BERTを用いた機械翻訳の自動評価

嶋中 宏希† 梶原 智之†‡ 小町 守†
†首都大学東京 ‡大阪大学
shimanaka-hiroki@ed.tmu.ac.jp, kajiwara@ids.osaka-u.ac.jp, komachi@tmu.ac.jp

1 はじめに

本研究では，参照文を用いた文単位での機械翻訳自動評価手法について述べる．人手評価との相関が高い文単位の評価ができることにより機械翻訳システムの細かい改善が可能になる．

我々の先行研究 [1] では，単語 N-gram などの局所的な素性に基づく従来手法 [2] では扱えない大域的な情報を考慮するために，大規模コーパスによって事前学習された文の分散表現を用いる機械翻訳自動評価手法 RUSE[1] (Regressor Using Sentence Embeddings) を提案した．RUSE は，機械翻訳自動評価手法の性能を競う WMT-2018 Metrics Shared Task [3] において，文単位の全ての to-English 言語対で最高性能を達成した．この結果は，事前学習された文の分散表現が機械翻訳の自動評価にとって有用な素性であることを示す．

このような文単位での表現学習に関する研究は近年急速に発展している．特に，BERT (Bidirectional Encoder Representations from Transformers) [4] が多くの応用タスクで最高性能を更新し，注目を集めている．BERT は，大規模な生コーパスを用いて双方向言語モデルおよび隣接文推定の事前学習を行った上で，タスクに応じた再訓練を行う．例えば，極性分類のような単一文の分類タスクを解く場合と含意関係認識のような文対の分類タスクを解く場合では，異なる方法で再訓練を行う．これによって，機械翻訳自動評価の類似タスクである文対の意味的類似度推定タスクにおいても，高い性能を発揮している．

そこで本研究では，BERT を用いた機械翻訳の自動評価を行う．WMT-2017 Metrics Shared Task [5] のデータセットにおける実験の結果，BERT は文単位の全ての to-English 言語対で RUSE を凌ぎ，最高性能を更新した．詳細な分析の結果，RUSE との主な相違点である**事前学習の方法，文対モデリング，符号化器の再訓練**の 3 点が，それぞれ BERT の性能改善に貢献していることが明らかになった．

[1] https://github.com/Shi-ma/RUSE

2 関連研究

本節では，WMT-2017 [5] および-2018 [3] の Metrics Shared Task において最高性能を達成した機械翻訳自動評価手法について説明する．本タスクでは，機械翻訳の翻訳文に対して人手で参照文および評価値が付与されたデータセットを利用する．各手法は，翻訳文と参照文の文対を入力として評価値を推定し，人手評価とのピアソンの相関係数によって評価される．本稿では，文単位の to-English 言語対について議論する．

2.1 Blend: 局所的な素性に基づく手法

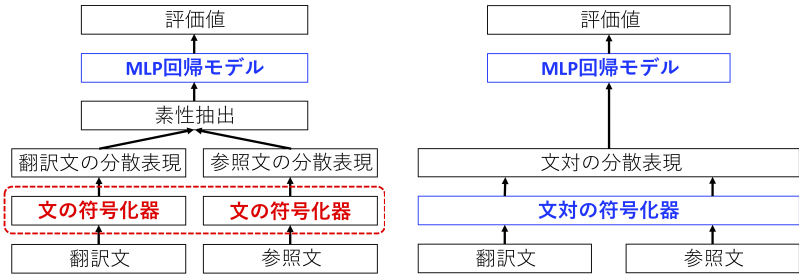
WMT-2017 において最高性能を達成した Blend[2] [2] は，機械翻訳の自動評価用ツールキット Asiya[3]の基本素性に 4 種類の他の機械翻訳自動評価手法を組み合わせたアンサンブル手法である．Blend は多くの素性を用いる手法であるが，文字単位の編集距離や単語 N-gram に基づく素性など，文全体を同時に考慮できない局所的な情報のみに頼っている．

2.2 RUSE: 文の分散表現に基づく手法

WMT-2018 において最高性能を達成した RUSE[1] [1] は，大規模コーパスによって事前学習された文の分散表現を用いる機械翻訳自動評価手法である．Blend などの従来手法とは異なり，RUSE は文全体の情報を分散表現として同時に考慮できるという利点を持つ．

