学修番号 18860628

修士論文

汎用的な文の分散表現を用いた機械翻訳自動評価

嶋中 宏希

2020 年 2 月 21 日

首都大学東京大学院
システムデザイン研究科 情報科学域

嶋中 宏希

審査委員：

小町 守 准教授　（主指導教員）

山口 亨 教授　　（副指導教員）

高間 康史 教授　（副指導教員）

汎用的な文の分散表現を用いた機械翻訳自動評価*

嶋中 宏希

修論要旨

本稿では，文単位での機械翻訳の**自動評価**および**品質推定**（参照文を利用しない自動評価）について述べる．機械翻訳の自動評価では，機械翻訳システムによる翻訳文について，参照文（原文を人手で翻訳した文）と比較して評価する．機械翻訳の品質推定では，機械翻訳システムによる翻訳文について，参照文を利用せずに原文と比較して評価する．本研究では，機械翻訳の自動評価および品質推定の両方に焦点を当て，手法の提案と分析を行う．

文単位での信頼性の高い自動評価により，機械翻訳システムの細かい分析が可能になる．また，信頼性の高い品質推定によって，機械翻訳システムのより幅広い分析が可能になる．文単位での機械翻訳の評価手法には，ある機械翻訳システムの翻訳文に対して他のシステムの翻訳文と比較して相対的に評価する手法と，翻訳文の品質を絶対的に評価する手法がある．本研究では，機械翻訳システムの文単位での定性的な分析，つまり，評価対象の機械翻訳システムがどのような文に対してどの程度の品質で翻訳できるのかについての分析を可能にするため，各翻訳文に対して絶対的な自動評価を行う．また，人手評価に近い絶対評価ができる手法を信頼性の高い自動評価であると捉え，その信頼性に基づいて各評価手法の性能比較や分析を行う．

機械翻訳に関する国際会議 Conference on Machine Translation（WMT）では，機械翻訳自動評価手法の人手評価との相関を比較する Metrics Shared Task が開催されており，これまでに多くの手法が提案されてきた．しかし，現在のデファクトスタンダードである BLEU をはじめとして，ほとんどの機械翻訳自動評価手法は文字 N-gram や単語 N-gram などの局所的な素性を利用しており，文単位での評価にとっては限定的な情報しか扱えていない．また，大域的な情報を考慮するために，

*首都大学東京大学院 システムデザイン研究科 情報科学域 修士論文, 学修番号 18860628, 2020 年 2 月 21 日.

文全体の特徴をベクトル空間上で表現することができる文の分散表現を用いた手法も存在するが，人手評価値付きのデータセットなどの比較的少量の教師ありデータのみを用いてモデル全体を学習するため，十分な性能を示せていない．

そこで本研究では，局所的な素性に基づく従来手法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に基づく，機械翻訳自動評価手法を提案する．我々の提案手法は，(a) 翻訳文と参照文を独立に符号化した文の分散表現を用いる手法と，(b) 翻訳文と参照文を同時に符号化した文の分散表現を用いる手法に大別できる．これらの 2 つの提案手法は，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素性として利用し，人手評価値付きのデータセット上で訓練された回帰モデルによって機械翻訳の自動評価を行うという点で共通している．これらの 2 つの提案手法に対して性能の評価を行い，文の分散表現の事前学習の方法，翻訳文と参照文の符号化器への入力方法，符号化器の再訓練の 3 点について詳細な分析を行った．

また本研究では，多言語の生コーパス上で事前学習した文の分散表現を用いた機械翻訳品質推定手法（参照文を利用しない自動評価手法）についても提案する．多言語コーパス上で事前学習した文の分散表現を用いることで，異なる言語である原文と翻訳文を用いた参照文を利用しない自動評価を可能にした．我々の提案手法は，多言語の大規模な生コーパス上で共通の文や文対の符号化器の事前学習を行い，原文・翻訳文・翻訳品質スコアの 3 つ組を用いて，原文および翻訳文の文対から翻訳品質を推定する回帰モデルを学習する．この提案手法に対して性能の評価を行い，言語横断的に文符号化器を再訓練することによる性能への影響について分析を行った．

本研究の主な貢献は以下の 4 つである．

- 事前学習された文の分散表現に基づく機械翻訳自動評価手法を提案し，事前学習された文の分散表現が機械翻訳自動評価において有用な素性であることを示した．
- 提案手法についての詳細な分析により，文の分散表現の事前学習の方法，翻訳文と参照文の符号化器への入力方法，符号化器の再訓練の 3 点が，それぞれ機械翻訳の自動評価における性能改善に貢献していることを明らかにした．

- 事前学習された多言語の文の分散表現に基づく機械翻訳品質推定手法を提案し，事前学習された多言語の文の分散表現が機械翻訳の品質推定において有用な素性であることを示した．
- 提案手法についての詳細な分析により，事前学習された多言語の文符号化器を言語横断的に再訓練することが，機械翻訳の品質推定における性能改善に貢献していることを明らかにした．

本稿の構成を示す．第 1 章では本研究の提案，貢献，概要について述べる．第 2 章では，機械翻訳の人手評価について説明し，続いて機械翻訳の自動評価手法および品質推定手法の関連研究について概説する．第 3 章では，事前学習された文の分散表現に基づいた機械翻訳の自動評価手法および品質推定手法を提案する．第 4 章では，WMT Metrics Shared Task の人手評価値付きデータセットを用いて，提案手法の評価実験を行う．第 5 章では，提案手法についての分析と考察を行う．最後に第 6 章で，本研究のまとめを述べる．

Metric for Automatic Machine Translation Evaluation Using Universal Sentence Representations*

Hiroki Shimanaka

Abstract

In this paper, we describe sentence-level methods of machine translation evaluation and quality estimation (translation evaluation without reference translation). In **machine translation evaluation (MTE)** task, the machine translation (MT) hypothesis is evaluated by comparing it with the reference translation. In **quality estimation (QE)** task, the MT hypothesis is evaluated by comparing it with the source sentence without using the reference sentence. In this study, we propose and analyze methods for these two tasks.

The MTE methods with a high correlation with human evaluation enable continuous detailed deployment of an MT system. The QE methods with a high correlation with human evaluation enable continuous extensive deployment of an MT system. There are two types of sentence-level MTE methods: one is to evaluate the translation of one MT system relative to the translation of another system, and the other is to absolutely evaluate the quality of the translation. In this research, we focus on absolute automatic evaluation to enable qualitative analysis of sentence-level in MT systems. In addition, we consider a method that can perform absolute evaluation close to human evaluation to be highly reliable automatic evaluation, and compare and analyze the performance of each evaluation method based on the reliability.

Various MTE methods have been proposed in the Metrics Shared Task of

*Master's Thesis, Department of Computer Science, Graduate School of System Design, Tokyo Metropolitan University, Student ID 18860628, February 21, 2020.

the Conference on Machine Translation (WMT). However, most MTE metrics, including the current de facto standard BLEU, are obtained by computing the similarity between an MT hypothesis and a reference based on the character or word N-grams. Therefore, they can exploit only limited information for the sentence-level MTE. There is also a method that uses sentence representations to consider global information. However, since the whole model is trained using only a relatively small amount of supervised data, it does not show sufficient performance.

Therefore, we propose a sentence-level MTE method using universal sentence representations capable of capturing global information that cannot be captured by local features. Our method can be roughly divided into (a) the method that uses sentence representations of an MT hypothesis and a reference translation which are independently encoded and (b) the method that uses sentence representations of an MT hypothesis and a reference translation which are jointly encoded. These two proposed methods have in common that they use sentence representations pre-trained on large-scale corpus as features and evaluate MT hypothesis using a regression model that is trained on datasets with human evaluation. We evaluated the performance of these two proposed methods and analyzed pre-training methods of sentence representations, input methods of an MT hypothesis and reference translation into an encoder, and fine-tuning methods of encoder in detail.

In this study, we also propose a QE method (an MTE method without reference translation) using sentence representations pre-trained on a raw multilingual corpus. It is possible to perform MTE without reference translation using a source sentences and an MT hypothesis in different languages by using sentence representations pre-trained on a multilingual corpus, Our method pre-trains a sentence or sentence-pair encoder on a large-scale multi-lingual raw corpus and trains a regression model that estimates translation quality score from source sentence and MT hypothesis. We evaluated the performance of the proposed method and analyzed the effect of cross-lingual fine-tuning on the sentence or

sentence-pair encoder.

The main contributions of the study are summarized below:

- We propose the MTE methods based on pre-trained sentence representations, and show that pre-trained sentence representations are useful features in MTE.
- A detailed analysis of the proposed methods revealed that pre-training methods of sentence representations, input methods of a MT hypothesis and reference translation into an encoder, and fine-tuning methods of encoder contributed to the performance improvement in MTE.
- We propose the QE methods based on pre-trained multi-lingual sentence representations, and show that pre-trained multi-lingual sentence representations are useful features in QE.
- A detailed analysis of the proposed methods revealed that cross-lingual fine-tuning on pre-trained multi-lingual sentence encoder contributed to the performance improvement in QE.

The structure of this paper is as follows. Chapter 1 describes the proposal, contribution, and outline of this research. Chapter 2 describes human evaluation of MT, followed by an overview of related work on MTE and QE task. Chapter 3 describes the proposed methods for MTE and QE based on pre-trained sentence representations. Chapter 4 describes an evaluation experiment of the proposed methods using datasets with human evaluation score of WMT Metrics Shared Task. Chapter 5 describes the analysis and consideration of the proposed methods. Finally, Chapter 6 describes the summary of this research.

目次

図目次 ix

第 1 章 はじめに 1

第 2 章 関連研究 5
2.1 機械翻訳の人手評価 5
2.2 機械翻訳自動評価 6
2.2.1 機械翻訳自動評価のための教師なし手法 6
2.2.2 機械翻訳自動評価のための教師あり手法 7
2.2.3 機械翻訳自動評価のための大域的な素性に基づく教師あり手法 8
2.3 機械翻訳品質推定 8
2.3.1 機械翻訳品質推定のための教師なし手法 9
2.3.2 機械翻訳品質推定のための教師あり手法 9

第 3 章 事前学習された文の分散表現に基づく機械翻訳の自動評価および品質推定 10
3.1 RUSE: 文の分散表現に基づく機械翻訳自動評価のための回帰モデル 10
3.1.1 事前学習された文の分散表現 10
3.1.2 機械翻訳自動評価のための回帰モデルと素性抽出 12
3.2 BERT による機械翻訳自動評価 12
3.2.1 BERT における事前学習 13
双方向言語モデル： 13
隣接文推定： 14
3.2.2 BERT における文対モデリング 14
3.2.3 BERT における符号化器の再学習 14
3.3 多言語 BERT による機械翻訳品質推定 14

第 4 章 評価実験 16

4.1 機械翻訳自動評価についての評価実験 16
4.1.1 実験設定 16
4.1.2 比較手法 17
RUSE with GloVe-BoW: 17
RUSE with IS: 17
RUSE with QT: 18
RUSE with USE: 18
RUSE with BERT: 18
BERT (w/o fine-tuning): 18
BERT: 18
4.1.3 実験結果 19
4.2 機械翻訳品質推定についての評価実験 21
4.2.1 実験設定 22
4.2.2 比較手法 22
4.2.3 実験結果 24

