文芸翻訳における機械翻訳の利用可能性の検討

藤井俊英 [1]　影浦峡 [2]
[1] 東京大学大学院学際情報学府　[2] 東京大学大学院教育学研究科
fujiit1234@g.ecc.u-tokyo.ac.jp kyo@p.u-tokyo.ac.jp

概要

本研究では，文芸翻訳，とりわけ推理小説の翻訳における機械翻訳の利用可能性を検討した．推理小説の短編を対象に，DeepL，みんなの自動翻訳@TexTra，GPT-5 の訳文を比較するために，ジャンル特性を反映し構築した人手評価スキームを適用した．その結果，GPT-5 ではエラー数が最も少なく一貫性も維持された一方，情報の再提示を省略する傾向が課題として確認された．DeepL では記号処理と訳抜け，TexTra では語義の誤りや不自然さ，代名詞の誤解が特徴的にみられた．算出された翻訳スコアを確認すると，現段階では人手訳の最低水準に届かず，利用には課題が残ることを示す結果となった．

1　はじめに

これまで，機械翻訳が対象とする領域は主として産業翻訳であり，文芸翻訳はほとんど対象とされてこなかった．しかし機械翻訳の品質が年々向上することを受けて，機械翻訳の文芸翻訳への利用可能性が検討され始めた．たとえば，2024 年の WMT General MT Shared Task[1] から，その翻訳対象に文芸テクストが使用され始めている．

このような潮流のなかで重要な課題となっているのは，文芸機械翻訳をいかに分析し，評価するかである．一般的な機械翻訳の人手評価には，多くの場合 MQM[2] が用いられてきた．しかし文芸機械翻訳に対して MQM を適用することには疑義が呈されている [3]．また文芸機械翻訳の人手評価には SCATE taxonomy[4] が用いられることも多いが，研究によっては改良を加えたうえで運用される場合もある [5]．以上を踏まえると，文芸機械翻訳の評価手法は現在なお多様なアプローチが検討されている段階にあると言える．

そこで本研究では，文芸翻訳領域，特に推理小説の翻訳への機械翻訳の利用可能性について明らかにすることを試みる．対象を推理小説に絞るのは，文芸機械翻訳の評価にはジャンルの観点が必要であると考えられるからである．

2　評価スキームの構築

本研究でははじめに文芸機械翻訳を評価するための人手評価スキームを構築した．

本研究で構築した評価スキームは，文芸機械翻訳について正確かつ多角的な分析を可能にすることで，機械翻訳が文芸領域で到達している品質をより明晰に示すことを主目的としている．そのために文芸小説と一括りにされる傾向にあったドメインに対し，ジャンル固有の要請が評価されていなかった点を補い，また日本語特有の表現についてのエラーについて捕捉可能であるように設定した．

評価項目は大きく三つの項目に分けられる．翻訳文単体の読みやすさを観察する Monolingual Categories，原文との差異としての翻訳エラーを観察する Multilingual Categories，ジャンルの観点からエラーを分析する Genre Cetegories の三つである[1)]．本研究では翻訳テクストに推理小説を用いたため Genre Categories には推理小説に特化したエラー項目を設定した．これらの下位項目にエラーの性格を示す大項目があり，さらにその下位項目に具体的なエラーを示す小項目がある．アノテーションに用いる項目として計 40 の項目を設定した．付録 (表 5) にスキームの全体を評価結果とともに示す．

本研究では再現性を担保する試みとして項目の付与に決定木を導入し，さらには評価フローを明確に示した．発見されたエラーには重大度 (minor/major/critical) を付与し，エラーごとに重み付けをおこなっている．そうすることにより MQM と同様の手順 [6] で文書全体の翻訳スコアを算出できるようにした．これによりシステム間の比較や改善点の抽出が可能になり，翻訳品質を定量と定性の両面で評価することができる．

1) 表 1 において SP〜RD までが Monolingual，SE〜MT までが Multilingual，SG と CB が Genre Categories である．

3 評価の実施

評価は本論の第一著者自身が実施した．評価対象テクストは，James Yaffe “Mom Knows Best” [7] とし，参照訳には「ママは何でも知っている」[8] を用いた．これは推理小説の短編であり，ジャンルの観点から評価する際にふさわしいテクストと考えられる．

