双方向翻訳モデルの相互学習におけるデータ多様化の適用

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では大量の学習データが必要となるが, 十分な量のデータを用意できないドメインにおいては高性能なモデルを作ることは難しい. この問題に対し, 単言語コーパスで学習データを拡張する Back Translation (BT) や, BT を 2 方向のモデルで相互に繰り返す Itarative Back Translation (IBT) が提案されている. 本研究では, 複数のモデルによる出力文を用いた学習とアンサンブル翻訳により, 通常の IBT の効果を更に向上させる手法を提案する. 英日・英独コーパスの実験を通して, 通常の IBT と比較して提案手法がより高い BLEU スコアを達成することを確認した.

1　はじめに

近年ニューラル機械翻訳 (NMT) が広く浸透している. NMT モデルの学習には大量の対訳コーパスが必要であるが, ドメインによっては十分な量の学習データを集めることが難しく, その場合品質が高い翻訳モデルを作ることは困難である. これに対し, 単言語コーパスを追加の学習リソースとして利用する Back Translation(BT) [1] 手法や, BT を順方向・逆方向のモデルで相互に反復することで, 両方向のモデルを複数回に渡って改善する Iterative Back Translation(IBT) [2, 3, 4] 手法が提案されている. BT や IBT はデータ拡張により翻訳モデルの精度を向上できるほか [1] , 単言語コーパスのドメインにモデルを適用させるという面でも高い効果を発揮することが報告されている [5, 6, 7]. 一方, 追加の学習リソースを用いずに複数のモデルで多様な疑似対訳文を生成するデータ拡張手法も存在する [11].

本研究では, IBT に対し複数のモデルによる出力文を用いた学習を行うことで従来の IBT 手法の効果を更に向上させる手法を提案する. 具体的には, 順方向・逆方向のモデルを初期パラメータの値を変えた上で複数用意し, それぞれのモデルで翻訳文を出力することで, 複数の疑似対訳文を生成する. 複数個のモデルによる異なる出力を混合して再学習のための学習データとして使うことで, 通常の IBT より高い精度でモデルを学習することできる. また, 複数個のモデルを利用してアンサンブル翻訳を行うことで, 更に翻訳精度を高めることもできる.

英日コーパスと英独コーパスでの実験を通して, 提案手法が通常の IBT と比較してより高い BLEU スコアを達成することを確認した.

1.1　関連研究

Sennrich ら [1] は NMT において Back Translation (BT) によりターゲット言語側の単言語コーパスをソース側言語に翻訳することで疑似的な対訳文を作り, データを拡張した上でモデルを再学習する手法を提案した. Hoang ら [2] や Zhang ら [3], 森田ら [4] は, Back Translation を順方向と逆方向のモデルで相互に繰り返し適用することで両モデルの翻訳精度を向上させる IBT 手法により BT を上回る BLEU を達成したことを報告した. 藤澤ら [5], Jin ら [6], 森田ら [7] は IBT 手法がドメイン適応の面でも効果的であることを示した. Nguyen ら [11] が提案した Data Diversification (DD) 手法では, 対訳コーパスから順方向・逆方向のモデルをパラメータの初期値を変更して k 個ずつ学習し, 学習に使った対訳コーパスの文を翻訳することで計 2k 個の疑似データを得て, それらを学習データに追加することでモデルの翻訳精度を向上させている.

2　提案手法

提案手法の具体的な流れについて説明する. また, 図 1 に図解したものを示す. ソース側言語を X, ターゲット言語を Y とする. 対訳コーパスである P_X と P_Y, 対訳関係にない単言語コーパス M_X と M_Y をそれぞれ用意し, 以下の操作を行う.