第 5 章 分析 26
5.1 機械翻訳の自動評価についての分析 26
5.1.1 学習データの文対数と性能の関係 26
5.1.2 from-English 言語対における性能 28
5.1.3 出力例 30
5.2 機械翻訳の品質推定についての分析 31
5.2.1 対象言語対のみで学習 31
5.2.2 Zero-shot 学習 33

第 6 章 おわりに 34

謝辞 35

参考文献 36

発表リスト 40

図目次

1.1 機械翻訳の自動評価および品質推定の概要. 3
1.2 各提案手法の概要．青色部分は学習し，赤色部分は固定する． . . . 4

3.1 InferSent の概要図 11
3.2 Quick Thought の概要図 11
3.3 BERT の文対モデリング（u, v : 入力トークン，T, T' : 各入力トークンに対する分散表現） 13

5.1 RUSE（左）と BERT（右）における学習曲線（人手評価とのピアソンの積率相関係数） 27
5.2 RUSE（左）と BERT（右）における学習曲線（人手評価とのスピアマンの順位相関係数） 27
5.3 RUSE（左）と BERT（右）における学習曲線（人手評価との平均 2 乗誤差） 28

第 1 章　はじめに

本稿では，文単位での機械翻訳の**自動評価**および**品質推定**（参照文を利用しない自動評価）について述べる（図 1）．機械翻訳の自動評価では，機械翻訳システムによる翻訳文について，参照文（原文を人手で翻訳した文）と比較して評価する．機械翻訳の品質推定では，機械翻訳システムによる翻訳文について，参照文を利用せずに原文と比較して評価する．本研究では，機械翻訳の自動評価および品質推定の両方に焦点を当て，手法の提案と分析を行う．

文単位での信頼性の高い自動評価により，機械翻訳システムの細かい分析が可能になる．また，信頼性の高い品質推定によって，機械翻訳システムのより幅広い分析が可能になる．文単位での機械翻訳の評価手法には，ある機械翻訳システムの翻訳文に対して他のシステムの翻訳文と比較して相対的に評価する手法と，翻訳文の品質を絶対的に評価する手法がある．本研究では，機械翻訳システムの文単位での定性的な分析，つまり，評価対象の機械翻訳システムがどのような文に対してどの程度の品質で翻訳できるのかについての分析を可能にするため，各翻訳文に対して絶対的な自動評価を行う．また，人手評価に近い絶対評価ができる手法を信頼性の高い自動評価であると捉え，その信頼性に基づいて各評価手法の性能比較や分析を行う．

機械翻訳に関する国際会議 Conference on Machine Translation（WMT）*では，機械翻訳自動評価手法の人手評価との相関を比較する Metrics Shared Task が開催されており，これまでに多くの手法が提案されてきた．しかし，現在のデファクトスタンダードである BLEU [27] をはじめとして，ほとんどの機械翻訳自動評価手法は文字 N-gram や単語 N-gram などの局所的な素性を利用しており，文単位での評価にとっては限定的な情報しか扱えていない．また，大域的な情報を考慮するために，文全体の特徴をベクトル空間上で表現することができる文の分散表現を用いた手法も存在するが，人手評価値付きのデータセットなどの比較的少量の教師ありデータのみを用いてモデル全体を学習するため，十分な性能を示せていない．

そこで本研究では，局所的な素性に基づく従来手法では扱えない大域的な情報を

*https://aclanthology.info/venues/wmt

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に基づく，機械翻訳自動評価手法を提案する．我々の提案手法は，（a）翻訳文と参照文を独立に符号化した文の分散表現を用いる手法（図 1.2(a)）と，（b）翻訳文と参照文を同時に符号化した文の分散表現を用いる手法（図 1.2(b)）に大別できる．これらの 2 つの提案手法は，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素性として利用し，人手評価値付きのデータセット上で学習された回帰モデルによって機械翻訳の自動評価を行うという点で共通している．これらの 2 つの提案手法に対して性能の評価を行い，文の分散表現の事前学習の方法，翻訳文と参照文の符号化器への入力方法，符号化器の再学習の 3 点について詳細な分析を行った．

また本研究では，多言語の生コーパス上で事前学習した文の分散表現を用いた機械翻訳品質推定手法（図 1.2(c)）についても提案する．多言語コーパス上で事前学習した文の分散表現を用いることで，異なる言語である原文と翻訳文を用いた参照文を利用しない自動評価を可能にした．我々の提案手法は，多言語の大規模な生コーパス上で共通の文や文対の符号化器の事前学習を行い，原文・翻訳文・翻訳品質スコアの 3 つ組を用いて，原文および翻訳文の文対から翻訳品質を推定する回帰モデルを学習する．この提案手法に対して性能の評価を行い，言語横断的に文対符号化器を再学習することによる性能への影響について分析を行った．

本研究の主な貢献は以下の 4 つである．

- 事前学習された文の分散表現に基づく機械翻訳自動評価手法を提案し，事前学習された文の分散表現が機械翻訳自動評価において有用な素性であることを示した．
- 提案手法についての詳細な分析により，文の分散表現の事前学習の方法，翻訳文と参照文の符号化器への入力方法，符号化器の再学習の 3 点が，それぞれ機械翻訳の自動評価における性能改善に貢献していることを明らかにした．
- 事前学習された多言語の文の分散表現に基づく機械翻訳品質推定手法を提案し，事前学習された多言語の文の分散表現が機械翻訳の品質推定において有用な素性であることを示した．
- 提案手法についての詳細な分析により，事前学習された多言語の文対符号化器を言語横断的に再学習することが，機械翻訳の品質推定における性能改善

に貢献していることを明らかにした．

本稿の構成を示す．第 1 章では本研究の提案，貢献，概要について述べる．第 2 章では，機械翻訳の人手評価について説明し，続いて機械翻訳の自動評価手法および品質推定手法の関連研究について概説する．第 3 章では，事前学習された文の分散表現に基づいた機械翻訳の自動評価手法および品質推定手法を提案する．第 4 章では，WMT Metrics Shared Task の人手評価値付きデータセットを用いて，提案手法の評価実験を行う．第 5 章では，提案手法についての分析と考察を行う．最後に第 6 章で，本研究のまとめを述べる．

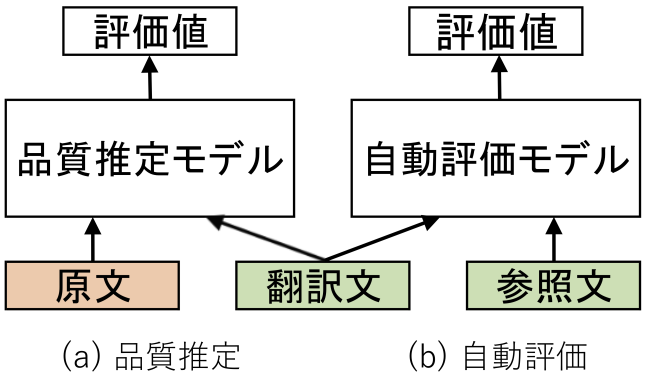
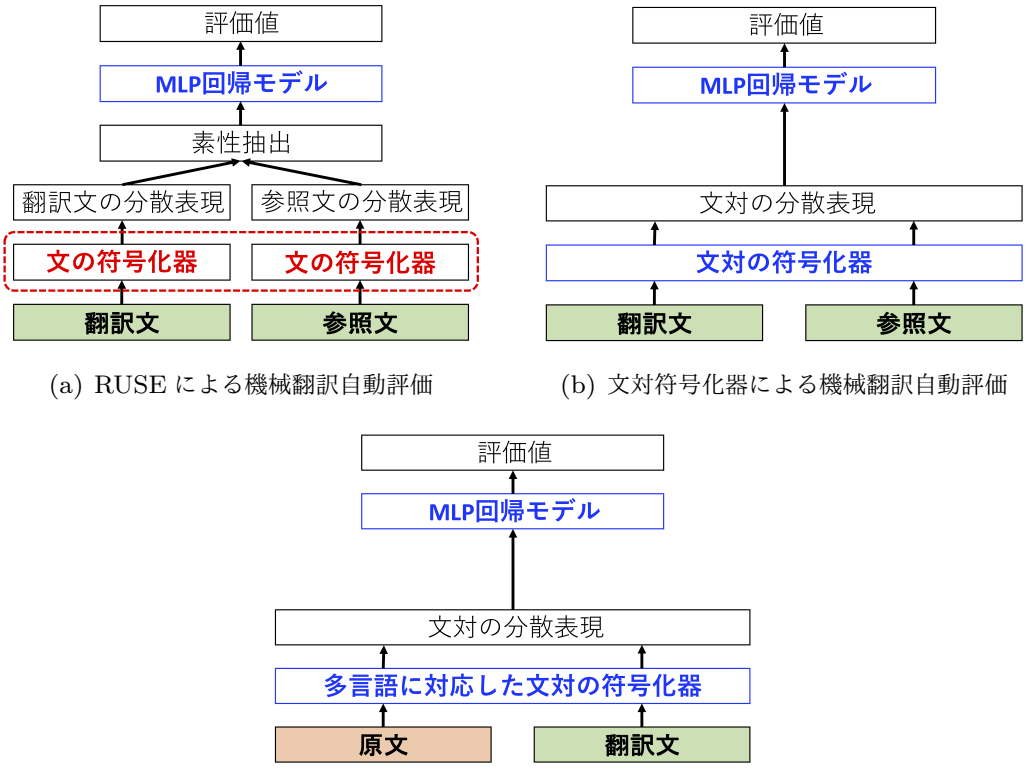


図 1.1　機械翻訳の自動評価および品質推定の概要．



(a) RUSE による機械翻訳自動評価

(b) 文対符号化器による機械翻訳自動評価

(c) 文対符号化器による機械翻訳品質推定

図 1.2　各提案手法の概要．青色部分は学習し，赤色部分は固定する．

第 2 章　関連研究

本章では，まず機械翻訳の人手評価について説明する．続いて機械翻訳自動評価手法の関連研究について概説し，最後に機械翻訳品質推定手法の関連研究について概説する．本稿では，人手評価値付きのデータセットを用いて学習する手法を教師あり手法，学習しない手法を教師なし手法として分けて説明する．

2.1　機械翻訳の人手評価

機械翻訳に関する国際会議 WMT では，参加者が提案した機械翻訳システムの性能を比較する News Translation Shared Task が開催されており，各システムの翻訳文を研究者やクラウドソーシングによって人手評価してきた．WMT Metrics Shared Task における機械翻訳の人手評価としては，各翻訳文に対する相対評価（RR: Relative Ranking） [2] と絶対評価（DA: Direct Assessment） [13, 14, 15] が行われてきた．

人手の相対評価では，ある原文と参照文に対して複数の機械翻訳システムによる翻訳文が与えられ，各翻訳文を順位付けする．しかし，このような相対評価では異なる原文に対する翻訳文同士の品質を比較できないという問題が存在する．そのため，WMT-2016 [4] からは人手の絶対評価が行われ始めた．*

