双方向翻訳モデルと反復的逆翻訳を用いた低資源言語に対するニューラル機械翻訳の性能向上

Bui Tuan Thanh　　秋葉 友良　　塚田 元

豊橋技術科学大学

{bui.tuan.thanh.mg, akiba.tomoyoshi.tk, tsukada.hajime.hl}@tut.jp

概要

本稿では、低資源言語に対して翻訳性能を改善する手法を提案する。提案法は、２つの手法を組み合わせて適用することで、追加の学習データなしに少量の対訳コーパスだけを用いて、翻訳モデルを学習する。まず、対訳コーパスを用いて、翻訳対象言語対の双方向の翻訳モデルを一つのモデルで構成する双方向翻訳モデルを構築する。次に、同じ対訳コーパスを単言語コーパス対として用いて、反復的逆翻訳 (IBT: Iterative Back Translation) を適用することで、翻訳モデルを更新する。IWSLT2015 の英語・ベトナム語翻訳タスクの低資源対訳コーパス (13 万文) を使用した英越及び越英の翻訳実験により、提案手法の有効性を示した。

1　はじめに

近年、ニューラルネットワークを用いたニューラル機械翻訳 (Neural Machine Translation:NMT) が機械翻訳の品質を飛躍的にに向上させている [1,2] 。ニューラル機械翻訳の翻訳モデルを学習するためには、大規模かつ品質が高い対訳コーパスが必要であるが、そのような対訳コーパスを作成するのはコストが高い。そのため、比較的入手が容易な単言語コーパスでデータ拡張する手法が提案されている。中でも、逆翻訳を用いる手法 [3] やそれを反復的に適用する手法 [4,5] は効果的な手法として知られている。また、Ding ら [6] は、翻訳対象の２言語間の双方向翻訳を単一の翻訳モデルとして学習することで、各方向の翻訳性能を向上できることを報告している。

本稿では、Ding ら [6] の双方向ニューラル翻訳 (BiNMT: Bidirectional Neural Translation) モデルと、対訳コーパスを単言語コーパスとして利用した反復的逆翻訳手法を組み合わせることにより、元の対訳コーパス以外の追加の学習データを使わずに翻訳性能を向上させる手法を提案する。本手法では、最初に対訳コーパスで BiNMT の初期モデルを学習する。次に、学習した BiNMT モデルで、対訳コーパスの各言語側を単言語コーパスとみなして逆翻訳を行い、２つの疑似対訳データを生成する。元の対訳学習データと生成した疑似対訳データを混合し、新たな学習データを構築する。新たな学習データを用いて、初期 BiNMT モデルを fine-tuning し、新しい BiNMT モデルを得る。新しいモデルは、再び擬似対訳データを生成するために利用され、元の対訳学習データと組み合わせて新規学習データを得て、これを用いて再度 BiNMT モデルを更新する。このプロセスを繰り返すことにより、翻訳モデルの性能を改善していく。本手法は、元の対訳コーパス以外の学習データを使用しないため、低資源言語の翻訳モデルの性能向上に利用することができる。

提案した手法の有効性を検証するために、英越言語対に対して実験を行った。英越言語対の単方向翻訳モデルと比較し、双方向翻訳モデルの利用のみでも翻訳性能を有意に改善できた。さらに反復的逆翻訳と組み合わせることで、さらに翻訳性能が向上することを確認した。

2　関連研究

2.1　ニューラル機械翻訳の逆翻訳手法

Sennrich らはデータ拡張手法として目的言語側の単言語データを利用するシンプルで効果的な逆翻訳法 (Back Translation) を提案した [3]。この手法では、最初に目的言語から原言語の方向の翻訳モデルを学習し、目的言語の単言語データを原言語に翻訳し, 疑似対訳コーパスを生成する。次に、擬似対訳コーパスを元の対訳コーパスと混合して原言語から目的言語の方向の翻訳モデルを学習する。この手法は目

