ニューラル機械翻訳における文脈情報の選択的利用

藤井 諒
東北大学 工学部 電気情報物理工学科

1 はじめに

近年，ニューラル機械翻訳（NMT）の登場および発展により翻訳品質は劇的に向上してきた．しかし，大量のデータに基づくニューラルネットワークの学習をもってしてもなお，代名詞の誤訳や省略，生成文間の語彙の一貫性などから生じる不適切な訳出が問題視されている．これらの多くは，現在の翻訳システムの多くが 1 文を見て対応する 1 文を翻訳する 1 文対 1 文（1-to-1）翻訳（図1（a））を行っているために生じると考えられ，機械翻訳における文外文脈利用の重要性が指摘されはじめている [1]．

先行文脈の与え方として原言語側，目的言語側の双方あるいは一方に直前の文を結合することによる 2 文対 2 文（2-to-2）翻訳（図1（b）），2 文対 1 文（2-to-1）翻訳（図1（c））が提案されており，談話関係や語彙間の依存関係を捉えることが示されている [2]．しかし，文脈情報の与え方には他にも多様な手法が考えられ，どのように文脈情報を加えることが有効なのかは未だ十分に検証されていない．

そこで，本研究では文脈情報の与え方が翻訳品質に与える影響について検証を行う．先行研究の 2-to-2, 2-to-1 翻訳に加えて，今回新たに異なる文脈幅を持つ学習データの混ぜ合わせ学習（図1（d），（e））を試みる．文脈を考慮しないベースラインモデルに加えて，文脈を考慮する 4 つの手法で翻訳モデルを学習し，それぞれのモデルがどのように優れているのか，およびどのような問題点を抱えているのかを明らかにすることを目指す．OpenSubtitles2018 コーパス [3] および Japanese-English Subtitle Corpus（JESC） [4] を用いた日英方向の翻訳において混ぜ合わせ学習によるモデルが先行手法の 2-to-2 翻訳に対し有意に高い BLEU スコア [5] を達成することを示す．また，1 文に対し，異なる文脈情報を与えた複数の翻訳結果を得ることで，必要に応じて文脈情報を使い分けることが更なる翻訳精度の向上に有効となる可能性を議論する．

図1: 各手法の学習過程の比較

2 先行研究

本節では，ベースラインおよび提案法で用いる sequence-to-sequence（seq2seq）モデル [6] と文脈情報を考慮した翻訳に関する先行研究について紹介する．

2.1 seq2seq モデル

seq2seq モデルは，入力として与えられる原言語文の系列を固定次元ベクトルに変換するエンコーダと，その情報を基に目的言語文を生成するデコーダからなる．エンコーダおよびデコーダには主に再帰型ニューラルネットワーク（RNN）が用いられる．入力系列 $x = x_1, x_2, ..., x_n$，出力系列 $y = y_0, y_1, y_2, ..., y_m, y_{m+1}$ に対し，seq2seq モデルは以下の条件付き確率をモデル化する．

$$\underset{y}{\operatorname{argmax}} \prod_{t=1}^{m+1} p(y_t | y_{<t}, x)$$

y_0，y_{m+1} は特殊記号であり，それぞれ開始記号，終端記号を表す．モデルは終端記号が出力されるまで，あるいは最大系列長として設定した単語数まで単語を出力する．

2.2 文外文脈を利用する翻訳モデル

ニューラル機械翻訳においてモデルに文脈情報を与える手法としては大きく分けて 2 つの観点から提案がされてきた．

ひとつは，文外文脈に対して別のエンコーダを設け，デコード時に複数のエンコーダからの情報を混ぜ合わせ文脈ベクトルとして用いる方法である．例えば，Voita ら [1] は Transformer [7] の最終層において，前文をエンコードするエンコーダからの出力と通常のエンコーダからの出力を混ぜ合わせることで，正しい代名詞の選択タスクにおいて文脈を使わないモデルを大きく上回る性能を達成した．これは，文脈情報の付加によりモデルが語彙間の依存関係を理解できることを示している．Wang ら [8] は階層的 LSTM を用いて，段階的に文，文章のベクトルを作成し，デコード時に追加文脈として文章ベクトルを与えることが翻訳精度の向上に有効であることを示した．また，Wang らは大域的な文脈情報の必要性が生成単語の曖昧性などに応じて変化することを指摘した．これらの手法では共通して追加文脈をゲート機構を用いて混ぜ合わせており，文脈情報を必要に応じて使い分けることの重要性が示唆されている．

