対訳文のみを用いた翻訳と言い換えのマルチタスク学習における翻訳精度

名村太一 [1] 村上仁一 [2]
[1] 鳥取大学大学院持続性社会創生学科学研究科
[2] 鳥取大学工学部
[1]m23j4038h@edu.tottori-u.ac.jp
[2]murakami@tottori-u.ac.jp

概要

機械翻訳の問題として学習データの対訳文不足がある．この問題への対策として単言語データを用いたデータ拡張などが行われている．提案手法は日本語と英語の対訳文から英日対・日英対・日日対・英英対の対訳を作成する．そしてそれぞれを英日翻訳，日英翻訳，日本語言い換え，英語言い換えタスクとしてマルチタスク学習を行う．提案手法の特徴としてアーキテクチャの変更，単言語データを用意する必要がない．実験の結果，自動評価と人手評価でベースラインを上回った．

1 はじめに

ニューラル機械翻訳（NMT）は統計的機械翻訳（SMT）と比較して高い精度を示した [1]．しかし，NMT が高い性能を示すには大量の対訳文が必要である．この問題への対策として，単言語データを利用したデータ拡張が行われている．Sennrich ら [2] はターゲット言語側の単言語データをソース言語に翻訳し疑似的な対訳を作成した．しかし，この手法ではターゲット言語の単言語データを必要とする．そこで，対訳文のみを利用したデータ拡張として，言い換えの利用 [3] や Bidirectional Training[4]，同一対の学習 [5] などの手法が提案されている．

本研究では翻訳と言い換えのマルチタスク学習を提案する．日英翻訳の一般的な学習では日本語から英語への翻訳のみ学習する．提案手法では，日本語と英語の対訳文から英日対・日英対・日日対・英英対を作成する．そしてそれぞれを英日翻訳．日英翻訳・日本語言い換え・英語言い換えタスクとしてマルチタスク学習する．特徴としてアーキテクチャの変更が必要なく，既存の対訳文のみを利用する．また，外部データを利用したデータ拡張と組み合わせることもできる．

実験の結果，自動評価と人手評価ではベースラインを上回った．

2 関連研究

Ding ら [4] は Bidirectional Training を提案した．Bidirectional Training では，まず src→tgt と tgt→src を同時に学習する．そして学習したモデルを src→tgt でチューニングする．8 つの言語ペアで翻訳精度が改善された．

Currey ら [5] は対訳文のターゲット言語データをコピーする手法を提案した．src→tgt の対訳文に tgt→tgt ペアを追加し学習する．オートエンコーダを追加することで固有名詞などの翻訳が成功する確率が上がった．

Johnson ら [6] はマルチタスク学習を利用した多言語翻訳を提案した．大量の多言語で学習することで低リソース言語の翻訳精度が向上した．

3 仮説

これらの手法は低頻出語の出現回数が増えることで翻訳精度が改善されていると考える．特に Currey ら [5] はターゲット言語のみに効果がある．ソース言語に着目した研究は少ない．その理由として，src→tgt の学習データに src→src ペアを学習データに追加すると，出力が src か tgt か決定できないからである．そこで src→src ペアを追加するためにマルチタスク学習を導入する．マルチタスク学習を用いて，出力文の言語を制御する．そして src→src ペアを追加することで，ソース側の低頻出語の出現回数が増えて翻訳精度が改善できると考える．

表 1　データ拡張の例

	src データ	tgt データ
1	Real ability will win in the long run . <en2ja>	結局 は 実力 が もの を 言う 。
2	結局 は 実力 が もの を 言う 。 <ja2en>	Real ability will win in the long run .
3	結局 は 実力 が もの を 言う 。 <ja2ja>	結局 は 実力 が もの を 言う 。
4	Real ability will win in the long run . <en2en>	Real ability will win in the long run .

