

# 逆翻訳によるデータ拡張に基づく 文脈考慮型ニューラル機械翻訳

杉山 普<sup>1,a)</sup> 吉永 直樹<sup>2,b)</sup>

**概要:** 会話や講演など省略表現の多いドメインや日英のように言語的に距離の遠い言語対の翻訳では、代名詞のゼロ照応や多義語など原言語では曖昧な情報を目的言語で詳細化する必要がある。そのため、文を越えた文脈の考慮が必要となる場合が多く、訳したい文の前後の文を追加情報として参照しながら翻訳する文脈考慮型の翻訳モデルが研究されている。しかしモデルの学習に必要な対訳コーパスの作成コストの高さもあり、現状では高精度のモデルの学習に必要な大規模な文脈付きの対訳コーパスが利用できない状況にある。そこで本研究では大量の目的言語の単言語コーパスを逆翻訳することで疑似対訳コーパスを生成し、これを利用して文脈考慮型ニューラル翻訳モデルの学習を行う。実験では IWSLT2017 TED Talk コーパスを対訳コーパス、BookCorpus を目的言語の単言語コーパスとして提案手法の学習と評価を行い、その有効性を検証した。

## 1. はじめに

ニューラル機械翻訳の発展により文単位の翻訳精度は大きく向上した。近年では翻訳時に文外の文脈を曖昧性解消の手掛かりとして参照することで翻訳性能の向上を図る文脈翻訳モデルが研究されている [1], [2], [3], [4]。翻訳に文脈を取り入れる簡素な手法として、原言語および目的言語それぞれにおいて注目文（翻訳したい原言語文とその翻訳結果である目的言語文）とその直前の文を特殊な記号 <CONC> を挟んで連結して一文とみなし、文単位の翻訳を行うものがある [1]。また注目文と文脈文を別々のエンコーダから入力する手法も提案されている [2], [3], [4]。

文脈考慮型モデルを学習する際の問題として学習データの不足がある。文脈考慮型モデルは文単位モデルと比較してタスクの入出力の空間が広く、写像の学習により大規模なデータが必要となる。現状利用できる大規模対訳コーパスは、既存の対訳文書間で文のアラインメントを取ってアラインメントの確率が高いペアを抽出したのことが多い [5], [6]。しかし、文脈翻訳の学習に有用な訳が文脈に依存する例はアラインメントが取りづらく、誤りが含まれやすくなるなど、信頼性の高い大規模対訳コーパスを得ることは難しい。

本研究では、目的言語から原言語への翻訳モデルを利用

できる対訳コーパスから学習し、これを用いて低コストで大量に入手可能な目的言語の単言語コーパスを逆翻訳して疑似対訳コーパスを生成した後、最終的にこの疑似対訳コーパスを追加で用いて文脈考慮型の順翻訳器を学習することを提案する。逆翻訳器は順翻訳と同一のドメインの既存の（比較的少量の）対訳コーパスにより文脈考慮型翻訳モデルとして学習する。評価ドメインで学習した翻訳モデルを用いて疑似対訳コーパスを生成することで、単言語コーパスのドメインと順翻訳のドメインのミスマッチが軽減されると同時に、文脈考慮の恩恵で順翻訳器の学習用途として質の良い疑似対訳コーパスが得られることが期待できる。

提案手法の効果を検証するため日英翻訳の実験を行った。実験では IWSLT2017 TED Talk コーパスにより英日の翻訳モデルを学習し、英語の大規模単言語コーパスである BookCorpus を日本語に逆翻訳して日英の疑似対訳コーパスを生成した。さらにこれを追加で用いて日英の文脈考慮型ニューラル機械翻訳を学習し、分析を行った。

## 2. 関連研究

### 2.1 文脈考慮型翻訳モデル

翻訳時に注目文の直前や直後の文を補助情報として参照する文脈考慮型翻訳モデルは大きく分けて文結合型とマルチエンコーダ型の 2 種類が提案されている。文結合型は原言語及び目的言語それぞれにおいて注目文とその直前の文（文脈文）を区切り記号 <CONC> で挟んで連結し、既存の

