NMT の双方向反復的教師なし適応手法における初期対訳コーパスサイズの影響と翻訳モデル獲得に関する調査

藤澤 兼太
豊橋技術科学大学
fujisawa.kenta.tv@tut.jp

秋葉 友良
豊橋技術科学大学
akiba.tomoyoshi.tk@tut.jp

塚田 元
豊橋技術科学大学
tsukada.hajime.hl@tut.jp

1 序論

ニューラル機械翻訳 (Neural Machine Translation) モデルは，十分な量の対訳コーパスがあれば高い性能を達成できるため，近年の機械翻訳に関する研究において広く用いられている．しかし，十分な量の対訳コーパスを用意することが困難な分野 (ドメイン) では，良質な翻訳モデルを学習することは難しく，機械翻訳研究における重要な課題となっている．

このような課題を解決するためのアプローチとして，比較的構築が容易な単言語コーパスを用いた手法が多く提案されている．単言語コーパスを用いた代表的な手法として，Sennrich ら [1] が提案した Back-Translation(逆翻訳) アプローチが広く知られている．この手法は，大規模な単言語コーパスを他ドメインの対訳コーパスで予め学習しておいた翻訳モデルにより翻訳し，その翻訳結果をソース側，単言語コーパスをターゲット側とする当該ドメインの対訳を学習データに追加して，高性能な逆方向の翻訳モデルを新たに学習するという手法である．様々な拡張手法が提案されており，Hoang ら [2] や Zhang ら [3]，森田ら [4] は，逆翻訳および翻訳モデルの学習を双方向かつ反復的に行うことで，通常の逆翻訳手法よりも高い性能を示せることを報告している．以後，本論文では森田らの手法を森田らに倣って「双方向反復的教師なし適応手法」と呼ぶ．

双方向反復的教師なし適応手法は高い翻訳性能を達成できることが示されているが，その性質は十分に明らかになっていない．そこで本研究では，以下の点に関して実験及び分析を行い，この手法の性質をより明らかにすることを目的とする．

1. 初期対訳コーパスが小規模な場合の効果
2. 翻訳モデルの獲得に焦点を当てた評価

英日・日英翻訳を対象に実験を行い，BLEU[5] により評価した結果，初期対訳コーパスのサイズが 2 万文の場合 (初期モデルでは英日は 1.41，日英は 2.92 の BLEU スコア) でも，ドメイン内単言語コーパス 150 万文を用いて手法を適用することで英日は 25.46(+24.05)，日英は 24.08(+21.15) まで BLEU スコアが向上した．また，初期対訳コーパスのサイズが翻訳性能に与える影響を，単なる学習データサイズの効果とは分離して，評価・分析を行った．単語再現率を用いた評価から，単言語コーパスより新たに翻訳モデルを獲得できていることを確認した．

2 双方向反復的教師なし適応手法

図 1 に，2 つの言語 X，Y 間の翻訳に関する双方向反復的教師なし適応手法の流れを示す．図 1 において，C_X^{in}，C_Y^{in} はそれぞれ言語 X，Y のドメイン内単言語コーパスを示し，(C_X^{out},C_Y^{out}) はドメイン外対訳コーパスを示す．逆翻訳により生成されたコーパスを C' と表記すると，単言語コーパス C_X とその逆翻訳結果 C'_Y を組み合わせた疑似対訳コーパスは (C'_Y,C_X) のように表すことができる．

これらの表記方法に従い，双方向反復的教師なし適応手法の流れを以下に記述する．

1. ドメイン外対訳コーパス (C_X^{out},C_Y^{out}) を用いて，初期モデルとなる $Model_{X\to Y}$ 0，$Model_{Y\to X}$ 0 を学習する．
2. 性能が収束するまで以下を反復する．(i←0)
   - 2.1. 単言語コーパス C_Y^{in} を $Model_{Y\to X}$ i により翻訳し，疑似対訳コーパス ($C'_X{}^{in}_i$,C_Y^{in}) を得る．疑似対訳コーパスとドメイン外対訳コーパスを結合し，それを用いて逆方向のモデル $Model_{X\to Y}$ (i+1) を学習する．
   - 2.2. 単言語コーパス C_X^{in} を $Model_{X\to Y}$ i により翻訳し，疑似対訳コーパス ($C'_Y{}^{in}_i$,C_X^{in}) を得る．疑似対訳コーパスとドメイン外対訳コーパ

