Vol.2025-NL-264 No.7
2025/7/6

WRIME-MT：日英・日中ソーシャルメディア対訳データセットの構築

東山 翔平[1,a)]　梶原 智之[2,b)]　内山 将夫[1,c)]

概要：自然言語処理・機械翻訳において，ユーザ生成テキストは，逸脱的現象への対処が課題となる難しいドメインの 1 つである．本研究では，日本語ユーザ生成テキストの機械翻訳タスクと，同タスクにおける逸脱的表記のテキスト正規化に焦点を当て，評価用対訳データセット WRIME-MT の構築を行った．WRIME-MT は，日本語ソーシャルメディア投稿の原文テキストに，英語訳・中国語訳と，逸脱的表記の正規化情報や固有名などの言語情報が付与されたデータセットであり，日本語ユーザ生成テキストの機械翻訳評価において既存データセットを補完する位置づけとなる．本データセットを用いて，多言語および日本語中心の翻訳特化モデル・汎用言語モデルの翻訳精度評価を行い，最先端の自動正規化モデルによる正規化適用について一定の有効性を確認した．

1. はじめに

ソーシャルメディア，レビューサイト，電子掲示板などに投稿されるユーザ生成テキスト（UGT）は，個人や消費者の発言・意見についての貴重な情報源である．機械翻訳では，2010 年代以降，UGT に焦点を当てたシェアドタスク [1–5] が開催されるなど，重要なドメインの一つとして注目を集め，研究が進められてきた．

UGT に特異な言語現象として，口語的表現，省略，ネットスラング，絵文字，誤記など，書き手による逸脱した言語使用（自然言語処理では「ノイズ」とも呼ばれる）がある．UGT ドメインに特化した注釈付きコーパスは通常限られており，こうした言語現象に対処することは，UGT を対象とする自然言語処理における共通の課題と言える．機械翻訳においても，大規模なドメイン内対訳テキストがないことから，逸脱的現象を含む UGT に十分に適応したモデルを構築することは容易ではない．そのため，逸脱的現象に起因する誤訳や，翻訳品質の低下が問題となる [6, 7]．

この問題に対する既存研究の取り組みとして，「ノイズを除去する」方法や，「(学習時に) ノイズを加える」方法が用いられてきた．前者の「ノイズを除去する」方法には，テキスト正規化 [8] がある．これは，逸脱的現象のうち特に表記揺れとみなせる現象（以降，「逸脱的表記」と呼ぶ）に焦点を当て，テキストを規範的・標準的な表記に変換してから機械翻訳モデルに入力することで，ノイズを含まないテキストと同等の翻訳品質を達成することを目指すものである [9–11]．後者の「ノイズを加える」方法では，モデル学習用の疑似対訳を生成する．目標言語のドメイン内単言語テキストを逆翻訳したり，クリーンな対訳テキストに人工的なノイズを注入するといった方法により，ドメイン内の大規模な疑似対訳を作成し，ノイズに頑健な機械翻訳モデルの学習を行う [7, 12, 13]．

本研究の焦点は二つある．一点目は，日本語を原言語とする UGT 機械翻訳における機械翻訳モデルの性能を評価することであり，二点目は，逸脱的表記への対処方法としてテキスト正規化の有効性を検証することである．

一点目に関して，日本語を含む言語方向についての UGT 対訳データセットには，MTNT [14] や PheMT [15] があるものの，日本語 UGT を対象とした機械翻訳研究の推進・発展のためには，多様なベンチマークデータセットが利用できることが望ましい．MTNT は，英語 ↔ 日本語方向を含む Reddit 投稿の学習・評価用対訳データセットであり，PheMT は，MTNT の日本語 → 英語対訳を高品質な事例にフィルタリングし，さらに 4 種類の言語現象の種別と正規化情報を付与した評価用対訳データセットである．本研究では，日本語ソーシャルメディア投稿に，逸脱的表記の正規化情報や固有名などの言語情報を付与し，さらに英語および中国語の翻訳を作成することで，2 言語方向の評価用