<sup>1</sup> 東京大学大学院情報理工学系研究科

<sup>2</sup> 東京大学生産技術研究所

<sup>a)</sup> sugi@tkl.iis.u-tokyo.ac.jp

<sup>b)</sup> ynaga@iis.u-tokyo.ac.jp

文単位の翻訳モデルにより翻訳する [1]。原言語、目的言語両方の文脈を使うモデルを 2-to-2、原言語側の文脈のみを使うものを 2-to-1 と呼ぶ。また、これらに対して通常の文単位のモデルを 1-to-1 と呼ぶ。これに対し、マルチエンコーダ型は文脈文を専用のエンコーダを通してネットワークに取り込むモデルであり、具体的なネットワークの構成は様々なものが提案されている [2], [3], [4]。文結合型、マルチエンコーダ型共に文単位のベースラインと比較して高い翻訳精度が得られることが報告されている。

また、文結合型はマルチエンコーダ型と比較して簡素な構成であるが、精度に大きな差はないことが知られており [2]、文結合型は評価の定まった既存のニューラル翻訳モデルをコアの系列変換システムとして流用できるという利点がある。本研究ではデータ拡張による既存のモデルの性能向上を確認することを目的としているため、単純な文結合型の文脈翻訳モデル [1] を用いて性能の向上を確認する。

## 2.2 逆翻訳による対訳コーパスの拡張

文単位の翻訳においても学習データの量は翻訳精度に大きく影響することが知られている [7]。そこで、豊富に得られる単言語コーパスを逆翻訳して疑似対訳コーパスを生成し、これを学習に利用する手法が提案されている [8], [9], [10]。

2つの言語  $L1$ ,  $L2$  について、最終的に  $L1 \rightarrow L2$  の翻訳器を学習したいとき、手順は次のようになる。まず、既存の（少量の）対訳コーパスから  $L2 \rightarrow L1$  の逆翻訳器を学習する。次に逆翻訳器を用いて  $L2$  の単言語コーパスを翻訳し、疑似対訳コーパスを得る。これを既存の対訳コーパスに追加したものを学習データとして、順翻訳器  $L1 \rightarrow L2$  を学習する。

本研究では順翻訳が文脈考慮型であるため、単言語コーパスとして互いに無関係の文を集めたものではなく、一定数以上の文を含む文章の集合を用いる。また、逆翻訳器として文脈考慮型のモデルを使用した場合に、文単位の逆翻訳と比較して順翻訳の性能に差が生じるかを検証する。

## 3. 疑似対訳コーパスによる文脈翻訳の学習

本節では逆翻訳で生成した疑似対訳コーパスによる文脈考慮型モデルの学習を提案する。提案手法の大まかな流れは 2.2 節で述べた文単位翻訳のための逆翻訳と同様である。

本研究では逆翻訳器として通常の文単位のモデルと 2-to-1 のモデルを比較する。逆翻訳に文脈考慮型モデルを使用することで期待できる効果として、逆翻訳時の適度な情報の省略がある。例えば英日の文脈考慮型翻訳では、英文で繰り返し登場し同一の対象を参照する代名詞を、日本語に訳す際にすべては明示的に訳さずに省略する（ゼロ照応として訳す）という、人手翻訳に近い挙動が期待できる。このような疑似対訳コーパスで日英の順翻訳を学習する場合、モデルは日本語文で省略された代名詞を英文で復元するた

データセット	モデル	文数	len(ja)	len(en)
IWSLT2017 train	1-to-1	211,152	22.7	23.3
	2-to-1	218,775	50.1	25.2
	2-to-2	217,523	49.7	50.9
IWSLT2017 dev	-	871	28.0	29.4
IWSLT2017 test2010	-	1549	24.5	26.3
BookCorpus 1M	1-to-1	998,871	16.0	16.7
	2-to-1	999,380	32.5	16.8
	2-to-2	999,237	32.5	33.9

表 1 順翻訳（日英）に用いたデータセットの文対数と平均サブワード数。BookCorpus の ja は逆翻訳により生成されたデータ。BookCorpus は 1M 文対を利用する場合の数値を記載。入力あるいは出力が文脈文を含む場合は注目文、<CONC>トークンおよび文脈文のサブワードの合計を記載。

めに文脈を利用することを学習することが期待される。

また、文脈考慮型のモデルは文単位の翻訳と比べてモデルの入出力のサイズが大きく、より複雑なタスクとなっている。そのため逆翻訳により学習データを拡張することで受ける恩恵もより大きいことが期待できる。

