目的言語側の文間文脈を考慮した文脈つきニューラル機械翻訳

山岸駿秀　小町守
首都大学東京
{yamagishi-hayahide@ed., komachi@}tmu.ac.jp

1 はじめに

ニューラル機械翻訳（NMT）[1, 9] は従来の機械翻訳システムに比べてより長距離の単語文脈を捉えることができることから，注目を集めている．しかし，ほとんどの NMT は他の手法と同様に文単位での処理を前提としているため，文を超える文間文脈を用いることはできていない．これを受けて，近年，NMT が捉えられる文脈を単語間から文間[1]へと拡張するための研究がいくつか報告されている．

現在提案されている**文脈つき NMT**[2]のモデルの多くは，従来の NMT に加えて文脈文を読み込むための Encoder を備える Multi-Encoder 型である．例えば，Bawden ら [2] は目的言語側の文脈文の情報を Encoder から得ることは，少なくとも英仏翻訳では有用ではないことを示した．他の Multi-Encoder 型 NMT では原言語側のみを対象としているため，目的言語側の文脈文をどう扱うかについての知見は少ない．さらに，先行研究の多くは同じ語族の言語対を使って実験を行っている．離れた言語対であるほど文書構造も変化していると仮定すれば，特に離れた言語対においては目的言語側の文脈も必要であると想定される．

本研究では文脈つき NMT で目的言語側の文脈が有用であるかどうかについて分析を行う．現在の枠組みで目的言語側の文脈が有用でない原因として，「目的言語側の情報は Decoder を通して得るべきであり，Encoder を通して得る現在の手法では不十分なのではないか」という仮説を立てた．この仮説を元に，文脈文を翻訳する際に計算された Decoder の隠れ層を保存しておき，現在の対象文の翻訳時に Attention 機構を介して情報を得るという，重み共有による手法を提案する．これにより，目的言語側の文脈文の情報を Decoder から取得することができる．この手法をいくつかの言語対で検証したところ，目的言語側の文脈は少なくとも原言語側の文脈と同程度に有用であり，言語によってはより有用であることがわかった．また，重み共有は原言語側に対しても有用であることがわかったため，Multi-Encoder 型の文脈つき NMT をコンパクトに実現できる可能性を示した．

[1]以後，文間に流れる意味のつながりを単に文脈とする．本稿では特に前の文脈に着目し，ある時刻 i で翻訳される文を対象文，$i-1$ の入力文または出力文を文脈文と呼ぶことにする．

[2]文脈を用いた一連の研究は Context-aware NMT と表記されることが多いが，本稿では文脈つき NMT と表記する．

2 文脈つきニューラル機械翻訳

Bawden ら [2] の提案したモデルをもとに，Multi-Encoder 型の文脈つき NMT を構築した．図 1 に提案手法の概略を示す．

文書を $D = (X^1, Y^1), ..., (X^i, Y^i), ..., (X^L, Y^L)$ と表す．ここで，L は文書に含まれる文対の数を示す．X^i や Y^i はそれぞれ入力文と出力文であり，それらはさらに文長 M^i の単語列 $X^i = x^i_1, ..., x^i_m, ..., x^i_{M^i}$ のように分解できる．目的関数は以下の確率を最大化することである．

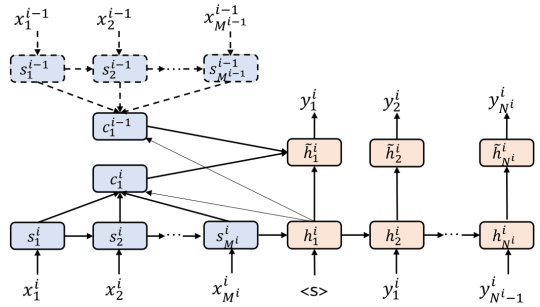
$$p(Y^i|X^i, Z^{i-1}) = \prod_{n=1}^{N^i} p(y^i_n|y^i_{<n}, X^i, Z^{i-1}) \quad (1)$$

ここで，Z^{i-1} は実験設定に応じて文脈文 X^{i-1} または Y^{i-1} が入るものとする．それぞれの確率 p は次のように計算する．

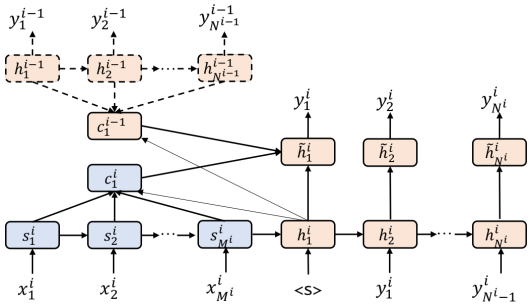
$$p(y^i_n|y^i_{<n}, X^i, Z^{i-1}) = \mathrm{softmax}(W_\mathrm{o}\tilde{\boldsymbol{h}}^i_n) \quad (2)$$

$$\tilde{\boldsymbol{h}}^i_n = W_\mathrm{h}[\boldsymbol{h}^i_n; \boldsymbol{c}^i_n; \boldsymbol{c}^{i-1}_n] \quad (3)$$

