逆翻訳を用いたデータ拡張と転移学習を利用した英日講演字幕翻訳の改善

山岸　勇輝　　秋葉 友良　　塚田　元
豊橋技術科学大学
{yamagishi.yuki.oj, akiba.tomoyoshi.tk, tsukada.hajime.hl}@tut.jp

1　はじめに

近年では、従来主流であったルールベース機械翻訳 (RBMT) や統計的機械翻訳 (SMT) の性能を超えるニューラル機械翻訳 (NMT) が確立され大きな成果をあげている。SMT では言語モデルや単語アライメント、フレーズテーブルなどいくつかの個別に学習された要素を組み合わせて翻訳モデルを実現していたが、NMT では一つの巨大なネットワークを通じて翻訳プロセス全体のモデル化を行っている。NMT で広く利用されているモデルは Sequence-to-Sequence ベースのアテンション機構付きエンコーダデコーダモデルである。エンコーダとデコーダは RNN を用いて実装されることが多い。エンコーダは与えられた入力文を一単語ずつ処理していき、隠れ状態としてベクトル空間にマッピングを行う。デコーダはエンコーダとアテンション機構の出力に条件づけられて、ターゲット単語を一つずつ生成する。アテンション機構はソース言語とターゲット言語の単語間のソフトアライメントを計算している。アテンション機構を用いることで、単語数が多い文章や複雑な文章をより正確に翻訳することが可能となる。

本論文では NMT を用いた TED(Technology Entertainment Design) の講演データの翻訳 [1] を扱う。近年では世界規模で開かれる講演会の数も増えてきており、このような場では広く英語を用いて講演が行われる。そこで英語以外の母語のみを扱う人々にとって講演データを他言語へ翻訳するというタスクの重要性が高まってきている。実際に、TED の講演データの翻訳を行ったところ英日翻訳が他言語対に比べて大きく性能が低いということがわかった。原因を考察したところ英日翻訳には他言語対の翻訳に比べ文法･文構造の違いが大きく、少量の講演データでは NMT の学習に対して十分な量でないことが考えられた。そのため本論文ではデータ拡張の手法に焦点をあて 2 つの手法を行うことで講演データの英日翻訳の性能改善を目指す。1 つ目の手法は同一ドメインの単言語コーパスを用いたデータの拡張である。単言語のコーパスを翻訳し疑似対訳コーパスを作成させることでデータの拡張を行った。2 つ目の手法はドメインの異なる対訳コーパスを用いた転移学習である。ドメインの異なる対訳コーパスで翻訳モデルを構築しこれを転移学習に用いた。本研究では実験に IWSLT[1] と CSJ[2]、ASPEC[3] の 3 つのコーパスを使用した。1 つ目の手法を IWSLT と CSJ で行い BLEU スコアで 0.33 ポイントの改善が見られた。2 つ目の手法では IWSLT と ASPEC を用いベースラインから 1.37 ポイントの改善が見られた。

2　関連研究

データ拡張の手法として、Sennrich ら [5] は目的言語の単言語コーパスを原言語へ逆翻訳して疑似対訳文を生成し、ベースとなる対訳コーパスと混合して訓練する方法を提案した。この手法の利点は疑似対訳文においてターゲット側には正しい文が使用されるため、デコーダーは正しく訓練されることである。そのため、単言語コーパスから言語モデルを構築する方法に比べて、安定した精度向上が可能である。

Fadaee ら [6] は、対訳文中の低頻度語を別の単語に置換して得られた文をベースとなる学習文に加えることで翻訳性能が向上することを示した。また Currey ら [7] は、目的言語の単言語コーパスの分をそのままコピーして原言語側のデータとして用いるだけでも翻訳精度が向上することを示している。

転移学習を利用した手法として、Firat ら [8] は大量のフランス語-英語対訳コーパスで学習した NMT モデルを親モデル、少量のスペイン語-英語対訳コーパスで学習した NMT モデルを子モデルとして親モデルのいくつかのパラメータを子モデルに転送することによって、少ない量の対訳コーパスを持つ言語間の翻訳精度を大幅に改善した。

表 1: ASPEC の詳細

データセット	文数	トークン数
訓練	1,000,000	25,907,147
開発	1,790	44,896
テスト	1,812	45,035

表 2: IWSLT2017 英日対訳コーパスの詳細

データセット	文数	トークン数
訓練	22,3108	455,3949
開発	871	1,9930
テスト	1,549	3,1591

3 拡張手法

3.1 データセット

本研究ではデータセットに ASPEC 英日対訳コーパスと IWSL2017 英独・英日対訳コーパス、CST 日本語単言語コーパスを使用した。ASPEC コーパスは科学技術論文抄録のコーパスで、文単位で様々な抄録から対訳文を抽出した対訳コーパスである。今回用いた ASPEC 英日対訳コーパスは文量が 100 万文と大量のパラレルコーパスのため 3.3 節の転移学習を用いた手法に利用した。
IWSLT コーパスは TED の講演書き起こしの話し言葉コーパスで講演単位で対訳を抽出した対訳コーパスである。IWSLT コーパスを翻訳モデルのベースとして利用した。
CSJ は IWSLT と同じく講演データに関して纏められた日本語話し言葉コーパスであり、ほぼ同一ドメインであると考えられる。この単言語コーパスを 3.2 節の手法で疑似対訳コーパスを作成する際に利用した。
これらのコーパスに関する詳細を表 1,2,3 に示す。

