文脈考慮型ニューラル機械翻訳における最適文脈文選択法*

木村 龍一郎† 飯田 頌平† 崔 鴻翌† 洪 博軒† 宇津呂 武仁† 永田 昌明‡
†筑波大学大学院 システム情報工学研究科 ‡NTT コミュニケーション科学基礎研究所

1 はじめに

ニューラル機械翻訳 (NMT) モデルは近年の研究によって大きく進歩している [13, 9, 16]. しかし, これら標準的な NMT モデルは原言語一文を目的言語一文に翻訳するよう設計されているため, 複数文にわたる文脈を考慮した翻訳は不可能である. この問題に対処するため, 文脈として追加で文を入力できる文脈考慮型の翻訳モデルがいくつか提案された [14, 7, 10, 11, 1, 17, 15]. これらのモデルは入力できる文脈の長さによって大別することが可能である. 最も典型的なモデルは, 直前の文を文脈とみなして入力するモデルで, すべての入力を一つのエンコーダで処理するモデル [14], および, 文脈用に別のエンコーダを用意するモデル [7, 1, 17] に分けられる. より広い文脈を考慮するモデルとしては, 直前 3 文を考慮するモデル [11], 文書中の先行文全体を考慮するモデル [15], および, 文書全体を考慮するモデル [10] に分けられる. いずれのモデルも, 原言語一文のみを考慮するモデル (以下, 1-to-1 翻訳モデルと呼ぶ) と比較してより高い翻訳精度を達成している. その中で, Tiedemann らの文脈考慮型 NMT モデルは, 直前の文を現言語文に連結した対訳対を用いて翻訳モデルを訓練するという簡易なもので, 2-to-2 翻訳モデルと呼ばれる [14]. 2-to-2 翻訳モデルは 1-to-1 翻訳モデルとほぼ同等の簡易なモデルである点において他の文脈考慮型モデルより優れるものの, GPU メモリサイズの制約のために 2 文より広い範囲を文脈とすることが難しく, 文脈として考慮できる範囲が限定されるという欠点がある.

本論文では, 前後 5 文までを文脈文とした 2-to-2 翻訳モデルによるオラクル翻訳の BLEU スコアを分析し, 2-to-2 翻訳モデルにおいて直前の一文よりも広い文脈を考慮する文脈考慮型 NMT モデルの有効性を示す. そして, 2-to-2 翻訳モデルを拡張するため前後 5 文までの中から適切な文脈を選択する方法を提案する. 評価実験の結果として, オラクル翻訳の BLEU スコアには及ばないものの, 直前のみを考慮する 2-to-2 翻訳モデルと比較して, 強制逆翻訳確率の最大化によって有意に BLEU が改善することを示す.

2 文脈考慮型 NMT のオラクル翻訳

2.1 拡張文脈

2-to-2 翻訳モデルにおいては, 原言語文の直前の文を文脈文とみなして, 連結記号 ⟨CONCAT⟩ を介して文脈文と原言語文と連結して翻訳する. 2-to-2 翻訳モデルの BLEU スコアは 1-to-1 翻訳モデルを上回るものの, 文脈として直前 1 文までしか考慮できない点が問題である. そこで, 本論文では, より広い範囲を文脈として考慮することにより, 2-to-2 翻訳モデルを拡張する. 具体的には, ベースラインである 1-to-1 翻訳モデルによる翻訳 $\boldsymbol{y_{11}}$ に加え, 直前 5 文を文脈文とする 2-to-2 翻訳モデルによる訳文 $\boldsymbol{y_{22}^{-1}}, \ldots, \boldsymbol{y_{22}^{-5}}$, および, 直後 5 文を文脈文とする 2-to-2 翻訳モデルによる訳文 $\boldsymbol{y_{22}^{+1}}, \ldots, \boldsymbol{y_{22}^{+5}}$ を生成し, これらを訳文候補集合とする [1]. 以上の 11 個の訳文候補集合の中から最適な訳文を選択する. 提案手法による訳文候補集合における翻訳精度の上限を示すため, 訳文候補集合中で BLEU スコア (文単位の翻訳精度を表す指標である sentence-BLEU) 最大となる訳文 (オラクル翻訳と呼ぶ) を選定し, オラクル翻訳の BLEU スコア, および, 文脈文位置の分布を分析する.

