高精度翻訳モデルのための自動評価手法の検討

今城 健太郎 *　平野 正徳 *
株式会社 Preferred Networks
imos@preferred.jp　research@mhirano.jp

概要

機械翻訳システムの品質向上に伴い、高精度なモデルにおいて、モデル間の性能差を識別することが難しくなっている。そこで、本研究では、高精度なモデルに対しても十分な分解能を持ち、再現性と計算効率に優れる表層マッチング型評価手法を検討する。まず、英和・和英それぞれ約 150 語程度の原文 50 件に対し、評価対象のモデルによって生成された翻訳文群を作成した。このデータセットを評価軸として、新たな表層マッチング指標について検討した。提案指標は、言語非依存性のため文字 *n*-gram 特徴量を用い、言い換えの多様性を吸収するため多数の参照訳を活用する。最大 1,000 件の生成参照訳を用いた評価では、参照訳の数の増加に伴い提案指標は一貫して性能向上を示し、BLEU および chrF を上回った。また、LLM を活用しつつ人手で作成した高品質参照訳を用いた実験では、提案指標を含む表層マッチング指標が、WMT24 で上位の指標である MetricX-24 より高い順位相関を示した。これらの結果から、表層マッチングに基づく軽量かつ安定した自動評価が可能であることが示唆された。

1　はじめに

ニューラル機械翻訳および大規模言語モデル (LLM) の発展により、機械翻訳の品質は人間の翻訳に迫る水準に達しつつある。一方で、高精度なシステム同士を比較する場面では、従来の自動評価指標では性能差を十分に識別できない [1, 2]。したがって、高精度領域での差分を捉えられる評価手法の整備が不可欠である。

翻訳品質の最終的な信頼基準として、人手評価は依然重要である。Multidimensional Quality Metrics（MQM）[3] のように誤りタイプを分類・集約する枠組みは、訳抜けや誤訳などの客観的な誤り検出に優れる。しかし、大規模な MQM 評価の分析からは、評価者間・評価方針間のばらつきや文脈を考慮した評価の難しさが報告されており [1]、人手評価は高コストでスケーラブルでなく、「良い翻訳」を一意に定めることにも限界がある。

コストと再現性の観点からは、参照訳との表層的一致に基づく自動評価指標が広く用いられてきた。BLEU[4] は単語 *n*-gram の一致度と長さに基づく古典的指標であり、chrF[5] は文字 *n*-gram に基づき分かち書きされない言語にも比較的頑健とされる。しかし、これらは局所的な *n*-gram 一致に依存するため、言い換えや語順の自由度を十分に扱えず、高精度システム同士の差分を正確に反映できないことが示されている [6, 2]。さらに、参照訳の品質と数量が指標性能に大きく影響し、高品質な参照を多数用意できれば性能を改善できる一方、人手のみでの構築はコスト面で困難である [6, 7]。

高精度な意味的評価を実現するため、評価に事前学習言語モデルを用いるモデルベース指標も広く研究されている。COMET [8] は原文・参照訳・候補訳を入力とする評価モデルを学習し、人手評価との高い相関を達成している。WMT Metrics Shared Task では、mT5 を基盤とする MetricX 系列 [9] が上位の性能を示す。また、LLM を評価者として用いる LLM-as-a-judge においても、Kocmi ら [10] は GPT 系モデルを用いた GEMBA により最先端指標と同等以上の相関を報告している。しかし、これらは大規模なモデルの推論を必要とし、プロンプトやモデル更新に起因する再現性の問題もあるため、日常的なベンチマークの基盤とするには課題が残る。

近年の LLM は、高品質な翻訳候補の大量生成や候補間の相対的な優劣判断に優れている [10] ため、その能力を効果的に活用するようなパイプラインを構築することで「多数の高品質参照訳」や「高精度だがコストの高い基準」を半自動的に構築すること

* Equal contribution
Full paper: https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/2428

の現実性が高まりつつある。多数の参照訳が利用可能であれば、軽量な表層マッチング指標であっても言い換えの多様性を吸収し、高精度翻訳モデル間の微小な差分を識別できる可能性がある。このような枠組みは、高コストな人手評価やモデルベース指標と、軽量で再現性の高い表層マッチング指標との間を橋渡しすることが期待される。