4　提案手法

日本語と英語の対訳文から，英日対，日英対，日日対，英英対を作成する．それぞれ英日翻訳 (en2ja)，日英翻訳 (ja2en)，日本語言い換え (ja2ja)，英語言い換え (en2en) タスクとしてマルチタスク学習する．各タスクは各タグで制御し，ソース側の学習データを「入力文 <tag>」の形とする．提案手法の学習データの例を表 1 に示す．

通常の英日翻訳では表 1 の 1 行目のみを学習する．提案手法では 1 から 4 行目まで利用して学習する．推論時には「入力文 <tag>」までを入力する．

5　実験

5.1　実験データ

電子辞書を中心に集められた単文対訳文 16 万対，複文対訳文 9 万対 [7] を合わせた 25 万対を train データとする．dev データとして 1 万対，test データとして 2 万対を用いる．

5.2　実験条件

英日翻訳においては表 1 の 1 行目のみを利用して学習したモデルを baseline とする．日英翻訳は 2 行目のみを学習したモデルを baseline とする．提案手法の 1 行目から 4 行目全てを学習したモデルを aug1234 とする．

encoder-decoder モデルとして Transformer[8] を利用する．ハイパーパラメータは Vaswani ら [8] を参考に決定する．fairseq[9] で実装する．

自動評価では BLEU，TER，COMET で評価する．BLEU，TER は sacreBLEU[10] を用いる．COMET[11] のモデルは Unbabel/wmt20-comet-da を用いる．

人手評価は，学生 3 人が英日翻訳においてランダムに抽出した 100 文で行う．提案手法の方が良かった，baseline が良かった，両方良かった，両方悪かったの 4 択で対比較評価する．

5.3　実験結果

表 2 に自動評価結果を示す．表上部の en2ja が日英翻訳，下部の ja2en が英日翻訳結果を示す．表 3 に英日翻訳の人手評価結果を示す．提案手法の aug1234 はベースラインと比較して BLEU と TER ではわずかに改善した．COMET ではより改善された．

表 2　ベースラインと提案手法の自動評価結果

(en2ja)	BLEU(↑)	TER(↓)	COMET(↑)
baseline	28.7	56.7	0.3843
aug1234	**28.9**	**56.5**	**0.4122**
(ja2en)	BLEU(↑)	TER(↓)	COMET(↑)
baseline	24.6	60.6	0.2479
aug1234	**25.0**	**58.8**	**0.2626**

表 3　baseline と aug1234 の英日翻訳の人手評価結果

(en2ja)	aug1234	baseline	○	×
学生 A	**15**	7	56	22
学生 B	**15**	10	58	17
学生 C	13	13	52	22

表 2 より，BLEU ではあまり変化が無かったが COMET は改善された．そのような例を表 4 に示す．

表 4　COMET が大きく改善された出力例

		BLEU	COMET
input	Nobunaga ’ s army went up to Kyoto again .		
refrence	信長勢は再び京へ上った。		
baseline	織田信長は京都へ逃げた。	0.482	0.063
aug1234	織田信長はまた京都へ向かった。	0.417	0.4966

aug1234 の翻訳精度が向上する理由として，低頻出語の精度改善が考えられる．例を表 5 に示す．入力の「acquaintance」がベースラインの学習データ中には 1 度しか出現しない．また「her own way」は 10 回の出現で翻訳が失敗していた．aug1234 では出現回数が増えたため翻訳精度が改善したと考える．

表 5　低頻出語の翻訳が改善された出力例

input	I gained acquaintance with rural life .
refrence	田園生活を知った。
baseline	私は田舎の生活に知り合いました。
aug1234	田園生活についての知識を得た。
input	She had her own way .
refrence	彼女は我を通した。
baseline	彼女はいつもの思いでやった。
aug1234	彼女は意地を通した。

他の提案手法が優れていた例を表 6 に示す．ベースラインが優れていた例を表 7 に示す．

表 6　aug1234 の優れた出力例

input	This bill is hard to pass.
refrence	今回の法案には無理がある。
baseline	この手形は期限が過ぎていない。
aug1234	この法案は通過し難い。
input	I shall let you know as soon as it is decided.
refrence	決まり次第お知らせします。
baseline	問題は決まったままお知らせします。
aug1234	この件が決まったらすぐにご通知します。

