単語単位で評価が可能な機械翻訳向け自動評価

高橋 洸丞　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{takahashi.kosuke.th0, sudoh, s-nakamura}@is.naist.jp

概要

機械翻訳の自動評価として広く用いられている BLEU[1] や COMET[2] などは、文単位で評価スコアを算出するように設計されている。しかし、文単位の評価は文内部の誤り箇所や誤り度合いを明示できないという欠点がある。そこで本研究では、文単位よりも細かい、単語単位で誤り評価が行われた人手評価データを使用して、単語単位での評価を行い、その評価結果を基に文単位での評価スコアを算出する自動評価モデルを提案する。

2021 年の WMT metrics タスクによる実験において、提案手法は、従来の文単位でのみ評価を行うモデルと同等もしくは言語対によっては上回る人手評価との相関が得られた。

1　はじめに

自動評価には、(a) 単語単位で誤り計算をした後に文単位の評価値を算出するモデルや、(b) 人手評価を教師信号として文単位の評価値を直接学習するモデルがある。前者のモデル (a) として広く使用されている BLEU、BERTscore[3] は、単語の一致または一致度合いをそれぞれの単語で計算した後に、一致数や一致度合いの合計を文単位の評価値としている。しかし、これらのモデルは本質的に文中に含まれる単語数やトークン数、また機能語の数によって評価結果が変動してしまう問題がある。さらに、評価素性が字面の一致のみであることもあり、WMT の metrics タスクにおいて (a) のようなモデルは (b) の自動評価モデルに比べて人手評価との相関が低い。後者のモデル (b) の BLEURT[4], COMET などは、XLM (Crosslingual Language Model) [5][6] という大規模な事前学習済の transformer[7] エンコーダを使用したモデルで、評価対象のシステム訳文・参照訳文・原言語文の三つを入力として、エンコーダを通して得た文ベクトルから文単位の人手評価と同じ評価値を予測するように学習をしている。そして、これらのモデルは WMT metrics タスクの 20-22 年度で人手評価との高い相関を記録している。

自動評価システムが文単位の評価性能を向上させていく一方で、人手評価は文単位から単語単位へと変容している。年々と機械翻訳システムの翻訳性能が向上している為、曖昧な文単位の評価から、より細かい評価が可能な単語単位の評価へと、WMT の人手評価が変更された。19 年度まで文単位で人手評価を行う DA (Direct Assessment) [8] を採用していたが、20 年度からは MQM (Multidimentional Quality Metrics) [9] という単語単位で誤り度合いや誤りの種類を記述する評価手法となり、人手評価の信頼性が向上した。

本研究は、こうした人手評価の変化を受けて、自動評価でもより細やかな評価を行えるように、単語単位でのエラーの有無、エラースコアを計算する。また同時に、文スコアに与える各単語の重要度を計算し、それらを基に文単位での評価値も算出するモデルを提案する。提案手法は単語単位での評価が可能な上に、実験の結果、文単位のみの評価手法 (ベースライン) と、同程度の人手評価との相関、あるいは言語対によってはより高い相関値を得ることができた。

2　関連研究

WMT の metrics タスクでは、DA と MQM の二種類の人手評価手法が取り入れられており、 DA が WMT15-21 の翻訳結果に、MQM が WMT20-22 の翻訳結果にアノテーションされている。DA は 0-100 の整数値で参照訳文に対して翻訳文の出来を評価する手法で、高性能な翻訳システムの評価には信頼性が欠けるとされている [10]。一方で、MQM は翻訳誤りの種類や程度を誤翻訳されている箇所にアノテーションし、その誤りの種類や程度によって各翻訳文の評価スコアを決定する。

翻訳文の自動評価というタスクでは、MQM の採用によって 21 年度より単語単位の評価が重要視さ

れ始めたが、品質推定のタスクではそれ以前より単語単位での評価が行われてきた。これらの二つのタスクの違いは、参照訳文の有無であり、本研究は参照訳文付きの翻訳文評価タスクでの使用を想定している。