本研究では、以上を踏まえ、高精度な翻訳モデル同士の性能差を、軽量かつ再現性の高い表層マッチング型自動評価により識別することを目指す。

2 提案手法

本研究の目的は、高精度な翻訳モデル同士の微小な性能差を、軽量かつ再現性の高い表層マッチング指標により識別することである。そのため、(i) 言語やトークナイザに依存しない文字 n-gram 特徴量を用いること、(ii) LLM により多数生成可能な参照訳集合を前提とし、参照訳の数の増加に伴って評価性能が効率的に向上するようスコアを設計すること、の 2 点を基本方針とする。以下では、文字 n-gram 特徴量を定義し、多数の参照訳を分布として扱う類似度スコアと長さに対する罰則項を説明する。

2.1 文字 n-gram に基づく特徴量

形態素単位やサブワード単位のトークナイザは言語や実装に依存し、評価指標の安定性・再現性を損なう要因となりうる。そこで本研究では、言語非依存な表現として文字 n-gram を採用する。

文章 x を $\ell(x)$ 文字からなる文字列とし、その中に現れる任意の文字 n-gram を w とする。実用上は長さ $n \le N_{\max}$ の n-gram のみを考え、本研究では $N_{\max} = 20$ とする。w が翻訳候補 x に出現する回数を $c_w(x)$ とおき、w が少なくとも i 回出現することを表す二値特徴量 $f_{w,i}(x)$ を

$$f_{w,i}(x) = \begin{cases} 1 & (c_w(x) \ge i) \\ 0 & (c_w(x) < i) \end{cases}, \quad i = 1, 2, \dots \tag{1}$$

と定義する。w を固定したときに $f_{w,i}(x) = 1$ となるのは、w 出現回数と一致するため、$f_{w,i}(x) = 1$ となる w、i の全ての組み合わせは $\ell(x) \times N_{\max}$ 個となる。下記では、この $f_{w,i}(x)$ を文章の特徴量とみなす。

2.2 類似度スコアと長さに対する罰則

評価対象の翻訳候補 x に対する最終スコア $S(x)$ は、参照訳集合 $R = \{R_1, \dots, R_M\}$ との特徴量の内積に、翻訳候補の文字列長に応じた罰則係数 C_{len} を掛け合わせて定義する。翻訳候補 x の文字列長を $\ell(x)$、n-gram w の長さを $\ell(w) = n$ とすると、

$$S(x) = C_{\text{len}}\big(\ell(x); \ell_R\big) \sum_{w,i,j} \frac{f_{w,i}(x)\, f_{w,i}(R_j)}{\ell(w)} \tag{2}$$

となる。ここで和は、$n \le N_{\max}$ を満たす全ての n-gram w、各 w の出現回数に対応する正の整数 i、および参照訳番号 $j = 1, \dots, M$ に渡って取る。

分母の $\ell(w)$ は、長い n-gram ほど情報量が高い一方で出現頻度が低いことを踏まえ、各 n-gram の寄与をその長さでスケーリングするための重みである。これにより、$N_{\max}$ を大きくしても参照訳とほぼ同一の翻訳候補に過大なスコアが与えられることを抑制できる。

一方、翻訳候補が過度に長い場合には、M 個の参照訳それぞれと多くの n-gram を共有しやすくなり、共通 n-gram に基づくスコアが実際の翻訳品質よりも過大になる恐れがある。この影響を抑えるため、参照訳集合 R における長さの中央値 $\ell_R = \text{median}_{r \in R}\big[\ell(r)\big]$ を用いて

$$C_{\text{len}}\big(\ell(x); \ell_R\big) = \min\left(1,\ \frac{\ell_R}{\ell(x)}\right) \tag{3}$$

と定義する。すなわち $\ell(x) \le \ell_R$ では $C_{\text{len}} = 1$ とし、$\ell(x) > \ell_R$ では $\ell(x)$ に反比例して係数を小さくすることで、不自然に長い翻訳候補のスコアを抑制する。なお、翻訳候補が極端に短い場合には、参照訳と共有する n-gram 自体が少なくなり、式 (2) の内積部分が小さくなることでスコアが低下する。

以上より、提案指標は（i）文字 n-gram に基づく言語非依存な特徴量によりトークナイザへの依存を排除し、（ii）多数の参照訳に含まれる n-gram を頻度分布として扱うことにより LLM が生成した多様な参照訳集合を活用しつつ、（iii）長さの罰則により冗長な表現による過大評価を防ぐよう設計した。

