、Vaswani ら [1] の base モデルと同じものを用いた．階層型 Transformer デコーダの文字レベル表現から単語レベルへの変換には、フィルタ幅 1 から 7 の畳み込み 7 層と Highway 層 2 層を用いた．単語レベル、文字レベルでの出力生成には、Vaswani ら [1] の base モデルと同様のものから、埋め込み層の次元を 512 から 256 に変更したものを使用した．

4.3　訓練時の実験設定

モデルの実装と訓練・翻訳には fairseq[2)]を用いた．学習には AdamW[11] を使用し、学習率は 5e-4 から inverse square root schedule にしたがって減衰させた．dropout を 0.3, weight-decay を 0.0001、label-smoothing を 0.1 に設定した．学習の際は、patience を 3 epoch とし、開発用データにおける損失を指標として early-stopping を行った．

1）　http://workshop2014.iwslt.org/
2）　https://github.com/pytorch/fairseq

表 2　各モデルの BLEU スコア

モデル	分割方法	ビーム幅	BLEU
Transformer		5	31.71
階層型 Transformer	4 文字毎	5	0.69
	単語境界	5	19.39
	単語境界	4	19.56
	単語境界	3	20.15
	単語境界	2	21.78
	単語境界	1	28.73

4.4　評価時の実験設定

テストデータにおいて翻訳を生成する際には、ビームサーチを用いた．階層型 Transformer デコーダを用いる際には、文字レベルデコーダでビームサーチを用いて生成する文字を推論し、生成された最も損失が少ない文字列を最終的な予測単語として確定させた．また、ビーム幅は、階層型 Transformer では 1 から 5 を試し、通常の Transformer ではビーム幅を 5 とした．また、評価用データを用いた機械翻訳の評価指標には BLEU スコア [12] を用いた．

5　結果と考察

5.1　結果

表 2 は、それぞれのモデルの評価データにおける BLEU スコアである．また、生成された翻訳文の例を付録 A.1 に示す．

5.2　考察

まず、分割方法を 4 文字毎とした際は、BLEU スコアが 0.69 と非常に低い値となった．文字数での分割方法の場合は、訓練時も損失が下がりにくく、タスク自体が困難であったことが考えられる．この一因として、単語分割の場合と異なり、実際の単語境界を跨いで違う単語の接頭辞と接尾辞が単語レベルの表現へと変換されるケースや、実際には同じ単語内であっても、不適切な部分語として分割されている場合が避けられないことが考えられる．以下の考察では、実際の単語境界を用いてトークン列の分割を行った場合について述べる．

階層型 Transformer デコーダによって生成された文の特徴の一つに、通常の Transformer デコーダでは途中で終了していた文を最後まで生成できていることが多いことが挙げられる． その一例を表 3 に示す．こ

れは、階層型 Transformer デコーダによって長距離間の文脈を保持しやすく、長文の復号を完遂しやすいことが考えられる．その一方で、表 4 に示すような、比較的短い文の場合でも、同じような語句が繰り返されるケースが散見され、階層型 Transformer デコーダでは近距離での単語出現を把握する能力が低いことが考えられる．また、階層型 Transformer でのビームサーチについては、ビーム幅が小さいほど BLEU スコアが高くなり、最もスコアが高かったのはビーム幅が 1 の場合となった．ビーム幅を変更した場合の生成例を表 5 に示す．この例にも観察できるように、ビーム幅を小さくすることで文章の単語ごとの一致だけでなく、文構造も一致するケースがある．このように、文字レベルデコーダにおけるビームサーチによって、単語レベルデコーダの担う文構造の把握が阻害されるケースが発生する場合があることがわかる．これは、単語デコーダで実際の単語を生成する際にエンコーダからの翻訳元の出力を利用していないため、翻訳元の文構造を把握する能力が低いことが一因として考えられる．文字レベルデコーダでのビームサーチを行うと、単語を生成する文字列としてのスコアが低いものがより選ばれやすい一方で、エンコーダの出力からの情報よりも、実際に Source-Target Attention で注視している単語レベルデコーダの出力からの情報の影響をより受けやすいことが想定される．

