言い換え文を用いた機械翻訳の学習データの増加

松本武尊 [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
m23j4054m@edu.tottori-u.ac.jp
murakami@tottori-u.ac.jp

概要

NMT において，言い換え文を利用して，学習文を増加する手法がある．本論文では，外部コーパスを利用せずに言い換え文を生成する．そして，生成された言い換え文を用いて学習データを増加させ，機械翻訳の精度向上を試みた．その結果，自動評価結果では，わずかに精度が向上した．しかし，人手評価結果では学習文に言い換え文を追加しても精度に変化が見られなかった．

1　はじめに

ニューラル機械翻訳の精度を向上させるためには，大量の学習データが必要となる．しかし，大量の学習データを収集するには，大きなコストがかかる．そこで，言い換え文を用いて学習データを増加させる手法がある．

本研究では，外部コーパスを利用せずに，折り返し翻訳を用いて言い換え文を生成し，学習データを増加させる．先行研究 [1] と同様に折り返し翻訳を行い元の文と折り返し翻訳をされた文を比較する．そして，2 の文が一致した場合のみ言い換え生成を行う．このことにより，翻訳誤りが減少し，高い精度の疑似データを作成することができると考えられる．そして，生成された言い換え文を用いて学習文を増加させ，機械翻訳の精度を調査する．

2　関連研究

2.1　言い換え生成

学習データを増幅させるために，言い換え生成を行う．言い換え生成には様々な手法がある．先行研究 [1] では，日本語の言い換え生成を行った．NMT を用いて折り返し翻訳を行い, 翻訳結果が一致した場合のみ言い換え生成をした．また，言い換えの取得数を増加させるために出力候補数を 4 候補にした 4Best，乱数による出力の違いを考慮した 4 システムで実験を行った．表 1 に生成された言い換え文とランダムに選択された言い換え文 100 文に対する人手評価の結果を示す．

表 1　生成された日本語の言い換え文と人手評価結果

	1Best	4Best	4system
言い換え文	25,003	134,574	362,971
100 文の正解率 (%)	85	85	81

2.2　単言語コーパスを用いた学習データの増加

学習データを増幅させるために単言語コーパスを用いる手法がある．Sennrich[2] らは，ターゲット言語の単一言語コーパスはフレーズベースの統計的機械翻訳の流暢性を高めるための重要な役割を果たしていることから，NMT でも単言語コーパスを利用した．

単一言語コーパスを逆翻訳と組み合わせることにより，追加の並列学習データとして扱った．その結果，機械翻訳の精度の大幅な改善が見られた．

3　提案手法

本研究では，日英翻訳を基本としている．提案手法として，学習データにターゲット言語の言い換え文を追加した翻訳モデルとする．つまり，英語の言い換え文を追加する．また，ベースラインは言い換え文を学習データに追加していない翻訳モデルとする．表 2 に学習データの追加例を示す．

なお，言い換え生成には，ベースラインの学習データを用いている．そのため，外部コーパスを利用していない．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2　学習文の追加例

ベースライン	この豆腐は悪くなっている	This tofu has gone bad .
追加する英語の言い換え文	この豆腐は悪くなっている	This tofu has become bad .
ベースライン	我々の努力は報われた。	Our efforts succeeded .
追加する英語の言い換え文	我々の努力は報われた。	Our efforts were rewarded .

4　実験

4.1　実験手順

本研究では以下の手順で実験を行う

1. ベースラインとして，言い換え文を利用しない翻訳モデルを作成する．
2. 提案手法として，英語の言い換え文を学習データに追加し，翻訳モデルを作成する．
3. 1,2 で作成した翻訳モデルを用いてテストデータで翻訳を行い自動評価を行う．

4.2　実験条件

本実験の実験条件を示す．NMT の学習 · 翻訳には OpenNMT-py[3] のバージョン 3.1.1 を用いて，デフォルトの設定で実験を行う．

4.3　対訳コーパス

本研究では，電子辞書などから抽出した日英単文対訳コーパスを用いる [4]．表 3 に日英対訳コーパスの一部を示す．また，追加する英語の言い換え文の一部を表 4 に示し，表 5 に生成された言い換え文とランダムに選択された言い換え文 10 文に対する人手評価結果を示す．

表 3　日英対訳コーパスの例

日本語文	英語文
警察はその犯人を追い詰めた。	The police tracked down the criminal .
私は毎夜 12 時まで英語の勉強をしています。	I study English every night until twelve .
評議会はこのほか軍改革委員会の創設などを決め、同日午後閉幕した。	The Council also decided to set up a military reform committee , and the meeting ended the same afternoon .

