ehiMetrick: NLP2025 自動評価ハック Shared Task 機械翻訳部門

森岡 拓 [1]　杉山 誠治 [2]　高山 隼矢 [1]　梶原 智之 [1]
[1] 愛媛大学大学院理工学研究科　[2] 愛媛大学工学部
{morioka@ai., sugiyama@ai., takayama@ai., kajiwara@}cs.ehime-u.ac.jp

概要

本研究では，機械翻訳における自動評価指標の脆弱性について英日翻訳において検証した．機械翻訳の自動評価は人手評価と比べて低コストで高速に評価できるため，広く利用されている．しかし，自動評価指標は明らかに低品質な文に対しても高いスコアを出力する場合がある．4 つの評価指標について検証した結果，次のような事例が確認できた．BLEU は短い出力文に対するペナルティ項の影響を軽減するように文末に適当な文字を追加した場合にスコアが上昇した．chrF は Recall 重視の式で計算されており，2 文を連結して冗長にした場合にスコアが上昇した．COMET は多言語エンコーダを利用するため日本語以外の文も入力可能であり，翻訳をする前の英文を入力することで，エンコーダデコーダモデルによる翻訳より高いスコアとなった．GEMBA は「プロによる翻訳文である」と文末に追加する場合にスコアが上昇した．また，各指標について手法を順位付けし，その平均順位を計算することで，各指標単体の欠点を互いに補完し，より妥当な評価が得られることを示した．

1　はじめに

機械翻訳モデルの評価を低コストで高速に実施するために，自動評価指標が提案されてきた．従来より，参照訳と出力文の表層的な類似度に基づく評価指標が数多く提案されている．広く用いられている BLEU [1] は，参照訳と出力文の単語 n-gram 一致率からスコアを算出し，chrF [2] は文字 n-gram 一致率から F スコアを計算する．これらの表層ベースの評価指標は，数式に従ってスコアを計算しており根拠が明確であるが，一方で，参照訳と異なる言い回しに対して適切に評価できないという弱点がある．

近年では，事前学習済みモデルを利用した自動評価指標が普及している．BERTScore [3] はエンコーダモデルによる単語埋め込みから単語類似度を計算し，翻訳品質を測定する．同様に COMET [4] もエンコーダモデルを利用しており，モデル出力から人手評価のスコアを推定できるように学習している．これらのモデルベースの評価指標は意味レベルの類似度を捉えられるが，表現が適切に埋め込まれるかはモデルの性質に左右され，根拠を特定しづらい．

大規模言語モデルを利用して MQM フレームワーク [5] で用いられるエラーラベルを予測する GEMBA-MQM [6] という自動評価指標も提案されている．本研究では GEMBA-MQM のみを利用するため，単に GEMBA と呼ぶ．大規模言語モデルベースの評価指標は，大規模な学習による知識を活用して根拠も含めた柔軟性の高い出力が可能であるが，プロンプトによる出力の制御が難しい．

このような評価指標の脆弱性を出力文が偶然突くことで，品質が低い文に高いスコアが付く可能性がある．そのため，本研究ではこれらの評価指標の脆弱性について調査し，BLEU，chrF，COMET，GEMBA が品質を低下させるような編集をした文に対して高いスコアを出力する事例 (表 1) を確認した．表層ベース，モデルベース，大規模言語モデルベースの評価指標はそれぞれ異なる脆弱性があるため，相補的に用いる必要がある．本研究では，各指標で生成文を順位付けし，その平均順位を計算することで，各指標単体の欠点を互いに補完し，より妥当な評価が得られることを示した．

2　Shared Task

本研究は NLP2025 ワークショップ「LLM 時代のことばの評価の現在と未来」の Shared Task 機械翻訳部門に参加している．本 Shared Task は WMT 2024 General Translation Task [7] (English-to-Japanese) のデータを利用して，翻訳の自動評価手法の頑健