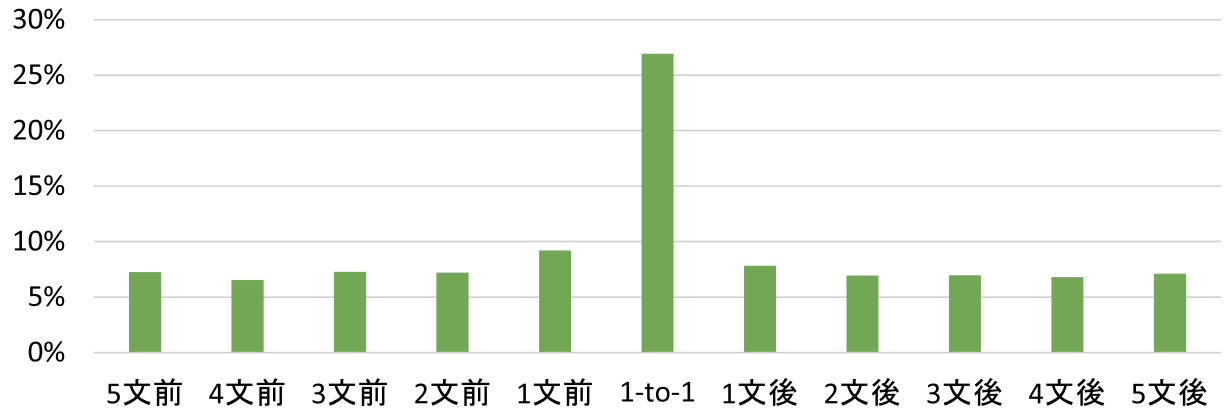
2.2 データセットと実験条件

映画字幕の公開データである Opensubtitles2018 [8] を使用し, 文脈付き対訳文の作成法 [14] に従い, 合計で 2,083,576 文の日英対訳を作成した. 映画等の作品単位で 90%を訓練用, 5%を開発用, 5%を評価用として無作為に分割し, 合計で 1,876,624 対訳文を訓練用, 104,379 対訳文を開発用, 102,573 対訳文を評価用とした. オラクル翻訳の作成および BLEU スコア評価にお

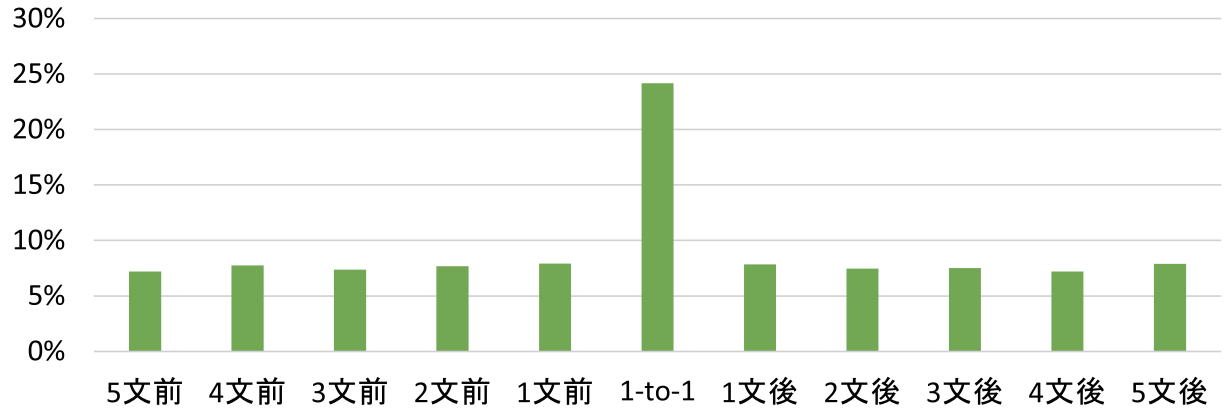
*A Method of Selecting Informative Context Sentences in Context based NMT