人手の絶対評価では，ある原文と参照文に対して単一の機械翻訳システムによる翻訳文が与えられ，各翻訳文に妥当性や流暢性についての品質スコアを付与する．WMT の人手評価では，原文は考慮せず，翻訳文と参照文の比較のみによって各翻訳文の妥当性や流暢性について絶対的な評価を行っている．ここでの翻訳文の妥当性とは，参照文との意味的な類似度のことであり，機械翻訳におけるターゲット言語側の単言語の評価タスクとなっている．WMT News Translation Shared Task における妥当性や流暢性の人手評価値の収集は下記の手順で行われる．

*WMT において，人手の絶対評価が採用され始めたのは WMT-2016 News Translation Shared Task からであるが，WMT-2016 Metrics Shared Task では学習用のデータセットとして WMT-2015 News Translation Shared Task [31] における翻訳文と参照文に対して人手で付与した絶対評価値付きデータセットが公開されている．

1. WMT News Translation Shared Task に参加した機械翻訳システムの翻訳文とそれに対応する参照文の対が 100 文対ずつ無作為抽出され，各評価者に割り振られる．
2. 各評価者は，翻訳文と参照文を比較し，0〜100 のアナログスケールにより各翻訳文の妥当性や流暢性を評価する．
3. 品質管理 [14] により，質の低い評価者による評価値を排除する．
4. 評価者ごとのスコアの偏りを均質化するため，評価者ごとに平均が 0，標準偏差が 1 となるように z-score を用いて評価値を標準化する．
5. 複数の評価者による標準化された評価値を平均し，最終的な評価値とする．

WMT Metrics Shared Task では，上記の方法で収集されたデータセットの中から妥当性についての質の高いデータ†を各言語対ごとに無作為抽出することにより，人手の絶対評価値付きデータセットを作成している．本研究では，この人手による妥当性についての絶対評価値付きデータセットを用いて，提案手法を学習および評価する．

2.2 機械翻訳自動評価

機械翻訳自動評価では，翻訳文と参照文を比較することにより翻訳文の意味的な品質を評価する．各自動評価手法は，2.1 節で述べた WMT Metrics Shared Task における人手評価との相関により性能を評価される．

2.2.1 機械翻訳自動評価のための教師なし手法

機械翻訳の自動評価におけるデファクトスタンダードである BLEU [27] は，単語 N-gram の一致率に基づくシステム単位の教師なし手法である．文単位での評価のためには，平滑化された SentBLEU‡が用いられる．SentBLEU は，WMT Metrics Shared Task におけるベースライン手法のひとつとして利用されている．

†評価者 15 人以上によって評価された翻訳文と参照文の対 [12]

‡`https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl`

chrF [29] は，文字 N-gram の一致率に基づく手法である．また，chrF+ および chrF++ [30] は，文字 N-gram とともに単語 N-gram の一致率も考慮する．chrF および chrF+ は，WMT-2018 [23] 以降の Metrics Shared Task においてベースライン手法のひとつとして利用されている．

MEANT 2.0[§] [21] は，逆文書頻度で重み付けされた単語 N-gram，単語分散表現に基づく単語類似度および意味役割付与 (SRL) に基づく構文類似度を用いる手法である．SRL を利用できない言語においては，MEANT 2.0-nosrl を適用することができる．MEANT 2.0 は WMT-2017 Metrics Shared Task において，文単位の to-English 言語対[¶]で高い性能を示している．また，MEANT 2.0-nosrl は WMT-2017 Metrics Shared Task において, 文単位の from-English 言語対[‖]で高い性能を示している．

これらの教師なし手法は，多くの言語対において一貫した評価ができるという利点を持つ．しかし，評価値のラベル付きデータが比較的多く存在する to-English 言語対においては，教師あり手法がより高い性能を示している．我々は，to-English 言語対を主な対象として，より人手評価に近い絶対評価ができる教師ありの機械翻訳自動評価手法を提案する．

2.2.2 機械翻訳自動評価のための教師あり手法

BEER[**] [32] は，文字 N-gram の一致率を素性として 2.1 節で述べた人手の相対評価値付きデータセット上で学習を行う教師あり手法である．この手法は，WMT-2017 の Metrics Shared Task において，文単位の from-English 言語対で高い性能を示している．

Blend[††] [24] は，機械翻訳の自動評価用ツールキット Asiya[‡‡] [11] の基本 25 素性に先述の BEER など 4 種類の他の機械翻訳自動評価手法 [32, 35, 37, 38] を組み合

[§] http://chikiu-jackie-lo.org/home/index.php/meant
[¶] 英語以外の言語から英語への翻訳の評価
[‖] 英語から英語以外の言語への翻訳の評価
[**] https://github.com/stanojevic/beer
[††] https://github.com/qingsongma/blend
[‡‡] http://asiya.lsi.upc.edu

わせたアンサンブル手法であり，2.1 節で述べた人手の絶対評価値付きデータセット上で学習する教師あり手法である．この手法は，WMT-2017 Metrics Shared Task において，文単位の to-English 言語対で最高性能を達成している．

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．本研究では，これらの教師あり学習に基づく従来手法では扱えない大域的な情報を考慮する手法を提案する．

2.2.3 機械翻訳自動評価のための大域的な素性に基づく教師あり手法

文全体の大域的な情報を考慮する手法として，文の分散表現に基づく ReVal[§§] [16] がある．ReVal は WMT Metrics Shared Task および文対の意味的類似度推定タスク [26] における人手の相対評価値付きデータセット上で Tree-LSTM [33] によって文の分散表現を学習する．しかし，小規模なラベル付きコーパスのみを用いるため十分な性能を達成できていない [4]．本研究では，大規模な生コーパス上で事前学習された文の分散表現を利用することで，文単位での表現学習における少資源問題を克服する．

2.3 機械翻訳品質推定

機械翻訳品質推定には，翻訳文と原文を比較し翻訳文がプロの翻訳者の修正をどの程度必要とするかを推定する手法と，翻訳文と原文を比較し翻訳文の意味的な品質を推定する手法が存在する．WMT Quality Estimation（QE） Shared Task では前者が，WMT の Metrics および QE Shared Task 内の QE as a Metric Task においては後者が提案されている．どちらの評価手法においても人手評価との相関によりその性能が評価される．

以下の 2 つの従来手法は対訳コーパスを用いて事前学習するが，提案手法では多言語の大規模な生コーパスを用いて事前学習するため，少資源の言語対の評価にも対応することができる．

[§§] https://github.com/rohitguptacs/ReVal

2.3.1 機械翻訳品質推定のための教師なし手法

LASIM [10] は，複数言語対の対訳コーパス上で事前学習することにより得られる文の分散表現である LASER[¶¶]を用いた教師なしの手法である．LASIM は，翻訳文と原文をそれぞれ LASER により文の分散表現へ符号化し，それらのコサイン類似度により翻訳文と原文の類似度を計算する．LASIM は，WMT の QE as a Metric Task においてベースラインのひとつとして利用されている．

2.3.2 機械翻訳品質推定のための教師あり手法

Predictor-Estimator [19] は，対訳コーパス上で目的言語文の各単語を原言語文と目的言語文の文脈から推定するように事前学習された Predictor と，Predictor により得られる素性から人手評価値を推定する Estimator から構成される教師ありの手法である．Predictor-Estimator は，翻訳文がプロの翻訳者の修正をどの程度必要とするかについての評価値である Human Translation Error Rate（HTER）が付与されたデータセットを教師データとする手法であり，WMT-2017 QE Shared Task [1] において最高性能を示している．

[¶¶] https://github.com/facebookresearch/LASER

第 3 章　事前学習された文の分散表現に基づく機械翻訳の自動評価および品質推定

従来手法に多く見られる文字や単語の N-gram 素性に基づく機械翻訳自動評価手法には，文全体の大域的な情報を考慮できないため，参照文と表層的には異なるが意味的には似ている翻訳文に対して正確な評価ができないという問題がある．一方で，2.2.3 節で説明した ReVal は文の分散表現を用いて大域的な情報を考慮するが，WMT Metrics Shared Task のデータセットなどの小規模なラベル付きコーパスのみを用いてモデル全体を学習するため，文単位での十分な表現学習ができていない．そこで本研究では，大域的な情報を考慮する際の少資源問題を解決するために，事前学習された文の分散表現に基づく機械翻訳自動評価手法を提案する．

我々の提案手法は，RUSE と BERT による機械翻訳自動評価および BERT による機械翻訳品質推定の 3 つである．まず 3.1 節では，文の分散表現を利用する機械翻訳自動評価のための回帰モデルである RUSE について説明する．次に 3.2 節では，文対を同時に符号化する BERT による機械翻訳自動評価について説明する．最後に 3.3 節では，多言語 BERT による機械翻訳品質推定について説明する．

3.1　RUSE: 文の分散表現に基づく機械翻訳自動評価のための回帰モデル

本節では，事前学習された文の分散表現を素性とする回帰モデル RUSE（Regressor Using Sentence Embeddings）について説明する．まず 3.1.1 節では，RUSE で使用する 3 種類の文の分散表現について説明する．続いて 3.1.2 節では，機械翻訳自動評価のための回帰モデルおよび素性抽出について述べる．

3.1.1　事前学習された文の分散表現

大規模なコーパスを用いて事前学習された文の分散表現は，文書分類や文対の意味的類似度推定など多くの応用タスク [7] において高い性能を発揮している．本研究では，教師あり学習に基づく InferSent [8]，教師なし学習に基づく Quick

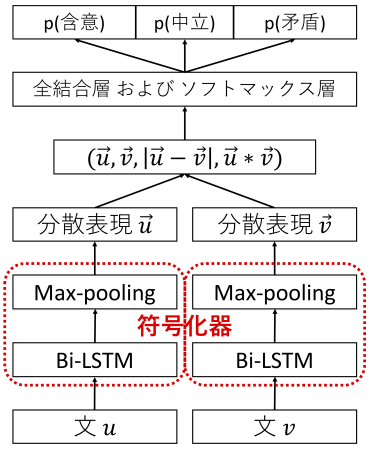


図 3.1　InferSent の概要図

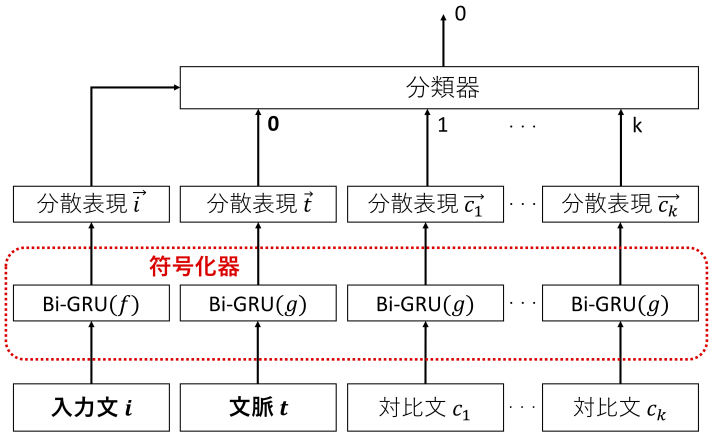


図 3.2　Quick Thought の概要図

Thought [22] およびマルチタスク学習に基づく Universal Sentence Encoder [6] の 3 手法を用いて文全体の大域的な情報を考慮する．