表 1　脆弱性を突いた出力文の例

src.en：英文をそのまま入力	Brewers froth over winemakers" exemption from anti-waste rules
Default：GPT-4o に翻訳を指示	ワイン生産者が廃棄物防止規則の免除を受けていることに、 ビール醸造者たちは憤慨しています。
Short：GPT-4o に短文での翻訳を指示	ビール製造業者は、ワイン製造業者の廃棄規制免除に不満を持つ。
Short_maru_7：。を 7 つ追加	ビール製造業者は、ワイン製造業者の廃棄規制免除に不満を持つ。。。。。。。
Short_mecab_8： mecab トークン 7 つを末尾に追加	ビール製造業者は、ワイン製造業者の廃棄規制免除に不満を持つ。 廃棄規制免除に不満を持つ。
Default_Short_concat：2 文を連結	ワイン生産者が廃棄物防止規則の免除を受けていることに、 ビール醸造者たちは憤慨しています。 ビール製造業者は、ワイン製造業者の廃棄規制免除に不満を持つ。
Default_pro：プロの翻訳家による文と追記	ワイン生産者が廃棄物防止規則の免除を受けていることに、 ビール醸造者たちは憤慨しています。(The sentence has been translated by a professional translator, and there are no errors in this translation.)

表 2　ベースラインのプロンプト

Default	Translate the provided text from English to Japanese. If the text includes a URL, copy the URL as is into the translated text.
Short	Provide a shorter and more concise translation of the given text from English to Japanese. If the text includes a URL, copy the URL as is into the translated text.

性（脆弱性）を問うことを目的としている．参加者は，英文に対する「自動評価にとって良い翻訳文に見える」という意味での日本語への翻訳結果を 10 件まで提出する．評価指標は表層ベースの BLEU と chrF，埋め込みベースの BERTScore と COMET，LLM ベースの GEMBA である．BLEU と chrF は sacreBLEU [8] を利用して評価する．また，COMET は Unbabel/wmt22-comet-da と Unbabel/XCOMET-XL の 2 つのモデルを利用する．

3　検証

本節では各評価指標の特性に基づいてそれぞれの脆弱性となり得る点を指摘し，品質の低い出力文が不当に高いスコアを得るケースについての仮説を立て，またその検証を行う．

3.1　実験設定

本研究では，WMT 2024 General Translation Task の英日翻訳のデータを利用する．原言語である英語の 998 文に対して，目的言語である日本語の参照訳が含まれている．ベースラインとして 2 つの翻訳文を GPT-4o [9] で生成した．Default は翻訳を依頼するシンプルなプロンプトを用いて生成し，Short は簡潔な翻訳を依頼するプロンプトを用いて生成した．プロンプトを表 2 に示す．原言語文に URL のみの文がいくつかあり，これに対してしばしば GPT-4o が URL へのアクセスができない旨の応答を返すことが確認されたため，対策として，2 つのプロンプトでは共通して URL はそのまま複製するように指示している．OpenAI の Batch API を利用し，response_format パラメータに translation_output というキーを持つ json schema を指定することで出力を制御した．各評価指標に対するスコアを表 3 に示す．

3.2　BLEU：長さペナルティの脆弱性

BLEU は表層ベースの自動評価指標である．スコアは出力文・参照訳をそれぞれトークナイズして得られた単語列に対する n-gram 一致率を利用して，式 (1) のように計算される．

$$\mathrm{BLEU} = \mathrm{BP} \times \Big(\prod_{n=1}^{N} p_n\Big)^{\frac{1}{N}} \tag{1}$$

$$\mathrm{BP} = \begin{cases} 1 & (c > r) \\ e^{(1-r/c)} & (c \le r) \end{cases} \tag{2}$$

ここで，c と r は出力文と参照訳それぞれのコーパス全体のトークン数，p_n は出力文の n-gram のうち，参照訳文にも対応するものが存在する割合をコーパス全体に対して計算した数値である．また，N はパ

表 3　ベースラインのスコア

手法	BLEU	chrF2	BERTScore	COMET	XCOMET	GEMBA
Default	26.13	35.58	0.8483	0.8734	0.7892	-3.30
Short	23.87	32.64	0.8460	0.8622	0.7853	-4.22