表 7　baseline の優れた出力例

input	I must brush up my English .
refrence	英語をやり直さねばならない。
baseline	英語に力をつけなければならない。
aug1234	英語にブラシをかけてやらねばならない。
input	Her heart thudded .
refrence	彼女の心臓はどきどきしていた。
baseline	彼女の心臓が揺らいだ。
aug1234	彼女の心臓がきらきらと音を立てた。

6　考察

6.1　アブレーションテスト

表 1 において，様々な組み合わせで実験を行う．自動評価結果を表 8 に示す．各行の特徴を下記に示す．

aug12　日英・英日
aug13, 24　ターゲット言語のデータ増加
aug14, 23　ソース言語のデータ増加
aug134, 234　逆方向の対訳を除外

実験の結果，どの組み合わせも単独では aug1234 には及ばない．baseline と比較しても大きな変化はない．しかし，全てを組み合わせることで大幅に翻訳精度が改善された．

表 8　アブレーションテストの自動評価結果

(en2ja)	BLEU(↑)	TER(↓)	COMET(↑)
1(baseline)	28.7	56.7	0.3843
aug12	28.6	57.0	0.3967
aug13	28.3	58.0	0.3876
aug14	28.6	56.8	0.3828
aug134	27.8	58.9	0.3777
aug1234	**28.9**	**56.5**	**0.4122**
(ja2en)	BLEU(↑)	TER(↓)	COMET(↑)
2(baseline)	24.6	60.6	0.2479
aug12	24.8	59.4	0.2475
aug23	24.9	60.7	0.2475
aug24	25.0	59.5	0.2444
aug234	23.1	63.4	0.2082
aug1234	**25.0**	**58.8**	**0.2626**

表 9 に出力例を示す．この結果では，表 1 の 2 行目を含まないモデルでは良い結果を示した．しかし，翻訳の逆方向を学習データに含まない aug134, aug234 が自動評価では翻訳精度が低下した．

表 9　日英対を含まないモデルが優れた英日翻訳例

input	We put on special programs .
reference	我々は特別番組を放送した。
baseline	特別計画を立てた。
aug12	我々は特別計画を行なった。
aug13	特別番組を設けた。
aug14	我々は特別番組を組んだ。
aug134	特別な番組を課した。
aug1234	我々は特別な計画を立てた。
input	The children ’ s description of the trip centered on the food .
reference	子供たちの旅行の話は食べ物に集中した。
baseline	子供たちの話で旅の中心になった。
aug12	その 旅行が食糧の中心になっている子供たちの人相書。
aug13	子供たちの旅の描写は食べ物を中心にしたものだ。
aug14	子供たちの旅行の話は食べ物を中心にしていた。
aug134	子供たちの旅行の人相書は食物に集中していた。
aug1234	その旅行の内容は子供達の料理中心になっていた。

6.2 test 入力文の追加

テスト文の入力を利用して表 1 の 3 行目のデータを作る．実験結果を表 10 に示す．テスト文は 2 万文で，学習データ 25 万文と比較して少ないが，翻訳精度が改善された．

改善された翻訳例を表 11 に示す．ベースラインの学習データでは「sewers」「Nobunaga」が低頻出語であった．test-src を追加することで出現回数が増えて翻訳精度が向上した．

表 10 test 文の入力を学習に追加した場合の自動評価結果

(en2ja)	BLEU(↑)	TER(↓)	COMET(↑)
1(baseline)	28.7	56.7	0.3843
1+test-src	**28.8**	**57.0**	**0.4052**
(ja2en)	BLEU(↑)	TER(↓)	COMET(↑)
2(baseline)	24.6	60.6	0.2479
2+test-src	**25.1**	**59.2**	**0.2617**