1. (P_X, P_Y) から, それぞれ異なるパラメータで初

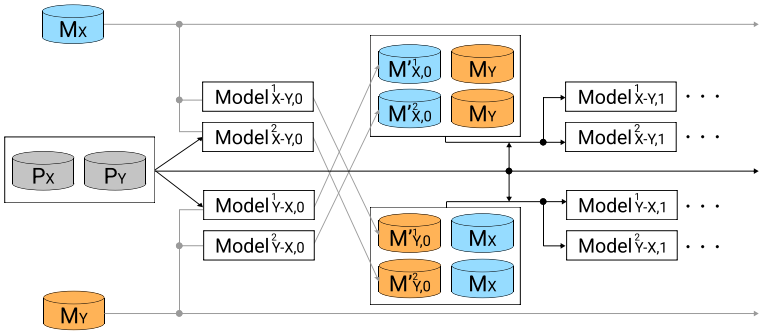

図 1: 提案手法の流れ (k=2 の場合)

期化した上で, X-Y 方向のモデルを k 個ずつ ($Model^1_{X-Y,0}$,⋯, $Model^k_{X-Y,0}$) と Y-X 方向のモデルを k 個ずつ ($Model^1_{Y-X,0}$,⋯, $Model^k_{Y-X,0}$) 学習する.

2. $i \leftarrow 0$ とする.
3. M_Y を ($Model^1_{Y-X,i}$, ⋯, $Model^k_{Y-X,i}$) で翻訳することで ($M'^1_{X,i}$, ⋯, $M'^k_{X,i}$) を生成し, 疑似対訳文 (($M'^1_{X,i}, M_Y$), ⋯, ($M'^k_{X,i}, M_Y$)) を得る.
4. ステップ 3 と同様の処理を逆方向で行い, 疑似対訳文 (($M'^1_{Y,i}, M_X$), ⋯, ($M'^k_{Y,i}, M_X$)) を得る.
5. (P_X, P_Y) と (($M'^1_{X,i}, M_Y$), ⋯, ($M'^k_{X,i}, M_Y$)) を混合した対訳コーパスを用い, ($Model^1_{X-Y,i}$, ⋯, $Model^k_{X-Y,i}$) のパラメータを引き継いで学習することで, ($Model^1_{X-Y,i+1}$, ⋯, $Model^k_{X-Y,i+1}$) を得る.
6. ステップ 5 と同様の処理を逆方向で行い, ($Model^1_{Y-X,i+1}$, ⋯, $Model^k_{Y-X,i+1}$) を得る.
7. $i \leftarrow i+1$ とし, ステップ 3 に戻る.

上記の流れで k=1 とおいた場合が通常の IBT の流れに相当する. 異なるパラメータで初期化した複数個のモデルを用いることで, 同じ翻訳元の文から異なる表現の文が出力される. よって, 1 翻訳方向に付き k 個のモデルを用いることで通常の 2k 倍の疑似対訳文を学習に用いることができる. 上記の流れを繰り返すことで, 計 2k 個のモデルが互いを改善しあうため, 通常の IBT と比較してさらなるモデルの精度向上が期待できる.

また, 複数の翻訳モデルの出力分布を平均して推論するアンサンブル翻訳を行うことで, 単独のモデルより高い精度で翻訳ができることが知られている [12]. 提案手法では同一の翻訳方向で常に k 個のモデルが存在するため, 提案手法で作成されたモデルの効果をより引き上げることが可能である.

3 実験

3.1 データセット

2 つの翻訳タスクにおいて提案手法の効果を実証するための実験を行った. いずれの実験でも dev/test データは単言語コーパス側のデータセット内のものを使用した. 1 つ目の英日実験では, 対訳コーパスとして京都に関する Wikipedia 記事から構成される KFTT[13] を, 単言語コーパスには科学技術論文の抄録から構成される ASPEC[14] を用いた. KFTT は英語側の文に日本語の文字が含まれている対訳ペアを除外した. ASPEC データセット内の対訳コーパス 100 万文を前半 50 万文と後半 50 万文に分割し, 前半の英語文と後半の日本語文をそれぞれ文の内容が異なる単言語コーパスとして使用した. 2 つ目は英独実験で, 対訳コーパスとしてニュース記事の文章から構成される News Commentary v12[15] を, 単言語コーパスには欧州議会議事録から構成される Europarl v7[15] を用いた. 英日実験と同様, Europarl データセット内の対訳コーパス約 190 万文を 2 分割し, 前半の英語文と後半のドイツ語文をそれぞれ単言語コーパスとして使用した. dev データと test データには Europarl データセット内に含まれる dev2006 と test2008 を使用した. 表 1 に各データの文数を示す.