3.2 同一ドメイン単言語コーパスを用いたデータ拡張

本研究で行った 1 つ目の手法である同一ドメイン単言語コーパスを用いたデータ拡張について図 1 を用いて説明する。

表 3: CSJ コーパスの詳細

言語	文数	トークン数
日本語	44,4620	720,3732

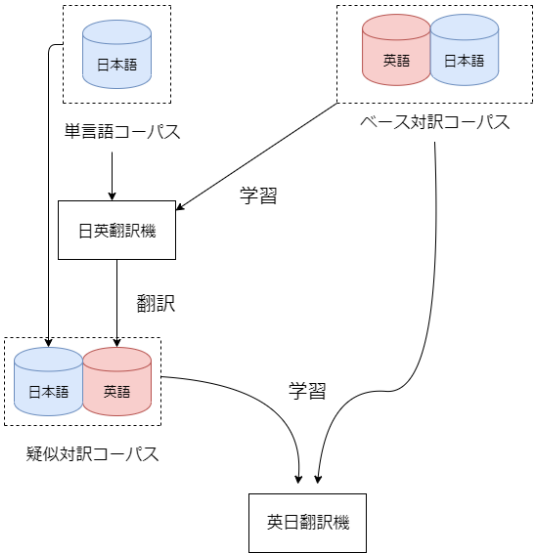

図 1: 同一ドメイン単言語コーパスを用いたデータ拡張

この手法は翻訳のベースとなる英日対訳コーパスを持つ IWSLT と日本語の単言語コーパスである CSJ の 2 つを用いて行う。
はじめに IWSLT のみを用いて日英翻訳モデルの構築を行う。翻訳モデルは 10epoch までの中で最も IWSLT の開発セットに対して BLEU スコアが高いものを選択する。この日英翻訳モデルで日本語の単言語コーパスである CSJ を翻訳し、疑似対訳コーパスを生成する。この疑似対訳コーパスをベースとなる IWSLT と混合させることで IWSLT のデータ拡張を行う。
最後に、この拡張したデータを用いて英日翻訳モデルを構築する。

3.3 ドメインの異なる対訳コーパスを用いた転移学習

本研究で行った 2 つ目の手法であるドメインの異なる対訳コーパスを用いた転移学習について図 2 を用いて説明する。

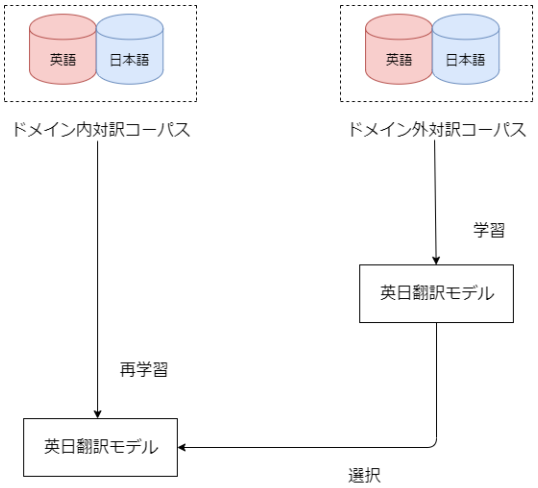

図 2: ドメインの異なる対訳コーパスを用いた転移学習

この手法では翻訳のベースとなる英日対訳コーパスを持つ IWSLT とドメインの異なる大量の英日対訳コーパスを持つ ASPEC を用いて行う。
はじめに ASPEC で学習を行い英日翻訳モデルの構築を行う。この時ボキャブラリは IWSLT のみから作成したものを利用し学習を行い。このモデルを翻訳のベースとなる IWSLT の開発セットを用いて評価を行う。BLEU スコアにて評価し、最も高い値を示すモデルを転移学習の初期モデルとして選択する。
IWSLT で学習を行う際 NMT のモデルの初期モデルとして選択したモデルを用いることで転移学習を行い英日翻訳モデルの構築を行う。

4 実験

4.1 モデルパラメータ

本研究では NMT のエンコーダは 1 層の双方向 LSTM を使用しデコーダは 1 層の単方向 LSTM を使用した。LSTM の隠れ層の次元数は 1,000 とし、ワードエンベディングの次元数は 500 とした。オプティマイザは Adam を使用し、学習率は 0.001 とした。ミニバッチサイズは 128 とし、使用する語彙の数はソース・ターゲット双方向の言語で 50,000 語を使用する。

4.2 実験結果

IWSLT コーパスのみを用いて英日翻訳モデルを作成した結果をベースラインとし、本実験で行った 2 つの手法の結果を示す。また、組み合わせとして転移学習の手法で作成した英日翻訳モデルのパラメータで CSJ を用いてデータ拡張を行ったコーパスの学習結果を示す。