InferSent*は，含意関係認識のための Stanford Natural Language Inference（SNLI）データセット [5] 上で Max-pooling を用いた双方向 LSTM ネットワークを学習する教師あり学習に基づく手法である．図 3.1 に示すように，文 u および v をそれぞれ符号化し，それらの分散表現 $\vec{u}$ および $\vec{v}$ から素性を抽出し，含意関係認識の 3 値分類を通して文の符号化器を学習する．含意関係認識とは，所与の文対の関係を含意／矛盾／中立に 3 値分類するタスクであり，意味の違いに敏感な文の分散表現が得られることが期待できる．

Quick Thought†は，大規模な生コーパス上で双方向 GRU ネットワークを用いて隣接文推定することにより，教師なしで文の表現学習を行う手法である．図 3.2 に示すように，文 i，その文脈 t，その他の文（対比文）$c_1, c_2, ..., c_k$ が与えられ，2 種類の文の符号化器 f および g がそれぞれ文を符号化する．そして，入力文の分散表現 $\vec{i}$ との最大の内積値を持つ分散表現に対応する文を隣接文として推定する分類器を用いて，隣接文推定の学習を行う．応用タスクでは，所与の文を 2 つの符号化器 f および g を用いてそれぞれ符号化し，各符号化器から得られる分散表現を連結することによって文の分散表現を獲得する．隣接文推定タスクを通して文の符号化器を学習することによって，文対の関係を考慮した分散表現が得られることが期待

*https://github.com/facebookresearch/InferSent
†https://github.com/lajanugen/S2V

できる．

Universal Sentence Encoder[‡]は，復号器を用いる Skip-Thought [20] のような隣接文推定，発話応答推定および含意関係認識の 3 タスクを用いて自己注意機構に基づくネットワーク [34] をマルチタスク学習する手法である．Universal Sentence Encoder では隣接文推定や発話応答推定のための学習データとして，Wikipedia，ニュース，QA サイト，議論サイトなどの多様な Web ソースを用いる．多様なドメインのコーパスに基づくマルチタスク学習によって，幅広い応用タスクにおいて有用な文の分散表現が得られることが期待できる．

3.1.2 機械翻訳自動評価のための回帰モデルと素性抽出

機械翻訳の自動評価は，翻訳文と参照文から翻訳文の人手評価値を推定する回帰タスクとして考えることができる．そこで RUSE（図 1.2(a)）は，所与の翻訳文 t と参照文 r から 3.1.1 節の符号化器を用いて分散表現 $\vec{t}$ および $\vec{r}$ を獲得し，InferSent [8] にならって以下の 3 つの方法で翻訳文と参照文の関係を抽出し，それら 3 つを連結したものを素性として多層パーセプトロン（MLP）に基づく回帰モデルを学習する．

- 連結：$(\vec{t}, \vec{r})$
- 要素積：$\vec{t} * \vec{r}$
- 要素差：$|\vec{t} - \vec{r}|$

回帰モデルには，これらの 3 種類の素性を連結した $4d$ 次元の素性が入力される．ただし，d は分散表現 $\vec{t}$ および $\vec{r}$ の次元数である．RUSE では回帰モデルのみを学習し，文の符号化器の再学習は行わない．

3.2 BERT による機械翻訳自動評価

文および文対単位の表現学習モデルである BERT (Bidirectional Encoder Representations from Transformers) [9] が，文対の意味的類似度推定など多くのタス

[‡]https://www.tensorflow.org/hub/modules/google/universal-sentence-encoder-large/2

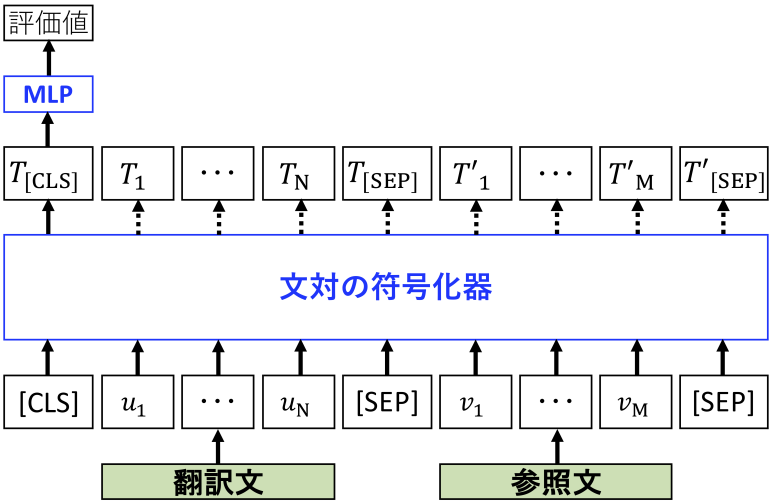


図 3.3　BERT の文対モデリング（u, v：入力トークン，T, T'：各入力トークンに対する分散表現）

クで最高性能を更新し，注目を集めている．本節では，BERT を用いて機械翻訳の自動評価を行う．BERT による機械翻訳の自動評価は RUSE と同じく，事前学習された文の分散表現を利用し，MLP によって人手評価値を推定する．ただし，図 1.2(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．以下では，RUSE との主な相違点であり 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，符号化器の再学習について詳細に説明する．

3.2.1　BERT における事前学習

BERT は，大規模な生コーパス上で双方向の自己注意機構に基づくネットワーク [34] 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．この教師なしの事前学習によって，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を 2 値分類する．この教師なしの事前学習によって，BERT の符号化器は文対の関係を学習する．

3.2.2 BERT における文対モデリング

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 3.3）．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.§

3.2.3 BERT における符号化器の再学習

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 を学習する際，文または文対の分散表現を得るための符号化器も再学習する．

3.3 多言語 BERT による機械翻訳品質推定

本節では，多言語 BERT¶を用いて機械翻訳の品質推定を行う．まず，多言語のそれぞれで大規模な生コーパスを用意し，共通のモデルで BERT の事前学習を行う．そして，原文・翻訳文・翻訳品質スコアの 3 つ組を用いて，原文および翻訳文の文対から翻訳品質を推定する回帰モデルを学習する（図 1.2(c)）．このとき，多言語 BERT の文対符号化器も同時に再学習する．

機械翻訳の品質推定タスクのために，3.2 節の BERT による機械翻訳自動評価

§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] に対応する最終の隠れ層が文の分散表現を表す.

¶https://github.com/google-research/bert/blob/master/multilingual.md

（図 3.3）から以下の 3 点を変更する．

- 多言語の大規模な生コーパス上で事前学習された多言語 BERT を用いる．
- 翻訳文と参照文の文対ではなく，原文と翻訳文の文対を用いて翻訳品質を推定する．
- 再学習の際には，対象言語対だけでなく利用可能な全言語対の人手評価値付きデータを用いる．

多言語 BERT では，多言語のコーパス全体でサブワードに基づく共通の語彙を構築する．共通の語彙と共通のモデルを用いて多言語のコーパス上で BERT の事前学習を行うため，多言語の情報を同一のベクトル空間上で符号化できる．これによって，品質推定タスクの再学習において，対象言語対以外の言語対のデータも対象言語対の性能改善に貢献することが期待できる．

第 4 章　評価実験

本章では，まず 3.1 節および 3.2 節で述べた提案手法についての機械翻訳自動評価における評価実験を行い，続いて 3.3 節で述べた提案手法についての機械翻訳品質推定における評価実験を行う．

4.1　機械翻訳自動評価についての評価実験

本節では，WMT Metrics Shared Task における人手の絶対評価値付きデータセットを用いて，文単位の to-English 言語対における機械翻訳自動評価についての提案手法の有効性を検証する．

4.1.1　実験設定

表 4.1 に，2.1 節の手順により作成された人手の絶対評価値付きデータセットの言語対*ごとの文対数を示す．これらのデータセットにおける人手の絶対評価値は，約 −1.95〜約 1.65 の実数値で示されている．本実験では，WMT-2015 [31] および WMT-2016 [4] の to-English 言語対の合計 5,360 文対を無作為に分割し，9 割を学習用，1 割を開発用に利用する．また，WMT-2017 [3] の文対は評価用に利用する．

RUSE の素性には，それぞれ著者らによって公開されている学習済みの InferSent，Quick Thought および Universal Sentence Encoder により得た文の分散表現を用いる．BERT には，著者らによって公開されている学習済みモデルのうち，$\mathrm{BERT}_{\mathrm{BASE}}$ (uncased)†を用いる．

各自動評価手法のメタ評価のために，人手の絶対評価値とのピアソンの積率相関係数，スピアマンの順位相関係数および平均 2 乗誤差を用いる．ピアソンの積率相関係数は，WMT Metrics Shared Task で用いられており，各手法が出力する評価値の絶対的なメタ評価ができる指標である．しかし，ピアソンの積率相関係数は外

*en : 英語，cs : チェコ語，de : ドイツ語，fi : フィンランド語，lv : ラトビア語，ro : ルーマニア語，ru : ロシア語，tr : トルコ語，zh : 中国語

†https://github.com/google-research/bert

表 4.1　WMT Metrics Shared Task の各言語対における人手の絶対評価値付き文対数

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

れ値が存在した場合に不当な値を示すという問題が存在するため，本実験ではスピアマンの順位相関係数によるメタ評価も行う．また本研究では，機械翻訳の自動評価を回帰問題として扱っているため，各自動評価手法がどれほど人手の評価値に近い値を出力しているかについても評価したい．そのため，本タスクを回帰問題として扱っている Blend，RUSE および BERT については，人手の評価値と各手法の評価値の平均 2 乗誤差によるメタ評価も行う．

4.1.2　比較手法

本実験では，WMT-2017 Metrics Shared Task におけるベースラインである SentBLEU および上位 3 手法を提案手法と比較する．比較手法のメタ評価には，WMT-2017 Metrics Shared Task[‡]で公開されている各手法の評価値を利用した．

提案手法については，事前学習された文の分散表現による貢献を明らかにするため，RUSE の素性として単語分散表現の平均ベクトルを用いた実験も行う．RUSE と BERT による機械翻訳自動評価を比較するため，最終的に以下の 7 つの設定で実験した．

■RUSE with GloVe-BoW:　図 1.2(a) の文の分散表現として，単語分散表現 GloVe [28] (glove.840B.300d[§]) の平均ベクトルを用いる．この 300 次元のベクトルを文の分散表現として，3.1.2 節の方法で素性を抽出する．

■RUSE with IS:　SNLI データセット [5] の 56 万文および MultiNLI データセット [36] の約 43 万文の両方を用いて事前学習された InferSent によって 4,096 次元

[‡]http://www.statmt.org/wmt17/results.html
[§]https://nlp.stanford.edu/projects/glove

の文の分散表現を獲得し，3.1.2 節の方法で素性を抽出する．

■RUSE with QT:　BookCorpus データセット [39] の 4,500 万文および UMBC WebBase [17] の約 1 億 3,000 万文の両方を用いて事前学習された Quick Thought によって 4,800 次元の文の分散表現を獲得し，3.1.2 節の方法で素性を抽出する．

■RUSE with USE:　Wikipedia，ニュース，QA サイト，議論サイトなどの多様な Web ソースを用いて事前学習された Universal Sentence Encoder によって 512 次元の文の分散表現を獲得し，3.1.2 節の方法で素性を抽出する．