的言語側の正しい文を使用することで、質の低い擬似的な原言語を用いても翻訳品質の改善が可能であることを示している。逆翻訳手法の有効性は他の論文 [7] でも報告されており、特に対訳コーパスが小規模なときに有効に働くことが示されている。

逆翻訳の関連研究として、Zhang ら [8] は原言語側の単言語データを翻訳し、疑似対訳コーパスを構成してモデルを学習するという自己学習 (self-learning) 手法を提案した。Hoang らと森田らは言語対の両方側の単語データを用いる反復的逆翻訳手法 (IBT: Iterative Back Translation) を提案した [4,5]。反復的逆翻訳手法では原言語側と目的言語側の単言語データを使用し、逆翻訳を繰り返すことで疑似対訳コーパスの生成、モデルの再学習を繰り返す。この手法は低資源言語対においても有効に働くことが示されている。Imamura ら [9] は、逆翻訳においてランダムサンプリングで逆翻訳を行う手法を調査し、ベストサンプリングより有効に働くことを報告した。

2.2 双方向ニューラル機械翻訳

Dinh ら [6] は、学習データを (原言語, 目的言語) のペアとそれを入れ替えた (目的言語, 原言語) のペアから成る 2 倍のサイズのデータに構成を変更し、翻訳対象の 2 言語間の翻訳を一つの双方向翻訳モデルで学習、最後に目的の単一方向学習データで fine-tuning することで、各方向の翻訳品質を改善できることを報告した。この手法では全学習時間の $\frac{1}{3}$ で双方向翻訳モデルを学習し、残る時間で単方向翻訳モデルを学習する。

3 提案手法

3.1 BiNMT モデル

対訳コーパスからソースとターゲットが逆方向の 2 つの学習データを構築し、ソース側の先頭に翻訳方向を示すタグを追加する。これらを連結して新たな学習データを構成し双方向翻訳モデルの学習に用いる。その際、単一方向の翻訳に学習が一方的に偏るのを避けるため、各方向の対訳を交互に配置するように学習データを構成する。双方向ニューラル機械翻訳モデルの構成を図 1 に示す。翻訳の対訳コーパスは X 言語と Y 言語から構成される。C_X は X 言語のコーパス、C_Y は Y 言語のコーパスである。

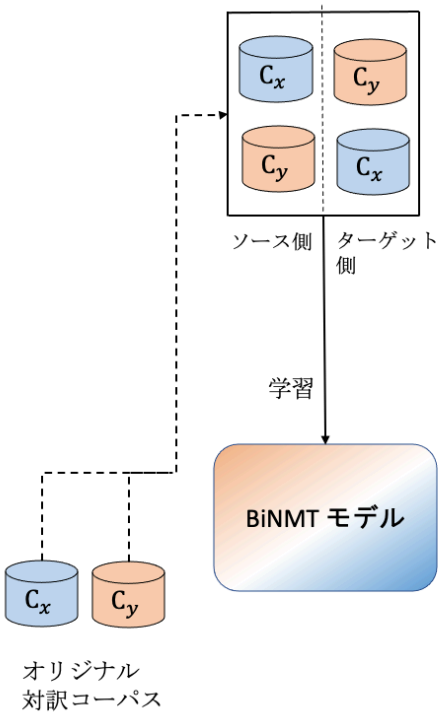

図 1 BiNMT モデルの構成

BiNMT の目的関数を次に式に示す。

$$\overleftrightarrow{\mathcal{L}}(\theta) = -\sum_{(\mathbf{x},\mathbf{y})} (\log p(\mathbf{y}|\mathbf{x};\theta) + \log p(\mathbf{x}|\mathbf{y};\theta)) \quad (1)$$

