

NMT による複文翻訳における単文対訳学習文の利用

丸山京祐¹ 村上仁一²¹ 鳥取大学工学部電気情報系学科² 鳥取大学工学部¹b19t2110u@edu.tottori-u.ac.jp²murakami@tottori-u.ac.jp

概要

NMT を使った複文翻訳においては複文対訳文を学習して翻訳する。しかし、多くの複文対訳文を集めることには限界がある。そこで、複文対訳文だけでなく、単文対訳文を学習に使用することで複文翻訳の精度を向上させることを考えた。

本論文では、複文翻訳の学習文に複文対訳文だけでなく、単文対訳文も同時に学習した結果を示す。

1 はじめに

NMT において翻訳する文と同じ分野の文を学習することが望ましいとされている [1]。そのため、複文翻訳においては複文対訳文を学習して翻訳する。しかし、多くの複文対訳文を集めることには限界がある。そこで、複文対訳文だけでなく単文対訳文を学習に使用することで複文翻訳の精度を上げることが考えた。複文は単文の組み合わせの構造になっている。そのため、単文の学習が翻訳の精度向上に影響すると考えている。

本論文では、NMT を使用して複文翻訳するときの単文対訳文の量の影響を調査する。

2 提案手法

複文は複数の単文を組み合わせた構造になっている。そのため、複文学習文に単文を追加することで複文翻訳の精度向上ができると考えた。

本研究では、以下の手法を提案する。

- 複文対訳学習文と単文対訳学習文を合わせたコーパスを作成
- 作成したコーパスを NMT で学習
- 学習した NMT で複文テスト文を翻訳

3 実験

本実験では、OpenNMT-py[2] を用いて翻訳実験を行う。

NMT に複文対訳学習文 9 万文のみを学習したモデルと半分の 4 万 6 千文のみ学習したモデルを作成し複文テスト文を翻訳する。そして、それぞれの複文対訳学習文に単文対訳学習文を 2 万文、4 万文、8 万文、16 万文を加えてモデルを作成し、それぞれのモデルで複文テスト文を翻訳する。

3.1 実験データ

実験には、電子辞書などの例文より抽出したコーパス [3] を用いる。使用するデータの内訳を表 1 に示す。

表 1 実験データ

単文対訳学習文	163,188 文
複文対訳学習文	92,427 文
複文テスト文	9,243 文

表 2 に単文対訳学習文の例文を示す。表 3 に複文対訳学習文の例文を示す。表 4 に複文テスト文の例文を示す。

表 2 単文対訳学習文の例文

1 ドル紙幣を崩した。	She broke a dollar bill into change.
あかりが一晩中ついている。	The lights have been burning all night.

表 3 複文対訳学習文の例文

毎日自転車に乗って通学しています。	I ride my bicycle to and from school every day.
毎日タイプライターで決まった分量だけの仕事をするという事は文筆家には優れた修業法だ。	A daily stint at the typewriter is excellent discipline for a writer.

表4 複文テスト文の例文

その国はこの戦いを前から予知してそれに備えてきた。
彼らは、わがXYZ社への訪問中、当社が完全にオープンで協力的であったことに同意するでしょう。

4 実験結果

4.1 自動評価結果

複文のテスト文 9243 文の翻訳結果の出力文に対して自動評価指標 BLEU[4] による自動評価を行った。表 5 に複文 9 万文に単文を加えた学習の自動評価結果を示す。表 6 に複文 4 万 6 千文に単文を加えた学習の自動評価結果を示す。

表5 複文 9 万文に単文を加えた学習の自動評価結果

学習文	BLEU
複文 9 万文	15.6
複文 9 万文 + 単文 2 万文	16.5
複文 9 万文 + 単文 4 万文	17.2
複文 9 万文 + 単文 8 万文	18.2
複文 9 万文 + 単文 16 万文	18.6

表6 複文 4 万 6 千文に単文を加えた学習の自動評価結果

学習文	BLEU
複文 4 万 6 千文	12.0
複文 4 万 6 千文 + 単文 2 万文	13.0
複文 4 万 6 千文 + 単文 4 万文	14.2
複文 4 万 6 千文 + 単文 8 万文	15.7
複文 4 万 6 千文 + 単文 16 万文	15.9