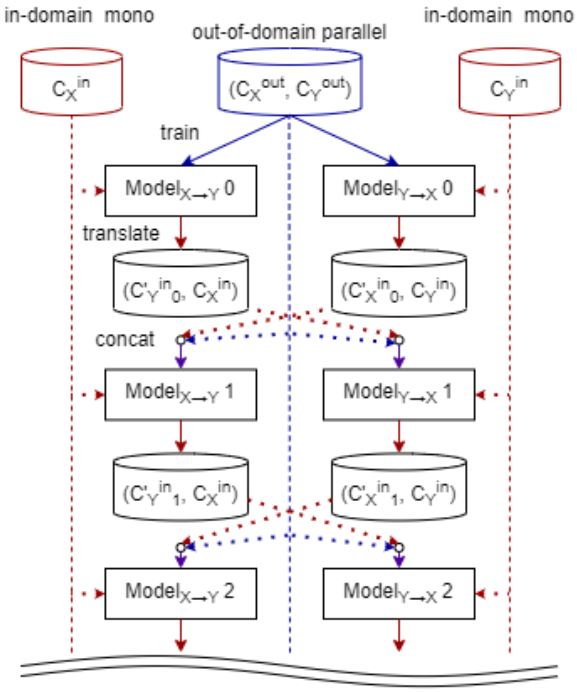

図 1　双方向反復的教師なし適応手法の流れ

スを結合し，それを用いて逆方向のモデル $\mathrm{Model}_{Y\to X}$ (i+1) を学習する．

2.3.i←i+1

3　実験

3.1　実験条件

ドメイン外の対訳コーパスとして，Asian Scientific Paper Excerpt Corpus(ASPEC)[6] の英日対訳コーパスから 100 万文を用いる．対訳コーパスのサイズを縮小する際には，データ数が多いコーパスが少ないコーパスを常に包含するという条件を保つようにデータセットを作成する．ドメイン内の単言語コーパスとしては，NTCIR8-PATMT コーパスから英語，日本語それぞれ 150 万文ずつ対訳にならないように抽出して用いる．バリデーションセットには NTCIR8-PATMT から 2000 対を，テストセットには同じく NTCIR8-PATMT から 899 対を用いる．

すべてのデータに対して NFKC 変換と小文字化を適用する．英語データに対しては Moses tokenizer[7] を用いて事前分割したあと，Sentence Piece[8] によるサブワード化を行う．日本語データに対しては事前分割せずに Sentence Piece によるサブワード化を行う．Sentence Piece は BPE ベースのものを使用し，分割する語彙の上限は 8000 とする．Sentence Piece のモデルの学習には，ドメイン外の対訳コーパスとドメイン内の単言語コーパス双方の全データを用いる．

モデルは，エンコーダを 1 層の双方向 LSTM，デコーダを 1 層の単方向 LSTM としたものを使用し，隠れ状態の次元数は 500 とする．最適化アルゴリズムには Adam を使用し，学習率を 0.001 とする．各モデルは 10 エポックずつ学習し，1 エポックごとのチェックポイントの中から最も accuracy が高かったモデルを採用する．モデルの翻訳精度の評価には単語単位の BLEU[5] を用いる．

表 1　NTCIR-8 PATMT 対訳コーパスのベースライン性能 (BLEU)

corpus size	En2Ja	Ja2En
1M	36.28	35.91
500K	34.97	33.07
100K	28.20	26.14

3.2　ベースライン

NTCIR-8 PATMT(ドメイン内) のオリジナルの対訳コーパスを用いて同様のモデル設計で学習を行い，それをベースラインとする．結果を表 1 に示す．

3.3　実験結果

表 2 に，双方向反復的教師なし適応手法を用いて学習した場合の Model 0，Model 1，Model 10 の BLEU スコアを示す．初期対訳コーパスサイズが最も大規模な 100 万文 (1M) の場合，Model 10 では英日で 34.60(Model 0 と比べて+21.08)，日英で 33.41(+15.15) という BLEU スコアを得た．対訳コーパスサイズを縮小していくと，BLEU スコアは低下していくが，手法による効果はほとんどのサイズで確認できた．対訳コーパスサイズが 2 万文の場合，Model 0 では英日が 1.41，日英が 2.93 とかなり低い値であるにも関わらず，Model 10 では英日が 25.46(+24.05)，日英が 24.08(+21.15) という BLEU スコアを得た．1 万文 (10K) の場合は反復学習による改善が小さく，Model 10 まで学習を行っても 20K 以上の結果と比べてかなり低い BLEU スコアとなった．

表 1 と比較すると，1M の結果はドメイン内の対訳コーパスを 50 万文 (500K) 用いた結果と同等の数値である (34.60 vs. 34.97, 33.41 vs. 33.07)．3 万文 (30K) の場合は，英日翻訳の性能はドメイン内の対訳コーパスを 10 万文 (100K) 使用した結果に比べて

