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

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

本稿では，機械翻訳における文単位の絶対的な**品質推定**（参照文を利用しない自動評価）について述べる．BLEU などの参照文に基づく自動評価は，ベンチマーク上での翻訳器の性能改善に貢献してきたが，翻訳器が実際に利用される場面では参照文が存在しないため翻訳品質を評価できない．人手評価と高い相関を持つ品質推定手法の開発により，翻訳結果を信じるか，後編集や他の翻訳器を利用するかの判断が可能になる．

機械翻訳に関する国際会議 WMT では，想定される後編集の割合 HTER [10] を推定する Quality Estimation (QE) タスク [5] および翻訳文の妥当性を相対評価する QE as a Metric タスク [8] という Shared Task が開催され，多くの品質推定手法が提案されている．一方，本研究では翻訳文の妥当性を絶対評価する品質推定を目的として，WMT-2017 Metric Shared Task [2] のためのデータセットを用いて実験する．

WMT の QE タスクでは，大規模な対訳コーパスを用いて学習する手法 [1, 4, 7] が高い性能を達成している．しかし，大規模な対訳コーパスを利用できない中品質な翻訳器こそ翻訳品質を慎重に推定したいため，本研究では対訳コーパスを利用せずに品質推定器を学習する．まず，多言語のそれぞれで生コーパスを用意し，これらを用いて多言語の文符号化器を事前学習する．そして，任意の言語対における原文・翻訳文・翻訳品質ラベルの 3 つ組を用いて，図 1 に示すように翻訳品質を推定する回帰モデルを再学習する．

文符号化器として多言語 BERT[1] を用いた評価実験の結果，提案手法が多くの言語対において既存の品質推定手法の性能を上回ることを確認した．また，SentBLEU[2] などの参照文に基づく自動評価手法と比較しても，提案手法は参照文なしで人手評価との高い相関を達成した．分析の結果，品質推定のための学習データは，対象言語対だけでなく他言語のデータも言語横断的に利用することで性能を改善できることが明らかになった．また，対象言語対の学習データを利用しない zero-shot の設定でも良好な結果が得られ，多言語の文符号化器を用いる品質推定の有効性を確認できた．

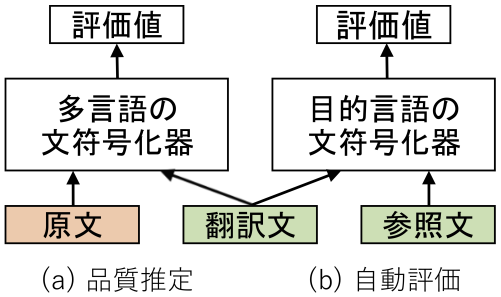

図 1: 文符号化器を用いた品質推定および自動評価

2 関連研究

2.1 機械翻訳の品質推定

Predictor-Estimator [7] は，大規模な対訳コーパス上で目的言語文の各単語を原言語文および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 によって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品質推定手法である．WMT-2017 の QE タスクにおいて，Predictor-Estimator は最高性能を示している．

LASER [1] は，複数言語の大規模な対訳コーパス上で事前学習された多言語の文符号化器である．WMT-2019 の QE as a Metric タスクにおいて，LASER はベースライン手法のひとつとして利用されている．LASER に基づく品質推定は，原文と翻訳文をそれぞれ文ベクトルに符号化し，それらの余弦類似度を品質スコアとして出力する教師なし品質推定手法である．

大規模な対訳コーパスを利用できない言語対における品質推定を実現するために，本研究では対訳コーパスなしで学習できる多言語 BERT を用いる．

[1] https://github.com/google-research/bert/blob/master/multilingual.md
[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl

表 1: WMT Metrics Shared Task の全言語対[3]における人手評価付き文対数

	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

2.2 機械翻訳の自動評価

参照文に基づく自動評価では，BERT [3] に基づく手法が成功を収めている．MoverScore [12] および BERTScore [11] は，BERT から得られる文脈化された単語分散表現を用いて翻訳文と参照文の類似度を計算する教師なし自動評価手法である．我々の先行研究 [13] では，BERT を WMT Metrics Shared Task のデータセット上で再学習することにより，参照文に基づく自動評価タスクにおいて最高性能を達成している．

本研究では先行研究 [13] に従いつつ，多言語 BERT を用いて原文と翻訳文を比較する品質推定に取り組む．

3 提案手法

本研究では，多言語の文符号化器を用いた機械翻訳の品質推定手法を提案する．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 [3] 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．このとき，BERT の文符号化器も同時に再学習する．我々の先行研究である BERT を用いた参照文に基づく自動評価手法 [13] に従う部分が多いが，品質推定タスクのために以下の 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[1] 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（図 1）．
- 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と期待できる．