[1] 情報通信研究機構
NICT, Seika-cho, Kyoto, 619-0289, Japan
[2] 愛媛大学／大阪大学
Ehime University, Matsuyama, Ehime, 790-8577, Japan
The University of Osaka, Suita, Osaka, 565-0871, Japan
[a)] shohei.higashiyama@nict.go.jp
[b)] kajiwara@cs.ehime-u.ac.jp
[c)] mutiyama@nict.go.jp

Dataset	Data Source	Lang pair	Train+Dev size	Test size
MTNT [14]	Reddit	en↔fr, en↔ja	17.3k–85.8k	1,002–1,022
WMT19 Reddit Test Set [1]	Reddit	en↔fr, en↔ja	–	1,111–1,401
WMT20 Reddit Test Set [2]	Reddit	en↔ja	–	997–1,376
WMT20 Wiki. Comments Test Set [2]	Wikipedia	en→{de, ja}	–	1,098–1,100
PheMT [15]	Reddit	ja→en	–	1,566
PFSMB [10]	Social media	fr→en	–	1,554
PMUMT [16]	Social media	fr→en	–	400
MMTC [17]	Twitter	{ar, de, es, fa, fr, hi, ko, ps, pt, ru, tl, ur, zh}→en	0–59,247	901–3,000
RoCS-MT (WMT23) [4, 18]	Reddit	en→{cs, de, fr, ru, uk}	–	1,922
WRIME-MT	Twitter	ja→{en, zh}	–	769

表 1 主な UGT 対訳データセット．サイズは文数または投稿数.

対訳データセット WRIME-MT を構築した．WRIME-MT は，PheMT と同様に言語現象に注目した詳細評価が可能なデータセットであり，ドメインや言語方向の点で既存データセットを補完する位置付けとなる．

二点目に関して，既存研究 [9–11, 19] と同様，自動正規化手法を適用した入力テキストを用いて，機械翻訳タスクにおける正規化の有効性を検証する．特に本研究では，正規化，機械翻訳の両方で，大規模言語モデル（Large Language Model，LLM）としても知られる高性能な decoder-only 言語モデルを使用し，ソーシャルメディアドメインにおける最先端のオープンモデルの機械翻訳精度と，逸脱的表記の翻訳精度への影響を調査する．

評価実験の結果，高精度な正規化モデル [20] を用いた場合に，自動正規化の適用により複数の翻訳特化モデル・汎用言語モデルで翻訳精度が向上することが示された．元から逸脱的表記に頑健であった言語モデルでは，自動正規化の有効性は確認できなかったものの，人手正規化は有効であったため，正規化精度に改善の余地があることを確認した．

2. 関連研究

2.1 UGT に関する機械翻訳シェアドタスク

UGT ドメインの機械翻訳研究を推進してきた取り組みとして，WMT（Conference on Machine Translation）における国際コンペティションの開催が挙げられる．WMT11 では，2010 年ハイチ大地震に焦点を当て，ハイチ・クレオール語の SMS メッセージを英語に翻訳するタスクが開催された [21].

近年では，WMT19，20 において，UGT を対象とした機械翻訳タスクである Machine Translation Robustness タスク [1, 2] が行われた．これら Robustness タスクでは，Transformer [22] 等の NMT システムが用いられ，複数の参加者が用いた有効な方法として，(i) 学習データから低品質な対訳文を除外するクリーニング，(ii) 翻訳不要文字列のプレースホルダ化，(iii) 目標ドメイン単言語テキストの逆翻訳で作成された疑似対訳データでの学習などが挙げられている．

その後の WMT22, 23, 24 では，General Machine Translation タスク [3–5] のテストセットの一部として UGT ドメインのデータが使用され，WMT23–24 では，test suites（特定の評価観点に焦点を当てた難しいテストセット群）の一つとして Reddit 投稿に由来する RoCS-MT [18] データセットが使用された．

2.2 UGT 対訳データセット

表 1 に示すように，これまでに構築・公開されている UGT ドメインの対訳データセットはいくつか存在する．

Michel ら [14] は，UGT ドメインにおける大規模な対訳ベンチマーク構築の先駆的な研究として，Reddit 投稿とその人手翻訳からなる MTNT[*1] データセット（英 ↔ 仏，英 ↔ 日方向）を構築した．同データセットは，前述の WMT19–20 Machine Translation Robustness タスクでも採用され[*2]，UGT ドメインの機械翻訳研究において重要な貢献を果たしたと考えられる．ただし，翻訳品質が低い訓練事例も存在していることが指摘されている [15].