## 4. 実験

### 4.1 実験設定

2言語対訳コーパスとして、TED Talks（英語の講演）の英語字幕および日本語字幕（人手翻訳）から作成された IWSLT2017 TED Talks コーパス\*1（以下 IWSLT2017 コーパス）を用いた。IWSLT2017 コーパスは学習、開発、テストデータそれぞれ約 22 万、900、1500 文対からなる。単言語コーパスとしては BookCorpus [11] を用いた。BookCorpus はウェブ上で閲覧できる電子書籍をクロールして得られたデータである\*2。本実験では BookCorpus に含まれる段落のうち、10 文以上からなるものを抽出することで得られた約 600 万文を利用した。

IWSLT2017 の日本語文は mecab-ipadic (ver. 2.7.0) を辞書に用いた MeCab\*3 により分かち書きした。IWSLT2017 の英語文および BookCorpus は Moses toolkit v4.0\*4 を用いてトークン化および truecase 処理を行った。さらに SentencePiece (ver.0.1.81)\*5 により、分かち書き済みの日本語文および英語文をサブワード分割した（語彙サイズは 16k で日英共通）。学習データ中で長文はノイズを含みやすいことから 1-to-1 の学習時には入力、出力どちらかの文が 64 個以上のサブワードを含む文対を学習データから除外した。また、2-to-1 および 2-to-2 の学習時には文脈文と注目文を <CONC> トークンで結合した文が入力、出力どち

\*1 <https://wit3.fbk.eu/mt.php?release=2017-01-trnted>

\*2 <https://github.com/soskek/bookcorpus> で公開されているクローラを用いた。

\*3 <https://taku910.github.io/mecab/>

\*4 <http://www.statmt.org/moses/>

\*5 <https://github.com/google/sentencepiece>

疑似対訳例数	1-to-1 back-translation				2-to-1 back-translation			
	1-to-1	2-to-1	2-to-2	2-to-2 (oracle ctx)	1-to-1	2-to-1	2-to-2	2-to-2 (oracle ctx)
0M	11.05	11.27	11.76	12.34	-	-	-	-
0.5M	11.88	12.46	12.24	12.75	11.45	11.62	12.28	12.62
1M	12.06	12.49	*12.84	13.21	11.41	*12.60	*12.82	13.24
2M	<b>12.35</b>	12.40	* <b>13.30</b>	<b>13.69</b>	<b>11.96</b>	*12.96	* <b>13.33</b>	<b>13.69</b>
6M	11.60	* <b>12.91</b>	12.46	13.08	11.83	* <b>12.97</b>	*12.91	13.31

表 2 BLEU による各逆翻訳モデル, 順翻訳モデル, 追加疑似対訳コーパス量に対する翻訳精度評価. 2-to-2 oracle ctx は 2-to-2 のテスト時に目的言語側の文脈文として, モデル自身が翻訳した 1 ステップ前の注目文ではなく, 正解データ中の文脈文 (オラクル文脈) を与えた場合の性能. 太字はモデルを固定し疑似対訳コーパス量を変えた場合に精度が最大のもの. \*は同じ逆翻訳モデルの 1-to-1 の最大精度と比較して統計的に有意差あり (2-to-1 と 2-to-2 についてのみ記載).

疑似対訳例数	1-to-1 back-translation				2-to-1 back-translation			
	1-to-1	2-to-1	2-to-2	2-to-2 (oracle ctx)	1-to-1	2-to-1	2-to-2	2-to-2 (oracle ctx)
0k	0.6723	0.6581	0.6699	0.6737	0.6656	0.6651	0.6679	0.6728
0.5M	0.6762	0.6770	0.6688	0.6795	0.6752	0.6778	0.6821	0.6853
1M	<b>0.6790</b>	0.6792	0.6807	0.6871	0.6766	0.6808	0.6826	<b>0.6956</b>
2M	0.6789	0.6776	<b>0.6848</b>	<b>0.6932</b>	<b>0.6856</b>	0.6844	<b>0.6874</b>	0.6945
6M	0.6786	<b>0.6833</b>	0.6768	0.6831	0.6829	<b>0.6899</b>	0.6849	0.6914

表 3 RIBES による各逆翻訳モデル, 順翻訳モデル, 追加疑似対訳コーパス量に対する翻訳精度.

らから 128 個以上のサブワードを含む場合その文対を学習データから除外した. この操作は逆翻訳, 順翻訳両方で行った. その結果, 順翻訳の学習とテストの際に実際に使用するデータとそのサイズおよび平均文長は表 3 に示すようになった.