ここでは $\mathbf{x}$ が X 言語の文、$\mathbf{y}$ が Y 言語の文である。

3.2 BiNMT モデルに基づく反復的な逆翻訳

提案法は BiNMT と反復的逆翻訳を組み合わせる。ただし、外部の単言語コーパスは使わずに元の対訳コーパスを単言語コーパス対と見なすことで、新たな疑似対訳コーパスの生成、BiNMT モデルをの再学習を繰り返す。また、Imamura ら [9] と同様に、ランダムサンプリングを用いて逆翻訳を行い、疑似対訳コーパスを生成する。本手法に用いる反復的逆翻訳の手順を以下に示す。言語 X を言語 Y に翻訳することは X→Y, 言語 Y を言語 X に翻訳することは Y→X, 言語 X と言語 Y の双方向翻訳は X↔Y とする。

1 文対応のとれた言語 X と Y のコーパス対 C_X と C_Y コーパスから X↔Y の双方向翻訳モデルを学習する。このモデルをモデル 0 とする。

2 以下の手順で X↔Y の双方向の翻訳モデルを再学習する. i の初期値は 0 である。

2.1 双方向の翻訳モデル i を用いて X→Y 方向で単言語コーパス C_x を翻訳し、C_y^i を得る。

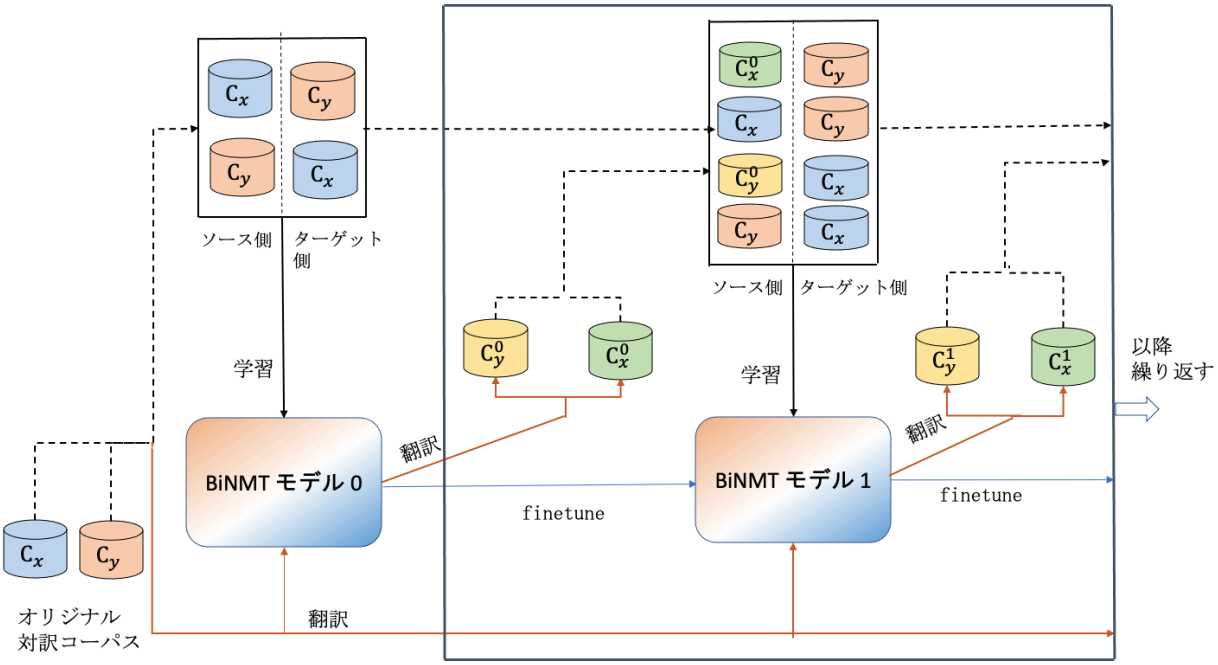

図 2　反復的な逆翻訳を用いた BiNMT モデル学習

2.2　双方向の翻訳モデル i を用いて Y→X 方向で単言語コーパス C_y を翻訳し、C_x^i を得る。

2.3　文対応の取れたコーパス対 C_X と C_Y で学習データを初期化する。C_x^i と C_y の組みを 1 つの疑似対訳コーパス (X→Y) として学習データに追加する。同様に C_y^i と C_x の組みを疑似対訳コーパス (Y→X) として学習データに加え、新しい学習データを構築する。

