2 文の連結を用いた機械翻訳におけるデータ拡張

今藤 誠一郎　甫立 健悟　平澤 寅庄　金子 正弘　小町 守

東京都立大学

{kondo-seiichiro, hotate-kengo, hirasawa-tosho, kaneko-masahiro}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

ニューラル機械翻訳は翻訳精度が高いことで知られているが，一方で様々な課題も残されている．その課題の一つに長い系列の文に対する翻訳精度の低下が挙げられる．先行研究では統計的機械翻訳との精度を比較した際に，入力文が ある文長までの範囲ではニューラル機械翻訳の精度が上回ることが示されている．しかし一定の文長を越えるとニューラル機械翻訳の精度は下回り，文長の増加に伴い，著しく翻訳精度が低下することが指摘されている [1]．

さらに，ニューラル機械翻訳において学習データの量が精度に大きく影響することが示されており [1]，低資源の言語において翻訳モデルを作成する上で大きな課題となる．そのため，これまでにも限られた対訳コーパスに対して様々なデータ拡張方法が研究されてきた．例えば，単言語コーパスの逆翻訳や，元の対訳データにおける特定の単語を別の単語に置き換える，などを行うことで擬似的に対訳文を生成し，学習データとして追加する方法が提案されている [2, 3, 4]．

本研究では低資源の対訳コーパスしか得られない言語対という条件で，長い系列の文の翻訳に対して有効であると考えられるデータ拡張方法を提案する．具体的には，長い系列の文の翻訳精度が低くなるのは学習データに長い系列の文が少量しか含まれないためであると考え，そのような文を学習データに増やすことで長文に対する翻訳精度の向上を試みた．

図 1 に提案手法の概要図を示す．学習データからランダムに 2 文を選び，それらを連結させたデータを拡張データとして，元の学習データに追加する．この手法は学習データ内で唯一となるような長い系列を増やしたいという動機である．

英日コーパスを用いた低資源な設定のもとでの実験の結果，2 文を連結させたものを拡張データとして追加することで非常に長い系列の文に対する翻訳精度が高くなり，全体的にもスコアの向上が見られた．さらに同じデータ拡張方法の一つである逆翻訳の手法 [4] と提案手法を組み合わせることでより精度が向上することを確認できた．

2 関連研究

ニューラル機械翻訳では様々なデータ拡張の手法が提案されている．例えば，目的言語側の単言語データに対して逆翻訳を行うことにより擬似的にデータセットを作成する手法 [2] がある．これは，まず対訳コーパスを用いて原言語と目的言語を逆向きにモデルを学習させる．そしてそのモデルを用いて目的言語の単言語コーパスの文を原言語に翻訳させることで擬似的に対訳コーパスを作成し追加の拡張データとする．一方，単言語コーパスを用いずに元の対訳コーパスを逆翻訳させることで学習データを拡張する手法 [4] も提案されており，このデータ拡張方法でも翻訳精度を高めることが可能であることが示されている．今回提案するデータ拡張の手法はこれらの先行研究で示された手法と組み合わせることが可能である．

またニューラル機械翻訳は長い系列の文の翻訳精度が低くなることが先の研究で示されている．文献 [1] ではニューラル機械翻訳における文長に関する翻訳精度の分析が統計的機械翻訳との比較を通して行われている．英語-スペイン語の翻訳において，入力文の文長が 60 を越えるとニューラル機械翻訳の精度は統計的機械翻訳の精度を下回り，80 を越えると大幅に翻訳精度が低下することが示された．この長い系列の文における翻訳精度の低下は，ニューラル機械翻訳の生成文の文長が非常に短く