表 5, 6 より、複文に単文を加えるほど精度は向上していくことがわかる。

4.2 出力結果の例文

複文対訳学習文 9 万文に単文対訳学習文を加えて学習する実験の出力結果の例文を示す。表 7 に翻訳精度が向上した例文を示す。表 8 に翻訳精度が低下した例文を示す。

表7 翻訳精度が向上した例

テスト文	矢印に従えば出口に出られます。
参照文	Follow the arrows and you will find the exit .
複文 9 万文	Fossils come to the exit if there is gliding .
複文 9 万文 + 単文 2 万文	If you touch the arrows , you will come to the exit .
複文 9 万文 + 単文 4 万文	When you come to the vein , you will come to the exit .
複文 9 万文 + 単文 8 万文	The arrows will be attended by the arrows .
複文 9 万文 + 単文 16 万文	If you follow the arrows , you will come to the exit .

テスト文	ご存じのように、自動車業界は法律の要求を満たすことができないと主張しています。
参照文	As you may know , the automotive industry claims that requirements of the Act can not be met .
複文 9 万文	As you know , the car industry insists that there can not be a legal demand for the law .
複文 9 万文 + 単文 2 万文	As you know , the car line is under the point of obtaining law requirements .
複文 9 万文 + 単文 4 万文	As you know , we insist on the legal demands of the law .
複文 9 万文 + 単文 8 万文	As you know , I insist that the auto business is unable to meet the needs of the law .
複文 9 万文 + 単文 16 万文	As you know , I insist that the automobile industry can not meet the demands of law .

表8 翻訳精度が低下した例

テスト文	A B C 社を去るのは、まことになごり惜しいことです。
参照文	It is with considerable sorrow that I leave ABC .
複文 9 万文	It is a great regret to leave ABC .
複文 9 万文 + 単文 2 万文	It is quite regret that you leave ABC .
複文 9 万文 + 単文 4 万文	It is a very regrettable thing to leave ABC .
複文 9 万文 + 単文 8 万文	It is a great regret to have left ABC .
複文 9 万文 + 単文 16 万文	It is a pity that we leave ABC to leave .

4.3 人手評価結果

複文テスト文 9243 文を翻訳した結果から 100 文を抽出し、人手による対比較評価を行った。評価者

は著者 1 名である。表 9 に複文対訳学習文 9 万文のみの結果と学習文に単文対訳学習文 16 万文を加えた結果を比較した結果を示す。

表 9 複文 9 万文, 複文 9 万文 + 単文 16 万文の対比較評価

評価基準	合計
複文 9 万文 + 単文 16 万文 > 複文 9 万文	29
複文 9 万文 > 複文 9 万文 + 単文 16 万文	18
互角でどちらも良い	14
互角でどちらも悪い	39

表 9 から、複文 9 万文のみの結果より単文 16 万文を加えた結果の方が良い結果となっていることが示された。

4.4 実験のまとめ

実験結果より、複文対訳学習文に単文対訳学習文を多く加えて学習するほど複文翻訳の精度が向上することがわかる。

5 考察

5.1 学習文が単文対訳学習文のみの複文翻訳の精度

複文対訳学習文の必要性を確認するため、単文対訳学習文のみを学習した場合の複文翻訳の精度を示す。

5.1.1 出力結果の例文

学習文が単文対訳学習文のみの出力結果の例文を表 10 に示す。

表 10 単文対訳学習文のみの出力例

テスト文	ご存じのように、自動車業界は法律の要求を満たすことができないと主張しています。
参照文	As you may know, the automotive industry claims that requirements of the Act can not be met.
単文 2 万文	As you know, the Environment industry is known as a result of reducing the law.
単文 4 万文	As you know, the car needs to prevent the needs of a law.
単文 8 万文	As you know, car trade does not preclude the demands of legal demands.
単文 16 万文	As you know, the automobile industry lacks legal demands.

