ラウンドトリップ翻訳を用いたニューラル機械翻訳のデータ拡張

紺谷 優志　秋葉 友良　塚田 元
豊橋技術科学大学
{kontani.yushi.qu,akiba.tomoyoshi.tk,tsukada.hajime.hl}@tut.jp

概要

ニューラル機械翻訳 (NMT) は学習に膨大な規模の対訳コーパスを必要とするが，ドメインによっては大量の学習データを用意することが難しい場合もある．本研究では，ラウンドトリップ翻訳を用いて学習データを追加のリソースを用いることなく拡張する手法を提案する．疑似対訳コーパスを用いて翻訳モデルの性能を向上させる手法である Iterative Back Translation に本手法を結合した実験を行った結果，本手法がモデルの翻訳性能を有効に向上させられることが分かった．

1　はじめに

近年，機械翻訳手法の一つであるニューラル機械翻訳（NMT）が従来手法である統計的機械翻訳を大きく上回る翻訳性能を持つことが報告されている．しかし NMT を構築するためには数十万〜数百万文という大量の対訳コーパスが必要であり，十分な量の対訳コーパスを用意できない場合は十分な性能の翻訳モデルを作ることはできない．また，翻訳モデルは学習データのドメイン（分野）にも強い影響を受ける．ドメインによっては大量のデータを用意することが困難な場合もあり，そのような場合においても良質な翻訳モデルを構築するのは困難となる．

本研究では，ある言語の文を別の言語の文に翻訳し，その結果を元の言語の文に逆翻訳するプロセスであるラウンドトリップ翻訳を用いて，対訳コーパスをデータ拡張する手法を提案する．そして本データ拡張手法を Iterative Back Translation と組み合わせることで，追加の学習データなしで翻訳精度を向上できることを示す．

2　関連研究

モデルの精度向上の方法として，対訳コーパスに何らかの操作を行うことで，データを拡張する手法が提案されている．Fadaee ら [1] は，大量の単言語データで学習した言語モデルを活用し，対訳文中の一部の単語を低頻度語に入れ替え学習データに加える方法を提案している．張ら [2] は，単語アライメント情報を用いて対訳コーパスから抜き出した部分的な対訳文を用いて学習データを拡張する方法を提案している．

また，別の学習データを活用するアプローチも取られている．Sennrich ら [3] は目的言語の単言語コーパスを原言語へと逆翻訳して疑似対訳コーパスを生成し，対訳コーパスと合成して翻訳モデルの再学習に使う手法を提案した．森田ら [4]，Hoang ら [5]，Zhang ら [6] は Sennrich ら [3] の手法を双方向かつ反復的に拡張することで，2 つの翻訳方向のモデルを相互に改善する方法である Iterative Back Translation(IBT) を提案した．IBT では対訳コーパスと，比較的入手が容易な単言語データを用意し，対訳コーパスを単言語データでデータ拡張し，各方向の翻訳モデルを反復更新する．IBT によって，Sennrich ら [3] の手法を大幅に上回る翻訳性能を達成できることが報告されている．

3　提案手法

3.1　データ拡張手法

本論文では random sampling を交えたラウンドトリップ翻訳を行うことで，データ拡張を行う手法を提案する．ここで，ラウンドトリップ翻訳とは，ある言語の文を別の言語の文に翻訳し，その結果を元の言語に逆翻訳するプロセスのことを指す．また，random sampling は文の翻訳を行う際，出力される単

語にランダム性を加える方法である．NMT モデルは対訳コーパスから入力文に対する出力文の事後確率分布を学習する．通常の翻訳では，この事後確率分布に従い，入力された単語列に対して最も出力される確率が高い単語列が出力される．それに対して，事後確率分布に基づきランダムに単語列を選択し出力するのが random sampling である．random sampling をラウンドトリップ翻訳と組み合わせることでパラフレージングに似た効果を期待できる．