一方で，もうひとつのアプローチとしてはデータ側の工夫が考えられる．Tiedemann ら [2] は原言語文および目的言語文の系列に前文を連結して学習を行うことで，モデルのアーキテクチャに変更を加えることなく BLEU スコアを向上させることができることを示した．この手法は非常に単純であり，ベースとなるモデルに依存しないという利点がある．

3 文脈利用翻訳モデル

本節では本研究で用いる文脈利用翻訳モデルについて述べる．本研究では 4 つのモデルについて検証を行った．

3.1 2-to-2，2-to-1 翻訳

ここでは，2.2節で述べた Tiedemann らが提案した 2-to-2 および 2-to-1 翻訳の詳細を述べる．

2-to-2 翻訳は，原言語側，および目的言語側の双方に文脈として前文を結合し，原言語側の 2 文が目的言語側の 2 文に翻訳されるように学習を行うモデルである（図1（b））．一方で，2-to-1 翻訳は，原言語側のみに前文を付加し原言語側の 2 文目を目的言語に翻訳するように学習を行うモデルである（図1（c））．これらのモデルには文脈と注目文の間に特殊トークンを置くことで，どこまでが文脈情報であるかを教えている．2-to-2 翻訳ではテスト時には出力文に含まれる特殊トークン（図1中の $\langle CONC \rangle$ トークン）で文を分割し，後ろの文を翻訳結果として用いる．

3.2 文脈窓幅の異なるデータに対する混ぜ合わせ学習

必要に応じて先行文脈を用いるかどうかをモデルに判断させることのできる文脈データの与え方は翻訳精度の向上に有効であると考えられる．そこで，学習時に 1 行 1 文のデータに加え，前文を特殊トークンで結合した 1 行 2 文のデータを混ぜ合わせて学習を行う事を試みた．文脈情報としては先行手法と同様に，双方の言語に与える場合，および原言語側のみに与える場合を考えることができる（図1（d），（e））．モデルは文脈がない条件，文脈を与えられた条件の 2 度学習を行う事になるため，一度学習した文に対して文脈情報の必要性を判断し再学習を行う．2-to-2，2-to-1 翻訳モデルでは特殊トークンにより文の区切りを教えられているとはいえ，モデル自身が文単位のアライメント（対応）を学習し推論を行う必要があった．1-to-1 のデータを学習データに加え，モデルに文単位のアライメントを陽に与えることは有用であると考えられる．また，混ぜ合わせ学習を行うことにより，デコード時にも 1 つの文に対し文脈を与えた場合，与えない場合の別々の翻訳結果を与えることができる．これにより，後述する文脈情報の選択的利用を可能にする．先行手法にならい，学習時に 2-to-2 のデータを混ぜたモデルの出力結果に対しては特殊トークンで分割した後ろの文を翻訳結果として用いる．

4 実験

4.1 データセット

機械翻訳における先行文脈活用の可能性を検証するにあたり，文脈情報を保持しているコーパスを対象にする必要がある．そのようなコーパスとしては OpenSubtitles2018 [3] や Visual Storytelling コーパス [9] など複数のコーパスが考えられるが，今回は利用可能な文対の多さから OpenSubtitles2018 を対象とした映画字幕翻訳に取り組む．

このコーパスは opensubtitles.org[*1] に投稿された字幕データのクローリングにより作成された約 210 万文程度からなる日英対訳コーパスとなっている．コーパスには区切り情報がついており，一つの映画に含まれる文のみを取り出すことができる．コーパス全体は約 2600 のストーリーから構成されるが，学習データの品質を保証するため，実験ではこのコーパスに対し JESC [4] に付

[*1] http://www.opensubtitles.org/

属のクリーニングコードを適用し，文対の除去比率が閾値として設定した 3%以下であったストーリーのみを使用した．このクリーニングコードは発話主を表す文頭の [発話主] などの記号を取り除くとともに，辞書を用いて英語文に英語として妥当な単語がどれだけ含まれているかを判別し，その割合により文を除去するかを決定する．今回は文脈情報を保持するため，除去すべきと判断された文に対してもその場で除去を行わず，一時的にタグ付けを行った．無作為に選択した 10 ストーリー約 6000 文ずつを開発用データと評価用データ，残りすべてとなる約 110 万文を学習データとし，学習データおよび開発用データには連続する 5 文に対しタグがつけられなかった文のみを使用した．また，評価用データとしては，同様に映画字幕から構成される JESC も用いた．JESC では用意されている約 2000 文の評価用データをそのまま評価に用いた．