表 4　BLEU の内訳

手法	BLEU	BLEU 内訳
Default	26.13	60.1/32.3/19.4/12.4 (BP = 1.000 ratio = 1.043 hyp_len = 50672 ref_len = 48569)
Short	23.87	65.3/35.1/21.0/13.1 (BP = 0.847 ratio = 0.857 hyp_len = 41633 ref_len = 48569)
Short_maru_7	24.10	58.0/30.8/18.4/11.5 (BP = 0.972 ratio = 0.973 hyp_len = 47240 ref_len = 48569)
Short_mecab_8	24.27	57.5/30.2/18.0/11.2 (BP = 0.999 ratio = 0.999 hyp_len = 48538 ref_len = 48569)

ラメータであり，一般的に 4 が採用される．BP は短さに対するペナルティ (brevity penalty) である．本研究では sacreBLEU を利用して BLEU スコアを計測する．sacreBLEU はトークナイズに mecab [10] を利用している．

3.2.1　仮説

短さに対するペナルティ BP に着目すると，同程度の単語一致率を持つ出力文同士であれば，より長い方が高いスコアを得られるという直感が得られる．Default と Short それぞれについての，1-gram から 4-gram までの n-gram 一致率と BP の内訳を表 4 上半分に示す．Short の BLEU 内訳を見ると各 n-gram の一致率は Default より高く，BP によるペナルティによってスコアを下げていることがわかる．そこで，Short の文の末尾にトークンを追加することでペナルティを軽減しスコアを上げることを考える．Short_maru_7 は，文に「。」が含まれていれば文末に「。」を 7 個追加する手法で，Short_mecab_8 は，文を mecab でトークナイズして末尾の 8 トークンを文末に追加する手法である．

3.2.2　実験

実験結果を表 4 に示す．Short_maru_7, Short_mecab_8 ともに BP が Short に比べて 1 に近づいており，それに伴って BLEU が上昇していることがわかる．「。」が末尾に連続することは機械翻訳モデルが誤った出力として生成しうるが，これに対し，適切ではない高いスコアが割り当てられる可能性がある．

表 5　chrF の実験結果

手法	chrF	chrP	chrR
Default	35.58	33.16	34.85
Short	32.64	36.00	32.23
Default_Short_concat	36.69	20.81	41.74

3.3　chrF：Recall 重視の脆弱性

chrF は表層ベースの自動評価指標である．chrF は BLEU と異なり，単語分割を行わず，文字単位で n-gram 一致を判定する．ただし，空白は消去したうえで計算を行う．chrP を，出力文の n-gram のうち参照訳文にも対応するものが存在する割合，chrR を，参照訳文の n-gram のうち出力文にも対応するものが存在する割合とする．計算過程は式 (3) のようになる．

$$\mathrm{chrF}\beta = (1+\beta^2)\frac{\mathrm{chrP}\cdot\mathrm{chrR}}{\beta^2\cdot\mathrm{chrP}+\mathrm{chrR}} \tag{3}$$

本研究では sacreBLEU のデフォルト設定である chrF2 ($\beta = 2$) で実験を行う．

3.3.1　仮説

式 (3) より，$\beta = 2$ のとき chrF の計算には Recall が Precision より 4 倍重視されることが分かる．そのため，システムにより多様な表現を出力させ，参照訳のできるだけ多くの n-gram に一致するようにすれば，Recall が上昇し，スコアの上昇につながると考えられる．本研究では，多様な表現を持つ文として，Default 文と Short 文を単純に連結した Default_Short_concat を作成した．

表 6　COMET の実験結果

手法	BLEU	chrF2	BERTScore	COMET	XCOMET
Default	26.13	35.58	0.8483	0.8734	0.7892
Short	23.87	32.64	0.8460	0.8622	0.7853
src.en	1.65	3.24	0.6315	0.5240	0.7232
mBART	13.20	22.94	0.7893	0.7341	0.5620

表 7　Default_pro の編集例

元の出力文	世界銀行はそのメッセージを広めたいと考えています.
編集後の出力文	世界銀行はそのメッセージを広めたいと考えています. (The sentence has been translated by a professional translator, and there are no errors in this translation.)