†Ryuichiro Kimura, Shohei Iida, Hongyi Cui, Po-Hsuan Hung, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Masaaki Nagata, NTT Communication Science Laboratories, NTT Corporation, Japan

[1] 各 $\boldsymbol{y_{22}^{i}}$ $(i = \pm 1, \ldots, \pm 5)$ は, 原言語文を起点として, 前後 5 文中の i 文目 $\boldsymbol{c^{i}}$ $(i = \pm 1, \ldots, \pm 5)$ を文脈文とする 2-to-2 翻訳モデルによる訳文を表す.



(a) 日英



(b) 英日

図 1: オラクル翻訳における最適文脈文位置の分布

いては，評価文のうち 10,000 文を使用した [2] [3].

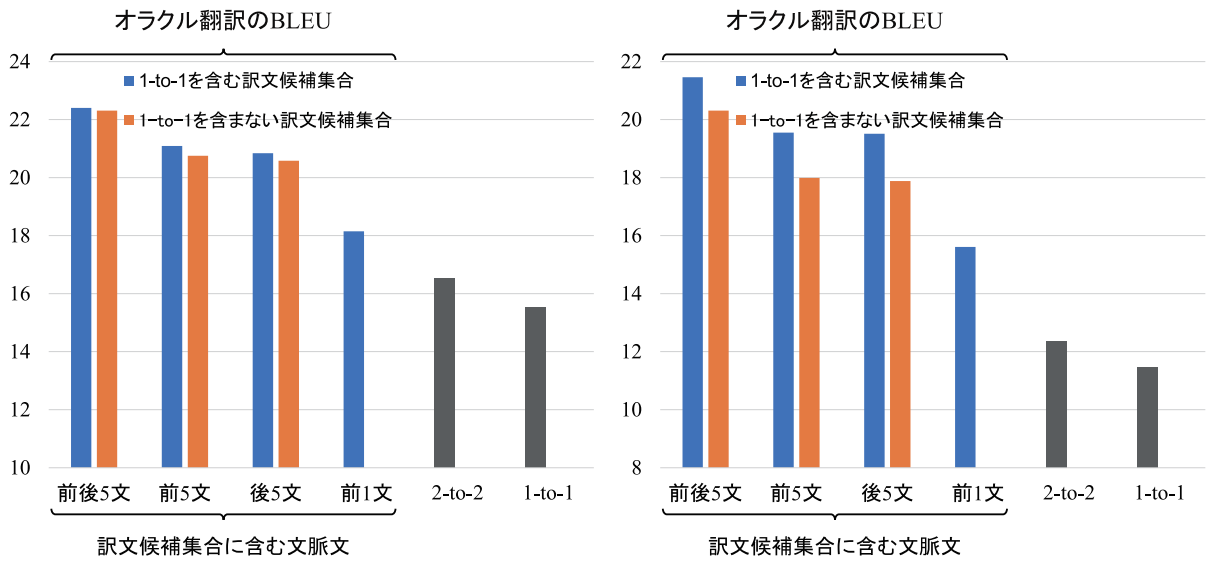
2.3 オラクル翻訳の分布

オラクル翻訳における訳文候補選択結果の分布を図 1 に示す．オラクル翻訳 10,000 文のうち，訳文候補集合の中から sentence-BLEU 最大である訳文が一意に定まったものは日英で 39%，英日で 46%であった．sentence-BLEU 最大の訳文が一意に定まらない評価文を集計から除いて最適文脈文位置の分布を求めた．1-to-1 翻訳モデルによる訳文が sentence-BLEU 最大である割合は日英で 27%，英日で 24%であった．また，2-to-2 翻訳が sentence-BLEU 最大となる文脈位置が 1 文前である割合は日英で 9%，英日で 8%，2 文前から 5 文前である割合は日英で 28%，英日で 30%，1 文後から 5 文後である割合は日英で日英で 36%，英日で 38%であった．この結果より，考慮する文脈として後方を追加する方式の有効性が示された．同様に，2-to-2 翻訳において 2 文前から 5 文前，または，1 文後から 5 文後のいずれかが最適文脈文となる割合は日英で 64%，英日で 68%であった．この結果より，2-to-2 翻訳モデルにおいて直前より広い文脈を考慮する方式の有効性が示された．