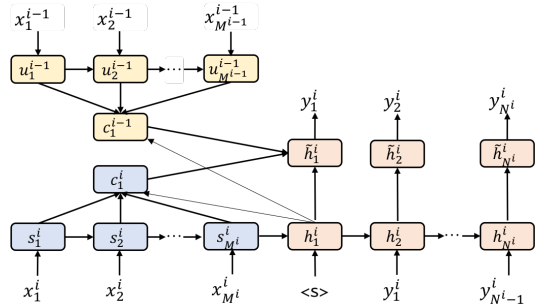
Multi-Encoder 型の文脈つき NMT は，文脈文のための Encoder の情報を扱うために追加の Attention 機構 $\boldsymbol{c}^{i-1}_t$ が備わる．その他の Decoder 及び Encoder の処理の詳細は Luong らの dot 型 global attention モデル [4] と同様である．



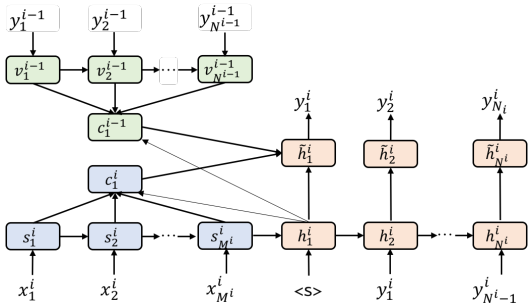
(a) Shared Source 型



(b) Shared Target 型



(c) Separated Source 型



(d) Separated Target 型

図 1: 提案手法の概略図

Separated 型　このモデルは，これまで提案された文脈つき NMT と同様に，対象文の前文を読み込むための文脈 Encoder を追加する．原理的には任意の数の文脈文を読み込むことができるが，今回は簡易化のために対象文の 1 つ前の入力文または出力文が文脈 Encoder に入力されるものとする．文脈 Encoder は入力文 Encoder と同じ手法であるが，それぞれ異なる重みを与えるものとする．したがって，この 2 つの Encoder はそれぞれ学習が必要である．本研究では，原言語側の文脈文を用いる場合は Separated Source 型，目的言語側の文脈文を用いる場合は Separated Target 型と表記する．

Shared 型　1 つ前の文を翻訳したときに計算した Encoder または Decoder の隠れ層を保存しておき，現在の文の処理時に，保存した隠れ層に対して Attention を計算する．特に文脈文として目的言語側を選択した際には，文脈文を入力したときの隠れ層ではなく出力したときの隠れ層の情報を得ることを想定している．さらに，このモデルは追加の重みや文脈文をもう一度読み込む処理を必要としないため，モデルの大きさや計算時間を削減できる．Separated 型と同様に，それぞれ Shared Source 型，Shared Target 型と表記する．

また，Shared 型では双方の文脈文を用いる Shared Mix 型の実験も行う．このときの Attention ベクトルは，$\boldsymbol{c}^{i-1} = \boldsymbol{c}^{i-1}_{\text{source}} + \boldsymbol{c}^{i-1}_{\text{target}}$ のように計算する．その他の機構は他の Shared 型と同様である．

3 実験

3.1 コーパス

実験には主に IWSLT2017 の独英，中英，日英のデータセット[3]を使用した．それぞれ TED の動画の字幕の対訳コーパスであり，実験では各動画（各講演）を 1 文書として処理した．日本語文は MeCab[4]（辞書として IPADic 2.7.0 を使用），中国語文は jieba[5]，その他の言語では Moses[6]に含まれる `tokenizer.perl` をそれぞれ用いて単語分割をした．100 語以上の文を含む文書を削除した結果，各コーパスの文数は表 1 のようになった．なお TED コーパスは複数のテストデータを含むが，今回は 2014 年度版を使用した．各文は，さらに Byte Pair Encoding（BPE）[7] を結合回数を 32,000 回[7]で実行し，文中の単語をサブワード化した．