MQM を使用した評価モデルには WMT22 で発表された、COMET22$\tilde{y}_{tag}$[11] や MaTESe[12] がある。COMET22$\tilde{y}_{tag}$ は、これまでの COMET と同様の文単位で評価値を出力する線形層に加えて、単語単位で OK/BAD という誤りの有無を出力する線形層を持ちマルチタスク学習をしている。そして、MaTESe は単語単位で誤りの有無や、誤り度合いが Major なのか Minor なのかを、BIO ラベルで系列ラベリング問題を解く形で予測するモデルであり、COMET から文単位のスコア予測の線形層が取り除かれた構造をしている。

本研究の提案手法は、これらのモデルと比較すると、以下の点で異なる。

1. 単語単位で誤りの有無と誤りのスコアを共に出力する点
2. 単語単位の誤り予測結果を基に文単位のスコアを計算する点

3 提案手法

従来の評価モデルは、文単位の評価しか行えないので、文中のどこがどの程度誤っているのかを提示できない。また、COMET22$\tilde{y}_{tag}$ などの単語単位評価モデルは、文単位スコアが単語単位スコアに依らずに決定されるため、文単位と単語単位で一貫性のない設計となっている。そこで本研究では、システム訳文の単語単位のエラーの有無、エラースコア、そして文単位のエラースコアをそれぞれ出力することができる手法を提案する。提案手法の評価モデルは、図 1 のように、システム訳文・原言語文・参照訳文の三つの入力文から、XLM を通したトークンごとの embedding を用いて評価を行う。また図 1 内の Transformer に対して、XLM のそれぞれの出力に、positional embedding と入力の種類によって異なる token type embedding をそれぞれ足し合わせたものを、[システム訳文, 原言語文, 参照訳文] の順に結合したものを入力する。(図 2)

単語単位のエラーの有無は IO ラベルによって予測され、cross-entropy ロスを用いた学習をする。また単語単位のエラースコアは回帰問題として予測され、MSE ロスを用いて学習する。そして文単位のエラースコアは、エラーの有無を表現するエラーラベルとエラースコア、そしてスコア重みの三つの掛け合わせによって決定され、MSE ロスを用いて学習も行う。

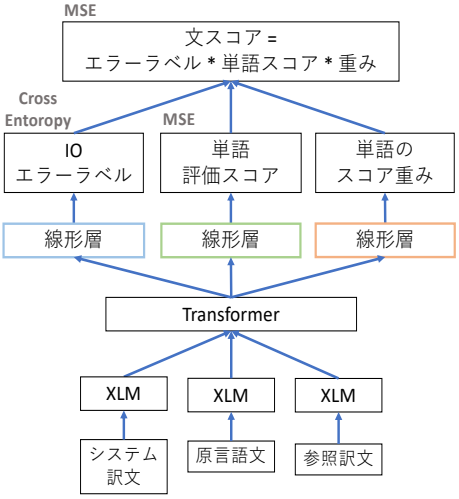

図 1 提案手法の全体構造

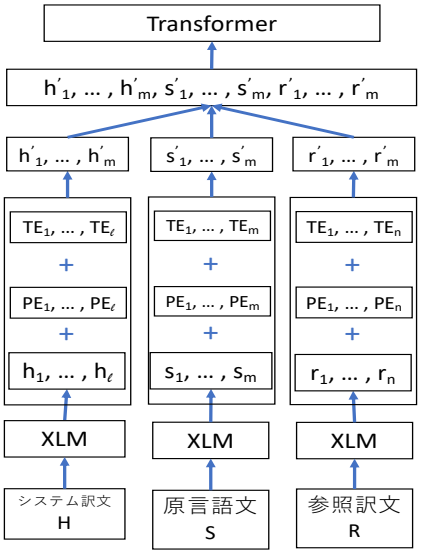

図 2 提案モデルの transformer への入力。XLM から出力された分散表現に PE(Positional Embedding) と TE(Token Type Embedding) を足し合わせて、システム訳文・原言語文・参照訳文の順に結合したものを入力する。

MQM データの前処理 MQM のデータは評価者によって誤りの位置や誤り度合い・誤りカテゴリも異なるので、評価者が誤りだと表記した単語ごとに google の MQM の重みスコア (表 1) に従ったスコアを割り振り、同一システム訳文の全評価者分で単語ごとのスコア平均を取る。このスコア平均を単語単位エラースコアの正解データとして扱い、その中でもスコアが 0 以上の単語についてはエラー有りラベルを、スコアが 0 の単語はエラー無しのラベルを付

