単言語データを用いた逆翻訳と順翻訳による データ拡張の効果の比較

美野 秀弥 [1]　衣川 和尭 [1]　伊藤 均 [1]　後藤 功雄 [1]　山田 一郎 [1]　田中 英輝 [2]
川上 貴之 [3]　大嶋 聖一 [3]　朝賀 英裕 [3]
[1] NHK 放送技術研究所　[2] NHK エンジニアリングシステム　[3] 時事通信社
{mino.h-gq, kinugawa.k-jg, itou.h-ce, goto.i-es, yamada.i-hy}@nhk.or.jp
tanaka.hideki@nes.or.jp, {kawakami, sohshima, asaka}@jiji.co.jp

1 はじめに

ニューラル機械翻訳（NMT）の精度を向上させるアプローチの 1 つに、単言語データを用いたデータ拡張がある。Sennrich ら [1] は、既存の対訳データで学習した目的言語から原言語へ翻訳する NMT モデルを用いて目的言語の単言語データを原言語に翻訳した擬似的な対訳データを生成し、それを既存の対訳データに追加して学習することで翻訳精度が向上することを示した。目的言語から原言語に翻訳することは「逆翻訳」と呼ばれており、逆翻訳によって生成された擬似対訳データは、目的言語側のデータの流暢性が担保されていることから、訳文の流暢性が向上することが知られている。一方で、原言語の単言語データを目的言語に翻訳すること (以下、順翻訳と呼ぶ) で得られる擬似的な対訳データを用いたデータ拡張の効果について報告した研究 [2, 3, 4] は順翻訳によるデータ拡張による翻訳精度の向上のみに言及しており、逆翻訳との比較や対訳データの規模による違いなどの詳細な分析はない。そこで、本稿では、規模の異なる対訳データと内容がほぼ均衡した日英の単言語データを用い、逆翻訳と順翻訳によって生成された異なる擬似対訳データによるデータ拡張の効果を日英、英日の NMT システムの性能比較を通じて検証した。検証の結果、順翻訳で生成した疑似対訳データと、逆翻訳で生成した疑似対訳データは共に、データ拡張による翻訳の精度向上に寄与することを確認した。また、両方の擬似対訳データを併用することでさらに翻訳精度が向上することを確認した。既存の対訳データの分量を小規模にして行った実験では、逆翻訳で生成された擬似対訳データの効果が大きいことを確認した。

表 1　時事通信社ニュースの各コーパスの文数

コーパス	文数
日英均衡（日英翻訳）	283,382
日英均衡（日本語修正）	128,823
日英自動アライメント	391,538

2 実験概要

2.1 データセット

本稿では、国立研究開発法人情報通信研究機構の委託で進められている機械翻訳の研究プロジェクト内で開発されたニュースの日英均衡対訳コーパス [5, 6, 7] と自動アライメント手法によって抽出された日英自動アライメントコーパスとを用いる。表 1 に各コーパスの文数を示す [1)]。日英均衡対訳コーパスは、日英翻訳コーパスと日本語修正コーパスの 2 種類に分けられる。日英翻訳コーパスは、時事通信社の日本語記事の文を過不足なく人手で英語に翻訳した対訳コーパスである。日本語修正コーパスは、時事通信社の日英記事に対して記事アライメントを自動で行い、英語記事を文単位で内容が等価になるように日本語文を修正した対訳コーパスである。日英自動アライメント対訳コーパス

1）田中らが構築したコーパス [5] の一部を利用。

は、文単位に分割された日本語と英語のニュース記事を入力として日英の文単位間が交差しないように結びつける文アライメントアルゴリズム [8] により抽出された対訳コーパスである。テストセット、開発用セットには、WAT2020 の JIJI タスク [9] のデータセットを用いた [2]。

2.2 実験設定

NMT システムには、Transformer モデル [10] が実装されている Sockeye 2 [11] を用い、日英、英日の 2 種類の翻訳モデルを学習した。英語については Moses のツール[3]、日本語については Kytea [12] をそれぞれ用いてトーカナイズした。語彙サイズは、Byte Pair Encoding (BPE) [13] を用いて 64,000 語に定めた。最大文長は 150 トークン、最大エポック数は 30 とし、チェックポイント間隔は 5,000 とした。連続して 15 回チェックポイントで開発用データのパープレキシティの値が改善しない場合は学習を終了した。学習時のその他の設定は Sockeye のデフォルト値を用いた。翻訳時は、ランダムシードを変えて学習した 5 つのモデルをアンサンブルしたモデルを用い、ビーム幅を 30 とした。全てのシステムは、BLEU [14] で評価した。