■RUSE with BERT:　単一文を入力とする BERT の [CLS] トークンに対応する隠れ層のうち，最終 4 層を連結したものを 3,072 次元の文の分散表現として 3.1.2 節の方法で素性を抽出する．ただし，BERT の符号化器の部分は再学習しない．

■BERT (w/o fine-tuning):　文対を入力とする BERT の [CLS] トークンに対応する隠れ層のうち最終 4 層を連結したもの（3,072 次元）を，図 1.2(b) の MLP の入力として用いる．ただし，BERT の符号化器の部分は再学習しない．

■BERT:　文対を入力とする BERT の [CLS] トークンに対応する最終隠れ層（768 次元）を図 1.2(b) の MLP の入力として用い，MLP とともに BERT の符号化器の部分も再学習する．

RUSE と BERT (w/o fine-tuning) の各パラメータは，以下の組み合わせの中からグリッドサーチにより，開発データにおける平均 2 乗誤差が最小のモデルを選択する．なお，全ての層において活性化関数は ReLU を使用する．

- 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 $\in \{0.1, 0.3, 0.5\}$
- MLP の隠れ層の数 $\in \{1, 2, 3\}$
- MLP の隠れ層の次元 $\in \{512, 1024, 2048, 4096\}$

BERT の各パラメータは，著者らによって提唱されている組み合わせの中からグリッドサーチにより，開発データにおける平均 2 乗誤差が最小のモデルを選択する．

表 4.2　WMT-2017 Metrics Shared Task（to-English 言語対）におけるピアソンの積率相関係数

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
SentBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
chrF++	0.523	0.534	0.678	0.520	0.588	0.614	0.593	0.579
MEANT 2.0	0.578	0.565	0.687	0.586	0.607	0.596	0.639	0.608
Blend	0.594	0.571	0.733	0.577	0.622	0.671	0.661	0.633
RUSE with GloVe-BoW	0.475	0.479	0.645	0.532	0.537	0.547	0.480	0.527
RUSE with IS	0.556	0.568	0.706	0.650	0.626	0.649	0.634	0.627
RUSE with QT	0.599	0.588	0.736	0.690	0.655	0.710	0.645	0.660
RUSE with USE	0.592	0.596	0.681	0.621	0.598	0.645	0.620	0.622
RUSE with BERT	0.622	0.626	0.765	0.708	0.609	0.706	0.647	0.669
BERT (w/o fine-tuning)	0.645	0.607	0.780	0.727	0.644	0.704	0.705	0.687
BERT	**0.720**	**0.761**	**0.857**	**0.828**	**0.788**	**0.798**	**0.763**	**0.788**

表 4.3　WMT-2017 Metrics Shared Task（to-English 言語対）におけるスピアマンの順位相関係数

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
SentBLEU	0.429	0.424	0.555	0.362	0.495	0.488	0.532	0.469
chrF++	0.495	0.518	0.655	0.474	0.579	0.593	0.570	0.555
MEANT 2.0	0.561	0.550	0.685	0.549	0.601	0.582	0.616	0.592
Blend	0.578	0.564	0.713	0.547	0.609	0.644	0.638	0.613
RUSE with GloVe-BoW	0.468	0.474	0.641	0.504	0.513	0.530	0.482	0.516
RUSE with IS	0.525	0.551	0.699	0.627	0.621	0.624	0.605	0.607
RUSE with QT	0.600	0.593	0.734	0.690	0.673	0.693	0.627	0.659
RUSE with USE	0.591	0.588	0.681	0.603	0.585	0.621	0.595	0.609
RUSE with BERT	0.637	0.622	0.759	0.701	0.609	0.692	0.644	0.666
BERT (w/o fine-tuning)	0.645	0.619	0.791	0.731	0.650	0.706	0.697	0.691
BERT	**0.733**	**0.760**	**0.854**	**0.824**	**0.777**	**0.793**	**0.755**	**0.785**

4.1.3　実験結果

表 4.2，表 4.3 および表 4.4 に WMT-2017 Metrics Shared Task における実験結果を示す．表 4.2 および表 4.3 より，BERT が全ての to-English 言語対におい

表 4.4　WMT-2017 Metrics Shared Task（to-English 言語対）における平均 2 乗誤差

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
SentBLEU	-	-	-	-	-	-	-	-
chrF++	-	-	-	-	-	-	-	-
MEANT 2.0	-	-	-	-	-	-	-	-
Blend	0.242	0.219	0.184	0.291	0.216	0.206	0.194	0.222
RUSE with GloVe-BoW	0.317	0.251	0.231	0.284	0.241	0.247	0.257	0.261
RUSE with IS	0.259	0.222	0.192	0.229	0.213	0.198	0.195	0.215
RUSE with QT	0.240	0.213	0.179	0.208	0.198	0.170	0.193	0.200
RUSE with USE	0.246	0.212	0.213	0.237	0.217	0.200	0.207	0.219
RUSE with BERT	0.229	0.200	0.174	0.187	0.214	0.174	0.200	0.197
BERT (w/o fine-tuning)	0.225	0.220	0.154	0.176	0.209	0.168	**0.174**	0.189
BERT	**0.222**	**0.194**	**0.105**	**0.117**	**0.194**	**0.123**	0.190	**0.164**

て人手評価との最高の相関を示す．同様に，表 4.4 より，BERT が zh-en 以外の言語対で最小の誤差を示す．これらの結果は，文対を同時に符号化する表現学習モデルである BERT が機械翻訳自動評価タスクにおいても有効であることを示す．

各表の下段を比較すると，事前学習された文の分散表現を素性として用いた全ての RUSE モデルが，単語分散表現の平均ベクトルを素性として用いた RUSE with GloVe-BoW よりも高い相関および小さな誤差を示していることがわかる．これらの結果は，文全体の大域的な情報を考慮できる文の分散表現に基づく素性が，機械翻訳自動評価にとって有用であることを意味する．

また，Quick Thought や BERT の素性を用いる RUSE with QT および RUSE with BERT が，単一の符号化器に基づく提案手法の中でも特に高い性能を示した．このことから，隣接文推定の教師なし学習によって得られる文の分散表現が，機械翻訳の評価において特に有効であると考えられる．Universal Sentence Encoder もマルチタスク学習の一部として隣接文推定を行っているが，これは Quick Thought や BERT における隣接文推定とは設定が異なる．Quick Thought や BERT における隣接文推定では，符号化器と単純な分類器を用いて文対が隣接するか否かを分類する．一方で Universal Sentence Encoder における隣接文推定では，符号化器と復号器を用いて入力文から隣接文を生成する．そのため，前者はタスクを解く

ための情報を符号化器が獲得するが，後者は符号化器と復号器の両方にタスクを解くための情報が散在すると考えられる．この違いのために，Quick Thought や BERT が有用性の高い文の分散表現を獲得できた．

さらに，隣接文推定のみによって事前学習された RUSE with 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 RUSE with BERT の方が，多くの言語対において高い性能を示していることがわかる．このことから，BERT 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．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効であると考えられる．

RUSE with BERT と BERT (w/o fine-tuning) を比較すると，BERT の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．RUSE では，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．

BERT (w/o fine-tuning) と BERT を比較すると，符号化器の再学習による性能への影響がわかる．事前学習された文対の符号化器から素性抽出を行い MLP のみを学習する BERT (w/o fine-tuning) よりも，文対の分散表現を素性とし MLP とともに符号化器を再学習する BERT の方が，全ての言語対において大幅に高い相関を示し，zh-en 以外の言語対で最小の誤差を示す．つまり，BERT の大きな特徴のひとつである符号化器の再学習は，機械翻訳の自動評価のためにも有効である．

4.2 機械翻訳品質推定についての評価実験

本節では，WMT Metrics Shared Task における人手の絶対評価値付きデータセットを用いて，文単位の機械翻訳品質推定についての提案手法の有効性を検証する．

4.2.1 実験設定

本節においても，4.1.1 節と同様の人手による絶対評価値付きデータセットを用いるが，本実験では機械翻訳品質推定（参照文を利用しない機械翻訳自動評価）についての評価実験を行うため，参照文を除いた，原文，翻訳文および人手評価値のみを用いる．表 4.1 における全言語対において，WMT-2015 および WMT-2016 の合計 6,420 文対を無作為に分割し，9 割を学習用，1 割を開発用に利用する．WMT-2017 の各言語対は評価用に利用する．

多言語 BERT には，著者らによって公開されている学習済みモデルのうち，$\text{BERT}_{\text{multi}}$ (Cased)[¶]を用いる．BERT の各パラメータは，著者らによって提唱されている組み合わせの中からグリッドサーチにより，開発データにおける平均 2 乗誤差が最小のモデルを選択するが，最大エポック数のみ 20 に変更した．

本実験においても 4.1 節と同様の理由から，各自動評価手法のメタ評価のために，人手の絶対評価値とのピアソンの積率相関係数，スピアマンの順位相関係数および平均 2 乗誤差を用いる．平均 2 乗誤差によるメタ評価については，本タスクを回帰問題として扱っている Predictor-Estimator および $\text{BERT}_{\text{multi}}$ についてのみ行う．

4.2.2 比較手法

本実験では，参照文を用いない機械翻訳自動評価手法として，WMT-2017 QE Shared Task で最高性能を達成した Predictor-Estimator [19] および WMT の QE as a Metric Task のベースライン手法のひとつである LASIM [10] と提案手法の結果を比較した．

Predictor-Estimator は，翻訳文がプロの翻訳者の修正をどの程度必要とするかについての評価値である HTER が付与されたデータセットを教師データとして用いる手法であるが，本研究では翻訳文の意味的な品質についての絶対評価に焦点を当てているため，人手による絶対評価値付きのデータセットを Predictor-Estimator の教師データとして用いて比較する．Predictor-Estimator における実験には，

[¶]https://github.com/google-research/bert/blob/master/multilingual.md

Kepker らによる再実装である OpenKiwi[‖] [18] を利用する．Predictor の事前学習には，WMT-2017 Translation Task[**]における各言語対[††]の News Commentary v12 対訳コーパスを利用し，Estimator の学習には，WMT-2015 および WMT-2016 の各言語対ごとの合計 1,060 文を無作為に分割し，9 割を学習用，1 割を開発用に利用する．Predictor の各パラメータは，エポック数および学習率のみ初期の設定値から以下に変更し，グリッドサーチにより開発データにおけるパープレキシティが最小のモデルを選択した．

- エポック数 $\in \{1, ..., 15\}$
- 学習率（Adam）$\in \{2\text{e-}3, 1\text{e-}3\}$

Estimator の各パラメータは，エポック数および学習率のみ初期の設定値から以下に変更し，グリッドサーチにより開発データにおける平均 2 乗誤差が最小のモデルを選択した．

- エポック数 $\in \{1, ..., 30\}$
- 学習率（Adam）$\in \{2\text{e-}3, 1\text{e-}3, 5\text{e-}4\}$