文の分散表現を用いる手法には，ReVal [6] もある．ReVal は WMT Metrics Shared Task および文対の意味的類似度推定タスクのラベル付きデータを用いて文の分散表現を訓練するが，少量のコーパスのみを用いるため十分な性能を達成できない．RUSE では，Quick Thought [7] などの大規模な外部データを用いて事前学習された文の分散表現を利用し，ラベル付きデータを用いて回帰モデルのみを訓練する．

[2] http://github.com/qingsongma/blend
[3] http://asiya.lsi.upc.edu



(a) RUSE による機械翻訳の自動評価　(b) BERT による機械翻訳の自動評価

図 1: 各手法の概要．青は訓練するが赤は固定する．

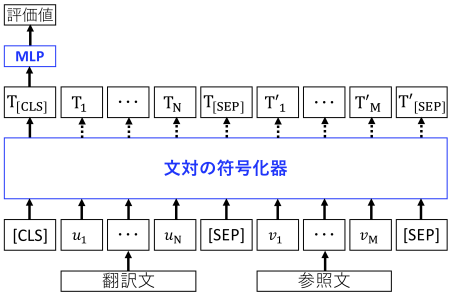


図 2: BERT の文対モデリング

図 1(a) に示すように，RUSE は翻訳文と参照文を文の符号化器でそれぞれ符号化する．そして，InferSent [8] にならって 2 つの文の分散表現を組み合わせて素性を抽出し，多層パーセプトロン (MLP) に基づく回帰モデルによって評価値を推定する．

3　BERT による機械翻訳自動評価

本研究では，BERT [4] を用いて機械翻訳の自動評価を行う．BERT は RUSE と同じく，事前学習された文の分散表現を利用し，MLP によって評価値を推定する．ただし，図 1(b) に示すように，BERT による機械翻訳の自動評価では翻訳文と参照文の両方を文対の符号化器で同時に符号化する．そして，文対の分散表現をそのまま MLP へ入力する．RUSE とは異なり，事前学習された符号化器も MLP とともに再訓練される．以下では，BERT による機械翻訳自動評価の特徴である事前学習の方法，文対モデリング，符号化器の再訓練について詳細に説明する．

3.1　事前学習

BERT は，大規模な生コーパスを用いて，以下の 2 種類の教師なし事前学習を同時に行う．

双方向言語モデル　生コーパスの一部のトークンを [MASK] トークンに置換した上で，双方向の言語モデルによって元のトークンを推定する．この教師なし事前学習によって，BERT の符号化器は文内におけるトークン間の関係を学習する．

隣接文推定　生コーパスの一部の文を無作為に他の文に置換した上で，連続する 2 文が隣接していた文対か否かの 2 値分類を行う．この教師なし事前学習によって，BERT の符号化器は文対の関係を学習する．

3.2　文対モデリング

BERT では，隣接文推定や含意関係認識などの文対を扱うタスクのために，各文を独立に符号化するのではなく，文対を同時に符号化する．文対に含まれる各文は，入力系列の先頭に一度のみ追加される [CLS] トークンおよび各文末に追加される [SEP] トークンによって区別される（図 2）．最終的に，[CLS] トークンに対応する最終の隠れ層が，文対の分散表現を表す[4]．

3.3　符号化器の再訓練

BERT では，符号化器で文または文対の分散表現を得た後，それを入力として MLP によって分類や回帰などの応用タスクを解く．なお，応用タスクのラベル付きデータを用いて MLP を訓練する際，文または文対の分散表現を得るための符号化器も再訓練する．

4　評価実験

WMT-2017 Metrics Shared Task [5] のデータセットを用いて，文単位の to-English 言語対における BERT の有効性を検証する．

4.1　実験設定

表 1 に，データセットの文対数を示す．WMT-2015 および WMT-2016 の合計 5,360 文対は無作為に分割し，9 割を訓練用，1 割を開発用に利用する．WMT-2017 の文対は評価用に利用する．