表 11 test-src を追加して改善した出力例

input	The sewers can ’ t cope with so much water .
refrence	下水管はこれだけ大量の水を処理しきれない。
baseline	下水管の中ではこれほど水がうまく流れない。
baseline+test-src	この下水管はそんなに大量の水では対応できない。
input	Nobunaga ’ s army went up to Kyoto again .
refrence	信長勢は再び京へ上った。
baseline	織田信長は京都へ逃げた。
baseline+test-src	織田信長はまた京都へ向かった。

6.3 チューニング

英日翻訳モデルとして aug1234 を表 1 の 1 行目，日英翻訳は 2 行目でチューニングする．自動評価結果を表 12 に示す．aug1234 と比較して日英翻訳では BLEU と TER は上昇した．しかし英日翻訳ではチューニングの結果翻訳精度が悪くなった．

表 12 チューニングの自動評価結果

(en2ja)	BLEU(↑)	TER(↓)	COMET(↑)
1(baseline)	28.7	56.7	0.3843
aug1234	**28.9**	**56.5**	**0.4122**
aug1234+tune	27.4	59.2	0.3843
(ja2en)	BLEU(↑)	TER(↓)	COMET(↑)
2(baseline)	24.6	60.6	0.2479
aug1234	25.0	58.8	**0.2626**
aug1234+tune	**25.4**	**58.4**	0.2204

6.4 言い換え

表 1 より，3 行目と 4 行目の <ja2ja> と <en2en> タグを利用することで，言い換え文を得ることができる．日本語言い換えは「入力文 <ja2ja>」を入力し，n-best で複数出力させる．多くの場合，第 1 候補は入力文と同じ文が出力される．そのため，第 2 候補以降を言い換え文とする．入力文を 20 文，第 4 候補まで出力させ，言い換え文を合計 60 文出力させた．言い換え成功か失敗かの二択で評価した．評価結果を表 3 に示す．言い換え成功率はおおよそ 70%だった．

表 13 日本語言い換え (<ja2ja>) 評価結果

	学生 A	学生 B
○	43	41
×	17	19

表 14 に日本語言い換え文の出力と評価の例を示す．言い換え文の特徴として，日本語言い換え文中に英単語が出現する．また，第 2 候補に意味が反転した文が出現する．

表 14 日本語言い換え (<ja2ja>) 出力と評価の例

		学生 A	学生 B
input	傘が裏返しになった。		
第 2 候補	かさが裏返しになった。	○	○
第 3 候補	umbrella が裏返しになった。	○	○
第 4 候補	傘がしまいになった。	○	×
input	彼は有罪の宣告を受けた。		
第 2 候補	彼は無罪の宣告を受けた。	×	×
第 3 候補	彼は有罪の判決を受けた。	○	○
第 4 候補	彼は guilty の宣告を受けた。	○	○

7 おわりに

本研究では英日翻訳，日英翻訳，日本語対，英語対のマルチタスク学習を通じて翻訳を行った．実験の結果，自動評価と人手評価においてベースラインを上回った．提案手法はシンプルな手法である．そのため，今後は逆翻訳などの他のデータ拡張手法などと組み合わせることでより翻訳精度が向上すると考える．詳しい言い換え文の精度についも今後調査を行う．

謝辞

評価に協力してくれた，松本武尊，丸山京祐に感謝する．

参考文献

[1] Maria Stasimioti, Vilelmini Sosoni, Katia Kermanidis, and Despoina Mouratidis. Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, **Proceedings of the 22nd Annual Conference of the European Association for Machine Translation**, pp. 441–450, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, **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.

[3] 松本武尊, 村上仁一. 言い換え文を用いた機械翻訳の学習データの増加. 言語処理学会第 30 回年次大会発表論文集, pp. 2348–52, 2024.

[4] Liang Ding, Di Wu, and Dacheng Tao. Improving neural machine translation by bidirectional training. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3278–3284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, **Proceedings of the Second Conference on Machine Translation**, pp. 148–156, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[6] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 339–351, 2017.

[7] 村上仁一. 日英対訳データベースの作成のための 1 考察. 言語処理学会第 17 回年次大会発表論文集, D4-5, pp. 979–82, 2011.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[9] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[11] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.