Fujii ら [15] は，MTNT の事例に対してルールおよび人手翻訳品質評価によるフィルタリングを行った上で，4 種類の言語現象ラベル（「固有名詞」，「名詞の省略」，「口語表現」，「異表記」）と逸脱した表現に対する正規化情報を付与した PheMT[*3]データセット（日 → 英方向）を構築した．

Rosales Nùñez らは，フランス語ソーシャルメディア投稿を原文とする PFSMB 対訳コーパス[*4] [10]（仏 → 英方向）

[*1] https://pmichel31415.github.io/mtnt/index.html
[*2] タスクの学習・開発データとして用いられた．テストデータは同様のプロセスにより別途新たに作成された.
[*3] https://github.com/cl-tohoku/PheMT
[*4] https://gitlab.inria.fr/seddah/parallel-french-social-mediabank

と，そのうち原文 400 文に対してスパンレベルでの言語現象種別と正規化情報を付与した PMUMT コーパス[*5] [16] を構築した．PMUMT コーパスはサイズが小さいものの，注目する現象以外を正規化後表現に変換したデータを生成することで，現象ごとの詳細分析が可能になっている．

McNamee ら [17] は，13 言語の Twitter 投稿テキストを英語へ翻訳した多言語の MMTC[*6]データセットを構築した．McNamee らの実験によると，一般ドメインの事前学習済み Transformer モデルをドメイン内訓練データで fine-tuning した際，各言語とも訓練データ 1,000 投稿程度で BLEU スコアの向上が概ね飽和したことが報告されている．

Bawden ら [18] は，英語の Reddit 投稿を，人手正規化の後，5 言語に翻訳した RoCS-MT[*7]データセットを構築した．同データセット原文の正規化箇所には現象種別も付与されている．同データセットは，前述のように WMT23 General Machine Translation タスク [5] でも使用され，商用翻訳システムおよび参加者システムの性能評価に用いられた[*8]．評価されたシステムの中で，GPT-4 [23]（5-shot）が自動評価スコア上も定性的にも高い翻訳品質を示したことが報告されている．

その他，ドメイン特化型の対訳コーパスとして，FIFA 2014 world cup に関する英語 Twitter 投稿をドイツ語に翻訳した FooTweets コーパス[*9] [24]，レストランレビューに関するフランス語 Foursquare 投稿を英語に翻訳したコーパス[*10] [25] なども構築されている（表 1 には非掲載）．

2.3 テキスト正規化と機械翻訳応用

テキスト正規化 [8] は，典型的には，（一般ドメイン向けの）自然言語処理技術・モデルにとって対処が難しいような，ある種の言語現象の表現を，処理しやすい表現に変換する基盤技術またはタスクを指す[*11]．特に UGT の正規化 [27,28]（以降，単に「テキスト正規化」や「正規化」と言及する）は，UGT 特有の逸脱的・非標準的な表現を一般的・標準的な表現へ変換する技術として多くの研究が行われてきた．

テキスト正規化を前段タスク，機械翻訳を後段タスクとして，機械翻訳における正規化の有効性を検証した研究には，[9–11] などがある．Wang ら [9] は，線形分類器による正規化モデルと，フレーズベース機械翻訳を用いて，中国語・英語間の SMS 正規化・翻訳に取り組んだ．Rosales Nùñez ら [10] は，Transformer に基づく grapheme-to-phoneme 変換型の正規化モデルと，フレーズベースおよび Transformer 機械翻訳モデルを用いて，フランス語 → 英語の UGT 正規化・翻訳に取り組んだ．Ahmadi ら [11] は，Transformer encoder-decoder モデルを用いて，ダイグロシア状態にあるペルソ・アラビア文字の少数言語における，支配的言語の影響を受けたソーシャルメディアの非標準的な表記を正規化する問題に取り組み，人工的なノイズを加えたテキストの機械翻訳タスクで，正規化モデルの有効性を示した．