LASIM における実験には，公開されている事前学習済みの LASER[‡‡]（bilstm.93langs.2018-12-26.pt）を利用した．

また，提案手法が参照文を利用する機械翻訳自動評価手法と比較してどの程度の性能を示せているかを示すため，以下の手法との比較も行った．WMT Metrics Shared Task のベースラインである SentBLEU [3] および chrF+[§§] [30] と WMT-2017 Metrics Shared Task において高い性能を示している Blend [24] および 4.1 節における BERT による機械翻訳自動評価である．SentBLEU，Blend における各言語対の評価値にはメタ評価には，4.1 節と同様に WMT-2017 Metrics Shared Task で公開されている各手法の評価値を利用した．chrF+ における実験では，著者等によって公開されている実装を用いて WMT-2019 Metrics Shared Task [25]

[‖] https://github.com/Unbabel/OpenKiwi

[**] http://www.statmt.org/wmt17/translation-task.html

[††] News Commentary v12 対訳コーパスが存在する，cs-en，de-en，ru-en および en-ru 言語対

[‡‡] https://github.com/facebookresearch/LASER

[§§] https://github.com/m-popovic/chrF

表 4.5　WMT-2017 Metrics Shared Task における各評価手法の人手評価とのピアソンの積率相関係数

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	en-ru	avg.
参照文なしの品質推定：									
Predictor-Estimator	0.337	0.163	-	-	0.272	-	-	0.441	0.303
LASIM	0.327	0.403	0.415	0.465	0.364	0.423	**0.467**	0.352	0.454
$\text{BERT}_{\text{multi}}$	**0.548**	**0.506**	**0.695**	**0.693**	**0.592**	**0.643**	0.460	**0.648**	**0.598**
参照文に基づく自動評価：									
SentBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.468	0.479
chrF+	0.523	0.531	0.677	0.529	0.592	0.609	0.595	0.612	0.584
Blend	0.594	0.571	0.733	0.577	0.622	0.671	0.661	0.578	0.626
$\text{BERT}_{\text{BASE}}$	0.720	0.761	0.857	0.828	0.788	0.798	0.763	0.741	0.782

表 4.6　WMT-2017 Metrics Shared Task における各評価手法の人手評価とのスピアマンの順位相関係数

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	en-ru	avg.
参照文なしの品質推定：									
Predictor-Estimator	0.327	0.176	-	-	0.286	-	-	0.451	0.310
LASIM	0.361	0.404	0.463	0.464	0.351	0.451	**0.482**	0.313	0.411
$\text{BERT}_{\text{multi}}$	**0.551**	**0.527**	**0.699**	**0.682**	**0.579**	**0.656**	0.457	**0.660**	**0.601**
参照文に基づく自動評価：									
SentBLEU	0.429	0.424	0.555	0.362	0.495	0.488	0.532	0.487	0.472
chrF+	0.493	0.513	0.656	0.484	0.581	0.592	0.571	0.604	0.562
Blend	0.578	0.564	0.713	0.547	0.609	0.644	0.638	0.563	0.607
$\text{BERT}_{\text{BASE}}$	0.733	0.760	0.854	0.824	0.777	0.793	0.755	0.743	0.780

と同様の設定を利用した．

4.2.3　実験結果

表 4.5，表 4.6 および表 4.7 に WMT-2017 Metrics Shared Task における実験結果を示す．表 4.5 および表 4.6 の上段の品質推定において，提案手法の $\text{BERT}_{\text{multi}}$ は zh-en 以外の言語対で比較手法の Predictor-Estimator および LASIM よりも人手評価との高い相関を示した．同様に，表 4.7 の上段の品質推定において，提案手法

表 4.7　WMT-2017 Metrics Shared Task における各評価手法の人手評価との平均 2 乗誤差

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	en-ru	avg.
参照文なしの品質推定：									
Predictor-Estimator	0.344	0.350	-	-	0.342	-	-	0.343	0.345
LASIM	-	-	-	-	-	-	-	-	-
$\text{BERT}_{\text{multi}}$	**0.292**	**0.270**	0.197	0.216	**0.241**	0.231	0.322	**0.267**	**0.255**
参照文に基づく自動評価：									
SentBLEU	-	-	-	-	-	-	-	-	-
chrF+	-	-	-	-	-	-	-	-	-
Blend	0.242	0.219	0.184	0.291	0.216	0.206	0.194	0.277	0.229
$\text{BERT}_{\text{BASE}}$	0.222	0.194	0.105	0.117	0.194	0.123	0.190	0.208	0.169

が全言語対で最小の誤差を示す．また表 4.5 および表 4.6 において，下段の参照文に基づく自動評価手法と比較すると，提案手法は多くの言語対において SentBLEU および chrF+ と同等以上の性能を達成した．これらの結果から，事前学習された多言語の文対符号化器が機械翻訳の品質推定のために有用なことがわかる．

なお，本実験において lv-en および zh-en の言語対には学習用データが存在しないが，lv-en の言語対では LASIM よりも $\text{BERT}_{\text{multi}}$ が高い性能を示す一方で，zh-en の言語対では $\text{BERT}_{\text{multi}}$ よりも LASIM が高い性能を示している．本実験で使用した多言語 BERT はサブワードに基づく語彙を言語間で共有しているが，漢字に基づく中国語よりもラテン文字に基づくラトビア語の方が学習データに含まれる他の言語との共通のサブワードを多く含むことが，この違いの要因のひとつであると考えられる．

第 5 章　分析

本章では，まず 3.1 節および 3.2 節で述べた提案手法についての機械翻訳自動評価における追加実験を行い，分析する．続いて，3.3 節で述べた提案手法についての機械翻訳品質推定における追加実験を行い，分析する．

5.1　機械翻訳の自動評価についての分析

本節では，3.1 節および 3.2 節で述べた提案手法について，機械翻訳自動評価における学習データの文対数と性能の関係，比較的学習データが少ない言語対である from-English 言語対における性能および提案手法の出力例について分析を行う．

5.1.1　学習データの文対数と性能の関係

本節では，WMT-2017 の lv-en 言語対の 560 文対を評価用データとして，RUSE と BERT について学習データの文対数と性能の関係を分析する．WMT-2015，WMT-2016 および WMT-2017 の to-English の lv-en 言語対以外の合計 8,720 文対を無作為に分割し，8,160 文対を学習用，560 文対を開発用に利用する．そして，学習用データを 510 文対，1,020 文対，2,040 文対，4,080 文対，8,160 文対の 5 つの大きさでそれぞれ無作為抽出し，開発用データと評価用データにおける人手評価値とのピアソンの積率相関係数，スピアマンの順位相関係数および平均 2 乗誤差を評価する．

RUSE の素性には 4.1.1 節で述べた Quick Thought を用いる．RUSE の各パラメータは 4.1.1 節の組み合わせの中からグリッドサーチにより，開発データにおける平均 2 乗誤差が最小のモデルを選択するが，バッチサイズのみ以下に変更した．

- バッチサイズ $\in \{16, 32, 64, 128, 256, 512, 1024\}$

BERT の各パラメータは，著者らによって提唱されている組み合わせの中からグリッドサーチにより，開発データにおける平均 2 乗誤差が最小のモデルを選択するが，バッチサイズとエポック数のみ以下に変更した．

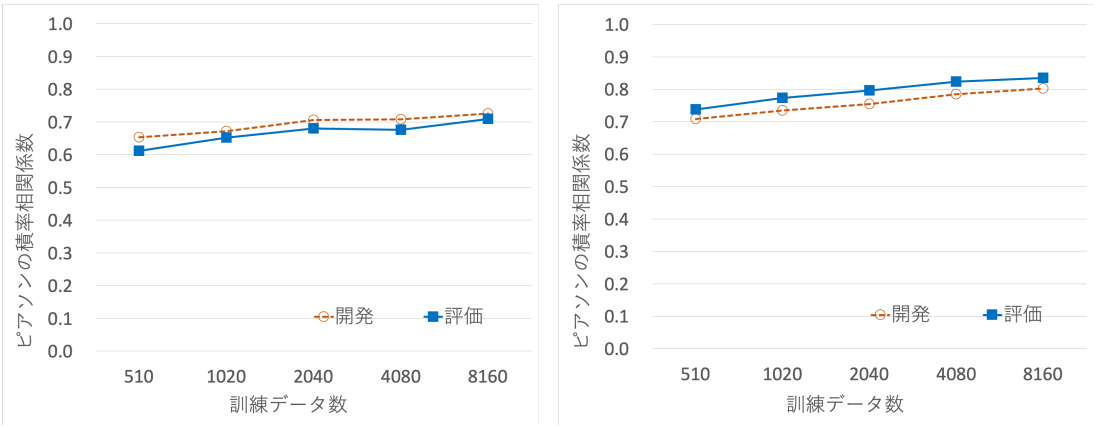


図 5.1　RUSE（左）と BERT（右）における学習曲線（人手評価とのピアソンの積率相関係数）

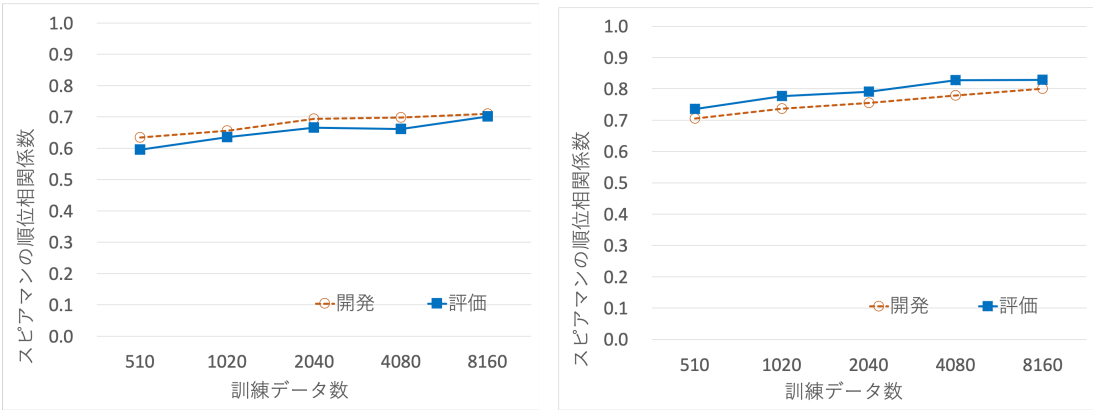


図 5.2　RUSE（左）と BERT（右）における学習曲線（人手評価とのスピアマンの順位相関係数）

- バッチサイズ $\in \{8, 16, 32\}$
- エポック数 $\in \{1, 2, ..., 6\}$

図 5.1，図 5.2 および図 5.3 に，RUSE と BERT のピアソンの積率相関係数，スピアマンの順位相関係数および平均二乗誤差に対する学習曲線をそれぞれ示す．RUSE の学習曲線および BERT の学習曲線において，学習データ数が 510 文対の場合と 8,160 文対の場合では評価時のピアソンの積率相関係数およびスピアマンの順位相関係数に約 0.1 の差があり，評価時の平均 2 乗誤差には約 0.05 以上の差があ