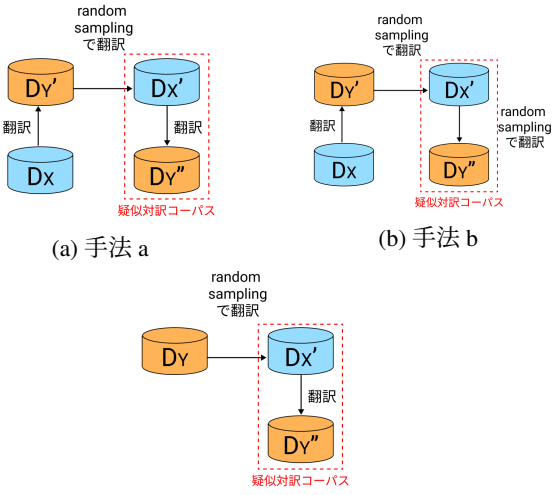
2 つの言語の文 D_X と D_Y から成る対訳コーパスを用いて X → Y 方向と Y → X 方向の翻訳モデルを学習したと仮定する．両方向の翻訳モデルを用いて，D_X を翻訳して言語 Y の文 D'_Y を生成し，D'_Y を翻訳して D'_X を生成したとき，対訳文 D_X-D_Y と疑似対訳文 D'_X-D'_Y は表現の大きく異なる対訳文になると考えられる．これが，本研究で提案するラウンドトリップ翻訳によるデータ拡張の基本的な流れである．

本論文では，ラウンドトリップ翻訳と random sampling を用いて目的言語の表現を多様化する 3 つのデータ拡張手法を提案する．なお，すべてのパターンにおいて，(D'_X, D''_Y) を疑似対訳コーパスとし，モデルの再学習に使用する．

(a)$D_X \rightarrow D'_Y \rightarrow D'_X$(random sampling で生成) → D''_Y（図??）

(b)$D_X \rightarrow D'_Y \rightarrow D'_X$(random sampling で生成) → D''_Y(random sampling で生成)（図??）

(c)$D_Y \rightarrow D'_X$(random sampling で生成) → D''_Y（図??）



(a) 手法 a　(b) 手法 b

(c) 手法 c

図 1: ラウンドトリップ翻訳で目的言語を多様化するデータ拡張手法（提案手法）

3.2 提案手法の双方向反復的適用

提案手法を用いて行う実験全体の流れを以下に示す．

1. 対訳コーパス D_X-D_Y を用いて X-Y 方向と Y-X 方向の翻訳モデルを学習する．
2. 両方向の翻訳モデルを用いて D_X に対してラウンドトリップ翻訳によるデータ拡張を行い，疑似対訳コーパス D^{Y-X}_{Xp}-D^{Y-X}_{Yp} を生成する．
3. 元の対訳コーパス D_X-D_Y と疑似対訳コーパス D^{Y-X}_{Xp}-D^{Y-X}_{Yp} を結合した学習データを用いて，Y → X 方向の翻訳モデルを更新する．

また，以上の手順を逆向きの言語方向に適用することで，X → Y 方向の翻訳モデルも構築できる．本実験ではこれらの手順を森田ら [4] の手法のように双方向に繰り返し適用することで両方向の翻訳モデルを更新していく．全体的な手順を図 2 に示す．

4 実験

4.1 データセット

学習用データセットとして TED の講演内容を書き起こした話し言葉のコーパスである IWSLT 2017 データセット [7] を使用した．training データは IWSLT2017 データセットに含まれる英日コーパス（223,108 文）を用いた．dev データと test データはデータセット内の 2010 年版のデータ（それぞれ 871 文と 1,549 文）を用いた．

前処理として英語文，日本語文ともに NFKC 正規化し，英語文は Moses[8] に付属するトークナイザーと truecaser でトークナイズと大文字小文字の表記統一を行った．学習前の事前処理として SentencePiece[9] によるサブワード化を行った．

4.2 実験設定

NMT システムには OpenNMT-py[10] の Transformer を使用した．エンコーダ，デコーダともに 6 層とし，隠れ層の次元を 512 とした．初期モデルは訓練データを 25000 ステップまで学習させて作成した．以降のモデルは 1000 ステップごとに保存した．性能評価には BLEU[11] を用いた．英日翻訳モデルを評価する際には，テストデータの翻訳結果をデトークナイズした後 MeCab[12] により分かち書きし評価を行った．また，本実験内で random sampling 翻訳

