バイエンコーダモデルを用いた参照訳なしの機械翻訳自動評価におけるランキングロスの適用

Application of Ranking Loss in Reference-free MT Automatic Evaluation using Bi-Encoder Model

北海学園大学　大学院工学研究科　教授

越前谷　博

1996 年北海学園大学大学院工学研究科修士課程修了。博士（工学）。2013 年～現在北海学園大学大学院工学研究科教授。機械翻訳の研究に従事。アジア太平洋機械翻訳協会（AAMT）／ Japio 特許翻訳研究会委員。

echi@hgu.jp　　TEL 011-841-1161（内線 :7863）

1 はじめに

現在、機械翻訳分野において様々なアプローチの研究が行われている。Transformer などのニューラル機械翻訳 [1] と共に、大規模言語モデル（LLM）や生成 AI を用いた機械翻訳研究 [2][3] も盛んに行われている。また、これらの機械翻訳研究を支えるために不可欠な技術が自動評価尺度である。日々進化している機械翻訳システムに対して、精度の高い自動評価を行うことで、さらなる機械翻訳研究の進歩に貢献できる。

このような状況において様々な自動評価尺度が提案されている。BLEU[4] に代表される MT 訳と参照訳間の表層情報に基づく手法は現在もデファクトスタンダードとして広く利用されている。これらの手法では MT 訳と参照訳を比較することでスコアを得る。また、COMET[5] のように文の意味情報に基づく手法も広く利用されている。意味情報に基づく手法の多くは事前学習済みのニューラルネットワークモデルをファインチューニングすることで自動評価に特化したシステムを構築する。また、参照訳を用いずにスコアを得るための様々な手法 [6][7][8] も提案されている。さらには生成 AI を利用することで訳文の評価を行うアプローチ [9][10] も提案されている。その際には有用なプロンプトを決定することが重要となる。

本報告では、意味情報に基づく手法、かつニューラルネットワークによるモデル構築の観点より、ランキングロスを取り入れた参照訳を必要としない新たな自動評価尺度を提案する。BLEU 等の表層情報のみに基づく手法では意味レベルの評価が困難である。また、参照訳を必須とする手法では、その収集のためのコストが問題となる。ChatGPT などの生成 AI を用いて評価を行う場合には、評価における再現性の担保が問題となる。提案手法では参照訳の収集コストの問題を回避するために参照訳を用いないことを前提としている。そして、意味レベルの評価を可能とするために、COMET と同様に事前学習済みのエンコーダモデルを活用することで原文と MT 訳を意味表現としてベクトルに変換する。その後、変換した原文と MT 訳のベクトル間のコサイン類似度が人手によるスコアと等価となるようにフィードフォワードによるニューラルネットワークを学習させることでモデルの構築を行う。提案手法は自動評価に特化した学習により構築されるため、生成 AI を利用した自動評価尺度に比べ、再現性が高くなると考えられる。さらに、損失関数には自動スコアの順位を反映させるためにペアワイズに基づくランキングロスを組み合わせたものとする。多くのエンコーダモデルを用いた自動評価尺度では損失関数には平均二乗誤差（MSE）が用いられている。しかし、その場合、自動スコアは個々のスカラー値として人手スコアに近づけることが目的となるため、自動スコア間の順序は反映されない。そこで、提案手法では MSE とランキングロスを組み合わせることで評価精度の向上を図る。

提案手法に対する性能評価実験では、WMT データの Metrics Task データを用いて、MT 訳に対する自動スコアを出力し、それらの評価を行った。その結果、損失関数に MSE のみを用いたベースラインに比べて、ラン

キングロスを導入した提案手法では人手スコアとの相関係数が向上し、提案手法の有効性を確認した。

2 提案手法

提案手法は、エンコーダモデルとフィードフォワードを組み合わせたモデルで構成され、原文と訳文、そして、人手スコアを用いて学習を行う。提案手法の概要を図 1 に示す。

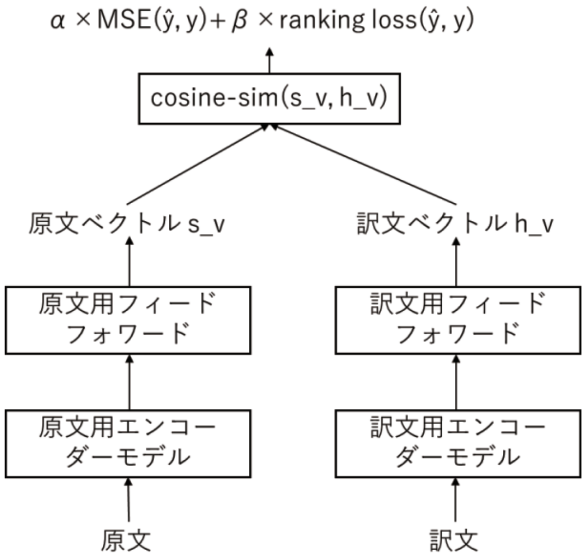

図 1　提案手法の概要

図 1 に示すように、提案手法ではバイエンコーダに基づく自動評価尺度となっている。すなわち、原文と訳文側で異なるエンコードモデルを用い、それぞれのエンコーダにより得られるベクトル間のコサイン類似度を求めることで自動スコアを得る。このようなバイエンコーダに対して、エンコーダモデルを 1 つとするクロスエンコーダを用いた自動評価尺度 [5][6][7][8] が存在する。クロスエンコーダはバイエンコーダよりも高い性能を示すことが知られているが、その一方で大量のデータを比較する際に処理時間がかかることが問題点として指摘されている [11]。そこで、本報告では、処理時間を重視し、バイエンコーダを用いることとした。