1-to-1 翻訳モデルによる訳文 ($\boldsymbol{y_{11}}$)，1 文前を文脈文とする 2-to-2 翻訳モデルによる訳文 ($\boldsymbol{y_{22}^{-1}}$)，および，オラクル翻訳の BLEU スコアを図 2 に示す．オラクル翻訳においては，訳文候補の数を増やすほど BLEU スコアが改善した．具体的には，$\boldsymbol{y_{11}}$ および $\boldsymbol{y_{22}^{-1}}$ に加えて $\boldsymbol{y_{22}^{-2}}, \ldots, \boldsymbol{y_{22}^{-5}}$，$\boldsymbol{y_{22}^{1}}, \ldots, \boldsymbol{y_{22}^{5}}$ を訳文候補とする

[2]翻訳モデルの訓練においては，2-to-2 翻訳時の文脈文連結後の原言語文および目的言語文のいずれかが 50 単語を超える文については除外した．

[3]詳細な実験条件は次の通り: 英語の tokenization には Moses tokenizer [4]，日本語の形態素解析には MeCab (`http://taku910.github.io/mecab/`) を使用．翻訳モデル作成には OpenNMT-py を使用．語彙は頻度上位 5 万語を使用．単語分散表現は 512 次元，エンコーダ・デコーダとも各 6 層，バッチサイズを 4,096，drop out rate を 0.3 として，100,000 エポックの訓練を行う．Adam optimizer を使用．ハードウェアは NVIDIA Tesla P100 16GB GPU 1 枚を使用．BLEU の測定及び有意差検定には MTEval Toolkit (`https://github.com/odashi/mteval`) を使用，sentence-BLEU の測定には Moses decoder の sentence-bleu.cpp を使用．



(a) 日英　　(b) 英日

図 2: オラクル翻訳・2-to-2 翻訳 (1 文前を文脈文とする)・1-to-1 翻訳の BLEU

ことにより，オラクル翻訳の BLEU スコアは英日で 6 ポイント，日英で 4 ポイント改善した (表 1)．この結果より，2-to-2 翻訳モデルにおいては，より広い範囲を文脈として考慮することで，翻訳精度の上限が改善すると言える．

3　最適文脈文選択法

2 節では訳文候補 $\boldsymbol{y_{11}}, \boldsymbol{y_{22}^{\pm 1}}, \ldots, \boldsymbol{y_{22}^{\pm 5}}$ の中から適切な訳文を選ぶことにより，BLEU スコアが大きく改善可能であることを示した．そこで本節では，この訳文候補の中から最適な訳文を選択する手法について述べる [4] [5]．

3.1　強制逆翻訳確率最大化

強制逆翻訳確率は，順方向の翻訳結果を原言語文に翻訳し直すという制約付きで逆方向に翻訳したときの生成確率として定義される．単後長 n の原言語文を $\boldsymbol{x}$，文脈文を $\boldsymbol{c}$ としたときの順方向の訳文を $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{c})$ とする．このときの原言語文の単語 x_j $(1 \leq j \leq n)$ の強制逆翻訳確率は次式となる．

$$b_j \quad = \quad -\log p(x_j | x_{<j}, \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{c}))$$

訳文 $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{c})$ の強制逆翻訳確率は b_j の和

$$\mathrm{B}(\boldsymbol{x}, \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{c})) \quad = \quad \sum_j b_j$$