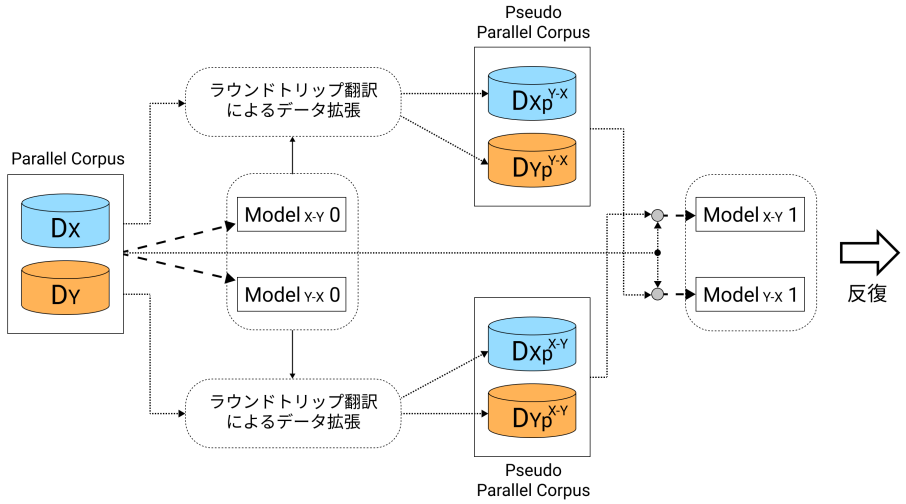


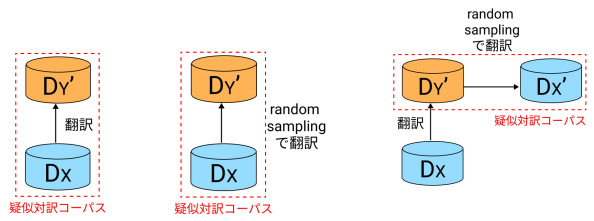
図 2: 提案手法を用いた双方向反復的データ拡張の流れ

を行う際は，いずれも事後確率の上位 10 語の中から語彙をランダムに選択し出力するオプションを指定した．

4.3 実験条件

3.1 節に示した 3 パターンのデータ拡張法で実験を行った．また，ラウンドトリップ翻訳による目的言語のデータ拡張の有効性を検証するため，図 3 に示すような比較手法の実験も行った．なお，すべてのパターンにおいて，最後の 2 つのデータを疑似対訳コーパスとする．また，図 1 と同様，図 3 中の内容は Y → X 方向の翻訳モデルを学習するための疑似対訳コーパスの作成内容である．X → Y 方向の翻訳モデルを学習するためには，図中の内容を逆向きの言語方向に行い，疑似対訳コーパスを作成する．

(d)$D_X \rightarrow D'_Y$（図??）
(e)$D_X \rightarrow D'_Y$(random sampling で生成)（図??）
(f)$D_X \rightarrow D'_Y \rightarrow D'_X$(random sampling で生成)（図??）



(d) 手法 d　(e) 手法 e　(f) 手法 f

図 3: 比較手法

手法 d は森田ら [4] の手法で単言語データから疑似対訳コーパスを作成する手順と同一である．また，手法 f は D'_Y と D'_X のペアを用いて Y → X 方向のモデルを学習している．全ての実験において baseline のモデル (モデル 0) には全て同一のモデルを使い，IBT で 4 回モデル更新を行った（モデル 1〜4）．

4.4 実験結果

手法 a〜f までの英日翻訳モデルと日英翻訳モデルの BLEU スコアの推移を図 4 と図 5 に示す．提案手法 a〜c はいずれもモデル 4 までの時点で，モデルの更新に従って BLEU スコアが順調に増加しており，中でも手法 a は突出したスコアを記録しており，英日方向の BLEU は baseline の 8.97 からモデル 4 で 10.09 まで，日英方向の BLEU は baseline の 9.53 から 10.77 まで増加した．それと比べ，比較手法 d〜f は BLEU スコアがごく僅かな増加に留まっており，モデル 1 やモデル 2 以降の段階で減少に転じているものがほとんどである．

以上のことから，提案法により元の対訳コーパスのみでモデルの性能を改善できること，比較手法よりも性能向上の効果が大きいこと，ラウンドトリップ翻訳によるデータ拡張の構成として