2.4　モデル i を新しい学習データでファインチューニングし、得られた双方向翻訳モデルをモデル i+1 とする。

　ファインチューニングし、新しい学習データで双方向翻訳モデルを再学習し、

3　$i = i + 1$ としてステップ 2 に戻る。

提案法のモデル学習の流れを図 2 に示す。

4　実験

4.1　データ

本稿では International Workshop on Spoken Language Translation (IWSLT-2015) の英語・ベトナム語翻訳タスクの小規模対訳コーパスを用い、英語 → ベトナム語及びベトナム語 → 英語の 2 つの方向の翻訳タスクで実験を行った。IWSLT 2015 コーパスの tst2012 を開発データとし、tst2013 をテストデータとして使用した。越英対訳コーパスの文数を表 4.1 に示す。英語文に対しては Moses キットの tokenizer で文をトークン化した。ベトナム語文に対してはは pyvi ライブラリの ViTokenizer で文をトークン化した。英語の文とベトナム語の文は Moses の truecaser を用いて処理を行った。また、原言語の文と目的言語の文を一括して部分単語単位を求める joint Byte Pair Encoding(BPE) [10] で 5000 のサブワードに分割した。

表 1　IWSLT2015 越英コーパスの対訳文体数

データ	ファイル	文数
訓練	train	133,317
開発	tst2012	1,553
テスト	tst2013	1,268

4.2　実験設定

ニューラル機械翻訳システムの実装には fairseq を使用した。各方向の翻訳モデルと双方向ニューラル機械翻訳モデルは fairseq の Transformer モデル [2] を用いた。実験の全てのモデルは同じハイパーパラメータの設定を使用した。学習率が 1×10^{-8}、ウォームアップが 4000 ステップ、学習率減衰が逆平方根、ラベル平滑化が 0.1、ドロップアウトが 0.1、重み減衰が 0.0001、損失関数がラベル平滑化クロスエントロピーである。モデル学習する際、Adam の最適化

アルゴリズム（$\beta_1 = 0.9, \beta_2 = 0.98$) を使用した。双方向翻訳モデルの訓練データのソース側には各文の先頭に翻訳方向を示すタグ $< e2v >$(英語 → ベトナム語), $< v2e >$(ベトナム語 → 英語) を追加した。逆翻訳を行う際、ランダムサンプリングで単言語データを翻訳した。反復的逆翻訳は、反復を 4 回まで繰り返した。ランダムサンプリングは予備実験により temperature=0.8 に設定した。

4.3 評価方法

機械翻訳モデルの自動評価手法として BLEU を用いた。越英対訳コーパスで各単方向ニューラル機械翻訳モデルを学習し、ベースラインとする。提案する双方向ニューラル機械翻訳モデルを学習し、ベースラインと比較し、双方向翻訳モデルが翻訳性能を改善できるかを調査する。さらに、反復的な逆翻訳手法を適用し、翻訳性能を改善できるか調査する。また、Koehn らが提案したブートストラップ法 [11] により 1000 サンプルで有意差検定を行う。

4.4 実験結果

ベースラインの単方向翻訳モデルと双方向翻訳モデルの翻訳性能を表 2 に示す。ベースラインと比較すると、双方向翻訳モデルが英越方向翻訳で+1.15、越英方向翻訳で+1.6 を向上した。また、ベースラインと双方向翻訳モデルとの間で BLEU スコアの有意差検定 ($p < 0.01$) を行い、両方向の翻訳性能には有意差が存在することが確認した。