表 4　英語の言い換え文の例

入力文	言い換え文
There is a park on the south of our school .	There is a park on the south of the school .
All of the machines are working smoothly .	All the machines are working smoothly .
The library has a great many books .	There are a lot of books in the library .

表 5　生成された英語の言い換え文と人手評価結果

	1Best	4Best	4system
言い換え文	14,818	188,256	485,936
10 文の正解率 (%)	90	90	80

5　結果

5.1　ターゲット言語の学習データの追加量

本実験では，言い換え文の量を 3 つに変化させて実験を行った．表 6 に日英翻訳におけるターゲット言語である英語の言い換え文を追加したときの学習データの数を示す．

表 6　日英翻訳における学習データの量

対訳コーパス	言い換え文
163,188	14,818
163,188	188,256
163,188	485,936

5.2　自動翻訳結果

表 7 にテストデータ 16,328 文における BLEU による自動評価結果を示す．

表 7　日英翻訳における自動評価結果

	BLEU	METEOR	TER	RIBES
ベースライン	0.190	0.471	0.594	0.771
163,188 + 14,818	0.194	0.474	0.619	0.776
163,188 + 188,256	0.193	0.475	0.620	0.775
163,188 + 485,936	0.194	0.475	0.622	0.776

表 7 より，英語の言い換え文を追加することにより BLEU 値は少し増加した．しかし，追加する言い換え文の量を増加させても BLEU 値の変化がわずかであることが確認できる．

5.3　人手評価結果

表 8 に英語の言い換え文 14,818 文を追加したときと，ベースラインをランダムに選択された 100 文で人手で比較を行った結果を示す．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 8　ベースラインとの比較結果

ベースライン ○	163,188 + 14,818 ○	差なし
17	18	65

表 8 より，ベースラインと英語の言い換え文 14,818 文を追加した場合を比較すると差がないことが確認できる．

5.4　出力例

表 9 に人手評価で提案手法のほうが良い例を示す．また，表 10 に言い換え文を追加したほうが良い例を示す．

表 9　提案手法が良かった出力例

入力文 1	政府 は この 事業 の 具体 的 な 内容 を 示した 。
参照文	The government presented some of the features of the project that are concrete .
ベースライン	The Government showed specific concrete material of this project .
163,188 + 14,818	The government indicated the specific details of this enterprise .
入力文 2	彼ら は その 考え方 に 迷わ され た 。
参照文	They were led astray by that way of thinking .
ベースライン	They were struck by the idea .
163,188 + 14,818	They were captured by the idea .
入力文 3	太陽 が 堂々 と 現われた 。
参照文	The sun appeared majestically .
ベースライン	The sun appeared full .
163,188 + 14,818	The sun appeared brilliantly .

表 9 より，言い換え文を追加することにより，間違って翻訳されていた単語が正しい単語に置き換わっていることが確認できる．

表 10　ベースラインが良かった出力例

入力文 1	捕虜 たち が 逃亡 を 図った 。
参照文	The captives attempted to make their escape .
ベースライン	The captives made their escape .
163,188 + 14,818	The captives went on the escape .
入力文 2	その 店 の 前 に 駐車 した 。
参照文	She parked her car in front of the store .
ベースライン	I parked in front of the store .
163,188 + 14,818	He parked himself in front of the store .
入力文 3	しばしば 寒暖計 が 零 度 を 下回る 。
参照文	The thermometer often goes below zero .
ベースライン	The thermometer is often below zero .
163,188 + 14,818	The mercury is often over zero .

表 10 より，言い換え文を追加することにより，ベースラインでは正しく翻訳されていた単語が別の単語へ置き換わっていたり，不要な単語が増加していることが確認できる．

6　考察

6.1　ソース言語の言い換えの追加

表 7 より，ターゲットである英語の言い換え文を追加しても，BLEU 値は，あまり増加しなかった．そこで，ソース言語である日本語の言い換え文を追加したときの日英翻訳の精度を調査する．