与し、これを単語単位のエラー識別の正解データとする。また、単語単位ではなく文単位でスコア平均を取ったものを文単位のエラースコアの正解データとして扱う。

表 1 MQM の誤り度合いとカテゴリによる重みスコア

誤り度合い	カテゴリ	スコア
Major	Non-translation	25
Major	その他	5
Minor	Fluency/Punctuation	0.1
Minor	その他	1
Neutral/No-Error	-	0

4 実験

実験は WMT metrics shared task の 21 年度のデータを用いて行なった。

4.1 実験設定

newstest20 と TED 共に en-de (英-独) と zh-en (中-英) の言語対から 9:1 の割合で訓練データと開発データに分けて使用し、newstest21 の en-de、zh-en の言語対において自動評価の性能評価を行なった。性能評価は、ピアソンの相関係数、ケンドールの相関係数を用いて文単位の人手評価との相関を計測し、単語単位での性能評価は予測エラーラベルの適合率、再現率、F1, エラースパンの適合率 (HSH)・再現率 (TSH)[12] を用いて行なった。また、モデルの性能比較のために、単語単位のエラー有無・エラースコアを一切予測せず、直接文スコアを予測するモデル (ベースライン) も用意した。さらに、提案手法と同様に単語単位の評価が行える COMET22 の公開されているモデル[1)]から, $\tilde{y}_{tag}$ モデル (COMET22 tagging) と文単位の出力を行う COMET22 のモデル (COMET22 segment) で評価実験を行なった。COMET22 の両方のモデルは、提案モデルよりも DA15-20 のコーパスと MQM20 en-ru 言語対のコーパス分、訓練データが多いため、比較対象ではなく、よりよい条件で訓練されたモデルの参考として用意した。そして COMET22 tagging は、OK/BAD の予想確率が各単語の出力となっていたので、その値が 0.5 以上ならエラー有り、未満ならエラー無しとして評価した。

1) https://github.com/Unbabel/COMET

4.2 実験結果

評価モデルごとの文単位、単語単位のメタ評価結果をそれぞれ表 2 と表 3 に示す。実験の結果、各言語対において、文スコアのみを直接予測するモデルよりも、提案モデルの方がピアソンの相関係数とケンドールの相関係数が高かった。また en-de、zh-en の個々の言語対では参考モデルの COMET22 segment に及ばないものの、2 つの言語対をまとめた状態でのメタ評価は COMET22 segment を上回った。そして単語単位のメタ評価では、提案モデルは適合率よりも再現率が高くなっており、それに伴ってエラースパンの適合率よりも再現率の方が高くなっていた。さらに、中国語を含んだコーパスにおいて提案モデルは COMET22 tagging よりもエラースパンの数値がよかった。

4.3 誤り度合いによる評価結果の違い

表 1 より、MQM の誤り度合い Major と Minor の境目は 5.0 と 1.0 である。本節ではより誤り度合いの大きい文スコア 5.0 に着目し、文単位スコア 5.0 以下と 5.0 超えにテストデータを分けたメタ評価を行なった (表 4, 5, 6, 7)。その結果、提案モデルは、文スコア 5.0 越えの zh-en 言語対を除いて、文スコアのみを直接予測するベースラインモデルを上回る相関係数値を示した。そして表 4 と 5 の相関係数の変化から、COMET22 よりも提案モデルは、より誤り度合いの高い文に対して相関係数が高くなる傾向があった。

また、単語単位のメタ評価やエラースパンについても、エラースコアが高い文に対してエラー予測の適合率や再現率が高くなった。

5 おわりに

本研究では、単語単位でエラーの有無・エラーの度合いを予測し、それらの予測結果から文単位のスコアを計算する機械翻訳向けの自動評価を提案した。WMT21 の評価タスク実験の結果、提案手法はベースラインと比較してピアソンやケンドールの相関係数値を高く維持したまま、単語単位での評価が行えていることがわかった。また、適合率・再現率や事例分析の結果、単語単位のエラー予測については実用段階には少し早いかもしれないが、エラースパンの予測という観点で見ると、半分程度のスパン予測の適合・再現が可能であることがわかった。