表 2 単方向翻訳モデルと双方向翻訳モデルの翻訳性能

モデル	英越方向	越英方向
ベースライン	26.58	20.08
BiNMT	27.73 (+1.15)	21.68 (+1.6)

反復的逆翻訳手法によって、各反復回数での翻訳性能を表 3 と図 3 に示す。反復的逆翻訳手法により、双方向翻訳モデルはそれぞれ英越翻訳方向で＋2.61、越英翻訳方向で+3.05 を改善した。英越翻訳方向について、BiNMT モデル 1 と BiNMT モデル 3 の差に対して有意差検定を行い、有意水準 $p < 0.01$ で差があることが確認できた。越英翻訳方向については、BiNMT モデル 1 と BiNMT モデル 2 の差に対して検定を行い、有意水準 $p < 0.05$ で差があることが確認された。

表 3 IBT を用いた双方向翻訳モデル

モデル	英越方向	越英方向
BiNMT モデル 0	27.73	21.68
BiNMT モデル 1	29.61	24.36
BiNMT モデル 2	29.97	**24.91**
BiNMT モデル 3	**30.34**	24.56
BiNMT モデル 4	29.92	24.87

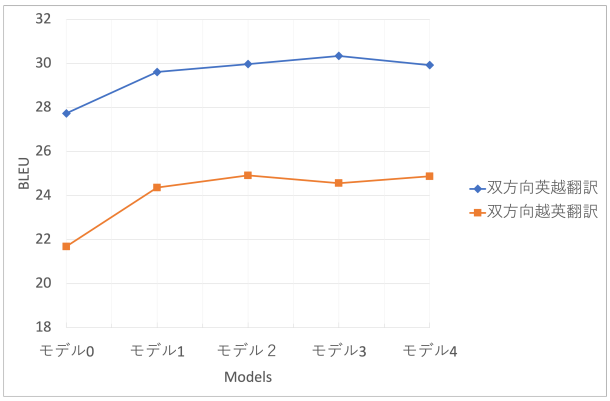

図 3 繰り返し学習の各モデルの BLEU スコア

5 おわりに

双方向翻訳モデルを学習し、元の対訳コーパスを単言語コーパスとみなして逆翻訳とモデル更新を繰り返す反復的逆翻訳手法により、両方向の翻訳性能を同時に改善する手法を提案した。IWSLT2015 の英越対訳コーパスで実験を行い、有効性を確認した。各翻訳方向の単方向翻訳モデルと比較し、越英双方向翻訳モデルは越英翻訳方向と英越翻訳方向を同時に改善でき、反復的逆翻訳で更なる双方向翻訳モデルの性能向上を達成した。本手法ではモデルの変更が必要なく、また追加の学習データを使うことなく元の学習データのみで両方向の翻訳性能を同時に向上させることができる。

本稿の実験では英越言語対の小規模対訳コーパスに行なったが、今後の研究として、他の言語対と大規模対訳コーパスに本手法の有効性を検証したい。コーパスサイズはどのように影響するかを調査したい。また、本手法と他の手法を組み合わせ、翻訳精度をさらに向上させる手法を研究したい。

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] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. CoRR, abs/1508.04025, 2015.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. CoRR, abs/1511.06709, 2015.

[4] Vu Cong Duy Hoang, Philipp 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, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[5] 森田 知熙, 秋葉 友良, and 塚田 元. 双方向の逆翻訳を利用したニューラル機械翻訳の教師なし適応の検討. 言語処理学会, 2018.

[6] Liang Ding, Di Wu, and Dacheng Tao. Improving neural machine translation by bidirectional training. CoRR, abs/2109.07780, 2021.

[7] Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. Investigating backtranslation in neural machine translation. CoRR, abs/1804.06189, 2018.

[8] Jiajun Zhang and Chengqing Zong. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, Austin, Texas, November 2016. Association for Computational Linguistics.

[9] Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 55–63, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. CoRR, abs/1508.07909, 2015.

[11] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.