入力側単言語資源と転移学習の利用による講演字幕を対象とした英日ニューラル機械翻訳の改善

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

概要

本論文では、”Technology, Entertainment, Design (TED)”の講演データに対する英日機械翻訳の性能向上のための取り組みとして、3 つの手法の利用を提案する。第 1 に入力側単言語資源で事前学習された BERT を機械翻訳に利用する手法、第 2 に事前学習モデルへ入力側単言語資源を利用し追加学習を行う手法、第 3 にドメイン外の大量の対訳コーパスを用いて転移学習を行う手法、を検討した。評価実験の結果、事前学習済みモデルを用いた翻訳モデルと転移学習を組み合わせることで、Trasformer ベースのベースラインモデルから 4.12Point の大幅な改善を達成した。

1　はじめに

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ことが必要となる。

研究で広く用いられている字幕付き講演データに TED[1] があるが、実際に TED の講演データの翻訳を行うと英日翻訳が他言語対に比べて大きく性能が低いという問題がある。原因としては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る。

学習データである対訳コーパスを補うために、Sennrich ら [2] は出力側単言語コーパスを逆翻訳して疑似対訳コーパスを作成する手法を提案した。対訳コーパスと比べ単言語コーパスの取得は比較的容易であるため、データ拡張の手法として広く用いられている。しかし本研究の対象である英日講演翻訳では、出力側単言語コーパスである日本語講演データは人手で翻訳され作成されているため学習データ以外の入手が困難である。

本研究では、TED 英日翻訳を対象に、利用可能な言語リソースを検討した上で、適用可能な手法を用いて性能改善を試みた。第 1 に、入力側単言語資源で学習された事前学習済みモデルを翻訳モデルに利用した。また事前学習済みモデルを適用した翻訳モデルへ、入力側単言語資源を利用した追加学習とドメイン外の対訳コーパスを利用した転移学習を適用した。

2　関連研究

2.1　BERT

BERT[3] は 2018 年に Google が発表した自然言語処理モデルである。モデルは双方向の Transformer をベースとしたエンコーダーで、あるトークン列 (単語またはサブワード) を入力として、入力の各トークンに対応する分散表現を出力する。事前学習に Masked Language Modeling(MLM) と Next Sentence Prediction(NSP) の 2 つのタスクを行っている。BERT は多くのタスクにおいて高い精度を記録していて、SQuAD や GLUE などの複数のベンチマークタスクにおいて最先端の性能を達成している。現在は、BERT を拡張した RoBERTa[4] や軽量版の ALBERT[5] など様々なモデルが提案されている。

2.2　BERT を利用した機械翻訳

高橋ら [6] は NMT の入力言語側の単語分散表現に BERT の出力を用いることで翻訳性能の改善を達成している。BERT の出力を分散表現に用いることで、従来の単語依存の分散表現ではなく、コンテキストも考慮された分散表現となり、より文脈に合わせた翻訳が行えるようになった。Zhu ら [7] は

Transformer のエンコーダ・デコーダにおいて BERT の出力への Attention を利用する BERT-fused NMT モデルを提案し、翻訳性能が改善することを示した。このモデルは IWSLT2014 の独英翻訳において最先端の性能を達成している。

2.3 転移学習

転移学習を利用した手法として、Firat ら [8] は大量のコーパスで学習した NMT モデルのパラメータを目的となる少量のコーパスで学習したモデルに転送することによって、少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。

3 提案手法

本研究では学習データ以外の日本語講演データの入手が難しい為、目的言語側の単言語コーパスを用いる Sennrich らの手法は困難と考えた。学習データが不足している場合、データ拡張の手法以外に事前学習済みモデルを用いる手法も広く知られている。本研究は、事前学習済みモデルである BERT を機械翻訳モデルに用いる。また利用可能なコーパスを検討し、英語側講演データとドメイン外の英日対訳コーパスが考えられた。この 2 つのコーパスを用いて BERT への追加学習と翻訳モデルへの転移学習の手法を行う。

3.1 事前学習済みモデルの利用

事前学習済みモデルである BERT を機械翻訳モデルへ適用した 2 つのモデルを利用する。1 つめに BERT 入力モデルとして高橋らが行った、BERT の出力を NMT の単語分散表現として利用するモデルを用いる。2 つめに Zhu らが提案する BERT-fused モデルを採用する。

3.2 事前学習済みモデルへの追加学習

本研究で用いた BERT は、英語版 Wikipedia と BooksCorpus を用いて事前学習を行ったモデルである。これらのデータは様々なドメインを含むものであるため、講演データを用いた BERT への追加学習を行うことで、より講演に適した翻訳が出来るようになると考える。具体的には英語側講演データを用いて MLM と NSP の 2 つのタスクを行う。

3.3 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 とドメイン外の大量の英日対訳コーパスを用いて行う。はじめにドメイン外対訳コーパス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う。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