機械翻訳には DeepL[2] のクラシック言語モデル，みんなの自動翻訳@ TexTra[3] の汎用 UM，GPT-5[4] を API から用いている．文書全体での一貫性の評価をおこないたいため，可能な限り文書全体を一括入力する．DeepL と GPT-5 は全文を，TexTra は段落ごと入力している．TexTra のみ段落ごとに入力したのは，入力可能文字数の制限があったためである．

4 評価結果

4.1 エラー分析

評価の結果を表 1 にまとめた．小項目単位での結果は付録内の表 5 に補足している．

まず総エラー数[5] を確認すると GPT-5 がエラー数の少なさで他 2 つを大きく上回っている．次いで TexTra，DeepL の順にエラー数が少なかったことが分かる．

エラー数が最も多かった DeepL では，SE が 190 と全体の大半を占め，訳抜けや体裁の乱れが品質を大きく押し下げていることが観察された．一方 TexTra は SE が 15 と少なく，DeepL とは逆に表層面においては比較的安定している．一方で RD が 47，MT[6] が 50 と突出して多く，文章としての自然さと意味の正確さに課題が残った．GPT-5 は SE の 18 と MT の 14 が主だったエラーであり，特定項目に極端に偏らず小さく分散している点が特徴的である．

表 2 に示すように，重大度においても示される品質の傾向は一致している．Critical は DeepL の 13，TexTra の 5 に対し GPT-5 は 0 で，Major も GPT-5 の 12 が最少，DeepL の 89 が最多となっている．

	Error Categories	DeepL	TexTra	GPT-5
SP	Spelling	0	0	0
IC	Inconsistency	14	21	0
RD	Readability	23	47	11
SE	Surface Error	190	15	18
UN	Untranslated	0	2	0
DN	Do Not Translate	0	0	0
LP	Logical Problem	11	14	4
GR	Grammer	3	12	1
RG	Register	2	2	1
MT	Mistranslation	25	50	14
SG	Single	0	0	0
CB	Combination	15	6	5
	Total	264	163	49

表 1: 大項目ごとの評価結果

Severity	DeepL	TexTra	GPT-5
Minor	152	117	37
Major	89	41	12
Critical	13	5	0

表 2: 重大度の結果

4.2 翻訳スコア

MQM におけるスコアリング [6] を参考に，重みづけされたアノテーション結果から翻訳スコアを算出した．その結果を表 3 に示す．翻訳スコアでも GPT-5 がと最良の結果を示しており，DeepL や TexTra のスコアを大幅に上回った．ここで CQS とは翻訳品質の基準点 (PT) を定めて規格化したスコアである．本研究では人手訳の最低水準を基準として PT = 80 [7] としている．すなわちもっともスコアが高かった GPT-5 であっても人手訳の最低水準には届かなかったことを意味している．

Metric	DeepL	TexTra	GPT-5
CQS (↑)	-132.23	-12.21	**75.65**

表 3: 翻訳スコアの算出

2） https://www.deepl.com/
3） https://mt-auto-minhon-mlt.ucri.jgn-x.jp/
4） モデルは gpt-5-2025-08-07 を用いている．プロンプトには “You are a professional translator. Please translate given English story into Japanese.” と入力した．
5） 表 1 において，CB は既にエラーが付与された箇所に対して付与するため total に計上していない．
6） 本稿では MT をエラーの項目名として用いる．
7） [9, p.24] ではこの基準点を翻訳者の 1 日当たりの作業で Critical が 1 つを超えたら不合格というかたちで定めている．

5 自動評価

本研究では人手評価スキームに加え，自動評価もおこなった．用いた自動評価指標は BLEU[10] ならびに COMET[11] である[8]．評価は文対ならびに段落対でおこなった．標準的には文対が用いられるが，複文単位でおこなわれることもあるためだ [14]．

結果は表 4 にまとめた．文対において BLEU は GPT-5 が最上位，COMET は TexTra が最上位で，DeepL は両指標で最下位となった．一方で段落対では BLEU は DeepL が最上位，COMET は GPT-5 が最上位で，TexTra が両指標で最下位となった．