逆翻訳器は IWSLT2017 の学習データを用いて学習した. 翻訳モデルは 2-to-1 の文脈考慮型モデルと 1-to-1 の文単位翻訳モデルを用い, それぞれのモデルで生成された逆翻訳を学習データに追加した場合の順翻訳の精度を比較した. 逆翻訳の精度 (BLEU) は 1-to-1 が 12.5, 2-to-1 が 12.9 であった. BookCorpus の逆翻訳はビームサイズ 5 のビームサーチにより行った.

順翻訳の翻訳モデルは 1-to-1, 2-to-1, 2-to-2 の 3 つを比較した. 学習データは IWSLT2017 の学習データに BookCorpus の疑似対訳コーパスを加えたものであり, 疑似対訳コーパスの追加量を 0 (追加なし), 500k, 1M, 2M, 6M の 5 段階で変化させて比較した. テスト時の翻訳はビームサイズ 8 のビームサーチにより行った.

全ての翻訳モデルにおいてコアの文単位翻訳器として Transformer [12] を用いた. 実装には Tensorflow\*6 (ver.1.12.0) を用いた. Transformer のエンコーダ, デコーダは共に 6 層とし, モデルサイズは 512 次元, フィードフォワード層は 2048 次元とした. 最適化手法は Adam [13] を

用いた. 学習率は [12] に従い, 学習ステップ数に応じて変化させた. また, バッチサイズは 1-to-1 は 256, 2-to-1 は 128 とした.

機械翻訳モデルの評価には BLEU [14] と RIBES [15] を用いた. BLEU は Mose toolkit 付属の multi-bleu.perl で, RIBES は著者らの配布しているスクリプト\*7で, それぞれ測定した.

## 4.2 結果

順翻訳の翻訳精度の BLEU と RIBES [15] による評価をそれぞれ表 2, 3 に示す. 表中の 2-to-2 (oracle ctx) は, 2-to-2 のテストにおいて目的言語側の文脈文としてモデル自身の出力ではなくテストデータの正解文脈文を使用した場合の性能であり, 2-to-2 の性能の上限に相当する. BLEU による評価では全てのモデルにおいて疑似対訳コーパスを学習に用いることで性能向上が得られた. ただし 2-to-1 以外のモデルでは疑似対訳コーパスの量が 2M から 6M になると逆に性能が低下する結果となった. また, 疑似学習データによる性能向上にさらに上乗せする形で, 文脈を使うことによる性能向上が確認された. 2-to-1 の逆翻訳で生成した疑似対訳コーパスにより文脈考慮型翻訳を行うことで, 疑似対訳コーパスを用いない文単位の翻訳と比較して, 2-to-1 で最大 1.92, 2-to-2 で最大 2.28 の BLEU の改善が得られ

\*6 <https://www.tensorflow.org/>

\*7 <http://www.kecl.ntt.co.jp/icl/lirg/ribes/>

Context	彼女の 20 代も困難なものでしたがそれ以前の人生はもっと困難に溢れていました
Source	診察中何度も涙を流しましたが「家族は選べないけど友達を選べる」とそのたびに言って気持ちを落ち着かせていました
Reference	she often cried in our sessions, but then would collect herself by saying, “You can’t pick your family, but you can pick your friends.”
1-to-1	I’ve had tears in my doctor’s office, and I’ve said, “I don’t have a family, but I’ve got a friend,” and I calmed down every time .
1-to-1 2M pseudo data	I cried a lot during my examination, but every time I said, “I can’t choose a family, but I can choose a friend,” I said calmly.
2-to-2	during my diagnosis, I ran a lot of tears, and I said, “No family can choose,” but every time I said, “I can choose a friend,” I kind of calmed down.
2-to-2 2M pseudo data	she cried many times during her examination, but each time she said, “I can’t c a family, but I can choose a friend,” she said calmly.

表 4 代名詞「彼女」が省略された日本語文の翻訳. 1-to-1 モデルおよび疑似対訳コーパスを用いずに学習した 2-to-2 モデルでは正しい代名詞 “she” の復元に失敗している. また, 代名詞の正しさ以外の点では疑似対訳コーパスを用いることにより翻訳の正確さが向上している.

た. RIBES による評価も BLEU と概ね同じ傾向を示したが, 1-to-1 と 2-to-1 や 2-to-2 の順位が一部で逆転した.