3 データセット

3.1 英和・和英翻訳データセット

英語から日本語（英和）および日本語から英語（和英）の 2 方向について、評価用の原文をそれぞれ 50 個ずつ用意した。各原文の長さに関しては、与えられた文章のみから背景が十分に理解できる必要最小限の長さであることが好ましいと考え、英語・日本語ともにおおよそ 150 語相当となるよう調整した。

翻訳難易度と文体の多様性を確保するため、ニュース、Wikipedia、Web テキスト、IT 技術文書、学術文書、法律関係の文書（法律、契約書、特許など）、専門分野（医療、金融など）特有の文書、質問応答、物語文、固有名詞を多く含む文など、10 ジャンルから各 5 文ずつ選定した。

3.2 高品質な参照訳

高品質な参照訳は、LLM を活用しながら人手で作成した。まず複数の LLM に、各原文に対する翻訳候補と改善案の提案を行わせ、最終的に人手でそれらを確認・統合・修正して参照訳を作成した。

3.3 生成参照文訳群

多数の参照訳を用いた表層マッチング評価を行うため、各原文に対して LLM による生成参照訳を用意した。具体的には、Preferred Networks が開発する事前学習モデル PLaMo 3 NICT 31B Base [1] に対して翻訳プロンプトを用い、温度 1 でサンプリングを行い、各原文ごとに多数の翻訳候補を生成した。

生成された翻訳候補には意味的に大きく逸脱した訳や文法的に崩れた訳が含まれる可能性があるため、文字 *n*-gram の出現頻度に基づきフィルタリングを行った。

4 実験と結果

実験においては、複数の LLM を利用する。ここで、評価者として活用する LLM を 7 つ用意する。これを評価モデルと呼ぶ。モデルの詳細については Appendix の表 2 を参照されたい。

さらに、評価の対象となる、つまり、ベンチマークスコアを計算する対象となる被評価モデルを 10 個用意し、それぞれのモデルで 50 個の原文に対する翻訳を作成する。この 50 個の翻訳された文章群を訳文群 A〜J と表記し、訳文群と呼ぶ。こちらの被評価モデルについては、本研究ではモデル自体の性能比較を目的としていないため、モデル名は本稿では言及しないこととする。

本実験において確認したいことは、より適切に訳文群 A〜J の翻訳性能の優劣を効率よく評価できるかどうかである。

その目的のために、まず複数の評価モデルを用いた全対比較により翻訳優劣データセットを構築し、評価モデル間の一貫性を分析する。そのうえで、構築したデータセットを「高精度だがコストの高い基準」とみなし、提案指標を含む各種自動評価指標との相関を比較する。

4.1 翻訳評価の正解データ構築と検証

翻訳の優劣を決定するため、各原文に対して 10 個の被評価モデルから得られた訳文群 A〜J を評価対象とした全対比較を実施する。7 個の評価モデルを用いて、原文と 2 つの翻訳候補を提示し、どちらが優れているかを選択させる全対比較設定を採用した。全ての組み合わせに対して提示順を入れ替えた 2 通りの比較を行い、1 原文あたり計 90 通りの比較を得た。

各原文について、訳文群が他の訳文群との比較で選好された割合を勝率と定義し、これを全対比較評価のスコアとした。つまり、10 訳文群のうち、自身の訳文群以外の 9 個の訳文群の翻訳に対する勝率をスコアとした。

この操作を 50 個の原文それぞれに対して実施し、その勝率のスコアを訳文群ごとに計算することで、訳文群の良さを示す評価値を計算できる。

訳文群のスコアの順位における、各評価モデルの Spearman の順位相関を計算したところ（詳しくは Appendix に掲載の表 2）、基準モデル群 (1) と公開モデル群 (2) の相関は英和で約 0.87、和英で約 0.85 と高く、個々の LLM 同士の相関も多くの組み合わせで 0.75 以上であった。

以上より、LLM による全対比較に基づき構築したデータセットは、高い一貫性を持っており、人手評価に代わる高信頼な評価基準として利用可能であると判断し、以降の実験では基準モデル群による順位を正解として自動評価指標の比較に用いる。