表 1: 各言語対の文数

言語対	対訳	単言語	dev	test
英日	432,076	500,000+500,000	1,167	1,161
英独	285.722	951,050+951,050	2,001	2,001

前処理として, 日本語文には NFKC 変換, それ以外の言語の文には加えて True-casing, Moses tokenizer によるトークナイズを行い, 最終的に全ての文を Sentence Piece[16] を用いて語彙数 16,000 でサブワー

表 2: 英日実験の結果 (英日方向)

	初期値	BT	IBT	提案手法 (k=2)	提案手法 (k=3)
モデル A	6.68	18.51	32.46	33.21	33.71
モデル B	6.89	-	-	33.19	33.33
モデル C	6.78	-	-	-	33.50
+Ens.	-	-	-	33.82	**34.17**

表 3: 英日実験の結果 (日英方向)

	初期値	BT	IBT	提案手法 (k=2)	提案手法 (k=3)
モデル A	7.64	12.39	22.18	22.50	22.75
モデル B	7.73	-	-	22.79	22.44
モデル C	7.62	-	-	-	22.56
+Ens.	-	-	-	**23.16**	23.09

表 4: 英独実験の結果 (英独方向)

	初期値	BT	IBT	提案手法 (k=2)	提案手法 (k=3)
モデル A	17.34	21.62	23.11	23.36	23.49
モデル B	17.25	-	-	23.52	23.58
モデル C	17.28	-	-	-	23.38
+Ens.	-	-	-	23.56	**23.70**

表 5: 英独実験の結果 (独英方向)

	初期値	BT	IBT	提案手法 (k=2)	提案手法 (k=3)
モデル A	18.44	27.65	29.30	29.56	29.84
モデル B	18.43	-	-	29.44	29.70
モデル C	18.34	-	-	-	29.62
+Ens.	-	-	-	30.02	**30.18**

ド分割した. サブワード分割モデルの学習にはそれぞれのタスクで各言語ごとに対訳コーパスと単言語コーパスを結合した文を用いた. BLEU の計測は翻訳した文をデトークナイズした後単語単位に分割した上で行った. 日本語文に関しては MeCab[17] の分かち書きを使用した.

3.2 実験条件

翻訳モデルの構築には OpenNMT-py を使用する. アーキテクチャとして Transformer を用い, Encoder, Decoder ともに 6 層とし, 隠れ層の次元には 512 を指定した. IBT の学習をする際, 対訳コーパスと疑似対訳文から取得するサンプルの比率が常に 1:1 になるように調整した. 学習の際モデルのチェックポイントから dev データの accuracy を毎回計算し, 最もスコアが高いモデルを採用した. 最終的なモデルの性能評価には BLEU[18] を用いた.

BT では, 翻訳タスクごとに, 単言語コーパスの翻訳時のデコード方法に beam search を用いる場合と random sampling(各時刻の出力分布を下にランダムに語彙を選択する) を用いる場合とで, 最終的な精度向上の効果に差があることが知られている [8, 9, 10]. 本実験では事前に BT の効果がより大きい方法を調べた上で, 英日実験には beam search (beam size=5), 英独実験には random sampling を用いた.

3.3 IBT 実験の構成

本稿では (1) baseline (通常の IBT), (2) 提案手法 (k=2), (3) 提案手法 (k=3) の 3 種類の IBT 実験を行い, 性能を比較する. ここで k は翻訳方向ごとのモデルの数を表す. 通常の IBT は k=1 の提案手法に相当する. 各翻訳タスクにおいて, 翻訳方向ごとに初期パラメータの値を変えた上で 3 種類ずつのモデルを作成し, 以降それぞれモデル A, モデル B, モデル C と呼称する. (1) ではモデル A を, (2) ではモデル A〜B を, (3) ではモデル A〜C をそれぞれ初期モデルとして IBT 実験を行った.