- $D_X \rightarrow D'_Y \rightarrow D'_X$(random sampling で生成) → D''_Y

という組み合わせが最も良い効果を発揮することが示された．本手法により，例えば表 1 の D'_X-D'_Y の

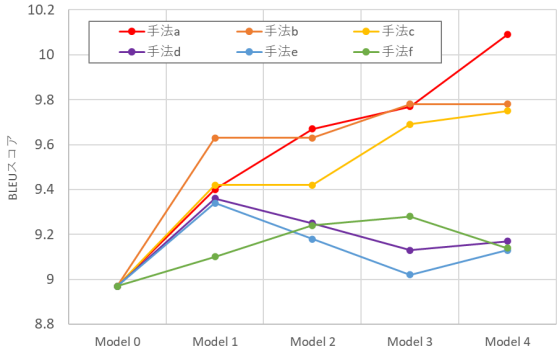


図 4: 英日翻訳における各手法の BLEU の推移

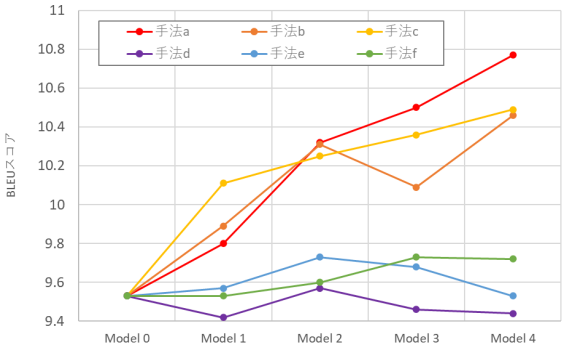


図 5: 日英翻訳における各手法の BLEU の推移

ような新たな対訳が学習データに追加される．

ラウンドトリップ翻訳を用いない手法 d〜f に関して，手法 d ではモデルの学習に使用した学習データをそのまま疑似対訳コーパスの作成に流用している．そのため，元の対訳コーパスと疑似対訳コーパスの内容がほとんど変化しないためデータ拡張としての意味をなさず，モデルの精度向上にもつながらなかった．

手法 e は手法 d と異なり原言語側 D'_Y の表現は多様化するが，目的言語側 D_X の表現は多様化しない．このことが，翻訳精度向上に結びつかない原因だと考えられる．手法 f は D'_Y が D_Y とほぼ同じになるため，random sampling を用いて目的言語側の表現を多様化させてはいるものの，self training と同等の処理となるため，翻訳精度向上に結びつかなかったと考えられる．

5 おわりに

本研究ではラウンドトリップ翻訳を用いたデータ拡張手法を提案し，提案手法と IBT の手法を結合した実験を行いその有効性を検証した．実験の結果，4 度のモデル更新を通して BLEU が英日方向で 1.12，日英方向で 1.24 の増加を達成し，本手法が元の対訳コーパスだけでモデルの精度を向上させられること

表 1: 手法 a での翻訳結果の一例

D_X（日本語文）	新しい 貝 殻 が いくつか 見つかります
D_Y（英語文）	every time the tide comes in and out , you find some more shells .
D'_X（日本語文）	潮 が 満 ち たとき より多くの 貝 が 見つかります
D''_Y（英語語文）	when the tide is full, you get more clams .

を示した．また，比較手法により目的言語の多様化が精度向上の鍵となっていることが示唆された．

謝辞

本研究は JSPS 科研費 19K11980 および 18H01062 の助成を受けた．

参考文献

[1] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resourse neural machine translation. In *Proc. 55th Annual Meeting of the Assoc. for Computational Linguistics (Volume 2: Short Papers)*, pp.567-573, Vancouver, Canada, 2017.

[2] 張津一, 松本忠博. ニューラル機械翻訳における長文分割によるコーパスの拡張. 言語処理学会第 25 回年次大会 発表論文集, pp. 683-686, 2019.

[3] 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, 2016.

[4] 森田知熙, 秋葉友良, 塚田元. 双方向ニューラル機械翻訳の反復的な教師なし適応の検討. 言語処理学会第 25 回年次大会 発表論文集, pp. 1451-1454, 2019.

[5] Vu Cong Duy Hoang, Phiilpp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24, 2018.

[6] Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. Joint training for neural machine translation models with monolingual data, In *AAAI*, 2018.

[7] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation(IWSLT)*, 2017.

[8] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual*

Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177-180, 2007.

[9] Taku Kudo, 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.

[10] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation, In *Proceedings of ACL 2017, System Demonstrations*, pp. 67-72, 2017.

[11] 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.

[12] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230-237, 2004.