表 2　双方向反復的教師なし適応の効果

	Model	En to Ja	Ja to En
10K	Model 0	0.00	0.45
	Model 1	0.82(+0.82)	1.00(+0.55)
	Model 10	4.50(+4.50)	2.42(+1.97)
20K	Model 0	1.41	2.93
	Model 1	7.39(+5.98)	9.03(+6.10)
	Model 10	25.46(+24.05)	24.08(+21.15)
30K	Model 0	2.48	4.59
	Model 1	11.56(+9.08)	11.25(+6.66)
	Model 10	27.33(+24.85)	26.70(+22.11)
50K	Model 0	4.59	5.47
	Model 1	13.39(+8.80)	15.55(+10.08)
	Model 10	30.14(+25.50)	28.27(+22.80)
100K	Model 0	7.44	9.28
	Model 1	21.00(+13.56)	19.30(+10.02)
	Model 10	30.91(+23.47)	29.35(+20.07)
500K	Model 0	11.64	14.77
	Model 1	27.82(+16.18)	27.52(+12.75)
	Model 10	33.54(+21.90)	33.25(+18.48)
1M	Model 0	13.52	18.26
	Model 1	29.14(+15.62)	29.90(+11.64)
	Model 10	34.60(+21.08)	33.41(+15.15)

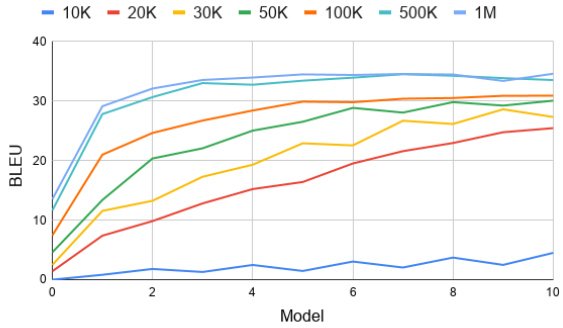

図 2　反復による BLEU スコアの推移 (英日翻訳)

0.87 劣っている (27.33 vs. 28.20) が，日英翻訳の性能は 0.56 優れている (26.70 vs. 26.14).

図 2 は，反復学習によりどのように BLEU スコアが上昇していくかを示したグラフである．対訳コーパスサイズが 20K 以上の場合は反復により BLEU が上昇していくことが確認できる．

以上の結果から，双方向反復的教師なし適応手法は，対訳コーパスのサイズが小規模な場合においても Back-Traslation アプローチ [1](Model 1 に相当) を大きく上回る性能を達成できることがわかった．

表 2 に示した翻訳性能は，サイズの異なる初期対訳コーパスを加えて学習したモデルの比較を行っている．例えば，100K の Model 10 は 100K の初期対訳コーパスと 1.5M の疑似対訳コーパスの合計 1.6M の対訳コーパスを用いて学習しているのに対し，1M の Model 10 は 1M の初期対訳コーパスと 1.5M の疑似対訳コーパスの合計 2.5M の対訳コーパスで学習している．従って，学習データサイズの違いもスコアに反映されているため，純粋に初期対訳コーパスサイズの効果が比較できない．そこで，Model 10 を学習する際，初期対訳コーパスを除いて学習し，学習データサイズを揃えて比較する評価も行った．

表 3　疑似対訳コーパスのみによる翻訳性能

	En to Ja	Ja to En
30K	27.57	26.71
50K	29.98	27.48
100K	30.41	29.95
500K	32.70	32.27
1M	32.77	33.49

具体的には，通常の双方向反復的教師なし適応手法により Model 9 まで学習を行った後，Model 10 を学習する際に疑似対訳コーパスのみで学習を行い，その翻訳性能を評価した．実験結果を表 3 に示す．初期対訳コーパスサイズが増えるに従い単調に翻訳性能が改善しており，初期対訳コーパスを増加する効果が確認できる．しかし，初期対訳コーパスサイズを 2 倍にしても BLEU の改善は+1 程度である．初期翻訳モデルのブートストラップに必要な初期コーパスサイズ (本研究では 20K 以上) が確保できれば，初期コーパスのサイズはこの手法にとってそれほど重要な要因ではないことがわかる．

3.4　翻訳モデル獲得の分析

前節の BLEU による翻訳性能の総合的な評価に加え，翻訳モデルが初期モデルに含まれない語彙を適切に獲得できているかという視点から評価を行った．双方向反復的教師なし適応手法では，単言語コーパスを繰り返し用いることになるため，ソース側，ターゲット側ともども初期モデルにない単語を獲得できる．しかしながら獲得された単語間の翻訳モデルも獲得されるのか明らかではない．この疑問を明らかにするために単語再現率を用いた分析を行った．