提案手法によって TED の英日講演字幕翻訳への性能が改善されるかを調査する。

4.1 データセット

英日対訳コーパスの学習データとして TED の講演データである IWSLT2017 を、転移学習のために ASPEC[9] と JESC[10] の 3 つのコーパスを利用する。IWSLT2017 の開発データには dev2010 をテストデータには tst2010 を使用する。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。JESC コーパスはインターネット上からクロールされた映画と TV 番組の字幕データで作成された対訳コーパスである。

表 1 英日対訳コーパス

コーパス	訓練データ	開発データ	テストデータ
IWSLT2017	223,108	871	1549
ASPEC	1,000,000	1,790	1,812
JESC	2,797,388	2,000	2,000

BERT への追加学習のための入力側単言語コーパスとして、学習に使用している IWSLT2017 の訓練データ、学習に使用していない IWSLT2012 から 2017 までの講演データ、JESC の訓練データ、IWSLT2017 のテストデータの 4 つを利用する。TED の講演データは英語がオリジナルのデータであるため、過去に公開されているデータを利用することが可能である。また、追加学習で用いるのは入力側

データとなるため、テストデータを直接追加学習に利用することが可能である。

表 2 追加学習に利用する入力側単言語資源

コーパス (入力側単言語)	文数
訓練データ (IWSLT2017)	223,108
学習データ以外の IWSLT	474,347
訓練データ (JESC)	2,797,388
テストデータ (IWSLT2017)	1,549

4.2 実験設定

ベースラインは、RNN に LSTM ＋ Attention 機構のモデルと Transformer のモデルの 2 つを用いた。翻訳の単位は両モデル共に単語単位での翻訳を行う。レイヤー数は RNN が 2 層、Transformer が 6 層で隠れ層の次元数は RNN が 768、Transformer が 512 となっている。Optimizer は両モデル共に Adam を使用しドロップアウトも共に 0.3 となっている。学習済みの BERT は Google が公開している BERT-Base モデル [3] を利用した。12 層の Transformer で構築され、隠れ層の次元数は 768 となっている。

4.3 実験結果

4.3.1 事前学習済みモデルの利用

事前学習済みモデルを利用した BERT 入力モデルと BERT-fused モデルの性能を調査する。BERT 入力モデルの NMT は LSTM を使用した為、ベースライン (RNN) からの性能を比較、BERT-fused モデルでは NMT は Transformer を使用する為、ベースライン (Transformer) との比較を行う。

表 3 事前学習済みモデルを利用した性能の比較 (BLEU)

model	dev	test
ベースライン (RNN)	9.80	10.14
ベースライン (Transformer)	10.14	10.58
BERT 入力モデル	12.21	12.32
BERT-fused	**12.26**	**12.48**

事前学習済みモデルを利用したモデルは共にベースラインモデルからの改善が確認できた。BERT 入力モデルでは、RNN から 2.18Point の改善が見られた。また BERT-fused モデルでは、Transformer から 1.90Point の改善が見られた。2 つのモデルを比較すると NMT に Transformer を利用する BERT-fused モデルがより高い性能を示した。

4.3.2 事前学習済みモデルへの追加学習

BERT-fused モデルにおける BERT への追加学習の効果を調べた。3 種類のコーパスを用いて、追加学習を行い 4 つの翻訳モデルを構築した。元の BERT-fused モデルと比較した。

Bf+trainIWSLT
BERT-fused モデルの BERT へ学習データである IWSLT2017 の英語側訓練データを用いて追加学習を行ったモデル

Bf+otherIWSLT
学習データを含まない IWSLT の英語側講演データを用いて追加学習を行ったモデル

Bf+trainJESC
追加学習に用いる IWSLT のデータは数十万文の少量のデータのため、ドメイン外ではあるが講演に近く大量のデータを持つ JESC の訓練データを用いて追加学習を行ったモデル

Bf+testIWSLT
追加学習は入力側データのみで行えることから、目的としている IWSLT2017 のテストデータを直接用いて追加学習を行ったモデル

表 4 追加学習の性能の比較 (BLEU)

model	dev	test
ベースライン (Transformer)	10.14	10.58
BERT-fused	**12.26**	**12.48**
Bf+trainIWSLT	12.10	12.21
Bf+otherIWSLT	11.90	11.89
Bf+trainJESC	11.54	11.53
Bf+testIWSLT	11.60	11.57

BERT へ追加学習を行った 4 つのモデル全てで、追加学習を行っていない元の BERT-fused モデルからの改善は見られなかった。原因として講演データを用いることでより特定のドメインに BERT を強化できると考えたが、本来 BERT が持つ汎化性能が低下し翻訳性能の低下に繋がったと考える。

4.3.3 転移学習

BERT-fused における転移学習の効果を調べた。転移学習には ASPEC と JESC の 2 つのドメイン外対訳コーパスを用いて 5 つの翻訳モデルを構築した。元の BERT-fused モデルと比較した。