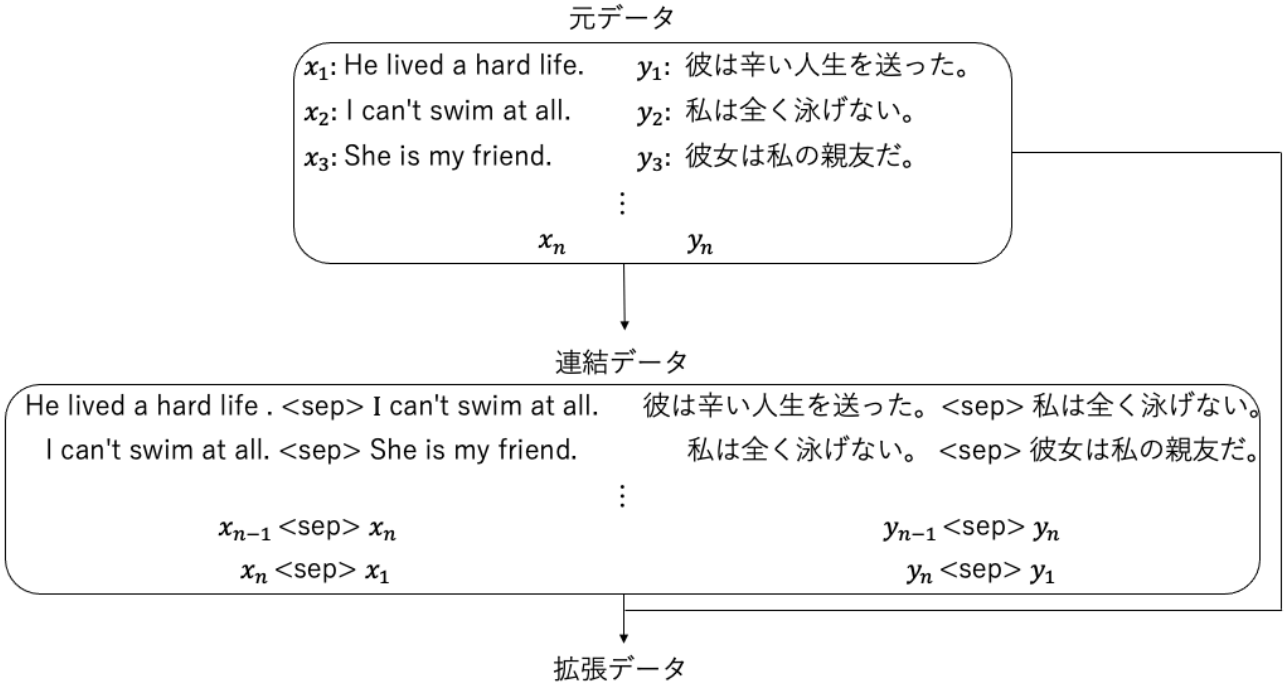

図 1　2 文連結によるデータ拡張（各言語対はランダムにサンプリングされているので文脈的なつながりはない）

なっているためであると述べられている．文献 [5] ではニューラル機械翻訳の中でも絶対的な位置情報を利用したモデルと相対的な位置情報を利用したモデルの比較が行われている．具体的には英語-日本語の翻訳において，学習データに文長の制限をかけて学習させたモデルを用いて分析されている．絶対的な位置情報を用いた Transformer モデルは他の相対的位置情報を用いたモデルに比べて，学習データに含まれない長い系列の文に対する翻訳精度が著しく低下することが示された．

3　2 文の連結を用いたデータ拡張

元の学習データ内にある原言語文 2 文とそれに対応する目的言語文 2 文をそれぞれ繋げたものを学習データとして追加する．つまり図 1 のように文の連結によって元のデータセットと同じ量の連結データを作成し，そのデータを元の学習データに加えることで学習データの文数を 2 倍にする．連結する際には文脈を考慮せずにランダムに 2 文を連結するが，学習データ全体としては 1 つの文がデータセット内に 3 度出現するようにする．

元のデータに連結したデータを追加するのは，単文の翻訳を元データで学習しながら，2 文の連結されたデータで長い系列の翻訳を学習させるためである．また，文を連結する際には，2 文の間に特殊トークンとして "<sep>" を挿入する．"<sep>" で区切ることで 2 文が別々の文であることを認識しながら，後方における位置エンコーディングを学習時に扱えるようになり，長い系列の文に対する翻訳精度の向上が期待できる．

学習時のデータには 2 文が連結したものが含まれているが，テスト時には 1 文のみを入力として扱う．[1)]

4　実験

4.1　実験設定

WAT17 [6] より ASPEC[2)] のデータを利用して英日翻訳を行なった．このデータセットは学習データ 2,000,000 文対，評価データ 1,790 文対，テストデータ 1,812 文対からなり，語彙サイズ 16,384 で Sentence Piece [7] によりサブワード化されている．この学習データからランダムサンプリングで 400,000 文対を抽出し，本実験での学習データとして利用した．