3.4 結果

表 2〜3 に英日実験の結果を, 表 4〜5 に英独実験の結果を示す. 英日実験は IBT の更新を 12 回まで, 英独実験は 6 回まで行い, 最も dev データでの BLEU が高いモデルでの結果を示している. 結果を見ると, 全ての実験において通常の IBT と比べ提案手法の方がより高い BLEU を記録している. 更に, 提案手法における複数のモデルでアンサンブル翻訳を行ったところ, 単一のモデルよりも更に精度が向上することが確かめられた.

図 2 と図 3 に英日実験・日英実験におけるモデル A の BLEU 推移を示す. また, 各更新段階におけるモデルのアンサンブル翻訳の BLEU も示している. 図中の”k=2”は提案手法 (k=2) のモデル, ”k=2(+Ens.)”は提案手法 (k=2) の各モデルのアンサンブル翻訳の BLEU である. 通常 IBT と比較して, 提案手法は一貫して高い BLEU で推移しており, アンサンブル翻訳の結果は一番高く推移している. また, 提案手法によりモデルの数を増やすほど最終的に BLEU がより高い値に収束する傾向にあることがわかる.

3.5 比較実験

3.5.1 単一モデルでの比較実験

提案手法で作成したモデルが通常の IBT よりも高い BLEU を記録することが確かめられたが, モデル

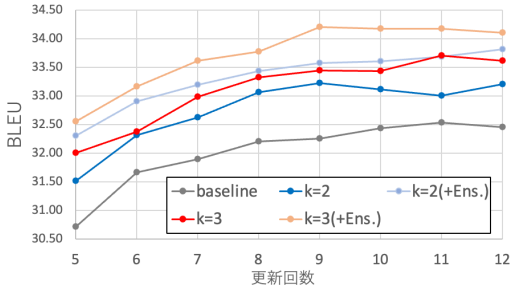

図 2: 英日実験のモデルの BLEU 推移

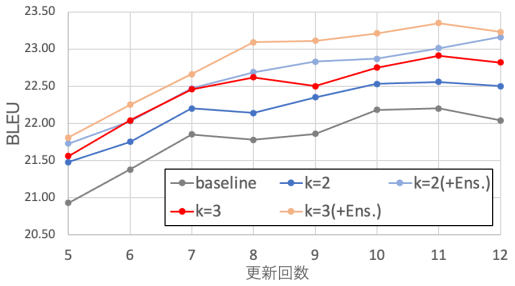

図 3: 日英実験のモデルの BLEU 推移

の数を増やしたことに意味があるわけではなく, 単純に疑似対訳文の量が増加したことで精度が向上したにすぎないということも考えられる. よって, 通常の IBT においてモデル 1 つにつき 3 対ずつの疑似対訳文を生成させて学習データのサイズを揃え, 提案手法 (k=3) とのモデルの精度を比較する実験を行った. 実験は英日・英独コーパスで行った. 3.2 節における baseline での設定に翻訳を 2 種類追加し, 疑似対訳文の量を増やした. 追加分の翻訳では単一のモデルから表現の異なる文を出力するため, random sampling でデコードを行う. 更新は 6 回まで行い, 提案手法 (k=3) での 6 回目までの最大値と比較した[1)].

表 6〜7 に結果を示す. いずれも提案手法 (k=3) が比較手法より高い BLEU を記録している. これにより, 提案手法が複数のモデルの出力を組み合わせて使うことで精度を向上させていることが示された.

3.5.2 通常 IBT でのアンサンブル翻訳実験

提案手法では, ある翻訳方向のモデル k 個に対して同一の単言語コーパスを入力することで k 種類の疑似対訳文を生成し, それらを混合した対訳文を逆方向のモデル更新時に学習データとして使用するという流れを繰り返す. しかし, 先述の通り複数個のモデルがあればアンサンブル翻訳によりそれぞれのモデルの知識を平均化した翻訳を行うことができるため, モデルの学習段階で複数モデルの出力を組み合わせて使わなくとも事足りる可能性もある. そこで本節では, 複数モデルの出力をまとめて逆方向のモデル更新に使用することが実際にモデルの改善に寄与しているのかを調べる実験を行った. 方法としてはモデル A〜C を初期モデルとして通常の IBT 実験を独立して 3 種類実施し, 各更新段階において 3 種類のモデルをアンサンブル翻訳した際の BLEU スコアを算出し, 提案手法 (k=3) によるモデル 3 種類のアンサンブル翻訳の BLEU スコアと比較する. IBT の更新は 6 回まで行った.