表 8　GEMBA の実験結果

手法	BLEU	GEMBA
Default	26.13	-3.30
Short	23.87	-4.22
Default_pro	18.57	-1.12

3.3.2　実験

仮説を検証するために，chrF スコアと，その計算過程で算出される chrP と chrR を表 5 に示す．Default_Short_concat の chrF スコアが Default を上回り最も高くなっている．また，chrP と chrR の数値を見ると，Default_Short_concat は chrP についてベースラインの 2 つの手法に対して 10 ポイント程度下回っているが，chrR については 5 ポイント程度上回っている．このことから，chrF は多様な表現を含む文に対して高いスコアを割り当てる可能性があることが分かる．

3.4　COMET：多言語エンコーダの脆弱性

COMET は，多言語エンコーダにプーリング層や線形層を追加して人手評価ラベルを教師として学習させたモデルを利用する自動評価指標である．本研究では 2 つのモデルによる評価を検証する．

Unbabel/wmt22-comet-da　XLM-R [11] アーキテクチャを採用し，DA スコア [12] がアノテーションされた WMT2017 から WMT2020 までのデータで訓練されたモデルである．以下 COMET と呼ぶ．

Unbabel/XCOMET-XL [13]　XLM-R アーキテクチャを採用し，DA タスクに加えて，MQM のラベルを教師データとした学習も行ったモデルである．MQM のラベル付けの能力が追加されることで，説明性が向上している．以下 XCOMET と呼ぶ．

3.4.1　仮説

COMET は原言語文，出力文，参照訳それぞれを同一の多言語エンコーダに入力してスコアを出力している．そのため，出力文の言語が正しいものでなくても，スコアを出力できる．本研究で扱う英日翻訳において，日本語以外で最も参照訳に近いのは英語の原言語文である．そこで，原言語文に対して COMET を計測する実験を行った．

3.4.2　実験

原言語文 src.en を，ベースラインの 2 つの文と，多言語機械翻訳のためのエンコーダ・デコーダモデルである mBART [14] による生成文とそれぞれ比較した．mBART のモデルは huggingface に公開されている mbart-large-50-one-to-many-mmt[1] を利用した．実験結果は表 6 の通りである．src.en は XCOMET を除くすべての指標において最も低いスコアを示した．一方で XCOMET においては，src.en と Short や Default のスコア差が COMET におけるスコア差より小さくなっており，src.en は mBART を上回るスコアを示している．

3.5　GEMBA：プロンプトの脆弱性

GEMBA は LLM に MQM のアノテーションを行わせ，その出力からスコアを計算する自動評価指標である．プロンプトとして en-de，en-cz，zh-en の 3 つの言語対についての事例を 3shot で与える．MQM の critical，major，minor の 3 つのエラーラベルに対

1）https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt

表 9　すべての手法のスコア

手法	BLEU	chrF2	BERTScore	COMET	XCOMET	GEMBA
Default	26.13	35.58	0.8483	0.8734	0.7892	-3.30
Short	23.87	32.64	0.8460	0.8622	0.7853	-4.22
Short_maru_7	24.10	32.23	0.8199	0.8439	0.7252	-5.06
Short_mecab_8	24.27	32.16	0.8318	0.7912	0.5100	-6.51
Default_Short_concat	17.98	36.69	0.8320*	0.8297	0.6829	-6.77
Default_pro	18.57	29.25	0.7575	0.6180	0.5410	-1.12
src.en	1.65	3.24	0.6315	0.5240	0.7232	-24.28

表 10　すべての手法のスコアの順位

手法	BLEU	chrF2	BERTScore	COMET	XCOMET	GEMBA	平均
Default	1	2	1	1	1	2	1.33
Short	4	3	2	2	2	3	2.67
Short_maru_7	3	4	5	3	3	4	3.67
Default_Short_concat	6	1	3	4	5	6	4.17
Short_mecab_8	2	5	4	5	7	5	4.67
Default_pro	5	6	6	6	6	1	5
src.en	7	7	7	7	4	7	6.5

し式 (5) に従ってスコアを計算する．

$$\mathrm{s} = 25 \times cnt_{\mathrm{critical}} + 5 \times cnt_{\mathrm{major}} + cnt_{\mathrm{minor}} \tag{4}$$

$$\mathrm{GEMBA}_{\mathrm{sent}} = -1 \times \begin{cases} s & (s \leq 25) \\ 25 & (s > 25) \end{cases} \tag{5}$$