表 2　WMT21 MQM コーパスでのメタ評価。各言語対のピアソン (ρ)、ケンドール (τ) の相関係数値を計測した結果

評価モデル	en-de		zh-en		en-de+zh-en	
	ρ	τ	ρ	τ	ρ	τ
COMET22 tagging	0.232	0.220	0.252	0.241	0.344	0.313
COMET22 segment	**0.361**	**0.266**	**0.509**	**0.387**	0.476	0.316
文スコアのみ直接予測するモデル	0.258	0.187	0.469	0.335	**0.534**	**0.367**
提案モデル	0.289	0.194	0.484	0.347	**0.539**	**0.369**

表 3　WMT21 MQM コーパスでの単語エラー予測を適合率 (P)、再現率 (R)、F1 スコア (F)、エラースパンの適合率 (HSH)、再現率 (TSH) で計測した結果

評価モデル	en-de					zh-en					en-de+zh-en				
	P	R	F	HSH	TSH	P	R	F	HSH	TSH	P	R	F	HSH	TSH
COMET22 tagging	**0.270**	0.331	**0.297**	**0.209**	**0.519**	**0.329**	0.596	**0.424**	0.253	0.583	**0.311**	0.495	**0.382**	0.237	0.563
提案モデル	0.220	**0.397**	0.283	0.199	0.500	0.258	**0.726**	0.381	**0.283**	**0.664**	0.248	**0.601**	0.351	**0.252**	**0.613**

表 4　文スコアが 5.0 以下、0.0 越えの誤りを含む文に対するピアソン (ρ) とケンドール (τ) の相関係数値

評価モデル	en-de		zh-en		en-de+zh-en	
	ρ	τ	ρ	τ	ρ	τ
COMET22 tagging	0.145	0.110	0.081	0.082	0.161	0.136
COMET22 segment	**0.181**	**0.142**	**0.146**	**0.109**	0.137	0.102
文スコアのみ直接予測するモデル	0.083	0.094	0.016	0.035	0.148	0.137
提案モデル	0.145	0.111	0.096	0.081	**0.186**	**0.149**

表 5　文スコアが 5.0 越えの誤りを含む文に対するピアソン (ρ) とケンドール (τ) の相関係数値

評価モデル	en-de		zh-en		en-de+zh-en	
	ρ	τ	ρ	τ	ρ	τ
COMET22 tagging	0.059	0.076	0.010	0.042	0.042	0.064
COMET22 segment	0.155	0.136	0.273	0.180	0.260	0.175
文スコアのみ直接予測するモデル	0.203	0.157	**0.302**	**0.213**	**0.308**	**0.212**
提案モデル	**0.230**	**0.184**	0.291	0.187	**0.302**	0.197

表 6　文スコアが 5.0 以下、0.0 越えの誤りを含む文に対する適合率 (P)、再現率 (R)、F1 スコア (F)、エラースパンの適合率 (HSH)、再現率 (TSH) で計測した結果

評価モデル	en-de					zh-en					en-de+zh-en				
	P	R	F	HSH	TSH	P	R	F	HSH	TSH	P	R	F	HSH	TSH
COMET22 tagging	**0.438**	0.314	0.366	**0.377**	**0.520**	**0.306**	0.550	**0.393**	0.250	0.523	**0.353**	0.412	**0.380**	0.305	0.521
提案モデル	0.394	**0.353**	**0.372**	0.372	0.489	0.254	**0.638**	0.363	**0.286**	**0.614**	0.301	**0.471**	0.367	**0.320**	**0.553**

表 7　文スコアが 5.0 超えの誤りを含む文に対する適合率 (P)、再現率 (R)、F1 スコア (F)、エラースパンの適合率 (HSH)、再現率 (TSH) で計測した結果

評価モデル	en-de					zh-en					en-de+zh-en				
	P	R	F	HSH	TSH	P	R	F	HSH	TSH	P	R	F	HSH	TSH
COMET22 tagging	**0.584**	0.379	0.460	0.499	0.516	**0.441**	0.617	**0.515**	0.360	0.616	**0.456**	0.571	**0.507**	0.378	0.601
提案モデル	0.540	**0.411**	**0.467**	**0.507**	**0.531**	0.383	**0.711**	0.498	**0.413**	**0.691**	0.397	**0.653**	0.494	**0.425**	**0.667**

謝辞

本研究は JSPS 科研費 22H03651 の支援を受けたものである。

参考文献

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

[2] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[3] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[4] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[5] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[8] 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, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[9] Arle Lommel, Hans. Uszkoreit, and Aljoscha Burchardt. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. **Tradumàtica**, No. 12, pp. 455–463, 2014.