実験は Fairseq [8][3)] で Transformer を実装して行なった．パラメータとして，最適化には Adam，dropout の確率は 0.3，max-update が 300,000，ミニバッチに含まれる token 数が 65,536 以下となるように設定した．

1）テスト時に 2 文を入力とした実験も行ったが，1 文を入力とした時の方が精度の向上が見られた．
2）http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2017/snmt/index.html
3）https://github.com/pytorch/fairseq

表 1　学習データ 400,000 文で学習した際のテストデータセットにおける文長ごとの BLEU スコア
（ただし vanilla+BT+concat は vanilla のデータと BT で得られたデータ，各々に対して concat したデータからなる．）

文長	all	1 ~ 10	11 ~ 20	21 ~ 30	31 ~ 40	41 ~ 50	51 ~ 60	61 ~ 70	71 ~
文数	1,812	73	529	600	341	164	74	18	13
vanilla (400K)	26.5	22.9	23.0	26.2	27.1	29.6	28.5	28.8	23.6
+ OS (+ 400K)	25.0	21.8	23.0	24.9	25.7	27.0	27.0	22.5	18.1
+ concat (+ 400K)	26.6	21.4	23.3	25.7	27.5	29.5	28.7	28.2	29.0
+ BT (+ 400K)	28.8	24.3	25.5	28.3	29.5	31.6	30.6	28.7	28.7
+ BT + concat (+ 1.2M)	**29.4**	**25.4**	**25.6**	**28.6**	**30.1**	**33.1**	**31.5**	**29.9**	**30.1**

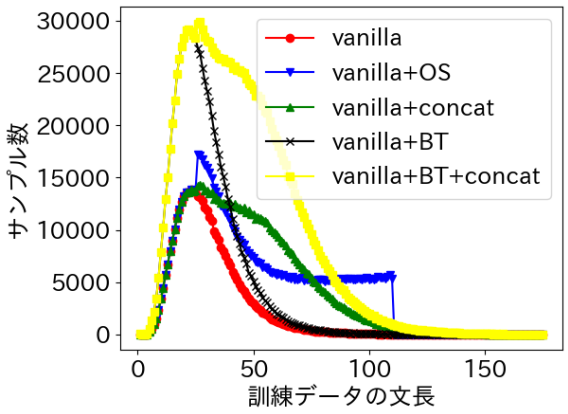

図 2　文長ごとの各データセットの分布

逆翻訳を行う際にも同じ設定の Transformer モデルを利用し，学習データは文献 [4] に従い，外部の単言語コーパスを用いずに行なった．この手法では目的言語から原言語への翻訳の際に学習に用いたデータをそのまま用いるため，一つの目的言語文に対して 2 つの原言語文が存在するようなデータセットとなる．

評価には BLEU スコア [9] を利用し，テストデータに対するモデルの出力全体のスコアと，テストデータを入力文の文長ごとに分類したものそれぞれに対するスコアを測定した．シードを変えて 3 回実験を回したときの BLEU スコアの平均を取った．

4.2　実験手順

400,000 文対の学習データから

1. 元のデータ (vanilla)
2. オーバーサンプリングで拡張したデータ (vanilla+OS)
3. 2 文の連結で拡張したデータ (vanilla+concat)
4. 逆翻訳で拡張したデータ (vanilla+BT)
5. 元のデータと逆翻訳したデータ，その双方に対して 2 文の連結を適用したものの複合データ (vanilla+BT+concat)

の 5 種類の学習データを用いて英日翻訳モデルを学習させ，テストデータに対する BLEU スコアで比較を行なった．

2 のオーバーサンプリングは，2 文連結と同様に長い文を増やすという動機で，提案手法の比較対象として実験を行なった．文長 25 以上 110 以下の文を対象にサンプリングを行い，学習データを 2 倍とした．3 でデータを連結する際には長い系列の文の翻訳精度の向上を目的とするため，連結した際に文長が 25 以下になるものは取り除いた．4 と 5 では低リソースの設定なので，先行研究でも行われている手法と組み合わせたときの有効性を調べるため，逆翻訳を利用した手法を利用した．

上記のデータ拡張手法で作成した原言語文 (英文) の文長ごとの学習データ数を図 2 に示す．vanilla+OS，vanilla+concat，vanilla+BT のいずれも同じサンプル数であることに注意．