4 評価実験

WMT-2017 Metrics Shared Task [2] のデータセットを用いて，多言語 BERT による文単位の品質推定の有効性を検証する．各手法の性能は，人手評価とのピアソンの相関係数を用いて評価する．

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 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 の文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)[1] 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 乗誤差が最も小さいモデルを選択するが，最大エポック数のみ 20 に変更した．

4.2 比較手法

本実験では，WMT-2017 QE タスクで最高性能を達成した Predictor-Estimator [7] および WMT-2019 QE as a Metric タスク [8] のベースライン手法である LASER [1] と提案手法を比較する．また，参考のために，参照文に基づく自動評価手法として，参照文との N-gram マッチングに基づく SentBLEU[2] および chrF++[4] [9]，BERT の文脈化された単語分散表現に基づく BERTScore [11] および MoverScore [12]，BERT の再学習に基づく自動評価手法 [13] の性能も示す．

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

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

表 2: 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
LASER	0.361	0.404	0.463	0.464	0.351	0.451	**0.482**	0.352	0.416
$\text{BERT}_{\text{multi}}$	**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 [2]	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
BERTScore [11]	0.657	0.680	0.823	0.712	0.725	0.718	0.711	0.657	0.710
MoverScore [12]	0.670	0.708	0.835	0.746	0.738	0.762	0.744	-	0.743
$\text{BERT}_{\text{BASE}}$ [13]	0.720	0.761	0.857	0.828	0.788	0.798	0.763	0.741	0.782

Predictor-Estimator の実装には，OpenKiwi[5] [6] を利用する．Predictor の事前学習には News Commentary v12[6]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 および Estimator の各パラメータは，エポック数と学習率のみ以下に変更し，それぞれ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および平均 2 乗誤差が最も小さいモデルを選択した．

- Predictor のエポック数 $\in \{1, ..., 15\}$
- Predictor の学習率 $\in \{2\text{e-}3, 1\text{e-}3\}$
- Estimator のエポック数 $\in \{1, ..., 30\}$
- Estimator の学習率 $\in \{2\text{e-}3, 1\text{e-}3, 5\text{e-}4\}$

LASER には，著者らによって公開されている学習済みモデル[7]（bilstm.93langs.2018-12-26.pt）を利用した．

4.3 実験結果

表 2 に WMT-2017 Metrics Shared Task における実験結果を示す．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 LASER よりも高い性能を示した．また，下段の参照文に基づく自動評価手法と比較すると，$\text{BERT}_{\text{multi}}$ 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示した．これらの結果から，事前学習された多言語の文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 LASER 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 LASER 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．

5 分析

提案手法の学習データによる性能の変化について分析するために，以下の 2 つの設定で実験する．

5.1 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

[5] https://github.com/Unbabel/OpenKiwi
[6] http://www.statmt.org/wmt17/translation-task.html
[7] https://github.com/facebookresearch/LASER

を行う．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 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 2 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．

5.2 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 2 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 LASER 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．

6 おわりに

本研究では，事前学習された多言語の文符号化器を用いて機械翻訳の品質推定に取り組んだ．WMT-2017 Metrics Shared Task における実験の結果，提案手法は zh-en 以外の言語対で他の品質推定手法を大幅に上回る性能を達成し，いくつかの参照文に基づく自動評価手法とも同等以上の性能を示した．また，Zero-shot 学習などの分析の結果，事前学習された多言語 BERT を言語横断的に再学習することの有効性を確認した．

謝辞

本研究の一部は JST（ACT-X，課題番号: JPMJAX1907）の支援を受けたものです．

参考文献

[1] Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *TACL*, Vol. 7, pp. 597–610, 2019.

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

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, pp. 4171–4186, 2019.

[4] Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, and Luo Si. "Bilingual Expert" Can Find Translation Errors. In *Proc. of AAAI*, pp. 6367–6374, 2019.

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

[6] Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An Open Source Framework for Quality Estimation. In *Proc. of ACL*, pp. 117–122, 2019.

[7] 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. *TALLIP*, Vol. 17, No. 1, pp. 1–22, 2017.

[8] 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 *Proc. of WMT*, pp. 62–90, 2019.

[9] Maja Popović. chrF++: Words Helping Character N-grams. In *Proc. of WMT*, pp. 612–618, 2017.

[10] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of AMTA*, pp. 223–231, 2006.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *Proc. of ICLR*, pp. 1–41, 2020.

[12] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proc. of EMNLP*, pp. 563–578, 2019.

[13] 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, 2019.