2.3 ベースライン

ベースラインとして、表 1 の日英均衡コーパス (日英翻訳コーパスと日本語修正コーパス) で学習した翻訳モデルを用意した。事前実験でタグを用いた適応化による効果が得られたため、日英翻訳コーパスには<J-JIJI>タグ（日本語が時事通信社のスタイルであることを示すタグ）を、日本語修正コーパスには<E-JIJI>タグ（英語が時事通信社のスタイルであることを示すタグ）をそれぞれ原言語側のデータの先頭に付与して翻訳モデルを学習した。翻訳時は、<E-JIJI>タグ（日英翻訳時）、または<J-JIJI>タグ（英日翻訳時）を入力文の先頭に付与した。

2） 自動アライメント手法により抽出された対訳データからノイズが多いものが人手で取り除かれている。

3） https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

表 2 各コーパスへの付与タグ (学習時)

コーパス	スタイルタグ	擬似対訳タグ
日英均衡		
日英翻訳	<J-JIJI>	<NOMT>
日本語修正	<E-JIJI>	<NOMT>
MonoEn	<E-JIJI>	<MT>
MonoJa	<J-JIJI>	<MT>

2.4 単言語データによるデータ拡張

日英均衡コーパスに、日本語と英語の単言語データを翻訳することで構築した 2 種類の擬似対訳コーパスを加えて翻訳モデルを学習した。日本語と英語の単言語データには、日本語側、英語側の内容がほぼ均衡している表 1 の日英自動アライメントコーパスを用いた。

単言語データの翻訳にはベースラインシステムを用い、日英自動アライメントコーパスの英語側を日本語に翻訳した擬似対訳コーパス（以下、MonoEn と呼ぶ）と日本語側を英語に翻訳した擬似対訳コーパス（以下、MonoJa と呼ぶ）、それぞれ 391,538 対を得た。内容がほぼ均衡している日本語、および英語の単言語データを用いることで、MonoEn と MonoJa はほぼ同じ情報を有したコーパスとなっている。本稿では、ベースラインシステムの学習で用いた日英均衡コーパスに、MonoEN を加えて学習した翻訳モデル、MonoJa を加えて学習した翻訳モデル、MonoEn と MonoJa の両方を加えて学習した翻訳モデル、の 3 種類の翻訳モデルを構築した。Caswell らは、学習時に擬似対訳コーパスと正規の対訳コーパスを区別するタグを付与して学習する手法 [15] を提案しており、本稿でもこの手法を用いた。原言語側、あるいは目的言語側が時事通信社のニュース文のスタイルであることを示すスタイルタグと、対訳データが擬似対訳か正規の対訳かを示す擬似対訳タグ、の 2 種類のタグを原言語側データの先頭に付与して翻訳モデルを学習した。表 2 に、各コーパスに付与するタグを示す。 日英翻訳時は “<E-JIJI> <NOMT>” を付与し、英日翻訳時は “<J-JIJI> <NOMT>” を付与した。

表 3　日英翻訳の実験結果

学習データ	文数	BLEU
日英均衡 (タグ未付与)	412,205	20.5
日英均衡 (タグ付与)	412,205	23.2
日英均衡+MonoEn(逆翻訳)	803,743	26.2
日英均衡+MonoJa(順翻訳)	803,743	26.4
日英均衡+MonoEn+MonoJa	1,195,281	26.7

表 4　英日翻訳の実験結果

学習データ	文数	BLEU
日英均衡 (タグ未付与)	412,205	30.3
日英均衡 (タグ付与)	412,205	31.0
日英均衡+MonoEn(順翻訳)	803,743	34.2
日英均衡+MonoJa(逆翻訳)	803,743	34.1
日英均衡+MonoEn+MonoJa	1,195,281	34.9

2.5　小規模対訳データ

小規模な対訳データにおけるデータ拡張の効果を確認するために、表 1 の日英翻訳コーパスと日本語修正コーパスから 75,000 データをそれぞれランダムに抽出した日英均衡コーパス 150,000 データを対訳データとし、同様に日英自動アライメントデータを用いてデータ拡張を行い、翻訳モデルを学習した。データ拡張には、日英均衡コーパス 150,000 データで学習した翻訳モデルを利用した。英日翻訳については、タグ付与の翻訳モデルよりもタグ未付与の翻訳モデルの翻訳精度が高かったため、データ拡張時もタグ未付与で学習した。英日翻訳でタグ付与の翻訳モデルの翻訳精度が下がった理由としては、翻訳時に付与する<J-JIJI>タグは学習データの原言語（英語）が時事通信社のニュース文ではない翻訳文であることに対応しているが、テストデータは英語、日本語共に時事通信社のニュース文であるため、その乖離が影響したためと考えられる。