4.2 実験設定

本研究では seq2seq モデルのツール `mlpnlp-nmt`[*2] を用いた．エンコーダは双方向 LSTM で構成され，デコーダと attention 機構は Luong らの seq2seq モデル [10] に準拠している．すべてのモデルに対し，同じハイパーパラメータ，語彙を用いて 20 エポックの学習を行なった．エンコーダおよびデコーダの層数は 2，単語埋め込み層，および LSTM 各層のユニット数は 512 とした．最適化手法には初期学習率 1.0 の確率的勾配降下法（SGD）を用い，13 エポックから 1 エポックごとに 0.7 倍とした．日本語文は半角，全角スペースをそれぞれ $\langle sp \rangle$，$\langle SP \rangle$ の特殊トークンにエスケープ処理し，mecab-ipadic-NEologd 辞書を適用した MeCab ver. 0.996[*3] を用いて分かち書きした．英語文の分かち書きには NLTK [11] を用いた．その後，結合ルール数が 8000 となるように Byte-Pair-Encoding（BPE）[12] を適用したところ語彙数は日本語が 14,305，英語が 8,049 となった．BPE の学習および分割には `subword-nmt`[*4] を用いた．

4.3 実験結果

文脈を考慮しないベースライン（1-to-1），先行手法の 2-to-2，2-to-1 翻訳，および混ぜ合わせ学習を行った時の BLEU スコアを表1に示す．ここで，表中の dec1，dec2 はそれぞれ，テスト時にモデルに対し 1 行 1 文の文脈情報を持たないデータを与えた場合，および前文を結合し文脈情報を与えた場合を表している．これらの手法に加え，前文の代わりに注目文を特殊トークンでつないだ疑似 2-to-2 データを 1-to-1 と混ぜ合わせた疑似データ拡張も行った．つまり，原言語文と目的言語文のペア (s_i, t_i) に対し，$(s_i \ \langle CONC \rangle \ s_i, t_i \ \langle CONC \rangle \ t_i)$ というペアを作成し 1-to-1 データに加えて学習を行った．疑似データ拡張に対するスコアは dec1 のものとなっている．また，表1中の各手法に対するスコアは乱数シードを変えて学習した 5 つのモデルに対するスコアの平均である．

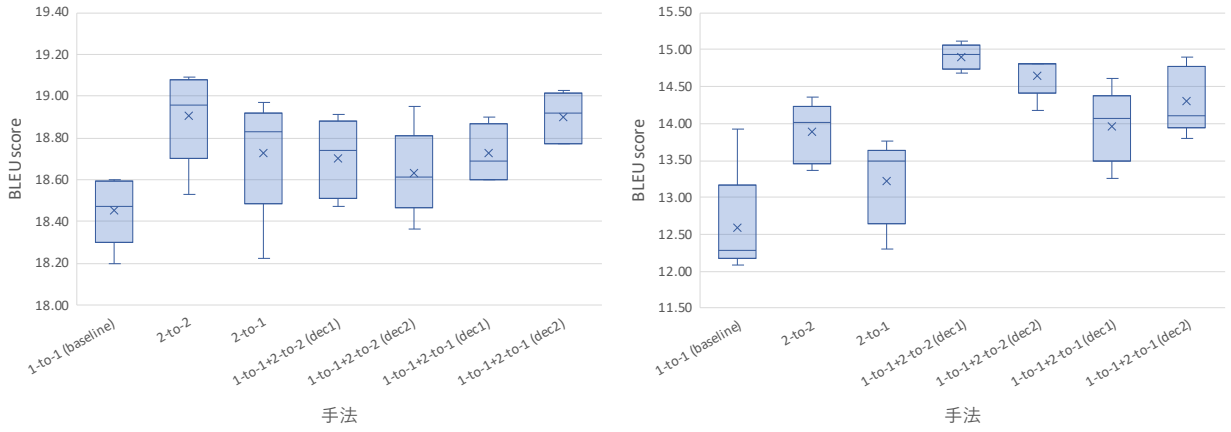
表1: ベースラインおよび提案手法による BLEU スコア（5 モデルの平均）

		OpenSubtitles	JESC
ベースライン	1-to-1	18.45	12.59
Tiedemann（2017）	2-to-2	18.90	13.89
	2-to-1	18.73	13.22
1-to-1+2-to-1	dec1	18.73	13.96
	dec2	18.90	14.30
1-to-1+2-to-2	dec1	18.70	14.91
	dec2	18.63	14.65
疑似データ拡張		18.40	13.98