また、提案手法では 2 つの誤差関数を導入することで精度向上を図っている。一つは平均二乗誤差（MSE）であり、これは個々の人手スコアと自動スコアを独立したものとして扱う場合に適している。もう一つはランキングロスである。提案手法におけるランキングロスではペアワイズに基づいているため、スコアの順序が人手スコアに近づくように学習を行う。すなわち、自動スコア間の相対的な関係を学習することが可能となる。この 2 つの誤差関数を以下の式を用いて組み合わせることで絶対的な観点と相対的な観点の両方の観点での学習が可能になると考えられる。

$$\alpha \times \mathrm{MSE}(\hat{y}, y) + \beta \times \text{ranking loss}(\hat{y}, y) \quad \cdots \quad \text{式 (1)}$$

上式の $\hat{y}$ は原文ベクトル s_v と訳文ベクトル h_v 間のコサイン類似度、すなわち、自動スコアを示し、y は人手スコアを示す。また、α と β はパラメータであり、スカラー値である。ランキングロスを用いた自動評価尺度として ReMedy[12] が提案されているが、ランキングロスのみを用いており、MSE を用いていない点が提案手法とは異なる。

3 性能評価実験

3. 1　実験データ

提案手法の有効性を検証するために、WMT の Metrics Task データを用いて実験を行った。学習データには WMT19[13] の原文とその訳文、そして、DA スコアをセットとした DA データ 327,314（17 言語ペア）、WMT20[14] の原文とその訳文、そして、DA スコアをセットとした DA データ 210,705（18 言語ペア）、WMT21[15] の to-English を除き、原文とその訳文、そして、DA スコアをセットした DA データ 313,418（9 言語ペア）、さらに、WMT21 の原文とその訳文、そして、MQM スコア [16] をセットとした MQM データ 27,141（3 言語ペア）の計 878,578 データを用いた。評価データには WMT22[17] の原文とその訳文、そして、MQM スコアをセットとした MQM データ 100,809 を用いた。評価データに用いた MQM データの内訳は en-de（英語からドイツ語）が 32,592、en-ru（英語からロシア語）が 32,592、zh-en（中国語から英語）が 35,625 である。また、WMT21 の to-English はモデルの性能を下げることが指摘されている [8] ため、実験データから除外した。MQM データにおいても WMT20 の MQM データも性能を低下させることが指摘されている [8] ため、学習データから除外した。なお、

DA データは 0 ～ 100 の raw スコア、MQM データは -25 ～ 0 の raw スコアを -1.0 から 1.0 に正規化した値を用いた。WMT データでは raw データ以外にも z スコアも使用可能であるが、DA スコアと MQM データの両方を混在して学習データとして用いる場合には、z スコアよりも raw スコアを用いる方が適切であることが指摘されている [8] ため、本実験においても raw スコアを用いた。

3. 2　実験方法

提案手法におけるエンコーダモデルには原文側と訳文側共に xlm-roberta-base を用いた。また、フィードフォワードには入力層が 768 ニューロン、中間層が 768 ニューロン、そして、出力層が 256 ニューロンの全結合層からなるニューラルネットワークを構築した。そして、式 (1) の α と β の値には 0.3 と 0.7 をそれぞれ用いた。

また、本実験では提案手法に対するベースラインとして誤差関数に MSE のみを用いた自動評価尺度を用いた。ベースラインのエンコーダモデルには提案手法と同様に xlm-roberta-base、フィードフォワードも提案手法と同様に入力層が 768 ニューロン、中間層が 768 ニューロン、そして、出力層が 256 ニューロンの全結合層を用いた。さらに参照訳を用いない先行研究の自動評価尺度との比較を行った。

提案手法とベースラインのモデルの学習時のパラメータとしてはバッチサイズに 128 を用いた。評価については自動評価尺度による自動スコアと人手評価間の相関係数を求めることで行った。具体的にはケンドール τ を用いた。

3. 3　実験結果と考察

表 1 に提案手法とベースライン、そして、先行研究の en-de、en-ru、zh-en それぞれの相関係数とそれらの相関係数の平均値（Avg.）を示す。

表 1 より、提案手法の相関係数はベースラインに対して、平均値と共にいずれの言語ペアにおいても高い値を示した。これは、損失関数において自動スコア間の相対的な順位を反映したランキングロスを導入したことが有効に働いたためと考えられる。損失関数において MSE のみを用いた場合、絶対値として値が人手スコアと近づくように学習されるが、その場合、自動スコア間の順序関係の学習が不十分となる。