4.2 提案手法の評価

本節では、前節で構築した評価基準（基準モデル群）との Spearman の順位相関を評価軸とし、提案指標と既存指標の性能を比較する。既存指標としては、表層マッチング型の BLEU、chrF、提案指標に加え、WMT24 Metrics Shared Task 上位の指標である MetricX-24 を用いる。

表 1 は、3 章で構築した高品質参照訳または生成参照訳群を用いた BLEU、chrF、提案指標の性能を示す。高品質参照訳を用いて評価する場合には、1 つしかない正解に対して評価値を計算するのに対し、生成参照訳群を用いて評価する場合には、前出

1） https://huggingface.co/pfnet/plamo-3-nict-31b-base

表 1　各スコア評価手法の翻訳評価性能

(a) 英和翻訳評価

参照訳	高品質	生成		なし
	単一	単一	複数	
MetricX	.656	-	-	.500
BLEU	.769	.577	.584	-
chrF	.789	.584	.634	-
提案手法	.778	.562	.653	-

(b) 和英翻訳評価

参照訳	高品質	生成		なし
	単一	単一	複数	
MetricX	.638	-	-	.412
BLEU	.618	.376	.486	-
chrF	.655	.390	.440	-
提案手法	.650	.400	.492	-

の通り、多数の参照訳を活用する余地がある。そのため、生成参照訳群を用いて評価する場合には、そのうちランダムに選んだ 1 つのみを使って評価する場合と、複数を用いて評価する場合の 2 パターンの実験を示している。

表 1 によれば、高品質参照訳を用いた場合、英和・和英のいずれでも提案指標および chrF は、同じ参照訳を与えた MetricX-24 を上回り、参照訳が単一であっても提案手法も含めた表層マッチング型指標がモデルベース指標と同等以上の精度を達成しうることが分かる。

また、生成参照訳を 1 件だけ用いる設定（表 1 の「生成・単一」列）では、高品質参照訳と比べ全ての指標で相関が大きく低下し、参照訳の品質が評価性能に強く影響することが再確認できる。一方、多数の生成参照訳を用いて提案指標を計算すると、参照訳なしの MetricX-24 を上回る相関が得られ、表層マッチングのみでも高い評価精度が実現できることが分かる。

次に、生成参照訳群の大きさの影響を検証する。生成参照訳は各原文あたり最大 1,000 件存在するため、そこからランダムに k 件を選択して評価指標を計算し、k を変化させたときの性能を調査した。図 1 は生成参照訳の数 k と順位相関の関係を示す。参照訳が少ない場合には chrF がやや優位なこともあるが、$k \geq 32$ では提案指標が BLEU、chrF のいずれよりも高い相関を維持した。

以上の結果より、(i) 高品質参照訳が利用できる場合には表層マッチング型指標が MetricX-24 と同等以

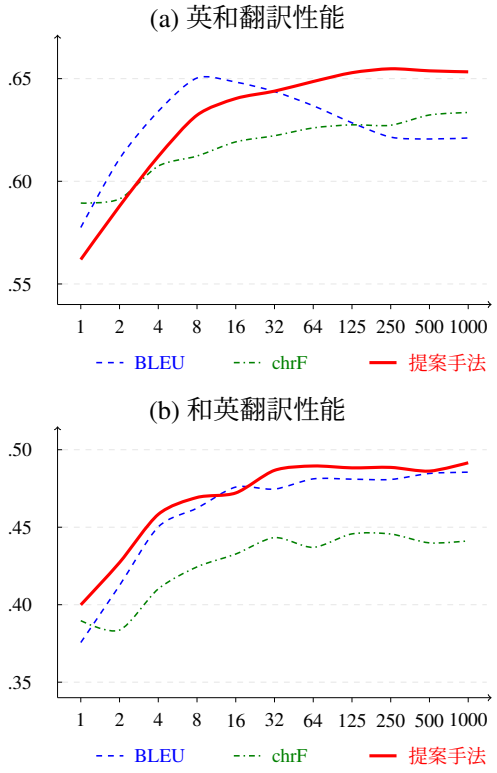

図 1　参照訳の数と翻訳性能

上の性能を示し、(ii) 高品質参照訳を用意できない場合でも LLM による生成参照訳を多数用いることで、提案指標が BLEU や chrF より安定して高い性能を達成することが示された。