OpenSubtitles コーパスにおいては，新規手法である 1-to-1+2-to-2 および 1-to-1+2-to-1 混ぜ合わせ学習を行う事による明らかなスコアの向上は見られなかった．一方で，JESC においては文脈を考慮した先行手法（2-to-2）と比較して BLEU スコアで最大 +1.02 ポイントのゲインを確認することができた．2 つのコーパスで異なる結果を示した理由として，JESC にはストーリーの区切り情報が与えられておらず，文脈情報が保持されていない例が散見されるため，混ぜ合わせ学習を行い単文の翻訳結果も学習することにより，モデルがより頑健に推論を行う事ができたと考えられる．

また，OpenSubtitles コーパスおよび JESC に対する BLEU スコアの分布はそれぞれ図2（a），（b）のようになった．図2から，テストデータとして用いたコーパスの種類に関わらず，モデル間のスコアには同様の関係性が見られることがわかる．混ぜ合わせ学習を行ったモデルについて dec1 および dec2 のスコアを比較してみると，学習時のデータの混ぜ方により異なる傾向を示していることがわかる．学習時に，2-to-1 のデータを混ぜ合わせたモデルでは，原言語文側に文脈を与えた場合でも

[*2] https://github.com/mlpnlp/mlpnlp-nmt
[*3] http://taku910.github.io/mecab/
[*4] https://github.com/rsennrich/subword-nmt



(a) OpenSubtitles2018　　**(b)** JESC

図2: 各手法による BLEU スコアの分布

出力される文が長くならない．dec1 と dec2 の違いは先行文の単語への注意をすることができるかどうかのみであるため，スコアの差分は文脈情報によるゲインであると考えられる．一方で，2-to-2 のデータを混ぜて学習した場合にはテスト時に文脈情報を与えた場合に性能の低下が見られた．これは主に，テスト時に前文の翻訳結果も出力する 2-to-2 のデコードは，出力系列が長くなることにより予測誤りの蓄積の影響を受けやすくなるためだと考えられる．

4.4　アンサンブルモデルによる実験結果

続いて，各手法について得られた 5 つのモデルのアンサンブルを行い BLEU を測定した結果，スコアは表2のようになった．

モデルのアンサンブルを行った結果，提案手法は文脈を考慮した先行手法との比較で OpenSubtitles コーパスでは最大 +0.41 ポイント，JESC では +1.90 ポイントの BLEU スコアを記録した．また，系列長が増大することによる計算誤差の影響が抑えられ，1-to-1+2-to-2 混ぜ合わせ学習を行ったモデルについても，テスト時に文脈情報を与える場合（dec2）のスコアが与えない場合（dec1）を上回ることが確認された．1-to-1+2-to-2 混ぜ合わせモデルによる dec2 のデコード結果と，文脈を考慮する先行手法である 2-to-2 翻訳のデコード結果に対し有意水準 $p = 0.05$ のブートストラップ検定を行ったところ，2 つのコーパスに対して，ともに提案手法は先行手法よりも有意に高い翻訳精度であることがわかった．JESC について，疑似データ拡張モデルについても 2-to-2 翻訳より高い結果が得られたことから，提案モデルは文脈情報とデータ量の増加の両方の影響を受けていると考えられる．

表2: ベースラインおよび提案手法による BLEU スコア（アンサンブル）

		OpenSubtitles	JESC
ベースライン	1-to-1	19.84	13.98
Tiedemann（2017）	2-to-2	20.47	15.45
	2-to-1	20.19	14.71
1-to-1+2-to-1	dec1	20.41	15.85
	dec2	20.74	15.99
1-to-1+2-to-2	dec1	20.88	17.27
	dec2	**20.85**	**17.35**
疑似データ拡張		20.13	16.35

4.5　文脈情報の選択的利用

コーパス中には文脈情報を参照できる方がよい場合と文内の情報のみで翻訳を行える場合が存在すると考えられる．混ぜ合わせ学習を行うことで dec1，dec2 の 2 通りの出力を得たのち，良い方の出力を選び続けることができる場合にどれだけ翻訳精度の向上が見込めるかを検証した．