Metrics	DeepL	TexTra	GPT-5
BLEU (Sent)	11.64	11.14	**11.99**
COMET (Sent)	0.8038	**0.8205**	0.8151
BLEU (Para)	**14.83**	12.72	13.46
COMET (Para)	0.8300	0.8269	**0.8314**

表 4: 自動評価の結果．(Sent) は文対を，(Para) は段落対を表す．

6 評価の分析

はじめに，算出された翻訳スコアをもとにそれぞれの機械翻訳システムについて全体的な評価をおこなう．

CQS を確認する限り，少なくとも本研究の評価対象であったテクストについては，GPT-5 による機械翻訳の品質が DeepL や TexTra と比較して良いことが示唆された．設定した基準点と比較しても，GPT-5 の機械翻訳はその他に比て明確に基準点に近い品質を出力している．このことは LLM を用いた機械翻訳の将来的な実用性を暗示しているようにも考えられる．一方で DeepL ならびに TexTra の CQS については，人手訳に対しては想定されていない負の値を取る結果となった．これは Critical なエラーが多かったことを表しており，あらためて人手翻訳の優位性を示すかたちとなった．

また翻訳スコアを用いることで，どの項目の品質を向上させれば，どの程度のスコアが上がるかというのが概算できる．たとえば，DeepL から記号のミスと Omission を除いた仮想計算では CQS は 54.31 まで改善する．これは，DeepL が意味の取り違えよりも欠落と形式で損をしている割合が大きいことを示唆している．一方で TexTra から定型的な訳出に関連するエラー[9]を取り除いた仮想計算をおこなっても CQS は 3.66 にしかならない．換言すれば，これは TexTra が特定のエラー項目に大きな弱点を抱えているわけではないことを示している．このように翻訳スコアを用いると，機械翻訳システムの改善をする際や，システムの利用における注意点として，どの部分にまず着目すればよいかを量的に理解することが可能になる．

次に人手評価スキームにおける大項目ごとの結果とその考察について述べる．項目を観察することによって，それぞれの機械翻訳システムによる翻訳出力の性格を確認することができる．

大項目ごとのエラーの内訳を見ると，DeepL は SE が 190 のエラーと突出しており，とりわけ記号のミスと Omission が多い．本研究で扱ったような推理小説では誰が誰になにを言ったかや情報提示の順序が小説としての骨格になるため，鍵括弧などの記号処理の乱れは単なる見た目の問題に留まらない．会話区間が曖昧になることで，読者は発話の主体と客体，証言の信頼性，手がかりの提示箇所などを誤って解釈しうる．また Omission のエラーは情報の欠落を示しており，こちらの小説としての瑕疵に直結する．事実ジャンルの観点からのエラーを示す CB のエラーは，そのほとんどが SE と共起していた．結果として DeepL では重大度が Major，Critical となるエラーが繰り返され，翻訳スコアでも極端に悪化したと考えられる．DeepL は特に当初訳文の「読みやすさ」や文体の「流暢さ」で一般に広まったシステムである．文芸翻訳も文体の「流暢さ」が重要なポイントとなるが，本評価では文芸翻訳でも，特に推理小説のジャンルでは，見かけ上の流暢さは翻訳の質を確保するものではないだけでなく，翻訳の質の低下と関係しうることが示唆される．

一方 TexTra は SE こそ 15 と比較的エラー数も少なかったが，RD が 47，MT が 50，GR が 12，LP が 14 とまんべんなくエラーが発生しているのが特徴的であった．RD の内訳では Lexical Choice と Awkward Style が多く，日本語の文章として不自然な出力が目立った．また MT におけるエラーの多さは単語の意味の取り違えの多さを表している．これは原文の語を機械的に対応させた結果によるものだと考えら