3　実験結果

日英、英日ニュース翻訳の実験結果を表 3,4 に示す。参考のため、タグを付与しない場合の結果も載せた。日英、英日ともに順翻訳データ、逆翻訳データの追加による翻訳精度の向上を確認した。また、順翻訳データと逆翻訳データの両方を追加することにより、さらに翻訳精度が向上することを確認した。英日翻訳については、テストセットの中の 100 文について、「日英均衡+MonoEn（順翻訳）」、「日英均衡+MonoJa（逆翻訳）」、「日英均衡+MonoEn+MonoJa」、で学習した翻訳モデル間の一対比較評価を評価者 1 名により実施した [4]。表 5 に評価結果を示す。BLEU の評価では、順翻訳データ (MonoEn) を加えて学習した翻訳モデルと逆翻訳データ (MonoJa) を加えて学習した翻訳モデルの差はなかった (表 4 の 3,4 行目) が、一対比較評価では逆翻訳データを追加した翻訳モデルよりも順翻訳データを追加した翻訳モデルの結果が良い評価を得た。また、順翻訳、逆翻訳の両方のデータを追加して学習した翻訳モデルは、どちらか一方のデータのみを加えた翻訳モデルと比較して良い評価を得られる傾向があった。本実験結果より、原言語側と目的言語側の両方の単言語データの利用の有効性が確認された。

表 5　英日翻訳の一対比較評価（100 文）

学習データ		同等
日英均衡+MonoEn(順翻訳)	38	38
日英均衡+MonoJa(逆翻訳)	24	
日英均衡+MonoEn(順翻訳)	15	60
日英均衡+MonoEn+MonoJa	25	
日英均衡+MonoJa(逆翻訳)	22	40
日英均衡+MonoEn+MonoJa	38	

表 6 に英日翻訳の翻訳例と評価結果を示す。#1 の例では、”crew member” の翻訳の違いにより、日英均衡+ MonoJa (逆翻訳) の結果（隊員）よりも日英均衡+MonoEn (順翻訳) を用いた翻訳モデルの結果（乗組員）が良いと評価された。#2 の例では、”moving service fees” の翻訳の違いにより、日英均衡+MonoEn (順翻訳) の結果（移動サービス料）よりも日英均衡+MonoJa (逆翻訳) を用いた翻訳モデルの結果（引越し料金）が良いと評価された。

逆翻訳によるデータ拡張により翻訳結果の流暢性が向上することは知られているが、BLEU および人手評価の結果から、順翻訳によるデータ拡張でも翻訳結果の流暢性が向上することがあることが分かる。

4）　2 つの出力のうち、どちらが翻訳として適切かどうかを評価した。同じ品質の場合は「同等」と評価した。

表 6　英日翻訳例

#1	英語文	A police officer playing the role of the crew member visited a building of a neighborhood community association, a supposedly private house.
	参照訳	乗組員役の警察官が、民家に見立てた自治会館を訪問。
	日英均衡+MonoEn(順翻訳)	**乗組員**の役を務める警察官が民家とみられる自治会の建物を訪れた。
	日英均衡+MonoJa(逆翻訳)	**隊員**役の警察官が民家とみられる町内会の建物を訪問。
	日英均衡+MonoEn+MonoJa	**乗組員**の役割を担う警察官が、民家とみられる町内会の建物を訪れた。
	BLEU	0.00, 27.69, 18.97
	一対比較評価	+MonoJa(逆翻訳) < +MonoEn(順翻訳) = +MonoEn+MonoJa
#3	英語文	Moving service fees are basically composed of transportation costs and loading and unloading charges.
	参照訳	引っ越し費用は、荷物を目的地まで運ぶ運賃と搬入・搬出作業にかかる料金で構成される。
	日英均衡+MonoEn(順翻訳)	**移動サービス料**は基本的に輸送費や積み荷料などで構成される。
	日英均衡+MonoJa(逆翻訳)	**引っ越し料金**は、輸送費や荷下ろし料などが基本。
	日英均衡+MonoEn+MonoJa	**引っ越し料金**は基本的に輸送費や積み荷の積み下ろし料などで決まっている。
	BLEU	16.69, 0.00, 0.00
	一対比較評価	+MonoEn(順翻訳) < +MonoJa(逆翻訳) = +MonoEn+MonoJa