6 まとめ

本研究では、文字レベル機械翻訳において、単語レベルと文字レベルのデコーダを組み合わせた階層型デコーダを Transformer に拡張し、通常の Transformer モデルと比較を行った．その結果、BLEU スコアでは劣るものの、長文を途切れることなく復号することに適しているということが判明した．また、階層型 Transformer デコーダを構成するにあたって、単語レベルへのトークン列の分割方法を複数検討したところ、単語境界で分割した場合に比べ、文字数で分割した場合は学習が適切に行われないということが明らかになった．今後の研究としては、文字レベルデコーダにおいて、エンコーダの情報を利用するアーキテクチャを考案することが挙げられる．また、文字レベル機械翻訳における復号時間の増大を解消するために、文字レベルデコーダを非自己回帰的な手法を用いた場合の検証を行う予定である．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[2] Marta R. Costa-jussà and José A. R. Fonollosa. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 357–361, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 365–378, 2017.

[4] Nikolay Banar, Walter Daelemans, and Mike Kestemont. Character-level transformer-based neural machine translation. *arXiv preprint arXiv:2005.11239*, 2020.

[5] Yingqiang Gao, Nikola I. Nikolov, Yuhuang Hu, and Richard H.R. Hahnloser. Character-level translation with self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1591–1604, Online, July 2020. Association for Computational Linguistics.

[6] Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1054–1063, Berlin, Germany, August 2016. Association for Computational Linguistics.

[7] Duygu Ataman, Orhan Firat, Mattia A Di Gangi, Marcello Federico, and Alexandra Birch. On the importance of word boundaries in character-level neural machine translation. *arXiv preprint arXiv:1910.06753*, 2019.

[8] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[9] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30, 2016.

[10] 渥美和大, 狩野芳伸. Hierarchical transformer によるストーリー生成. 言語処理学会 第 26 回年次大会, 2020.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 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, 2002.

A 付録

A.1 翻訳文の生成例

表 3 通常の Transformer では文が途中で途切れる例

ソース文	aber tatsächlich war es eine ausgabe auf den frühen 80ern , als ich gerade mit der grundschule anfing und damit begann ,mein selbstbild außerhalb meines familiären umfelds aufzubauen und zu formen , auch in bezug auf andere kinder und zur übrigen welt um mich herum .
ターゲット文	but , in fact , the print date was the early 1980s , when i would have been starting primary school and forming an understanding of myself outside the family unit and as related to the other kids and the world around me .
Transformer	but in fact , it was an early ' 80s copy when i was just starting with the elementary school and beginning to build my self-image outside of my family environment , in terms of other children and
階層型 Transformer	it was actually a copy on the early ' 80s when i was just starting to school with elementary school , and i was starting to build and shape my self-assembly set outside my family sphere , also in terms of other children and the rest of the world around me .

表 4 階層型 Transformer で繰り返しが生じるケース

ソース文	aber die wahrheit ist weit davon entfernt .
ターゲット文	but the reality is far from it .
Transformer	but the truth is far away from it .
階層型 Transformer	but the truth is that the truth is far away from it .

表 5 階層型 Transformer でビーム幅を変化させた例

ソース文	wir haben tausende organismen die das tun können .
ターゲット文	we have thousands of organisms that can do this .
階層型 Transformer（ビーム幅 5）	we have these people who can do this for hundreds of thousands of organisms .
階層型 Transformer（ビーム幅 4）	we have these thought agencies who can do it .
階層型 Transformer（ビーム幅 3）	we have hundreds of thousands of organisms that can do this .
階層型 Transformer（ビーム幅 2）	we have these thousand organism that can do that .
階層型 Transformer（ビーム幅 1）	we have thousands of organisms that can do that .