比較手法には，WMT Metrics Shared Task のベースラインである SentBLEU [5]，WMT-2017 にて最高

[4]極性分類などの単一文を扱うタスクのために，文対ではなく文を符号化することもできる．この場合，文頭と文末に [CLS] トークンと [SEP] トークンが一度ずつ追加され，[CLS] トークンに対応する最終の隠れ層が文の分散表現を表す．

	cs-en	de-en	fi-en	lv-en	ro-en	ru-en	tr-en	zh-en
WMT-2015	500	500	500	-	-	500	-	-
WMT-2016	560	560	560	-	560	560	560	-
WMT-2017	560	560	560	560	-	560	560	560

表 1: WMT Metrics Shared Task の to-English 言語対[5]における人手評価付き文対数

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
SentBLEU [5]	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
Blend [5]	0.594	0.571	0.733	0.577	0.622	0.671	0.661	0.633
RUSE [1]	0.614	0.637	0.756	0.705	0.680	0.704	0.677	0.682
$\text{BERT}_{\text{BASE}}$	**0.732**	**0.751**	**0.856**	**0.829**	**0.795**	**0.811**	**0.763**	**0.791**

表 2: WMT-2017 Metrics Shared Task（文単位，to-English 言語対）におけるピアソンの相関係数

性能を達成した Blend [2]，WMT-2018 にて最高性能を達成した RUSE [1] を用いる．各手法は，ピアソンの相関係数を用いて人手評価との相関を評価される．

BERT には，著者らによって公開されている訓練済みモデルのうち，$\text{BERT}_{\text{BASE}}$ (uncased)[6]を用いる．BERT のパラメータは，以下の組み合わせの中からグリッドサーチによって選択する．

- バッチサイズ $\in \{8, 16, 32\}$
- 学習率（Adam）$\in$ {5e-5, 3e-5, 2e-5}
- エポック数 $\in \{1, ..., 10\}$
- ドロップアウト率 $\in \{0.1\}$
- MLP の隠れ層の数 $\in \{0\}$
- MLP の隠れ層の次元 $\in \{768\}$

4.2 実験結果

表 2 に実験結果を示す．BERT は，全ての言語対で他の手法を大幅に上回る性能を達成した．5 節では，RUSE と BERT を比較しつつ，詳細な分析を行う．

5 分析：RUSE と BERT の比較

RUSE と BERT の主な相違点である**事前学習の方法，文対モデリング，符号化器の再訓練**の 3 点に関して分析するために，以下の設定で実験を行う．

- RUSE with GloVe-BoW: 図 1(a) の文の分散表現として，単語分散表現 GloVe [9] (glove.840B.300d[7]) の平均ベクトルを用いる．
- RUSE with Quick Thought: 図 1(a) の文の符号化器として，隣接文推定によって事前学習された Quick-Thought [7] を用いる．
- RUSE with $\text{BERT}_{\text{BASE}}$（文）: 図 1(a) の文の符号化器として，双方向言語モデルと隣接文推定によって事前学習された単一文入力の BERT を用いる．ただし，文の符号化器は再訓練しない．
- RUSE with $\text{BERT}_{\text{BASE}}$（文対）: 図 1(a) の MLP の入力として，文対を入力とする BERT の出力を用いる．ただし，文対の符号化器は再訓練しない．

RUSE の素性として BERT を用いる場合，[CLS] トークンに対応する隠れ層のうち最終 4 層を連結し，文または文対の分散表現として用いる．RUSE のパラメータは先行研究 [1] にならって，以下の組み合わせの中からグリッドサーチによって選択する．

- バッチサイズ $\in \{64, 128, 256, 512, 1024\}$
- 学習率（Adam）$\in$ {1e-3}
- エポック数 $\in \{1, 2, ..., 30\}$
- ドロップアウト率 $\in \{0.1, 0.2, 0.3\}$
- MLP の隠れ層の数 $\in \{1, 2, 3\}$
- MLP の隠れ層の次元 $\in \{512, 1024, 2048, 4096\}$

これらの実験結果を表 3 に示す．

[5]en: English, cs: Czech, de: German, fi: Finnish, ro: Romanian, ru: Russian, tr: Turkish, lv: Latvian, zh: Chinese