Bf+transASPEC
ASPEC を用いて BERT-fused モデルに転移学習を

行ったモデル

Bf+transJESC

JESC を用いて BERT-fused モデルに転移学習を行ったモデル

Bf+transASPEC/JESC

ASPEC と JESC を結合したコーパスを用いて BERT-fused モデルに転移学習を行ったモデル

Bf+transJESC'

JESC の 2 文を 1 文に連結したコーパス JESC' を用いて BERT-fused モデルに転移学習を行ったモデル

Bf+transASPEC/JESC'

ASPEC と JESC' を結合したコーパスを用いて BERT-fused モデルに転移学習を行ったモデル

表 5 転移学習の性能の比較 (BLEU)

model	dev	test
ベースライン (Transformer)	10.14	10.58
BERT-fused	12.26	12.48
Bf+transASPEC	13.72	14.19
Bf+transJESC	13.57	14.01
Bf+transASPEC/JESC	13.21	14.18
Bf+transJESC'	**13.70**	**14.70**
Bf+transASPEC/JESC'	13.72	13.70

転移学習の手法では ASPEC と JESC の 2 つのコーパスで共に改善が見られ、ASPEC を用いた場合がより大きな改善が見られた。科学技術論文の抄録である ASPEC に比べ字幕データである JESC は TED の講演データに近いドメインを持つが、ASPEC より大きな改善が見られなかった為、各コーパスの特徴を調査した。各コーパスでの平均文長を調べると JESC は IWSLT2017 と ASPEC に比べ半分以下の文長であることが分かった。そこで JESC の 2 文を 1 文に連結処理した JESC' を作成し転移学習を行った。

表 6 各コーパスの平均文長

コーパス	文数	単語数	平均文長
IWSLT2017	223,108	4,553,949	20.4
ASPEC	1,000,000	25,915,972	25.9
JESC	2,797,388	23,897,271	8.54

JESC' を転移学習に用いたモデルは ASPEC のみを用いる場合よりも高い性能を示した。文長をより、IWSLT2017 の学習データに近いものとしたため、性能の改善に繋がったのではないかと考える。ASPEC と JESC' を混ぜたデータを用いる転移学習では更なる性能改善は見られなかった。

表 7 はベースラインである Transformer と最も良い性能を示した Bf+transJESC' の翻訳例である。Bf+transJESC' は Transformer で翻訳出来ていなかった入力文の単語を正しく翻訳出来るようになっていることが分かる。

表 7 翻訳の改善例

英語 (入力文)	it 's law , it 's morality , it 's patent stuff .
正解文	法律 や 倫理 や **特許** の **問題** です
Transformer	法律 や 道徳 観 の こと です
Bf+transJESC'	法律 や 道徳 や **特許** の **問題** です

英語 (入力文)	can 't we just cut it in half or a quarter ?
正解文	半分 とか **1 / 4** で は ダメ な の か ?
Transformer	半分 に カット し て も いい です か ?
Bf+transJESC'	半分 か **1 / 4** で は どう でしょ う ?

5 おわりに

本研究では TED の講演データである IWSLT2017 を用いた英日翻訳性能の改善を目的に、入力側単言語資源で学習された事前学習済みモデルと入力側である英語データによる事前学習済みモデルへの追加学習、また大量のドメイン外の英日対訳コーパスを利用した転移学習の手法を行った。BERT を用いた翻訳モデルは既存の RNN、Transformer ベースのモデルからより性能が向上することが確認できた。BERT-fused モデルにおける BERT への追加学習は性能の改善は確認できなかった。BERT-fused における転移学習の効果は大きく、講演データに近いドメインを含む JESC コーパスを連結処理し用いた転移学習は最も良い性能が確認できた。最終的にベースラインである Transformer の BLEU スコア 10.58 から 4.12Point の大幅な改善を達成し 14.70 まで向上が見られた。提案法で得られた BLEU スコア 14.70 は、IWSLT2018 で報告されているベストスコア 10.88[11] と比べてもかなり高いものとなっている。今後の課題としては、追加学習に用いた入力側単言語資源を疑似的な対訳コーパス作成のために利用しデータ拡張の手法を試みる事や、現在提案されている様々な BERT モデルを用いることなどが考えられる。

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

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding", In North American Association for Computational Linguistics (NAACL),2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov."RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv e-prints, arXiv:1907.11692,2019.

[5] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", International Conference on Learning Representations ,2020.

[6] 高橋 竜, 秋葉 友良, 塚田 元,"汎用分散表現 BERT を用いたニューラル機械翻訳の検討", 言語処理学会第 27 回年次大会,2020

[7] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, Tie-Yan Liu."Incorporating BERT into Neural Machine Translation",International Conference on Learning Representations,2020.

[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.

[9] 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.

[10] Reid Pryzant, Youngjoo Chung, Dan Jurafsky, Denny Britz. "JESC: Japanese-English Subtitle Corpus",European Language Resources Association,2018.

[11] Yuto Takebayashi, Chu Chenhui, Yuki Arase, Masaaki Nagata. "Word Rewarding for Adequate Neural Machine Translation",In Proceedings of the 15th International Workshop on Spoken Language Translation,2018.