さらに，ユーザの投稿したレシピの日英対訳コーパスである Recipe Corpus[8]を追加的に用いた．BPE の結合回数を 8,000 回に設定した以外については，全て TED コーパスと同じ処理を行った．

3.2 実験設定

RNN 型の文脈なし NMT をベースラインとした．Encoder は 2 層 bi-LSTM，Decoder は 2 層 uni-LSTM を

[3]https://wit3.fbk.eu/mt.php?release=2017-01-trnted
[4]http://taku910.github.io/mecab/
[5]https://github.com/fxsjy/jieba
[6]http://www.statmt.org/moses/
[7]実装の都合上，NMT の語彙サイズは 32,000+文字種数である．
[8]http://lotus.kuee.kyoto-u.ac.jp/WAT/recipe-corpus/

実験	文数 Train	Dev	Test	Baseline	Separated 型 Source	Target	Shared 型 Source	Target	Mix
TED 独英	203,998	888	1,305	26.55	26.15	26.59	*27.14	***27.31**	*27.29
TED 英独				21.26	20.82	20.82	21.66	**21.93**	21.47
TED 中英	226,196	879	1,297	12.54	12.41	12.76	*13.13	***13.57**	*13.20
TED 英中				8.97	8.92	8.81	9.38	***9.62**	9.46
TED 日英	194,170	871	1,285	5.84	*6.76	*6.31	*6.98	***7.02**	*6.76
TED 英日				8.40	8.55	8.26	**8.64**	8.63	8.59
Recipe 日英	108,990	3,303	2,804	25.34	*26.56	*26.78	***26.96**	*26.87	*26.75
Recipe 英日				20.81	*21.94	*21.48	*21.94	***21.97**	*21.86

表 1: 各実験で使用したデータの統計量と，得られた BLEU スコア

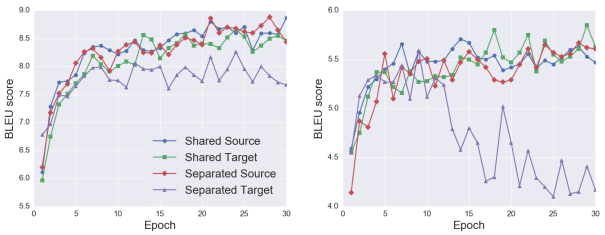
それぞれ用いた．隠れ層や埋め込み層の次元は 512 に設定した．ドロップアウトを確率 0.2 で適用した．最適化には AdaGrad を初期学習率 0.01 で用いた．各バッチは 128 文書を含む．バッチ内の文書の同じ文番号ごとにミニバッチを形成し，文番号の昇順にミニバッチ処理を行った．このため，同じ文書集合を処理中のミニバッチの大きさは 128 を最大値として減少する．以上の設定は全てのモデルの実験で同様である．文脈文に対する Attention には入力文に対する Attention と同じく dot 型の global attention を用いた．文番号が 1 のミニバッチを処理する際には文脈文が存在しないため，$c^0 = \mathbf{0}$ として計算した．

全ての文脈つき NMT はベースラインのモデルで事前学習を行う．ベースライン，提案手法ともに 30 epoch の学習を行った．Test 時にはビーム探索を窓幅を 5 にして使用した．全ての結果は BLEU [6] で評価した．全ての文脈つき NMT の実験は乱数のシードを変えて 2 回ずつ行い，得られた 2 つのスコアの平均を示す．また，文脈つき NMT の結果は bootstrap resampling を用いてベースラインとの統計的有意差を測定する．

3.3 結果

表 1 に結果を示す．ここで，*は 2 回の実験結果がともに統計的有意差（$p < 0.05$）があることを表す．Shared Target 型では全ての言語対において性能が向上した．いくつかの言語対では Separated Target 型でもベースラインを上回る性能を出しているが，その増分は Shared Target 型と比べると小さい．したがって，仮説の通り，目的言語側の文脈は Decoder から取り込むことが良いと考えられる．

原言語側の文脈については，どちらの手法でも Encoder から取り込むことができるため，性能にあまり違いが出ないと想定していた．しかし，Separated 型と比べてパラメータ数が少ないにもかかわらず，Shared 型がより高い性能を発揮することがわかった．したがって，どちら側の文脈を用いる場合であっても，Shared 型によって計算コストを抑えつつ高性能なモデルを構築できる．



(a) TED 中英翻訳　(b) TED 英中翻訳[9]