[6]https://github.com/google-research/bert

[7]https://nlp.stanford.edu/projects/glove

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
RUSE with GloVe-BoW	0.475	0.479	0.645	0.532	0.537	0.547	0.480	0.527
RUSE with Quick Thought [1]	0.601	0.587	0.737	0.685	0.661	0.692	0.647	0.658
RUSE with $\text{BERT}_{\text{BASE}}$（文）	0.622	0.626	0.765	0.708	0.609	0.706	0.647	0.669
RUSE with $\text{BERT}_{\text{BASE}}$（文対）	0.645	0.607	0.780	0.727	0.644	0.704	0.705	0.687
$\text{BERT}_{\text{BASE}}$	**0.732**	**0.751**	**0.856**	**0.829**	**0.795**	**0.811**	**0.763**	**0.791**

表 3: WMT-2017 Metrics Shared Task（文単位，to-English 言語対）における RUSE と BERT の比較

事前学習　表 3 の上から 3 行を比較すると，文の符号化器における事前学習の方法による性能への影響がわかる．まず，単語の分散表現に基づく GloVe-BoW よりも，文の分散表現に基づく Quick Thought の方が，一貫して高い性能を持つ．次に，隣接文推定のみによって事前学習された Quick Thought よりも，双方向言語モデルと隣接文推定の両方によって事前学習された $\text{BERT}_{\text{BASE}}$（文）の方が，多くの言語対において優れた性能を発揮する．つまり，BERT の大きな特徴のひとつである双方向言語モデルによる事前学習は，機械翻訳の自動評価のためにも有用である．

文対モデリング　RUSE with $\text{BERT}_{\text{BASE}}$（文）と RUSE with $\text{BERT}_{\text{BASE}}$（文対）を比較すると，文対モデリングによる性能への影響がわかる．多くの言語対において，翻訳文と参照文を独立に符号化する前者よりも，同時に符号化する後者の方が高い性能を持つ．我々は InferSent にならって 2 つの文の分散表現を組み合わせる素性抽出を行ったが，これが機械翻訳の自動評価に適した素性抽出の方法であるとは限らない．一方で，BERT の文対モデリングは，素性抽出を陽に行うことなく文対の関係を考慮した分散表現を得ている．BERT では，隣接文推定による事前学習の際に，上手く文対の関係を学習できている可能性がある．

符号化器の再訓練　表 3 の下から 2 行を比較すると，符号化器の再訓練による性能への影響がわかる．全ての言語対において，MLP のみを訓練する RUSE with $\text{BERT}_{\text{BASE}}$（文対）よりも，MLP とともに符号化器を再訓練する $\text{BERT}_{\text{BASE}}$ の方が大幅に優れた性能を発揮する．つまり，BERT の大きな特徴のひとつである符号化器の再訓練は，機械翻訳の自動評価のためにも有用である．

6　おわりに

本研究では，BERT を用いた機械翻訳の自動評価を行った．実験の結果，BERT は文単位の全ての to-English 言語対で他の手法を大幅に上回り，最高性能を更新した．また，先行研究の RUSE との比較に基づく分析の結果，事前学習の方法，文対モデリング，符号化器の再訓練の 3 つの点が，それぞれ BERT の性能改善に貢献していることを示した．

[謝辞]　本研究の一部は JSPS 科研費（研究活動スタート支援，課題番号: 18H06465）の助成を受けたものです．

参考文献

[1] Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In *Proc. of WMT*, pp. 764–771, 2018.

[2] Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. Blend: a Novel Combined MT Metric Based on Direct Assessment —CASICT-DCU submission to WMT17 Metrics Task. In *Proc. of WMT*, pp. 598–603, 2017.

[3] Qingsong Ma, Ondřej Bojar, and Yvette Graham. Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. In *Proc. of WMT*, pp. 682–701, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*, 2018.

[5] Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 Metrics Shared Task. In *Proc. of WMT*, pp. 489–513, 2017.

[6] Rohit Gupta, Constantin Orasan, and Josef van Genabith. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proc. of EMNLP*, pp. 1066–1072, 2015.

[7] Lajanugen Logeswaran and Honglak Lee. An Efficient Framework for Learning Sentence Representations. In *Proc. of ICLR*, 2018.

[8] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proc. of EMNLP*, pp. 670–680, 2017.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proc. of EMNLP*, pp. 1532–1543, 2014.