となる．強制逆翻訳確率最大化法においては，訳文の強制逆翻訳確率が高いほど意味的に原言語文に近くなるという仮説に基づき，強制逆翻訳確率を最大化する訳文を選択する [6]．

3.2　逆翻訳 sentence-BLEU 最大化

Rapp らは逆翻訳後の自動評価結果を尺度として順方向の翻訳結果を評価する手法を提案した [12]．逆翻訳 sentence-BLEU 最大化法においては，この手法に基づき，訳文候補の逆翻訳 $\boldsymbol{x_{11}}, \boldsymbol{x_{22}^{\pm 1}}, \ldots, \boldsymbol{x_{22}^{\pm 5}}$ のうち，原言語文 $\boldsymbol{x}$ との間の sentence-BLEU を最大化する訳文を選択する．

4　評価

強制逆翻訳確率最大化，および，逆翻訳の sentence-BLEU 最大化による文脈文選択後の BLEU スコアを表 1 に示す．文脈文選択の対象となる訳文候補集合としては，(1) 1-to-1 翻訳 $\boldsymbol{y_{11}}$，および，1 文前を文脈文とする 2-to-2 翻訳 $\boldsymbol{y_{22}^{-1}}$ の組，(2) $\boldsymbol{y_{11}}$，および，1 文前から 5 文前を文脈文とする 2-to-2 翻訳 $\boldsymbol{y_{22}^{-1}}, \ldots, \boldsymbol{y_{22}^{-5}}$ の集合，(3) $\boldsymbol{y_{11}}$，および，1 文後から 5 文後を文脈文とする 2-to-2 翻訳 $\boldsymbol{y_{22}^{+1}}, \ldots, \boldsymbol{y_{22}^{+5}}$ の集合，(4) $\boldsymbol{y_{11}}$，およ

[4] 本節で述べる手法以外の関連研究として，Li らは，順方向翻訳時の生成確率，強制逆翻訳時の生成確率，および，目的言語側の言語モデルによる生成確率，および，訳文長の線形和尺度によって訳文候補をリランキングする手法を提案している [6]．この手法を一部導入した手法によって最適文脈文選択を行った結果の BLEU スコアは，本論文で述べる手法とほぼ同等であった．

[5] 本論文の手法は，文脈考慮型 NMT モデルの中でも，1 文前を文脈文とする 2-to-2 翻訳モデルよりも広い範囲の文脈を考慮する点においては，[11, 15, 10] に近い試みであると言える．

[6] 後藤らは，NMT モデルにおける訳抜け検出において，強制逆翻訳確率比を用いる手法を提案した [3]．本論文は，文脈考慮型 NMT において強制逆翻訳確率を用いて最適文脈を選択する点が [3] とは異なる．

表 1: BLEU 評価 (強制逆翻訳確率最大化 / 逆翻訳の sentence-BLEU 最大化) (** は 1 文前を文脈文とするベースライン 2-to-2 翻訳モデルに対して有意差あり ($p < 0.01$) を示す)

訳文候補集合	BLEU		オラクル翻訳 BLEU	
	英日	日英	英日	日英
1-to-1	11.48	15.52	—	—
2-to-2 (1 文前)	12.36	16.52	—	—
1-to-1 + 2-to-2 (1 文前)	13.24** / 12.87**	17.12** / 16.61	15.61	18.15
1-to-1 + 2-to-2 (1 文前 ∼ 5 文前)	13.32** / 13.44**	17.20** / 16.65	19.55	21.09
1-to-1 + 2-to-2 (1 文後 ∼ 5 文後)	13.24** / 13.20**	17.10** / 16.45	19.51	20.84
1-to-1 + 2-to-2 (5 文前 ∼ 5 文後)	12.75** / 13.09**	17.16** / 16.50	21.46	22.40

び，前後 5 文を文脈文とする 2-to-2 翻訳 $\boldsymbol{y_{22}^{\pm 1}}, \ldots, \boldsymbol{y_{22}^{\pm 5}}$ の集合，の 4 通りを評価した．