図 2: 中英・英中翻訳の学習時における BLEU の変化

4 考察

言語対の影響　結果は言語対によって大きく異なる．例えば，日英･英日翻訳では TED･Recipe ともに Shared 型の中での性能差が小さい．したがって，この言語対ではどちら側の文脈も同程度に必要であると言える．日本語文は省略される情報が多いため，日英翻訳では文脈を考慮することで不足した情報を補うことができ，英日翻訳では重複する情報を削除することで流暢な生成が可能となる．また Separated 型の結果から，原言語側の文脈が有用なのではなく英語側の文脈が有用である可能性を指摘する．

学習の収束　図 2 に，TED 中英・英中翻訳の 1 回目の学習時における Dev データの BLEU 推移を示す．Shared 型や Separated Source 型の学習はとても安定していると言える．Shared 型は文脈 Encoder を学習する必要がなく，事前学習によって学習初期から一定の質が担保された隠れ状態ベクトルを用いることができる．Separated Source 型は文脈 Encoder を学習する必要があるが，今回の実験では，学習時も検証・評価時も同程度の質の入力文を使うことができる．したがって，学習の収束の観点から考えると，従来手法の拡張であ

[9] 凡例については左図と同様である．

る Separated Source 型と提案手法である Shared 型は扱いやすい手法であると言える．

一方，Separated Target 型の学習は不安定であり，特に目的言語が英語でない場合に顕著になる傾向があった．Separated Target 型の学習時には正解の文脈文を与えられているが，検証・評価時には実際の出力結果が文脈文として与えられている．このため，Decoder が目的言語らしい文を生成できるようになる前に文脈 Encoder が目的言語文の文法を学習してしまい，正しく読み込める Encoder で間違った文を読み込んだ結果，有用な文脈情報にならなかったのではないかと考えられる．したがって，Separated Target 型では生成しにくいかつ読み込みやすい言語を目的言語にした場合に学習が収束しにくい可能性がある．

重み共有の影響　NMT では，多層にすることで性能が向上することが知られている．Dabre ら [3] は，n 層にする際に 1 つの重みを用いて再帰的に n 回計算しても性能が低下しないことを発見した．また，重みが学習時の層数を記憶し，1 つの重みを用いても各層の計算時に層ごとの役割を模倣している可能性を示し，層方向の重みを共有できることを示した．

Shared 型は，Encoder または Decoder と文脈 Encoder の間で重み共有をしている．したがって，Shared 型は Dabre らの研究を文書の時系列方向に対して拡張したものとして扱える可能性がある．

5　関連研究

Tiedemann ら [8] は隣接する 2 文を特殊記号で 1 文に結合することで，従来の NMT の枠組みで文脈を扱えるか検証した．著者らは結合した文から結合した文を生成する実験によって，代名詞翻訳では文脈が必要になることを示した．Müller ら [5] は，それまでに提案された文脈つき NMT を独自に作成した英独翻訳用の代名詞翻訳のテストデータで再検証した．彼らは Tiedemann らの手法を再実験した結果から，目的言語側の文脈が必要である可能性を示した．本研究の Shared Target 型は，目的言語側の文脈文を Multi-Encoder で用いる方法を示すものである．

Voita ら [10] は Transformer [9] を用いて文脈つき NMT を試した．実験結果から文脈つき NMT が照応解析をしている形跡を発見し，さらに既存の照応解析ツールと同程度の性能を得られる可能性を示した．Zhang ら [11] も同時期に同様の手法を提案した．こちらでは，大規模な文脈なしコーパスで事前学習したモデルを小規模な文脈ありコーパスで調整することで性能が向上した．これにならい，本研究でも事前学習を行った．これらの研究とは Transformer を用いた点や目的言語側の文脈を用いない点などが異なる．

6　おわりに

本研究では，文脈つきニューラル機械翻訳における目的言語側の文脈の有用性について調査した．先行研究では目的言語側の文脈を Encoder を通して扱っていたが，Decoder を通して扱う手法を提案した．実験から，目的言語側の文脈は提案手法によって原言語側の文脈と同程度かそれ以上に有用であることがわかった．また，提案手法は原言語側の文脈を用いる場合であっても，より低資源な計算環境で文脈つきニューラル機械翻訳を実現できる．今後は，さらに多様なデータセットや，1 文以上の文脈文の有用性を調査したい．

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In *NAACL*, 2018.

[3] Raj Dabre and Atsushi Fujita. Recurrent stacking of layers for compact neural machine translation models. In *AAAI*, 2019.

[4] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[5] Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *WMT*, 2018.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016.

[8] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *DiscoMT*, 2017.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*. 2017.

[10] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In *ACL*, 2018.

[11] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the Transformer translation model with document-level context. In *EMNLP*, 2018.