[10] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation, 2021.

[11] Ricardo Rei, José Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Craig Stewart, Luisa Coheur, and André Martins. COMET-22:unbabel-ist 2022 submission for the metrics shared task. In **Proceedings of the Seventh Conference on Machine Translation**. Association for Computational Linguistics, 2022.

[12] Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. MATESE: Machine translation evaluation as a sequence tagging problem. In **Proceedings of the Seventh Conference on Machine Translation**. Association for Computational Linguistics, 2022.

A 付録

A.1 文スコア 1.0 の上下での評価結果の違い

4.3 節では文スコア 5.0 の上下で評価結果の違いを分析したが、本節では文スコア 1.0 の場合の結果を示す (表 8、9、10、7)。文スコア 5.0 の時と同様に、提案モデルはより誤り度合いの高い文に対してメタスコアがよくなっている。

表 8 文スコアが 1.0 以下、0.0 越えの誤りを含む文に対するピアソン (ρ) とケンドール (τ) の相関係数値

評価モデル	en-de		zh-en		en-de+zh-en	
	ρ	τ	ρ	τ	ρ	τ
COMET22 tagging	0.019	0.018	**0.116**	**0.072**	0.018	0.033
COMET22 segment	0.020	0.025	0.007	0.006	0.066	0.075
文スコアのみ直接予測するモデル	0.025	0.012	0.073	0.033	**0.244**	**0.193**
提案モデル	0.014	0.009	0.026	0.004	0.161	0.105

表 9 文スコアが 1.0 越えの誤りを含む文に対するピアソン (ρ) とケンドール (τ) の相関係数値

評価モデル	en-de		zh-en		en-de+zh-en	
	ρ	τ	ρ	τ	ρ	τ
COMET22 tagging	0.125	0.101	0.089	0.110	0.155	0.162
COMET22 segment	**0.226**	0.117	**0.373**	**0.264**	0.349	0.226
文スコアのみ直接予測するモデル	0.167	0.073	**0.379**	**0.260**	**0.414**	**0.287**
提案モデル	0.215	**0.122**	**0.375**	0.254	**0.410**	**0.281**

表 10 文スコアが 1.0 以下、0.0 越えの誤りを含む文に対する適合率 (P)、再現率 (R)、F1 スコア (F)、エラースパンの適合率 (HSH)、再現率 (TSH) で計測した結果

評価モデル	en-de					zh-en					en-de+zh-en				
	P	R	F	HSH	TSH	P	R	F	HSH	TSH	P	R	F	HSH	TSH
COMET22 tagging	**0.401**	0.280	0.330	**0.340**	**0.493**	**0.238**	0.465	**0.315**	0.203	0.463	**0.315**	0.333	**0.324**	0.276	0.481
提案モデル	0.355	**0.323**	**0.338**	0.329	0.467	0.199	**0.572**	0.295	**0.253**	**0.558**	0.268	**0.395**	0.319	**0.291**	**0.505**

表 11 文スコアが 1.0 超えの誤りを含む文に対する適合率 (P)、再現率 (R)、F1 スコア (F)、エラースパンの適合率 (HSH)、再現率 (TSH) で計測した結果

評価モデル	en-de					zh-en					en-de+zh-en				
	P	R	F	HSH	TSH	P	R	F	HSH	TSH	P	R	F	HSH	TSH
COMET22 tagging	**0.515**	0.360	0.423	0.446	**0.534**	**0.410**	0.608	**0.490**	0.333	0.598	**0.427**	0.533	**0.474**	0.358	0.582
提案モデル	0.473	**0.393**	**0.430**	**0.450**	0.518	0.352	**0.699**	0.468	**0.376**	**0.677**	0.371	**0.607**	0.460	**0.391**	**0.638**