表 8 に比較した結果を示す. 通常の IBT3 種のアンサンブル翻訳を baseline(+Ens.), 提案手法 (k=3) のアンサンブル翻訳を提案手法 (+Ens.) として示している. 両者を比較すると, 提案手法の方が BLEU が高いことがわかる. よって, 提案手法において学習の段階で複数モデルの出力をまとめて逆方向のモデル更新に使用することによる有効性が示された.

表 6: 単一モデルでの比較実験の結果 (英日)

翻訳方向	baseline (疑似対訳文 ×3)	提案手法 (k=3)
英日	31.64	**32.01**
日英	20.81	**21.56**

表 7: 単一モデルでの比較実験の結果 (英独)

翻訳方向	baseline (疑似対訳文 ×3)	提案手法 (k=3)
英独	23.35	**23.49**
独英	29.30	**29.84**

表 8: 通常の IBT+Ens. との比較実験の結果

翻訳方向	baseline×3 +Ens.	提案手法 (k=3) +Ens.
英日	32.53	**33.17**
日英	21.90	**22.25**

4 おわりに

本研究では, Itarative Back Translation 手法において複数のモデルによる出力文を用いた学習とアンサンブル翻訳により精度を向上する手法を提案した. 具体的には, 初期パラメータが異なる順方向・逆方向のモデルを複数用意し, それぞれのモデルで翻訳文を出力することで, 複数のモデルが互いを改善し合う. 英日実験・英独実験の結果から, DD により通常の IBT を大きく改善できること, さらに, 複数のモデルによるアンサンブル翻訳を組み合わせることで, 更に翻訳精度を向上させられることを確かめた.

1) 更新回数が違うため表 2〜5 の提案手法の BLEU と値が異なることに注意.

謝辞

本研究は JSPS 科研費 JP23K11118 の助成を受けたものです.

参考文献

[1] 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, Berlin, Germany. August 2016. Association for Computational Linguistics.

[2] 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, Melbourne, Australia. July 2018. Association for Computational Linguistics.

[3] Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. Joint training for neural machine translation models with monolingual data, In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Apr. 2018.

[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] 藤澤兼太, 秋葉友良, 塚田元. ニューラル機械翻訳における双方向反復的教師なし適応の改善. 言語処理学会 第 26 回年次大会 発表論文集, pp. 744–747, 2020.

[6] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. A simple baseline to semi-supervised domain adaptation for machine translation, CoRR, abs/2001.08140. 2020.

[7] 森田知熙, 秋葉友良, 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 第 25 回年次大会 発表論文集, pp. 1451–1454, 2019.

[8] 今村賢治, 藤田篤, 隅田英一郎. サンプリング生成に基づく複数逆翻訳を用いたニューラル機械翻訳. 人工知能学会論文誌, Vol. 35, No. 3, pp. A–JA9_1–9, 2020.

[9] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, 2018. Brussels, Belgium. October-November 2018. Association for Computational Linguistics.

[10] Jiahao Xu, Yubin Ruan, Wei Bi, Guoping Huang, Shuming Shi, Lihui Chen, and Lemao Liu. On Synthetic Data for Back Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp 419–430, Seattle, United States. 2022. Association for Computational Linguistics.

[11] Xuan-Phi Nguyen, Shafiq Joty, Wu Kui , Ai Ti Aw. Data Diversification: A Simple Strategy For Neural Machine Translation. Advances in Neural Information Processing Systems, 33:10018–10029. 2020.

[12] 今村賢治, 隅田英一郎. 双方向リランキングとアンサンブルを併用したニューラル機械翻訳における複数モデルの利用法. 情報処理学会研究報告, 2017-NL-233, No.9, 2017.

[13] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[14] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2204–2208, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[15] Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp.2214–2218, Istanbul, Turkey. 2012. European Language Resources Association (ELRA).

[16] 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.

[17] 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.

[18] 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.