表 4: 実験結果

手法	BLEU
baseline	10.15
データ拡張の手法	10.48
転移学習の手法	11.52
組み合わせ	11.43

CSJ を用いてデータを拡張した手法ではベースラインから BLEU スコア 0.3 以上の上昇が見られた。ASPEC を用いて転移学習を行った手法ではベースラインから BLEU スコア 1.3 以上の上昇が見られた。組み合わせの結果では BLEU が 11.43 と baseline から 1 以上の改善が見られた。しかし最も BLEU が高いのは転移学習のみで構築した英日翻訳モデルであり 2 つを組み合わせた結果は転移学習のみの結果を上回ることは見られなかった。

4.3 転移学習を利用した同一ドメイン単言語コーパスを用いたデータ拡張

同一ドメイン単言語コーパスを用いた拡張には疑似対訳コーパスの質が Baseline からの改善に関係すると考えられた為、CSJ を翻訳する際の日英翻訳モデルを転移学習を利用して構築した。またこの翻訳モデルを用いて作成した疑似対訳コーパスでデータの拡張を行い学習を行った英日翻訳モデルの結果を示す。

表 5: 日英翻訳モデル

転移学習なし	転移学習あり
9.56	10.95

表 6: データ拡張の手法

baseline	転移学習あり
10.48	11.05

転移学習を用いたことで日英翻訳モデルの BLEU 値が 9.56 から 10.95 へと上昇し、その結果 CSJ を翻訳して作成する疑似対訳コーパスの質が高まる為データを拡張した後の学習で構築する英日翻訳モデルにも

10.48 から 11.05 への上昇がみられた。最後にこの英日翻訳モデルを利用して組み合わせ実験を再度行った。

表 7: 組み合わせの結果

baseline	転移学習あり
11.43	11.48

日英翻訳機に転移学習を用いた場合でデータの拡張手法の結果では大きな改善が見られたが組み合わせの結果では 11.43 とほぼ同程度の 11.48 という結果だった。全てのの手法と比較すると最も BLEU が高い値を示したのは 2 つ目の手法の ASPEC と IWSLT で転移学習を行った時の結果だった。これに関して CSJ のコーパスサイズが ASPEC より小さかったことや疑似対訳コーパスを作成する際の日英翻訳モデルの質があまり高くなかったことなどが 2 つの手法での改善の差となったのではないかと考えられる。

5 おわりに

本研究では TED の講演データである IWSLT コーパスに CSJ と ASPEC の 2 つのコーパスを用いてデータ拡張を行う 2 種類の方法で英日翻訳の性能改善を目指した。実験の結果、CSJ を逆翻訳し疑似対訳コーパスを作成するデータの拡張と ASPEC を用いた転移学習の 2 つの手法は共に IWSLT コーパスの Baseline を上回り、有効性があるということが確認できた。転移学習を用いた方法では Baseline から 1.37 ポイントの改善が見られ本実験の中で最も良い改善を示した。逆翻訳を用いた拡張では IWSLT コーパスのみを用いて逆翻訳モデルの作成を行うと疑似対訳コーパスの質が低く Baseline から 0.3 ポイントほどの改善しか見られなかった。逆翻訳モデルを転移学習を利用して作成し、疑似対訳コーパスを作成した場合では Baseline から 0.9 ポイントの改善が見られた。また今回疑似対訳コーパスを作成するにあたり使用した CSJ は約 44 万文と 100 万文の ASPE と比べ半分にも満たない量であったため転移学習と比べ小さい改善に収まったと考えられる。今回の手法では Baseline からの改善は見られたがまだ他言語対の翻訳の質と比べると翻訳の性能は低くさらなる改善が望まれる。今後の課題としては転移学習におけるモデル選択の最適化や CSJ より大きな同一ドメイン単言語コーパスでの拡張などを検討したい。

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] 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.

[2] K. Maekawa and H. Koiso and S. Furui and H. Isahara. “Spontaneous Speech Corpus of Japanese”, In Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp.947–952, 2000.

[3] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. “ASPEC: Asian scientific paper excerpt corpus”, In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), pp. 2204–2208, 2016.

[4] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, “OpenNMT:Open-sorce toolkit for neural machine translation”, CoRR, vol. abs/1701.02810, 2017.

[5] R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. In Proc. of ACL-2016 (Volume 1: Long Papers), pages 86-96, 2016.

[6] M. Fadaee, A. Bisazza, and C. Monz, “Data augmentation for low-resource neural machine translation”, Proc. 55th Annual Meeting of the Assoc. for Computational Linguistics (Volume 2: Short Papers), pp.567-573, Vancouver, Canada, 2017.

[7] A. Currey, A. V. M. Barno, and K. Heafield, “Copied monolingual data improves low-resource neural machine translation”, In Proceedings of the Conference on Machine Translation, pp.148156, 2017.

[8] O. Firat, K. Cho, and Y. Bengio, “MultiWay, Multilingual Neural Machine Translation with a Shared Attention Mechanism”, ArXiv e-prints, pp.866-875, 2016.