6.1.1　ソース言語の学習データの追加量

表 11 に日英翻訳におけるソース言語である日本語の言い換え文を追加したときの学習データの量を示す．

表 11　日本語の言い換え文を追加した学習データ量

対訳コーパス	言い換え文
163,188	25,003
163,188	134,574
163,188	362,971

6.1.2　自動評価結果

表 12 にテストデータ 16,328 文における自動評価結果を示す．

表 12 より，日本語の言い換え文を追加すると，

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 12 日本語の言い換え文を追加した自動評価結果

	BLEU	METEOR	TER	RIBES
163,188 + 25,003	0.191	0.468	0.623	0.770
163,188 + 134,574	0.189	0.468	0.614	0.774
163,188 + 362,971	0.186	0.463	0.621	0.770

BLEU 値がわずかに減少したことが確認できる．

6.1.3 出力例

表 13 に出力例を示す．

表 13 ソース言語の言い換えを追加したときの出力例

入力文 1	私 は 電車 事故 で 足留め を 食った 。
参照文	I was stranded as a result of the train accident .
ベースライン	I was killed in the train accident
163,188 + 25,003	I was misled in the train accident
163,188 + 134,574	I was detained in a train accident in a train accident .
163,188 + 362,971	I was moved by a train accident in a train accident .
入力文 2	富士山 は 昔 から たくさん の 絵描き に 描か れ た 山 だ 。
参照文	Mount Fuji has been painted by numerous artists since ancient times .
ベースライン	Mount Fuji is a long reflection of many (unk) from olden countries .
163,188 + 25,003	Mount Fuji has been the resort of a lot of senders-off from olden times .
163,188 + 134,574	Mount Fuji has been (unk) in many mountains since long ago .
163,188 + 362,971	Mount Fuji is a mountain responsible for many people from olden times .

6.2 精度が向上しない原因

表 7 と 12 より，日本語，英語の言い換え文を追加しても機械翻訳の精度にはあまり差がなかった．以下に精度が向上しなかった原因として考えられることを示す．

6.2.1 未知語

本研究では，外部コーパスを利用していない．そのため，ベースラインと比較すると，学習データにおける単語の出現回数は変化する．しかし，単語の種類は増加しない．よって，ベースラインで未知語となっている単語は言い換え文を追加しても未知語のままである．例として，表 13 の入力文 2 ではベースラインで未知語として出力されている部分は，言い換え文を追加しても未知語として出力されていたり，不適切な単語に置き換わっている．

6.2.2 不適切な言い換え文

学習文に追加した言い換え文には，不適切な言い換え文が含まれている．そのため，ベースラインで翻訳が成功していた文が，不適切な言い換え文が追加することにより翻訳が失敗する場合がある．例として，表 10 の入力文 2 では，言い換え文を追加すると ” parked “ の目的語が不適切な単語になっている．

6.3 今後の課題

本研究では，外部コーパスを用いていないため，翻訳の情報量は増加していない．そのため，精度が向上していないと考える．そこで今後として，単語の情報量を増加させるために，外部の単一コーパスを用いて言い換え文を作成して，学習データに追加することを検討したい．

7 おわりに

本研究では，学習データに言い換え文を追加することにより，学習データを増加させ，機械翻訳の精度向上を試みた．結果より，ターゲット言語である英語の言い換え文を追加した場合では，BLEU 値はわずかに向上した．

精度がわずかしか向上しなかった大きな原因として，未知語が影響していると考えられる．そこで今後として，外部の単一コーパスを用いることにより，機械翻訳の精度の調査を行いたい．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] 松本武尊, 村上仁一. 折り返し翻訳を利用した言い換え生成. 言語処理学会第 29 回年次大会, 2023.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. **arXiv preprint arXiv:1511.06709**, 2015.

[3] OpenNMT. https://opennmt.net/OpenNMT-py/.

[4] 村上仁一, 藤波進. 日本語と英語の対訳文対の収集と著作権の考察. 第一回コーパス日本語学ワークショップ, pp. 119–130, 2012.

[5] 田中慎太郎, 飯間ほか. 往復翻訳を教師とした言い換え生成モデルによる高速テキストデータ拡張. 人工知能学会全国大会論文集 第 37 回 (2023), pp. 2E5GS604–2E5GS604. 一般社団法人 人工知能学会, 2023.

[6] Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5075–5086, 2021.

[7] 杉山普, 吉永直樹ほか. 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. 研究報告自然言語処理 (NL), Vol. 2019, No. 14, pp. 1–5, 2019.

[8] 矢野貴大, 村上仁一. ニューラル機械翻訳に乱数が与える影響. 言語処理学会第 27 回年次大会, 2021.

[9] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. **arXiv preprint arXiv:1508.04025**, 2015.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).