5　まとめと今後の課題

本研究では、LLM で参照訳が多数生成できることを前提とし、その参照訳を活用した文字 n-gram 特徴量ベースの表層マッチング型指標を提案した。また、LLM を評価者とする全対比較により翻訳優劣データセットを構築し、これを人手評価に代わる評価基準として用いることで、提案手法を含む表層マッチング型指標とモデルベース指標の性能を比較した。その結果、参照訳の品質と数量が評価性能に大きく影響すること、多数の生成参照訳を前提とした場合には提案指標が BLEU や chrF を上回り、参照訳なしの MetricX-24 と同等以上の水準に達することが分かった。

一方で、本手法は多数の参照訳の生成を前提としており、参照訳の生成・選別に要するコストやバイアスの分析は今後の課題である。最終的には、多数の高品質参照訳を前提とした表層マッチング型指標のみで、モデルベース指標と同等以上の評価精度を目指すことが今後の課題である。

参考文献

[1] 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.

[2] Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. Results of the WMT20 metrics shared task. In **Proceedings of the Fifth Conference on Machine Translation**, Online, 2020. Association for Computational Linguistics.

[3] Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Technical report, DFKI, 2014.

[4] 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, USA, 2002. Association for Computational Linguistics.

[5] 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, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[6] Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be guilty but references are not innocent. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 61–71, Online, 2020. Association for Computational Linguistics.

[7] Vilém Zouhar and Ondřej Bojar. Quality and quantity of machine translation references for automated metrics. **arXiv preprint**, Vol. arXiv:2401.01283, , 2024.

[8] 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**, Online, 2020. Association for Computational Linguistics.

[9] Google Research. Metricx-24: The google submission to the WMT 2024 metrics shared task. In **Proceedings of the Ninth Conference on Machine Translation**, 2024. Author list omitted; see ACL Anthology id 2024.wmt-1.35 for full metadata.

[10] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In **Proceedings of the 24th Annual Conference of the European Association for Machine Translation**, pp. 193–203, Tampere, Finland, 2023. European Association for Machine Translation.

A 翻訳評価モデルが 10 個の翻訳について優劣をつけたときの順位相関

表 2 に示す。

表 2　翻訳評価モデルが 10 個の翻訳について優劣をつけたときの順位相関

(a) 英和翻訳

評価モデル	(1)	(1-1)	(1-2)	(1-3)	(2)	(2-1)	(2-2)	(2-3)	(2-4)
(1) 基準モデル群 (平均)	1	.939	.930	.890	.868	.843	.809	.771	.691
(1-1) - Claude Opus 4.5	.939	1	.857	.799	.837	.812	.785	.746	.644
(1-2) - Gemini 3 Pro	.930	.857	1	.750	.814	.840	.746	.731	.605
(1-3) - GPT-5.1	.890	.799	.750	1	.815	.745	.762	.702	.720
(2) 公開モデル群 (平均)	.868	.837	.814	.815	1	.905	.903	.878	.812
(2-1) - DeepSeek V3.2	.843	.812	.840	.745	.905	1	.820	.786	.651
(2-2) - Qwen3 235B	.809	.785	.746	.762	.903	.820	1	.769	.703
(2-3) - Kimi K2	.771	.746	.731	.702	.878	.786	.769	1	.656
(2-4) - gpt-oss-120b	.691	.644	.605	.720	.812	.651	.703	.656	1

(b) 和英翻訳

評価モデル	(1)	(1-1)	(1-2)	(1-3)	(2)	(2-1)	(2-2)	(2-3)	(2-4)
(1) 基準モデル群 (平均)	1	.936	.906	.856	.850	.763	.761	.761	.690
(1-1) - Claude Opus 4.5	.936	1	.818	.747	.802	.704	.707	.738	.649
(1-2) - Gemini 3 Pro	.906	.818	1	.704	.802	.770	.710	.766	.611
(1-3) - GPT-5.1	.856	.747	.704	1	.779	.685	.719	.658	.696
(2) 公開モデル群 (平均)	.850	.802	.802	.779	1	.876	.869	.846	.786
(2-1) - DeepSeek V3.2	.763	.704	.770	.685	.876	1	.750	.738	.606
(2-2) - Qwen3 235B	.761	.707	.710	.719	.869	.750	1	.721	.609
(2-3) - Kimi K2	.761	.738	.766	.658	.846	.738	.721	1	.579
(2-4) - gpt-oss-120b	.690	.649	.611	.696	.786	.606	.609	.579	1