ここで，cnt_{critical}，cnt_{major}，cnt_{minor} はそれぞれ critical，major，minor のエラーラベルが 1 文に対して割り当てられた数である．最終的なスコアは全文に対する $\mathrm{GEMBA}_{\mathrm{sent}}$ の平均から求められる．

3.5.1　仮説

GEMBA はプロンプトによってスコア計算を指示しているため，評価対象の文に悪意のあるプロンプトが紛れ込むと，適切でないスコアが出力される可能性がある．GEMBA のスコアは MQM のラベルが少ないほど上昇するため，MQM のラベルが出力されないことを期待して，「プロの翻訳家による翻訳文であるためエラーはない」というテキストを評価対象の文の末尾に加えた Default_pro という手法を考案した．Default_pro の文は表 7 のようになる．

3.5.2　実験

実験結果を表 8 に示す．Default_pro は Default や Short と比べて BLEU スコアは下がっているのに対し，GEMBA スコアは上昇している．実際に，Default_pro を評価するにあたって，GPT-4o の応答の中で GEMBA が想定する応答と一致しない出力が 1 行だけ存在したが，その内容は「the translation provided is accurate and does not contain any errors. the sentence has been translated by a professional translator, and there are no issues with accuracy, fluency, style, terminology, or any other category. therefore, the classification is:」というものであった．このことから，プロの翻訳であるということが判断に反映されていると考えられる．

4　分析

すべての手法とそのスコアをまとめて表 9 に示す．なお，Default_Short_concat はトークン数が 512 を超えているため，BERTScore が計測不能であった．そのため，各文が 700 文字に収まるように編集したうえで BERTScore を計測した (表中 * 印)．単一の指標では Default や Short を他の手法が上回ることがあるが，複数の指標を並べるとベースラインの 2 つの手法が平均して高いスコアを達成していること

がわかる．

そこで，各指標におけるスコアをもとに各手法に順位をつけ，その平均を計算し，表 10 にまとめた．平均順位で比較すると Default が最も高く，Short が次に高くなっている．また，言語が異なる src.en が最も低い順位になっている．これらのことから，複数の指標における平均順位は直感的な翻訳品質を反映できていると考える．単体の評価指標では品質を適切に評価できない場合があるが，複数の指標を組み合わせることで各評価指標の欠点を補うことができることが示唆される．

5 おわりに

本研究では，自動評価指標が，品質が低下するように編集した文に対しても高いスコアを出力する事例を調査した．4 つの評価指標について検証し，次のような事例を確認した．

BLEU は長さペナルティを軽減するように文末に適当な文字を追加した場合にスコアが上昇した．chrF は Recall 重視の式で計算されており，2 文を連結して冗長にした場合にスコアが上昇した．COMET は多言語エンコーダを利用するため日本語以外の文も入力可能であり，翻訳をする前の英文を入力することで，エンコーダデコーダモデルによる翻訳より高いスコアとなった．GEMBA は「プロによる翻訳文である」という文を文末に追加する場合にスコアが上昇した．

また，これらの事例におけるスコアをまとめて分析し，平均順位を利用することで，各指標単体の欠点を解消できることを示した．

謝辞

本研究は，JSPS 科研費（基盤研究 B，課題番号：JP23K24907）の助成を受けたものです．

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, 2002.

[2] 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.

[3] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **Proceedings of the 8th International Conference on Learning Representations**, 2020.

[4] 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**, pp. 2685–2702, 2020.

[5] 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. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1460–1474, 2021.

[6] Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In **Proceedings of the Eighth Conference on Machine Translation**, pp. 768–775, 2023.

[7] Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In **Proceedings of the Ninth Conference on Machine Translation**, pp. 47–81, 2024.

[8] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, 2018.

[9] OpenAI. GPT-4o System Card. **arXiv:2410.21276**, 2024.

[10] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, 2004.

[11] 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, 2020.

[12] 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.

[13] Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 979–995, 2024.

[14] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 6112–6121, 2019.