1 文前を文脈文とするベースライン 2-to-2 翻訳 $\boldsymbol{y_{22}^{-1}}$ に対して，強制逆翻訳確率最大化によって英日・日英方向で BLEU スコアが有意に改善した．一方，逆翻訳の sentence-BLEU 最大化によって英日方向のみ BLEU スコアが有意に改善した．これらの結果から，強制逆翻訳確率を最大化する文脈文を選択することで翻訳精度が改善することがわかった．しかし，前後 5 文の範囲でこれらを文脈文とする翻訳候補を追加したとき，オラクル翻訳の BLEU は大きく改善するにも関わらず，強制逆翻訳確率最大化では大きな改善が見られなかった．この結果は提案手法が広い文脈を十分考慮できていないことを示しており，さらなる改善を要する．

5 おわりに

本論文では，2-to-2 翻訳モデルにおいても，文脈文の選択において従来手法よりも広い文脈を考慮することにより BLEU が改善することを明らかにした．また，前後 5 文の範囲で最適な訳文を求める文脈文の選択手法として，強制逆翻訳確率最大化法，および，逆翻訳の sentence-BLEU 最大化法を提案した．1-to-1 翻訳モデル，および，1 文前を文脈とする 2-to-2 翻訳モデルと比較して，強制逆翻訳確率最大化によって英日・日英方向ともで BLEU スコアが有意に改善した．今後の課題として，原言語文と文脈文との間の意味的つながりの強さを利用する手法として，BERT [2] 等の言語モデルの生成確率を組み合わせる手法，および，共参照関係 [5] を文脈文選択に導入する手法を検討する．

参考文献

[1] R. Bawden, R. Sennrich, A. Birch, and B. Haddow. Evaluating discourse phenomena in neural machine translation. In *Proc. NAACL-HLT*, pp. 1304–1313, 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *CoRR*, Vol. abs/1810.04805, 2018.

[3] 後藤功雄, 田中英輝. ニューラル機械翻訳での訳抜けした内容の検出. 自然言語処理, Vol. 25, No. 6, pp. 577–597, 2018.

[4] P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

[5] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *Proc. EMNLP*, pp. 188–197, 2017.

[6] J. Li and D. Jurafsky. Mutual information and diverse decoding improve neural machine translation. In *CoRR*, Vol. abs/1601.00372, 2016.

[7] J. Libovický and J. Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proc. 55th ACL*, pp. 196–202, 2017.

[8] P. Lison, J. Tiedemann, and M. Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proc. 11th LREC*, pp. 1742–1748, May 7-12, 2018 2018.

[9] M. Luong, I. Sutskever, O. Vinyals, Q. V. Le, and W. Zaremba. Addressing the rare word problem in neural machine translation. In *Proc. 53rd ACL*, pp. 11–19, 2015.

[10] S. Maruf and G. Haffari. Document context neural machine translation with memory networks. In *Proc. 56th ACL*, pp. 1275–1284, 2018.

[11] L. Miculicich, D. Ram, N. Pappas, and J. Henderson. Document-level neural machine translation with hierarchical attention networks. In *Proc. EMNLP*, pp. 2947–2954, 2018.

[12] R. Rapp. The back-translation score: Automatic mt evaluation at the sentence level without reference translations. In *Proc. 47th ACL and 4th IJCNLP*, pp. 133–136, 2009.

[13] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural machine translation. In *Proc. 27th NIPS*, pp. 3104–3112, 2014.

[14] J. Tiedemann and Y. Scherrer. Neural machine translation with extended context. In *Proc. 3rd DiscoMT*, pp. 82–92, 2017.

[15] Z. Tu, Y. Liu, S. Shi, and T. Zhang. Learning to remember translation history with a continuous cache. *Transactions of ACL*, Vol. 6, pp. 407–420, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. 30th NIPS*, pp. 5998–6008, 2017.

[17] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov. Context-aware neural machine translation learns anaphora resolution. In *Proc. 56th ACL*, pp. 1264–1274, 2018.