8） BLEU の評価を実行する際には，SacreBLEU[12] を用いた．COMET のモデルには wmt22-comet-da[13] を使用した．

9） RD5，RD10, ならびに MT2-5 にあたるエラーのことを指している．

れる．特に意味の取り違えの際には，よく用いられている意味に取り違えてしまうことが多く，これは定型表現に出力が偏っていることを示唆している．CB と共起したエラー項目を観察すると，GR のなかでも代名詞に関するエラーが目立った．特に TexTra は間接話法を上手く訳すことができておらず，主語と述語の対応に難がみられた．この結果は TexTra が基本的に技術的な文書を対象としたシステムとして想定されていることと対応していると考えることができる．上でエラーの観点から TexTra は特定のエラー項目に大きな弱点抱えているわけではないことを示していると述べたが，適用する文書の性質は評価において広い範囲で影響を与えることが示唆される．

最後に GPT-5 は IC のエラーが 0 であり，すなわち用語や文体の統一に関する破綻が観測されなかった．このことは文芸翻訳に採用する機械翻訳システムとして，その他の機械翻訳システムに対して大きな優位性を示している．また CB の下位項目を確認すると，GPT-5 は推理部分におけるエラーがないことが分かる．この結果は機械翻訳で懸念されていることのひとつであった，長距離での文脈を保持した翻訳の実現可能性を示唆しているといえよう．さらにはこのような結果に加え Critical が 0 であったことから，GPT-5 は翻訳を破綻をさせない方向に強いということがうかがえる．

ただし，GPT-5 も SE のエラーとして Omission を一定数含み，MT や RD のエラーも 0 ではない．すなわち細部の翻訳や語感の選択といった面で改善余地が残っている．加えて CB と共起したエラーを観察すると，情報の再提示に弱いことが分かった．文芸小説という分野では，情報を繰り返し提示することそれ自体が技巧のひとつとなる場合もあるため，情報として分かりきっていることだから省略するといった処理は全く望ましくない．

本節の最後に自動評価の結果との比較をおこなう．評価単位によって結果が異なっているため決定的なことはいえないが，いずれの設定においても GPT-5 は安定して評価が高かった．この結果は人手評価の結果と合致するものである．一方でスコアの差に着目すると自動評価では差があまり見られなかったが，人手評価による翻訳スコアでは大きな差がみられた．この乖離は，自動評価だけでは文芸翻訳における重要なエラーを十分に反映できないことを示唆している．

7 おわりに

本研究では文芸翻訳，なかでも推理小説の翻訳に対する機械翻訳の利用可能性を検討するため，複数の機械翻訳システムの出力を比較，分析した．そのためにジャンル特性を踏まえて評価可能な，文芸機械翻訳に特化した人手評価スキームを構築した．

具体的には，推理小説短編の “Mom Knows Best"を対象に，DeepL，TexTra，GPT-5 の翻訳を比較した．その結果，GPT-5 は総エラー数およびエラーの重大度の両面で最も良好であり，文体と用語の一貫性も文書全体で維持されていた．一方で情報の再提示を省略する傾向が観察され，文芸翻訳における課題が確認された．DeepL では，記号の処理と訳抜けが主要な課題として認められた．とりわけ訳抜けは情報を完全に欠落させるため，エラーの重大度が大きくなりやすい．TexTra は SE が比較的少なかったが，語義の取り違え，定型的な訳出に起因する不自然さ，代名詞処理の問題が顕著であった. これらは文芸作品の性格と TexTra の性格のミスマッチの反映であることが推測される．

翻訳スコアを確認すると，最も高かった GPT-5 でも設定した基準点には到達しなかった．基準点は人手訳の最低水準として定義しているため，現段階では文芸翻訳への機械翻訳の活用には依然として複数の課題が残ると結論づけられる．これは基本的に機械翻訳の結果を直接使うことを想定した結論である．一方で本稿では十分に解析できなかったが，MT+PE の枠組みを考慮したときにも，PE で対応しなくてはならない問題の広がりや重なりを考えると，産業翻訳よりも不透明な部分は大きいと考えられる．近年，文芸翻訳の賞を機械翻訳を利用した翻訳が受賞するといったニュースがあり [15]，これは文芸翻訳における品質とは何かをめぐる問題が十分に明確でないことを反映しているものであろう．本稿では推理小説においてではあるが，一定程度，品質とは何かの明確化を前提とした評価を行った点で，この明確化にも資するものである．

今後の展望として，第一に翻訳テクストと機械翻訳システムのサンプル数を増やすことが挙げられる．さらに構築した人手評価スキームを第三者にも使用してもらい，評価の再現性や妥当性をより頑健に検証することも重要である．

参考文献

[1] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In **Proceedings of the Ninth Conference on Machine Translation**, pp. 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[2] Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In **Proceedings of Translating and the Computer 35**, London, UK, November 28-29 2013. Aslib.