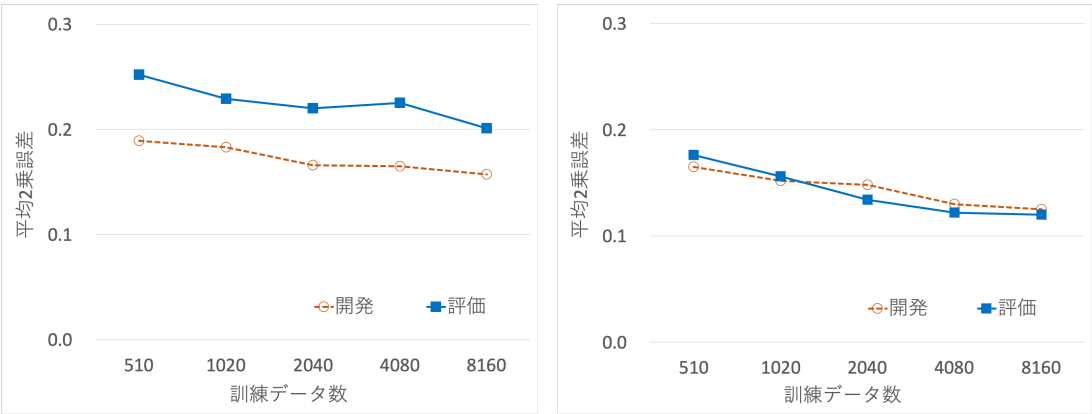


図 5.3　RUSE（左）と BERT（右）における学習曲線（人手評価との平均 2 乗誤差）

る．このことから，RUSE および BERT による機械翻訳自動評価の両手法において，学習データ数による性能の変化が大きいことがわかる．また，図 5.1 と図 5.2 および図 5.3 より，BERT による機械翻訳自動評価は学習データ数が 510 文対の場合でも，RUSE の 8,160 文対の学習データ数での性能を上回っていることがわかる．510 文対のデータで学習された BERT による機械翻訳自動評価の性能は，表 4.2 と表 4.3 および表 4.4 におけるいずれの比較手法よりも同等もしくは高い性能を示している．以上の分析から，BERT は少量のラベル付きコーパスを用いる学習でも高い性能を発揮することがわかり，学習データ数を増やすことが可能であれば，更に信頼性の高い機械翻訳自動評価手法になると考えられる．

5.1.2　from-English 言語対における性能

本節では，機械翻訳自動評価における to-English 言語対以外の設定として，from-English 言語対の中で最も多くのラベル付きコーパスが存在する en-ru 言語対における性能を調査する．WMT-2015 の 500 文対および WMT-2016 の 560 文対の合計 1,060 文対を無作為に分割し，9 割を学習用，1 割を開発用に利用する．評価用には，WMT-2017 の 560 文対を用いる．

RUSE の素性には，ロシア語の Wikipedia 上で事前学習した Quick Thought を用いる．RUSE の各パラメータは 4.1.1 節の組み合わせの中からグリッドサーチに

表 5.1　WMT-2017 Metrics Shared Task（en-ru 言語対）における人手による絶対評価とのピアソンの積率相関係数，スピアマンの順位相関係数および平均 2 乗誤差

	en-ru		
	ピアソン	スピアマン	平均 2 乗誤差
SentBLEU	0.468	0.487	-
chrF+	0.609	0.601	-
MEANT 2.0-nosrl	0.636	0.637	-
Blend	0.578	0.563	0.277
RUSE with QT	0.594	0.603	0.268
BERT	**0.741**	**0.743**	**0.208**

より，開発データにおける平均 2 乗誤差が最小のモデルを選択するが，バッチサイズと MLP の隠れ層の次元のみ以下に変更した．

- バッチサイズ $\in \{16, 32, 64, 128, 256\}$
- MLP の隠れ層の次元 $\in \{128, 256, 512, 1024, 2048, 4096\}$

BERT には，著者らによって公開されている多言語の大規模コーパス上で事前学習された $\text{BERT}_{\text{multi}}$ (Cased) を用いるを用いる．この学習済みモデルは，機械翻訳の品質推定における提案手法でも用いるが，参照文を用いた機械翻訳自動評価における本実験では，翻訳文と参照文の文対を入力しているため単言語の文しか入力されない．BERT の各パラメータは，5.1.1 節と同様に選択する．

比較手法として，WMT Metrics Shared Task のベースラインである SentBLEU，WMT-2017 Metrics Shared Task における上位 3 手法である，chrF+ [30]，MEANT 2.0-nosrl [21] および Blend [24] を用いる．4.1.2 節と同様に，各比較手法の公開されている評価値を用いて各比較手法のメタ評価を行った．人手評価値とのピアソンの積率相関係数，スピアマンの順位相関係数および平均 2 乗誤差によって各手法を評価する．

評価の結果を表 5.1 に示す．教師あり学習に基づく Blend と RUSE を比較すると，RUSE の方が高い性能を示しており，to-English 言語対以外の設定においても

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効な素性であると言える．しかし，RUSE は教師なし学習に基づく chrF+ および MEANT 2.0-nosrl の性能には及ばない．前節の分析から，RUSE は学習データ数による性能の変化が大きいことが確認されており，en-ru 言語対においても同様の理由で性能が低下していると考えられる．

一方で BERT は，他の手法よりも大幅に高い性能を示している．同じく前節の分析から，BERT は少量のデータでも高い性能を発揮することが確認されており，en-ru 言語対においても最高性能を達成した．このことから BERT は，少量のラベル付きコーパスが利用できれば，様々な言語対に対応した機械翻訳自動評価手法になると考えられる．

5.1.3 出力例

WMT-2017 Metrics Shared Task において文単位の to-English 言語対で最高性能を達成した Blend と，提案手法である RUSE および BERT による機械翻訳自動評価の出力を比較する．人手評価値と各手法の評価値を比較するために，5.1.1 節と同様に WMT-2017 における lv-en 言語対 560 文対に対する人手評価値および各手法の評価値を用いた．表 5.2 に，翻訳文と参照文に対する人手評価値および各手法の評価値を示す．

成功例 1 において，参照文と語彙や構文は異なるが意味が似ている（人手評価値が高い）翻訳文に対して，Blend では低い値をつけてしまっているのに対し，提案手法である RUSE や BERT による評価では Blend より高い値を示しており，正しい評価が行えている．また，成功例 2 において，参照文と語彙や構文が似ているが意味が異なる（人手評価値が低い）翻訳文に対して，Blend では高い値をつけてしまっているのに対し，RUSE や BERT による評価では低い値を示しており，正しい評価が行えている．このことから，RUSE や BERT は，局所的な素性に基づく手法である Blend では扱えない大域的な情報を考慮した評価ができていると考えられる．

失敗例には言い換えの問題があると考えられ，どの手法でも人手評価より低い値をつけてしまっている．Blend による評価では，言い換えによる表層の不一致の影

表 5.2　Blend，RUSE および BERT による機械翻訳自動評価の出力例

	成功例 1
参照文	'Needless to say, there was no disrespect intended and I'm very sorry.'
翻訳文	Needless to say I hadn't meant to show disrespect, and I was really sorry that it happened.
評価値	人手評価：**0.901**，Blend：**-0.186**，RUSE：**0.321**，BERT：**0.798**
	成功例 2
参照文	Vaino comes from a family of Soviet party elite.
翻訳文	Blame comes from the family in the Soviet party elite.
評価値	人手評価：**-0.634**，Blend：**0.298**，RUSE：**-0.366**，BERT：**-0.446**
	失敗例
参照文	Information providers will remain anonymous.
翻訳文	the information providers anonymity guaranteed.
評価値	人手評価：**0.418**，Blend：**0.0544**，RUSE：**-0.116**，BERT：**-0.219**

響により低い値がつけられていると考えられる．RUSE や BERT による評価では，文全体の言い換えの関係を上手く捉えられていないため，低い値をつけてしまっていると考えられる．

5.2　機械翻訳の品質推定についての分析

本節では，3.3 節で述べた提案手法の機械翻訳品質推定における学習データによる性能の変化について分析するため，以下の 2 つの設定で実験する．

5.2.1　対象言語対のみで学習

他言語のデータを利用する効果について検証するために，対象言語対のデータのみを用いて品質推定の再学習を行う．具体的には，WMT-2015 および WMT-2016 のデータセットの中から，cs-en・de-en・fi-en・ru-en の言語対においては 1,060 文対ずつ，tr-en の言語対においては 560 文対を使用して品質推定の再学習を行う．

表 5.3　WMT-2017 Metrics Shared Task における提案手法の学習データによる性能比較（ピアソンの積率相関係数）

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	en-ru	avg.
参照文なしの品質推定：									
Predictor-Estimator	0.337	0.163	-	-	0.272	-	-	0.441	0.303
LASIM	0.327	0.403	0.415	0.465	0.364	0.423	**0.467**	0.352	0.454
$\text{BERT}_{\text{multi}}$	**0.548**	**0.506**	0.695	**0.693**	**0.592**	**0.643**	0.460	**0.648**	**0.598**
$\text{BERT}_{\text{multi}}$（w/o 他言語）	0.474	0.442	0.638	-	0.424	0.533	-	0.599	0.518
$\text{BERT}_{\text{multi}}$（Zero-shot）	0.512	0.482	**0.697**	-	0.552	0.631	-	0.530	0.567
参照文に基づく自動評価：									
SentBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.468	0.479
chrF+	0.523	0.531	0.677	0.529	0.592	0.609	0.595	0.612	0.584

表 5.4　WMT-2017 Metrics Shared Task における提案手法の学習データによる性能比較（スピアマンの順位相関係数）

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	en-ru	avg.
参照文なしの品質推定：									
Predictor-Estimator	0.327	0.176	-	-	0.286	-	-	0.451	0.310
LASIM	0.361	0.404	0.463	0.464	0.351	0.451	**0.482**	0.313	0.411
$\text{BERT}_{\text{multi}}$	**0.551**	**0.527**	**0.699**	**0.682**	**0.579**	**0.656**	0.457	**0.660**	**0.601**
$\text{BERT}_{\text{multi}}$（w/o 他言語）	0.487	0.454	0.643	-	0.436	0.526	-	0.607	0.526
$\text{BERT}_{\text{multi}}$（Zero-shot）	0.520	0.484	**0.699**	-	0.556	0.634	-	0.549	0.574
参照文に基づく自動評価：									
SentBLEU	0.429	0.424	0.555	0.362	0.495	0.488	0.532	0.487	0.472
chrF+	0.493	0.513	0.656	0.484	0.581	0.592	0.571	0.604	0.562

なお，lv-en および zh-en の言語対は評価用データしか存在しないため本実験の対象外とする．再学習用のデータは，4.2 節と同じく，それぞれ 9 割を学習用，1 割を開発用に無作為分割して用いる．

表 5.3，表 5.4 および表 5.5 の実験結果より，他言語のデータも含めて品質推定の再学習を行う $\text{BERT}_{\text{multi}}$ が，対象言語対のデータのみで学習する $\text{BERT}_{\text{multi}}$（w/o 他言語）よりも常に高い性能を示した．この分析から，事前学習された多言語の文対符号化器を言語横断的に再学習することの有効性が確認できた．