4.3　結果

5 種類のデータを用いて学習を行った結果を表 1 に示す．テストデータ全体の BLEU スコアと文長ごとにテストデータを分類して測った BLEU スコアを記載している．

vanilla+OS の結果を見ると vanilla の結果に比べて全体的な精度の低下が確認でき，データ数を増やしたはずの長い系列の文に関しても精度が悪化していることが確認できる．これはオーバーサンプリングによって追加されるデータのうち元のデータが少ない部分，つまり極端に長い系列の文が多くサンプリングされるため，コーパス内に同一の文が大量に含まれており，それらがノイズになってしまったためであると考えられる．

vanilla+concat の結果を見ると全体としての精度が僅かに上回っていることが確認できる．文長ごとに

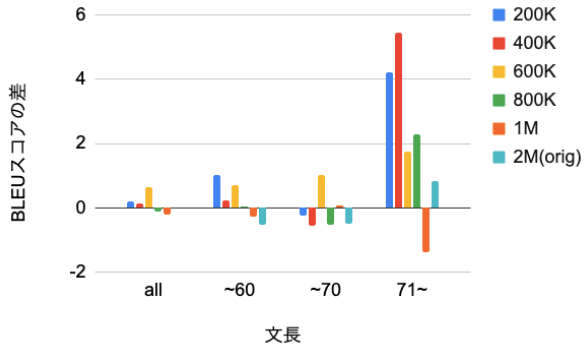

図 3 文長別に見たデータサイズごとの提案手法の有効性（縦軸は vanilla+concat の BLEU スコアから vanilla の BLEU スコアを引いた値）

精度を確認すると，vanilla と比べて文長が 61-70 に分類される部分の BLEU は微減であるものの，2 文の連結によって拡張されたデータの多くである文長 50 以上の翻訳に対して強くなっている．一方で特に短い文における翻訳精度が大きく下がってしまっていることも確認できた．

vanilla+BT と vanilla+BT+concat を比較すると，全体では 0.6 ポイントの向上が見られた．文長ごとの精度を見ると特に文長 41 以上のスコアが大きく改善されており，提案手法を他のデータ拡張の手法と組み合わせることで，長い文における翻訳精度をより一層高めることができることを確認できた．さらに，短い文における翻訳精度も低下していないことから，短い文の翻訳精度を維持しながら長い文の翻訳精度を高めることができるという点において，他の手法と組み合わせることの有効性が示せた．

4.4 データサイズと提案手法の分析

図 3 に各データサイズにおける今回の提案手法による BLEU スコアの変化を文長別に示す．全体としてはデータ量が 800,000 文を越えると，この提案手法による精度の改善が見られないことが確認できる．また今回の提案手法は長文の翻訳精度の向上を目的に提案したが，文長 51 以上に着目するとこちらもデータ量が 200,000 文〜600,000 万文の時は精度の向上が見られるものの，1,000,000 文を越えると精度の低下が見られる．今回の提案手法が学習データが豊富なときには適しておらず，低資源のデータを扱う際に有効であることが言える．

4.5 実例分析

本実験において提案手法が有効に働いたと考えられる事例を付録の表 2，表 3 に，今回の提案手法の目的に反して翻訳が悪化したと思われる事例を付録の表 4 に示す．

表 2 の例を見ると vanilla では本来出力すべき文よりも短い文を出力しており，翻訳にあたって必要な情報が欠落していることが確認できる．一方で提案手法を用いてデータ拡張を行なった vanilla+concat の出力では，より長い文の出力が実現され，情報の欠落を低減できていることを確認できる．

表 3 の例は BT と組み合わせた際に翻訳が改善された例となっている．こちらも先の例同様，vanilla+BT の結果では文前半の情報が完全に失われてしまっているが，vanilla+BT+concat では文全体の情報を捉えた翻訳文が生成されている．他の手法に提案手法を組み合わせた際にも，提案手法の効果が現れていることが確認できる．

しかし表 4 の例のように，提案手法を用いたデータ拡張を施す前のデータで学習したモデルの出力には見受けられなかったような繰り返し出力が，提案手法を用いたデータ拡張を行った際の出力結果に見られる場合も存在した．このような出力は例のような文長の短い事例においてより多く見られた．このことから長い文の出力が可能なゆえに，長い文を生成しようとしてしまうため，不自然な繰り返し出力を行ってしまうことがある可能性が考えられる．