[3] Ran Zhang, Wei Zhao, and Steffen Eger. How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs, 2025.

[4] Arda Tezcan, Joke Daems, and Lieve Macken. When a 'sport' is a person and other issues for NMT of novels. In **Proceedings of the Qualities of Literary Machine Translation**, pp. 40–49, Dublin, Ireland, August 2019. European Association for Machine Translation.

[5] Lieve Macken, Bram Vanroy, Luca Desmet, and Arda Tezcan. Literary translation as a three-stage process: machine translation, post-editing and revision. In **Proceedings of the 23rd Annual Conference of the European Association for Machine Translation**, pp. 101–110, Ghent, Belgium, June 2022. European Association for Machine Translation.

[6] Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. The Multi-Range Theory of Translation Quality Measurement: MQM scoring models and Statistical Quality Control. In **Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)**, pp. 75–94, Chicago, USA, September 2024. Association for Machine Translation in the Americas.

[7] James Yaffe. My Mother, The Detective: Enlarged Edition. chapter Mom Knows Best, pp. 14–26. Crippen & Landru Pub, 2016.

[8] ジェイムズ・ヤッフェ. ママは何でも知っている. ママは何でも知っている, pp. 7–32. ハヤカワ文庫, 2015. 訳者：小尾芙佐.

[9] 日本翻訳連盟. JTF 翻訳品質評価ガイドライン. 一般社団法人 日本翻訳連盟, 第 1 版, 2018.

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

[11] 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.

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

[13] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[14] Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In **Proceedings of the Tenth Conference on Machine Translation**, pp. 355–413. Association for Computational Linguistics, November 2025.

[15] National Museum of Taiwan Literature. 2025 Taiwan/Ireland Poetry Translation Competition — Results Announced. National Museum of Taiwan Literature (NMTL) website, News, November 2025. Accessed: 2026-01-09.

	Error Categories	DeepL	TexTra	GPT-5
SP1	文字化け	0	0	0
SP2	誤字	0	0	0
IC1	Term	4	5	0
IC2	Style	10	16	0
RD1	慣用表現の形式的な誤り	0	0	0
RD2	Cultural Specific	0	0	1
RD3	係り受けの曖昧さ	1	1	1
RD4	文末表現	1	3	2
RD5	Lexical Choice	13	28	6
RD6	助詞の選択	2	2	0
RD7	語順	0	0	0
RD8	文の長さ	0	0	0
RD9	句読点の補足	1	3	0
RD10	Awkward Style	5	10	1
SE1	記号のミス	113	3	1
SE2	Addition	1	3	2
SE3	Omission	71	9	16
SE4	Repetition	5	0	0
GR1	単数形・複数形の誤訳	0	0	0
GR2	Pronoun resolution error	2	7	1
GR3	Part of speech	1	0	0
GR4	Prepositional error	0	2	0
GR5	助詞の選択	0	3	0
GR6	時制の誤訳	0	0	0
LP1	接続詞の誤訳	1	0	0
LP2	句読点の誤訳	1	2	0
LP3	Semantically Unrelated	8	12	3
LP4	言葉遊び	1	0	1
MT1	Non-existing Word	3	1	0
MT2	俗語・口語の誤訳	4	8	5
MT3	イディオムの誤訳	1	5	0
MT4	起点テクスト側における多義性の誤解	6	14	3
MT5	Wrong term	10	15	4
MT6	話し手の誤訳	0	0	0
MT7	文脈理解	1	5	1
SG1	多義性	0	0	0
SG2	省略	0	0	0
SG3	説明の補足	0	0	0
CB1	手がかり	6	4	5
CB2	推理	9	2	0

表 5: エラー分類体系と DeepL，TexTra，GPT-5 のエラー項目別評価結果