表 5.5　WMT-2017 Metrics Shared Task における提案手法の学習データによる性能比較（平均 2 乗誤差）

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	en-ru	avg.
参照文なしの品質推定：									
Predictor-Estimator	0.344	0.350	-	-	0.342	-	-	0.343	0.345
LASIM	-	-	-	-	-	-	-	-	-
$\text{BERT}_{\text{multi}}$	**0.292**	**0.270**	0.197	0.216	**0.241**	0.231	0.322	**0.267**	**0.255**
$\text{BERT}_{\text{multi}}$（w/o 他言語）	0.324	0.297	0.225	-	0.295	0.248	-	0.286	0.279
$\text{BERT}_{\text{multi}}$（Zero-shot）	**0.292**	0.306	**0.194**	-	0.278	**0.206**	-	0.296	0.262
参照文に基づく自動評価：									
SentBLEU	-	-	-	-	-	-	-	-	-
chrF+	-	-	-	-	-	-	-	-	-

5.2.2　Zero-shot 学習

前節の分析から，対象言語対以外のデータを用いて品質推定の再学習を行うことの有効性が明らかになった．そのため，対象言語対のデータを用いない zero-shot 品質推定への期待が持てる．そこで本節では，WMT-2015 および WMT-2016 のデータセットの中から，対象言語対以外のデータのみを用いて，それぞれ 9 割を学習用，1 割を開発用に無作為分割して zero-shot 品質推定の実験を行う．なお，lv-en および zh-en の言語対はそもそも学習用データが存在しないため本実験の対象外とする．

表 5.3，表 5.4 および表 5.5 の実験結果より，対象言語対のデータも含めて品質推定の再学習を行う $\text{BERT}_{\text{multi}}$ には劣るものの，zero-shot 学習の $\text{BERT}_{\text{multi}}$（Zero-shot）が Predictor-Estimator および LASIM の比較手法よりも常に高い性能を示した．また，$\text{BERT}_{\text{multi}}$（Zero-shot）は参照文に基づく自動評価手法である SentBLEU と比較しても常に高い性能を示した．この分析から，事前学習された多言語の文対符号化器は，対象言語対のためのラベル付きデータが存在しない状況でも，他の言語対のラベル付きデータ上での再学習によって高性能な品質推定を実現できると言える．

第 6 章　おわりに

本研究では，信頼性の高い文単位での絶対的な自動評価を行うため，事前学習された文の分散表現に基づく機械翻訳自動評価手法を提案した．我々は，大規模な生コーパスを用いる隣接文推定や双方向言語モデルの教師なし事前学習によって，機械翻訳の自動評価のために有用な文の符号化器が得られることを示した．我々の提案手法は，局所的な素性に基づく従来手法では扱えない大域的な情報を考慮することができ，翻訳文と参照文の間の表層的な一致率にとらわれない正確な自動評価を可能にした．また，多言語の大規模な生コーパスを用いた事前学習によって得られる文の符号化器を用いることで，機械翻訳品質推定（参照文を利用しない機械翻訳自動評価）が可能であることを示した．

RUSE による機械翻訳自動評価では，WMT-2017 Metrics Shared Task のデータセットを用いた評価実験において，文単位の to-English 言語対で当時のどの従来手法よりも高い性能を示した．また，BERT による機械翻訳自動評価では，WMT-2017 Metrics Shared Task のデータセットを用いた評価実験において，文単位の全ての言語対で RUSE を凌ぎ，最高性能を更新した．多言語 BERT による機械翻訳品質推定では，WMT-2017 Metrics Shared Task のデータセットを用いた評価実験において，文単位の多くの言語対で他の手法を大幅に上回り最高性能を更新し，参照文を用いた機械翻訳自動評価におけるベースライン手法を上回る性能を示した．

詳細な分析の結果，事前学習された文対符号化器による機械翻訳の自動評価は，事前学習の方法，文対モデリング，符号化器の再学習の 3 点がそれぞれ性能改善に貢献しており，少量のラベル付きコーパスのみを用いても高い性能を発揮することを示した．また，事前学習された文対符号化器による機械翻訳の品質推定は，多言語の大規模コーパスにより事前学習するだけでなく，再学習の際も複数言語対のラベル付きデータを言語横断的に用いることでより高い性能を発揮することを示した．

本研究で，事前学習された文の分散表現の機械翻訳自動評価タスクおよび品質推定タスクへの転用が可能であることが示されたので，今後はこれらのタスクに特化したモデル構造や再学習方法について研究を行いたいと考えている．

謝辞

研究活動において丁寧な指導および研究環境の整備など，非常に多くのことでお世話になりました小町守准教授に深く感謝します．研究生活を通して，国内だけでなく海外での学会発表やメンターとしての後輩の指導など，非常に多くの貴重な経験をすることができました．また，学部 4 年生の頃からメンターとして指導してくださった梶原さんには，研究におけるアドバイスや論文の書き方など非常に多くのことについて丁寧に指導していただき，本当に感謝しています．研究室の先輩方や同期および後輩の皆さんには，様々な場面で相談に乗っていただいたり助けていただき，ありがとうございました．最後に，副査を引き受けていただいた山口亨教授と高間康史教授に感謝します．

参考文献

[1] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 169–214, September 2017.

[2] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pp. 131–198, August 2016.

[3] Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pp. 489–513, 2017.

[4] Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pp. 199–231, 2016.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015.

[6] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174, 2018.

[7] Alexis Conneau and Douwe Kiela. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 1669–1704, 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 *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pp. 670–680, 2017.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[10] Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation*, pp. 1–12, 2019.

[11] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. *The Prague Bulletin of Mathematical Linguistics*, pp. 77–86, 2010.

[12] Yvette Graham, Timothy Baldwin, and Nitika Mathur. Accurate Evaluation of Segment-level Machine Translation Metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1183–1191, 2015.

[13] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 33–41, 2013.

[14] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Is Machine Translation Getting Better over Time ? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics 2014*, pp. 443–451, 2014.

[15] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Can Machine Translation Systems be Evaluated by the Crowd Alone. *Natural Language Engineering*, Vol. 23, No. 1, pp. 3–30, 2017.

[16] Rohit Gupta, Constantin Orasan, and Josef van Genabith. Machine Translation Evaluation using Recurrent Neural Networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 380–384, 2015.

[17] Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Second Joint Conference on Lexical and Computational Semantics, Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pp. 44–52, 2013.

[18] Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An Open Source Framework for Quality Estimation. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations, pp. 117–122, 2019.

[19] Hyun Kim, Hun-Young Jung, HongSeok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Vol. 17, No. 1, pp. 1–22, 2017.

[20] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28*, pp. 3294–3302, 2015.

[21] Chi-Kiu Lo. MEANT 2.0: Accurate Semantic MT Evaluation for Any Output Language. In *Proceedings of the Second Conference on Machine Translation*, pp. 589–597, 2017.

[22] Lajanugen Logeswaran and Honglak Lee. An Efficient Framework for Learning Sentence Representations. In *Proceedings of the 6th International Conference on Learning Representations*, pp. 1–16, 2018.

[23] Qingsong Ma, Ondřej Bojar, and Yvette Graham. Results of the WMT18 Metrics Shared Task: Both Characters and Embeddings Achieve Good Performance. In *Proceedings of the Third Conference on Machine Translation*, pp. 682–701, 2018.

[24] 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 *Proceedings of the Second Conference on Machine Translation*, pp. 598–603, 2017.

[25] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation*, pp. 62–90, 2019.

[26] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 216–223, 2014.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[28] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[29] Maja Popović. chrF: Character N-gram F-score for Automatic MT Evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, 2015.

[30] Maja Popović. chrF++: Words Helping Character N-grams. In *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, 2017.

[31] Miloš Stanojević, Philipp Koehn, and Ondřej Bojar. Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 256–273, 2015.

[32] Miloš Stanojević and Khalil Sima'an. BEER 1.1: ILLC UvA Submission to Metrics and Tuning Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 396–401, 2015.

[33] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 1556–1566, 2015.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. 2017.

[35] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTER: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation*, pp. 505–510, 2016.

[36] Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 1112–1122, 2018.

[37] Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 417–421, 2015.

[38] Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. An Automatic Machine Translation Evaluation Metric Based on Dependency Parsing Model. *arXiv preprint arXiv:1508.01996*, 2015.

[39] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. *2015 IEEE International Conference on Computer Vision*, pp. 19–27, 2015.

発表リスト

論文誌

1. 嶋中宏希, 梶原智之, 小町守. **事前学習された文の分散表現を用いた機械翻訳の自動評価**. 自然言語処理, Vol. 26, No. 3, pp. 613-634. September, 2019.

国際会議

1. Hiroki Shimanaka, Tomoyuki Kajiwara, Mamoru Komachi. **Metric for Automatic Machine Translation Evaluation based on Universal Sentence Representations.** In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (NAACL-SRW 2018), pp.106-111. New Orleans, Louisiana, USA. Jun, 2018.
2. Hiroki Shimanaka, Tomoyuki Kajiwara, Mamoru Komachi. **RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation.** In Proceedings of the Third Conference on Machine Translation: Shared Task Papers (WMT 2018), pp.751-758. Belgium, Brussels. October, 2018.
3. Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi, Mamoru Komachi. **Filtering Pseudo-References by Paraphrasing for Automatic Evaluation of Machine Translation.** In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1) (WMT 2019), pp. 521-525. Florence, Italy. August, 2019.

国内会議

1. 嶋中宏希, 山岸駿秀, 松村雪桜, 小町守. **クロスリンガルな単語分散表現を用いた機械翻訳自動評価手法の検討**. NLP 若手の会第 12 回シンポジウム (YANS 2017). 那覇. September. 2017.
2. 嶋中宏希, 梶原智之, 小町守. **汎用的な文の分散表現を用いた文単位の機械翻訳自動評価**. 言語処理学会第 24 回年次大会 (NLP 2018). pp. 580-583. 岡山. March, 2018.
3. 嶋中宏希, 梶原智之, 小町守. RUSE: **文の分散表現を用いた回帰モデルによる機械翻訳の自動評価**. NLP 若手の会第 13 回シンポジウム (YANS 2018). 高松. August, 2018.
4. 嶋中宏希, 梶原智之, 小町守. BERT **を用いた機械翻訳の自動評価**. 言語処理学会第

25 回年次大会 (NLP 2019). 名古屋. pp. 590-593. March, 2019.

5. 中澤真人, 嶋中宏希, 黒澤道希, 小町守. **中日機械翻訳における事前学習された言語モデリングの利用に関する考察**. NLP 若手の会第 14 回シンポジウム (YANS 2019). 札幌. August, 2019.
6. 平尾礼央, 新井美桜, 嶋中宏希, 勝又智, 小町守. **ニューラルネットワークを利用した日本語学習者の複数項目作文能力推定**. NLP 若手の会第 14 回シンポジウム (YANS 2019). 札幌. August, 2019.
7. (発表予定). 嶋中宏希, 梶原智之, 小町守. **事前学習された多言語の文符号化器を用いた機械翻訳の品質推定**. 言語処理学会第 26 回年次大会 (NLP 2020). 水戸. March, 2020.
8. (発表予定). 平尾礼央, 新井美桜, 嶋中宏希, 勝又智, 小町守. **複数項目の採点を行う日本語学習者の作文自動評価システム**. 言語処理学会第 26 回年次大会 (NLP 2020). 水戸. March, 2020.

その他発表論文

1. Hiroki Shimanaka, Tomoyuki Kajiwara, Mamoru Komachi. **Machine Translation Evaluation with BERT Regressor.** In arXiv e-prints, 1907.12679. July, 2019.