1-to-1/2-to-2 と疑似対訳コーパスの有無で区別される 4 つのモデル設定による, 代名詞「彼女」が省略された日本語文の翻訳結果を表 4 に示す. なおここでの疑似対訳コーパス量は 2M 文対で, 2-to-1 の逆翻訳により生成されたものである. 彼女に対応する英語の代名詞 “she” を正しく復元できているのは疑似対訳コーパスを用いて学習された 2-to-2 モデルのみである. また, 疑似対訳コーパスの有無で比較すると疑似対訳コーパスを用いた場合に代名詞以外の翻訳内容の正確さが向上している.

### 4.3 議論

逆翻訳モデルとして 2-to-1 を用いることで 1-to-1 を用いる場合と比較して高品質な疑似対訳コーパスが得られ, より高性能な順翻訳器が学習されることが期待されたが, BLEU および RIBES による定量評価では両者の性能に明らかな差は見られなかった. 単言語コーパスを逆翻訳して生成された疑似対訳コーパスを順翻訳の学習に用いる際に一部の単語を削除したり入れ替えたりするなど, 文にノイズを加えることで翻訳モデルの性能が向上することが先行研究 [8] で報告されていることも踏まえ, 疑似対訳コーパスのどのような特性が順翻訳性能の向上に寄与するのか, より詳しい分析が必要である.

## 5. おわりに

本稿では単言語コーパスの逆翻訳によって得られた疑似対訳コーパスを文脈考慮型のニューラル機械翻訳の学習に利用し, その有効性を検証した. BookCorpus よび IWSLT2017 コーパスを用いた実験により, 疑似対訳コーパスを文脈考慮型翻訳モデルの学習に用いることで, 文単位翻訳の性能向上に加え, モデルが文脈を捉える性能も向

上し, 翻訳品質が複合的に改善することを示した.

謝辞 本研究成果の一部は, 独立行政法人情報通信研究機構 (NICT) の委託研究「環境×交通データの連携によるモビリティリスク情報生成・流通プラットフォームの実証的研究開発」により得られたものです.

### 参考文献

- [1] Tiedemann, J. and Scherrer, Y.: Neural Machine Translation with Extended Context, *Proceedings of the Third Workshop on Discourse in Machine Translation (DiscoMT 2017)*, pp. 82–92 (2017).
- [2] Bawden, R., Sennrich, R., Birch, A. and Haddow, B.: Evaluating Discourse Phenomena in Neural Machine Translation, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL 2018)*, pp. 1304–1313 (2018).
- [3] Voita, E., Serdyukov, P., Sennrich, R. and Titov, I.: Context-Aware Neural Machine Translation Learns Anaphora Resolution, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 1264–1274 (2018).
- [4] Maruf, S. and Haffari, G.: Document Context Neural Machine Translation with Memory Networks, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 1275–1284 (2018).
- [5] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus.
- [6] Pryzant, R., Chung, Y., Jurafsky, D. and Britz, D.: JESC: Japanese-English Subtitle Corpus, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).
- [7] Koehn, P. and Knowles, R.: Six challenges for neural machine translation, *arXiv preprint arXiv:1706.03872* (2017).
- [8] Edunov, S., Ott, M., Auli, M. and Grangier, D.: Understanding Back-Translation at Scale, *Proceedings of*

- the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 489–500 (2018).
- [9] Imamura, K., Fujita, A. and Sumita, E.: Enhancement of encoder and attention using target monolingual corpora in neural machine translation, *Proceedings of the Second Workshop on Neural Machine Translation and Generation (WNMT 2018)*, pp. 55–63 (2018).
- [10] Sennrich, R., Haddow, B. and Birch, A.: Improving Neural Machine Translation Models with Monolingual Data, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 86–96 (2016).
- [11] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. and Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, *Proceedings of the IEEE International Conference on Computer Vision (ICCV2015)*, pp. 19–27 (2015).
- [12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is all you need, *Advances in Neural Information Processing Systems (NIPS2017)*, pp. 5998–6008 (2017).
- [13] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
- [14] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics (ACL 2002), pp. 311–318 (2002).
- [15] Isozaki, H., Hirao, T., Duh, K., Sudoh, K. and Tsukada, H.: Automatic evaluation of translation quality for distant language pairs, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pp. 944–952 (2010).