それに対して、提案手法は MSE のみではなく、ランキングロスを組み合わせた損失関数を用いている。ランキングロスを用いることで自動スコア間の相対的な順序関係についての学習も可能となった。しかし、提案手法は先行研究の自動評価尺度に対して、十分な評価精度が得られたとはいえない。平均値が 0.3 以上の自動評価尺度も複数存在しており、また、個々の言語ペアにおいても提案手法の相関係数は十分とはいえない。そのため、今後、さらなる精度向上のための改良が必要と考えられる。

表 1　実験結果

	en-de	en-ru	zh-en	Avg.
COMETKiwi	0.290	0.359	0.364	0.338
COMET-QE	0.281	0.341	0.365	0.329
UniTE-src	0.287	0.342	0.343	0.324
Cross-QE	0.263	0.310	0.378	0.317
MS-COMET-QE-22	0.233	0.305	0.287	0.275
MATESE-QE	0.244	0.229	0.337	0.270
HWTSC-Teacher-Sim	0.155	0.143	0.272	0.190
KG-BERTScore	0.129	0.111	0.219	0.153
HWTSC-TLM	0.092	0.121	0.086	0.100
REUSE	0.065	0.078	0.130	0.091
ベースライン	0.190	0.232	0.242	0.221
提案手法	0.217	0.275	0.298	0.263

4　まとめ

本報告ではより精度の高い自動評価尺度の実現に向け、新たな自動評価尺度の提案とその性能評価について述べた。提案手法では、モデル構築時の学習における損失関数を先行研究の多くで用いられる MSE だけでなく、ランキングロスを導入した。MSE のみでは個々の自動スコアを絶対値としてロスを求めることになるが、ランキングロスを組み合わせることで自動スコア間の相対的な順序関係についても学習可能となる。

性能評価実験の結果、損失関数に MSE のみを用いたベースラインの自動評価尺度よりも提案手法は高い相関係数を示した。したがって、提案手法の有効性が確認された。しかし、先行研究との比較においては不十分であるため、改良が必要である。今後は損失関数の改善やフィードフォワードにおける最適なハイパーパラメータの検討を行う予定である。

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention Is All You Need” , Proceedings of the 31st Conference on Neural Information Processing Systems, pp. 6000-6010 (2017).

[2] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu Shujian Huang1, Lingpeng Kong, Jiajun Chen1, Lei Li, “Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis” , Proceedings of the Association for Computational Linguistics: NAACL 2004, pp. 2765-2781, (2024).

[3] Yafu Li, Yongjing Yin, Jing Li, Yue Zhang, “Prompt-Driven Neural Machine Translation” , Findings of the Association for Computational Linguistics: ACL 2022, pp. 2579- 2590 (2022).

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation” , Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318 (2002).

[5] Ricardo Rei, Craig Stewart, Ana C Farinha and Alon Lavie, “COMET: ANeural Framework for MT Evaluation” , Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 2685–2702 (2020).

[6] Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha1, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, André F. T. Martins, “COMETKIWI: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task” , Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 634- 645 (2022).

[7] Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag, “MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task” , Proceedings of the Eighth Conference on Machine Translation (WMT), pp. 756–767 (2023).

[8] Juraj Juraska, Daniel Deutsch, Mara Finkelstein and Markus Freitag, “MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task” , Proceedings of the Ninth Conference on Machine Translation, pp. 492–504 (2024).

[9] Tom Kocmi and Christian Federmann, “GEMBA-MQM:Detecting Translation Quality Error Spans with GPT-4” , Proceedings of the Eighth Conference on Machine Translation (WMT), pp. 768–775 (2023).

[10] Marcin Junczys-Dowmunt, “GEMBA-MQMV2: TenJudgmentsAreBetter Than One” , Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pp. 926–933 (2025).

[11] Nils Reimers and Iryna Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks” , Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3982–3992 (2019).

[12] Shaomu Tan and Christof Monz, “ReMedy: Learning Machine Translation Evaluation from Human Preferences with Reward Modeling” , Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 4370–

4387 (2025).
[13] 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” , Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2, pp. 62-90 (2019).
[14] Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma and Ondřej Bojar, “Results of the WMT20 Metrics Shared Task” , Proceedings of the 5th Conference on Machine Translation (WMT), pp. 688–725 (2020).
[15] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie and Ondřej Bojar, “Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain” , Proceedings of the Sixth Conference on Machine Translation (WMT), pp. 733–774 (2021).
[16] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, Wolfgang Macherey, “Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation” , Transactions of the Association for Computational Linguistics, vol. 9, pp. 1460–1474 (2021).
[17] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie and André F. T. Martins, “Results of WMT22 Metrics Shared Task: Stop Using BLEU– Neural Metrics Are Better and More Robust” , Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 46–68 (2022).