表 7　日英翻訳の実験結果（小規模対訳データ）

学習データ	文数	BLEU
日英均衡 (タグ未付与)	150,000	19.0
日英均衡 (タグ付与)	150,000	20.5
日英均衡+MonoEn(逆翻訳)	541,538	22.7
日英均衡+MonoJa(順翻訳)	541,538	20.6
日英均衡+MonoEn+MonoJa	933,076	22.8

表 8　英日翻訳の実験結果（小規模対訳データ）

学習データ	文数	BLEU
日英均衡 (タグ未付与)	150,000	26.0
日英均衡 (タグ付与)	150,000	24.6
日英均衡+MonoEn(順翻訳)	541,538	27.8
日英均衡+MonoJa(逆翻訳)	541,538	27.8
日英均衡+MonoEn+MonoJa	933,076	27.9

小規模対訳データを用いた日英、英日ニュース翻訳実験結果を表 7,8 に示す。日英翻訳については、順翻訳と比較して逆翻訳のデータ拡張による翻訳精度の向上が大きかった。英日翻訳については、順翻訳、逆翻訳のデータ拡張による翻訳精度の向上に差はなかった。一方で、日英翻訳、英日翻訳ともに、順翻訳と逆翻訳の両方のデータ拡張による翻訳精度の向上は小さかった。

4　おわりに

本稿では、単言語データによるデータ拡張の効果を、時事通信社のニュースコーパスを用いて検証した。原言語側の単言語データによる順方向の翻訳（順翻訳）と、目的言語側の単言語データによる逆方向の翻訳（逆翻訳）の 2 種類のデータ拡張を用いて日英、英日の翻訳実験を行い、両方のデータ拡張を併用することで翻訳精度が向上することを確認した。さらに、対訳データを小規模にして同様の実験を実施したところ、逆翻訳によるデータ拡張の効果が大きくなることを確認した。「逆翻訳」によるデータ拡張は対訳データが小規模のときに有用であることが知られており [1]、これを裏付ける結果となった。小規模の対訳データの実験では、順翻訳と逆翻訳の 2 種類のデータ拡張の併用による BLEU の向上は小さかったが、本稿でも指摘しているように、BLEU の評価で差が小さい場合でも人手による一対評価で差が大きくなる場合があるので、今後人手評価を実施したい。

謝辞

本研究成果は独立行政法人情報通信研究機構 (NICT) の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」により得られたものです。

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. Data diversification: A simple strategy for neural machine translation. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2020.

[3] Maja Popović, Alberto Poncelas, Marija Brkic, and Andy Way. Neural machine translation for translating into Croatian and Serbian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 102–113, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL).

[4] Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. OPPO's machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 282–292, Online, November 2020. Association for Computational Linguistics.

[5] 田中英輝, 中澤敏明, 美野秀弥, 伊藤均, 後藤功雄, 山田一郎, 川上貴之, 大嶋聖一, 朝賀英裕. 時事通信社ニュースの日英均衡対訳コーパスの構築-第 3 報. 言語処理学会 第 27 回年次大会, 2021.

[6] 田中英輝, 中澤敏明, 美野秀弥, 伊藤均, 後藤功雄, 山田一郎, 川上貴之, 大嶋聖一, 朝賀英裕. 時事通信社ニュースの日英均衡対訳コーパスの構築-第 2 報. 言語処理学会 第 26 回年次大会, pp. 545–548, 2020.

[7] 田中英輝, 美野秀弥, 後藤功雄, 山田一郎, 川上貴之, 大嶋聖一, 朝賀英裕. 時事通信社ニュースの日英均衡対訳コーパスの構築–第 1 報. 言語処理学会 第 25 回年次大会, pp. 371–374, 2019.

[8] Masao Utiyama and Hitoshi Isahara. A japanese-english patent parallel corpus. In *In proceedings of the Machine Translation Summit XI*, 2007.

[9] Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pp. 1–44, Suzhou, China, December 2020. Association for Computational Linguistics.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008. Curran Associates, Inc., 2017.

[11] Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pp. 110–115, Virtual, October 2020. Association for Machine Translation in the Americas.

[12] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 529–533, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[13] 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)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[14] 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[15] Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 53–63, Florence, Italy, August 2019. Association for Computational Linguistics.