リファレンスを参照できる条件下で，dec1 または dec2 の出力のうち BLEU スコアが高くなる出力を必ず選ぶことができると仮定した場合のオラクルスコアを確かめた．提案法の 1-to-1+2-to-2 混ぜ合わせ学習モデルについて，OpenSubtitles コーパスに対するオラクルスコアは BLEU で 21.96 ポイントとなり，片方の出力結果のみを用いる場合よりもさらに +1.08 ポイントのゲインを見込めることがわかった．この時，dec2 の出力が選ばれた例は 1076 文であった．これはテスト文全体の 6340 文に対して約 17%であり，文脈の理解を必要とする文の割合として妥当な割合であるといえる．これらの結果

は，混ぜ合わせ学習と出力文のバリエーションの利用，および文脈情報の必要性を測る指標による出力の適切な事後選択がさらなる翻訳精度の向上に貢献する可能性を示唆している．

5 事例研究

本節では，ベースラインの 1-to-1 翻訳，先行手法の 2-to-2 翻訳と提案手法の出力を比較し，既存手法がどのような問題点を抱えているのか，また混ぜ合わせ学習を行う事でどのように改善が見られたのかについて述べる．

表3（a）は，先行文脈に含まれる単語が注目文において省略されている場合の翻訳例である．先行手法の 2-to-2，および混ぜ合わせ学習を行いテスト時に文脈を与えられた場合（dec2）では，前文に含まれる「道がない」という情報を利用できているように思われる．このように，文脈情報を与えられる条件では省略された単語の情報を補って翻訳を行う例が散見された．

表3（b）には先行手法の 2-to-2 翻訳に比べ混ぜ合わせ学習の出力結果が良くなった例を示す．この例では 2-to-2 翻訳モデルが文単位アライメントを間違え前文の内容を翻訳結果として提示してしまっている．一方で，混ぜ合わせ学習モデルでは学習時に明示的に文の対応を教えられているため，訳出する文を正しく選択できたと考えられる．また，この例では電話をかける対象を前文の情報から参照し正しい目的語を出力できていた．

6 まとめ

本研究では，先行研究で行われていた 2-to-2 翻訳，2-to-1 翻訳の手法に加え，新たな文脈情報付加の手法として，異なる文脈幅を持つデータ系列の混ぜ合わせ学習を提案し，OpenSubtitles2018 コーパスおよび JESC に対する日英翻訳において，文脈を考慮する先行手法を上回る性能を示した．また，必要に応じて 1-to-1 と 2-to-2 の出力を使い分けることによる翻訳精度のさらなる向上の可能性について確認し，機械翻訳における文脈情報の重要性を再確認した．

謝辞

本研究の一部は株式会社日立製作所の支援を受けて行った．

表3: 各手法による出力の比較（エスケープ処理は元に戻されている）

(a)		
先行文脈		倉庫に戻り　違う道を探したほうがいいかも。 他の通路がない。
注目文		これしか。
参照訳		there is no other way .
ベースライン	1-to-1	this is it .
Tiedemann（2017）	2-to-2	this is the only way .
1-to-1+2-to-2	dec1	this is it .
	dec2	this is the only way .
(b)		
先行文脈		アンデゥルー・ブラウナーの退職の　ファクスを見て驚きました。 退職の手紙？
注目文		そう、電話もつながらない。
参照訳		a resignation ？ yeah , i ca n't get him on the phone
ベースライン	1-to-1	yeah , no phone calls .
Tiedemann（2017）	2-to-2	a retirement letter .
1-to-1+2-to-2	dec1	yeah , i ca n't call him .
	dec2	yeah , i did n't call him .

参考文献

[1] Elena Voita et al. “Context-Aware Neural Machine Translation Learns Anaphora Resolution”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 1264–1274.

[2] Jörg Tiedemann and Yves Scherrer. “Neural Machine Translation with Extended Context”. In: *Proceedings of the Third Workshop on Discourse in Machine Translation*. 2017, pp. 82–92.

[3] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. “OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[4] Reid Pryzant et al. “JESC: Japanese-English Subtitle Corpus”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[5] Kishore Papineni et al. “BLEU: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems 27*. 2014, pp. 3104–3112.

[7] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 5998–6008.

[8] Longyue Wang et al. "Exploiting Cross-Sentence Context for Neural Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2826–2831.

[9] Ting-Hao (Kenneth) Huang et al. "Visual Storytelling". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 1233–1239.

[10] Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 1412–1421.

[11] Edward Loper and Steven Bird. "NLTK: The Natural Language Toolkit". In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 2002.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1715–1725.