5 おわりに

本研究では，長い系列の文の翻訳精度を高められるようなデータ拡張方法を提案し，実験を行なった．結果として，2 文の連結によるデータの拡張方法が非常に簡単なデータ拡張方法でありながら，特に非常に長い系列の文の翻訳に対して有効に働くことが確認できた．しかし，短い文における翻訳精度を犠牲にしてしまい，全体としての精度にあまり変化が見られないような傾向も見受けられた．一方，他のデータ拡張の手法と組み合わせることで，短い文での精度を下げることなく，全体の精度の改善が見込めることも判明した．

今後は他の手法と組み合わせる際に，組み合わせる手法によって提案手法がどのように作用するのか，本研究のように長い文に対して翻訳精度を向上させられるのかを調査したい．

参考文献

[1] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39, 2017.

[2] 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*, pp. 86–96, 2016.

[3] Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 856–861, 2018.

[4] Kenji Imamura and Eiichiro Sumita. NICT selftraining approach to neural machine translation at NMT-2018. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 110–115, 2018.

[5] Masato Neishi and Naoki Yoshinaga. On the relation between position information and sentence length in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 328–338, 2019.

[6] Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. Overview of the 4th workshop on Asian translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pp. 1–54, 2017.

[7] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018.

[8] 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, 2019.

[9] 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.

A　付録

表 2　出力例 1（提案手法が有効に働いた例，原言語文の文長 71〜の事例）

src	Myanma is behind in market economization together with Laos, Canbodia, Vietnam, and the GDP per one person is the lowest in the 4 countries, and it remains $ 180, but Myanma is thought to remarkably develop if political problems are solved, because flatland occupies 7 × 10% of the land and natural resources are rich, and because personnel expenses are extremely cheap.
tgt	ミャンマーは自国とともに後発ＡＳＥＡＮ４カ国といわれるラオス，カンボディア，ベトナムと比較しても市場経済化が遅れ，一人あたりのＧＤＰは最低で１８０ドルにとどまっているが，平地が７割で天然資源もあり，人件費が極端に安価なので，政治的問題が解決されれば著しく発展すると見られる。
vanilla	ミャマは，陸上と自然資源の７割を占めるため，平地は土地と自然資源の７割を占めるので，人件費が極端に安く，４か国で１人当たりＧＤＰが最低である。
vanilla +concat	ミャンマーはラオス，カンボジア，ベトナムと共に市場経済化に遅れ，４国ではＧＤＰが１人あたり最低であるが，国土の７割を占める平坦な土地と自然資源が豊富で 人件費が極端に安く政上の問題が解決されれば，顕著に発展すると考えられる。

表 3　出力例 2（提案手法が有効に働いた例，原言語文の文長 61〜70 の事例）

src	Results of the analysis shows high accuracy properties, such as the reproducibility of relative standard deviation 0.3〜0.9% varified by repetitive analyses of ten times, the clibration curves with correlation coefficient of 1 verified by tests of standard materials in using six kinds of acetonitrile dilute solutions, and the formaldehyde detection limit of 0.0018 μ g/mL.
tgt	結果は，相対標準偏差０．３〜０．９%の再現性（１０回の繰返し分析），相関係数１の検量線（６種類のアセトニトリル希釈溶液による標準資料の検定），０．００１８μｇ／ｍＬのホルムアルデヒド検出限界，など高い精度を得た。
vanilla +BT	６種のアセトニトリル希薄溶液を用いた標準物質の試験及びホルムアルデヒド検出限界は０．００１８μｇ／ｍＬであった 。
vanilla +BT +concat	分析の結果は１０回の繰り返し解析で相対標準偏差０．３〜０．９%の再現性，６種のアセトニトリル希薄溶液を用いた標準物質の試験により検証された１の相関係数 を持つクライテリア曲線，及び０．００１８μｇ／ｍＬのホルムアルデヒド検出限界など高い精度を示した。

表 4　出力例 3（提案手法がエラーを引き起こしたと考えられる例，原言語文の文長 11〜20 の事例）

src	These seemed to be noticeable complications in case of extracorporeal circulation for umbilical hernia repair .
tgt	さい帯ヘルニア修復術における体外循環の合併症として注目すべきと思われた 。
vanilla	さい帯ヘルニア修復術における体外循環の合併症として注目すべきと思われた 。
vanilla +concat	以上の所見より，さい帯ヘルニアに対する体外循環では，特に合併症として特に合併症として，特に，さい帯ヘルニアでは体外循環がより注意を要すると考えられた。