5.1.2 自動評価結果

複文のテスト文 9243 文を入力文として翻訳実験を行い、出力文に対して自動評価指標 BLEU による自動評価を行った。

表 11 単文対訳学習文のみの自動評価結果

学習文	BLEU
単文 2 万文	3.8
単文 4 万文	6.6
単文 8 万文	9.1
単文 16 万文	11.3

表 11 より、単文対訳学習文のみでは複文翻訳の精度がかなり低いことがわかる。表 5, 6 の結果と比較すると、複文対訳学習文の量が単文対訳学習文よりも精度に対する影響が大きいことがわかる。

5.2 BLEU 値の考察

複文対訳学習文のみの学習の結果と単文対訳学習文を加えた結果において、自動評価の有用性を調べる。そのために、BLEU 値の近い複文対訳学習文 9 万文の結果と複文対訳学習文 4 万 6 千文に単文対訳学習文 8 万文を加えた結果を人手で比較する。

5.2.1 出力結果の例文

複文対訳学習文 9 万文の結果と複文対訳学習文 4 万 6 千文に単文対訳学習文 8 万文を加えて実験したときの出力結果の例文を示す。複文対訳学習文 9 万文の結果が良い例を表 12 に示す。また、複文対訳学習文 4 万 6 千文に単文 8 万文を加えた結果が良い例を表 13 に示す。

表 12 複文対訳学習文 9 万文の結果が良い例

テスト文	真を尽くして説明したら、相手も分かってくれました。
参照文	I explained it to him in all sincerity, and he understood.
複文 9 万文	I tried to explain it, and the other person was understood by me.
複文 4 万 6 千文 + 単文 8 万文	I have tried to explain the truth, and I would like to know the other person
テスト文	俺の目の黒いうちはお前たちに不自由はさせぬ。
参照文	You shall not want for anything as long as I am alive.
複文 9 万文	You shall want for nothing as long as I live.
複文 4 万 6 千文 + 単文 8 万文	You shall have nothing with me in my eyes.

表 13 複文対訳学習文 4 万 6 千文 + 単文 8 万文の結果が良い例

テスト文	そのような危機に直面しても彼女が冷静なのに驚いた。
参照文	He wondered at her composure in such a crisis .
複文 9 万文	We were surprised at the face of such a crisis in the face .
複文 4 万 6 千文 + 単文 8 万文	I was surprised that she could face a crisis like that .
テスト文	摩耗や損傷がないか全部品を調べてください。
参照文	Examine all parts for wear or damage
複文 9 万文	Check the whole parts and damages there if there are no sunglasses .
複文 4 万 6 千文 + 単文 8 万文	Examine all parts for wear or damage

5.2.2 人手評価結果

複文テスト文 9243 文を翻訳した結果から 100 文を抽出し、人手による対比較評価を行った。評価者は著者 1 名である。表 14 に複文対訳学習文 9 万文のみの結果と複文対訳学習文 4 万 6 千文に単文対訳学習文 8 万文を加えた結果を対比較評価した結果を示す。

表 14 複文 9 万文、複文 4 万 6 千文 + 単文 8 万文の対比較評価

評価基準	合計
複文 9 万文 > 複文 4 万 6 千文 + 単文 8 万文	24
複文 4 万 6 千文 + 単文 8 万文 > 複文 9 万文	15
互角でどちらも良い	13
互角でどちらも悪い	48

表 14 から、人手評価では複文 9 万文のみの結果の方が少し良い結果となっていることがわかる。表 12 の 1 つ目に示す例文のように出力単語は良く見える。しかし、文全体では意味がおかしい。このような出力結果が複文 4 万 6 千文 + 単文 8 万文では比較的多く見られた。このことが自動評価と人手評価の結果の違いになっていると考えられる。

6 おわりに

本研究では、NMT による複文翻訳において複文対訳学習文と単文対訳学習文それぞれを学習した翻訳と複文対訳学習文に単文対訳学習文を加えて学習し翻訳をした。

その結果、複文翻訳において以下のことがわ

かった。

- 複文対訳学習文に単文対訳学習文を加えることで翻訳精度が向上する
- 複文対訳学習文の量は単文対訳学習文よりもさらに翻訳精度に大きく影響する
- 複文対訳学習文のみと単文対訳学習文を加えた学習文の比較において、単語よりも文全体の意味が間違えることが比較的多くなるので、自動評価が参考になりにくい

今後の研究では、本研究で使用した NMT のモデルとは異なる翻訳モデルについての調査を行う。

参考文献

- [1] Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Sentence selection and weighting for neural machine translation domain adaptation. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 26, No. 10, pp. 1727–1741, 2018.
- [2] Opennmt - open-source neural machine translation, 2017. <https://opennmt.net/>.
- [3] 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.
- [4] 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, 2002.