3.4.1　単語再現率

本論文では正解文に含まれる単語がどの程度出力文に現れているかを単語再現率で評価する．単語再現率は式 1 によって計算される．

表 4　正解文データに含まれる各属性の単語数

	English words	Japanese Words
full	29877	31565
out	28937	30951
in, not out	841	507
neither	99	107

$$Word\_Recall = \frac{\sum_{S \in Refs} \sum_{w \in g_X(S)} Count_{match}(w)}{\sum_{w \in g_X(Refs)} Count(w)} \quad (1)$$

ここで，S は正解文，Refs は正解文の集合，w は単語，$Count(w)$ は正解文に含まれる単語の個数，$Count_{match}(word)$ は正解文と対応する出力文の間で一致する単語の個数を示す．また，$g_X(W)$ は単語集合 W からある属性 X を有する単語だけを抽出する関数である．考慮する属性 X は次節で導入する．この式では，各文ごとに単語再現数を計算した後に，それらの総和を正解文に含まれる全単語数で割り，平均を計算している．

3.4.2　分析方法と結果

テストデータの各正解文の各単語について，ドメイン外対訳コーパスに含まれるもの (“out” と表記)，ドメイン内単言語コーパスにのみ含まれるもの (“in,not out” と表記)，どちらにも含まれないもの (“neither” と表記) に分類し，これらを式 1 で考慮する属性 X として単語再現率を計算する．反復により初期モデルに含まれない単言語コーパスの単語の翻訳モデルが獲得されるのであれば，“in,not out” に関する単語再現率が上昇していくと考えられる．

本節では，初期モデルの対訳コーパスサイズが 5 万文の場合について実験を行った．テストデータの正解文に含まれる各属性の単語数を表 4 に示す．ここで，“full” は正解文データ全体を示す．

単語再現率による評価結果を図 3 と図 4 に示す．“out” の単語再現率は，“full” とほとんど同じように推移している．これは，“out” が “full” に含まれる単語をほぼすべて含んでいるためである．“in, not out” の推移をみると，グラフの形に差はあるものの英日，日英それぞれ反復により上昇していることがわかる．このことから，ドメイン内の単言語コーパスを繰り返し学習に用いることで，新たに単語翻訳モデルを獲得できていると考えられる．また，どのデータにも含まれていないはずの “neither” について

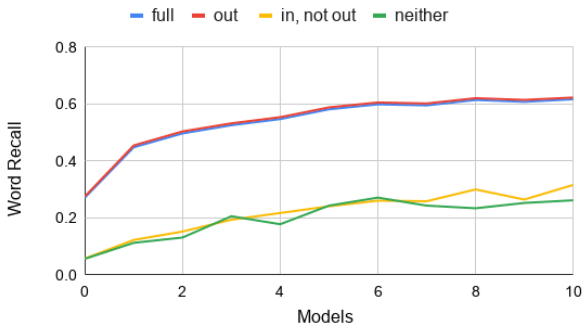

図 3　反復による単語再現率の推移 (英日翻訳)

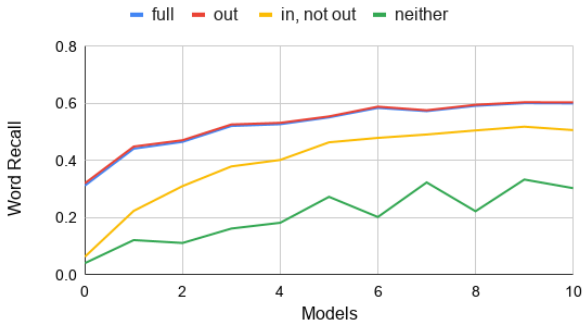

図 4　反復による単語再現率の推移 (日英翻訳)

も反復による単語再現率の上昇がみられる．以上の結果から，双方向かつ反復的な逆翻訳により新たに単語翻訳モデルを獲得でき，このことが性能向上に貢献していることがわかる．

4　結論

本論文ではまず，ドメイン外の対訳コーパスサイズが小規模な場合の双方向反復的教師なし適応手法の効果を検証した．実験の結果，対訳コーパスが 2 万文程度しかない場合でも，ドメイン内単言語コーパスが十分にあればドメイン内対訳コーパスが 10 万文存在する場合と同等の性能を達成することが確認できた．続いて，単語再現率という指標においても性能が向上していることを確認し，ドメイン内の単言語コーパスにしか含まれていない単語の翻訳モデルを単言語コーパスしか用いていないにもかかわらず獲得できていることを確認した．

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] 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, 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*, 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. *CoRR*, Vol. abs/1803.00353, , 2018.

[4] 知熙森田, 友良秋葉, 元塚田. 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. Technical Report 3, 豊橋技術科学大学, 豊橋技術科学大学, 豊橋技術科学大学, dec 2018.

[5] 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[6] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2204–2208, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[7] 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, and 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, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[8] 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, Brussels, Belgium, November 2018. Association for Computational Linguistics.