2.4 UGT 機械翻訳手法の評価

2010 年代前半頃の UGT 機械翻訳の研究では，オンラインフォーラムや SMS メッセージ等の UGT に対し，統計的機械翻訳システムの翻訳品質評価が行われた [21,29,30]．近年は，LLM を対象とした UGT ドメインにおける翻訳品質評価も行われている [31,32]．

Popovic ら [31] は，商品レビュー対訳コーパス[*12]（英語 → クロアチア語，フィンランド語，フランス語）を用いて，人間の翻訳者，商用サービスなどの機械翻訳システムおよび ChatGPT（GPT-3.5）[33] に対し，原文中のノイズが翻訳結果に与える影響を分析した．ChatGPT はノイズを訂正した翻訳結果を出力することが多く，他の MT システムと比べてノイズに頑健であったことが報告されている．

Pan ら [32] は，LLM を用いた翻訳で，文脈内学習によりノイズへの頑健性が向上することを示した．具体的には，ノイズを含む原文とその翻訳文のペアを few-shot 事例とした文脈内学習により，人工的なノイズを加えた非 UGT（中国語 → 複数言語），自然なノイズを含む UGT（インドネシア語 → 中国語）の双方についての翻訳精度が向上することを報告した．

3. 対訳データセット WRIME-MT の構築

ソーシャルメディアの日本語投稿テキストを収録した既存のデータセットに，WRIME[*13] [34,35] および WRIME 正規化データセット [36] がある．WRIME は，感情分析タスク向けに感情ラベルが付与された 35,000 投稿のデータセットであり，WRIME 正規化データセットは，WRIME のうち 6,000 投稿に対し，正規化情報（正規化テキストと種別）が付与されたデータセットである．本研究では，WRIME および WRIME 正規化データセットの原文と正規化情報を利用し，日本語投稿 769 件とそれらの英訳および中国語訳からなる，機械翻訳評価のための WRIME 対訳

[*5] https://github.com/josecar25/PMUMT_annotated_UGC_corpus
[*6] https://pmcnamee.net/research/mmtc/mmtc.html
[*7] https://github.com/rbawden/RoCS-MT
[*8] WMT24 [5] では，RoCS-MT の拡張版データセット（en→{cs, de, es, hi, is, ja, ru, uk, zh} 方向）が使用された．ただし，同データセットは本稿投稿時点で未公開のようである．
[*9] https://github.com/HAfli/FooTweets_Corpus
[*10] https://europe.naverlabs.com/research/natural-language-processing/machine-translation-of-restaurant-reviews/
[*11] ただし，正規化結果の直接的な「利用者」はモデルに限らない．テキスト正規化によって，当該言語の第二言語学習者にとっての可読性が向上することを示した研究もある [26]．
[*12] https://fedora.clarin-d.uni-saarland.de/dihutra/index.html
[*13] https://github.com/ids-cv/wrime

原文	正規化後文	補足情報	翻訳文
毎 日 <1>FB</1>開いてるよーー	毎 日 <1>FB</1>開いてるよ	FB：Facebook	I open <1>FB||Facebook</1> every day.

表 2 翻訳作業文の例

データセット（WRIME-MT）を構築した[*14].

本データセットの構築は，(1) 作業対象投稿の選択，(2) 言語情報アノテーション作業，(3) 翻訳作業，の手順で行った．詳細を以下に述べる．

3.1 作業対象投稿の選択

所定予算内で翻訳が可能な件数として，WRIME 正規化データセット（学習・開発・テストセット全体）の投稿のうち 769 件を選択した．これら作業対象投稿の選択は，第一著者が，「UGT 特有の崩れた表現，固有名，ネットスラング，文化依存表現を含むなど，他言語への翻訳が容易ではないような投稿である」という観点で行った．したがって，本データセットのこれらの投稿は，原文の文字通りの語彙的な直訳では誤訳になりやすい投稿を多く含むと考えられる．

3.2 言語情報アノテーション作業

本データセットを用いて多角的な評価が可能となることを意図し，第一著者により 3 種類の言語情報—正規化情報，固有名，略語等の補足情報—を作業対象投稿に付与した．

言語情報アノテーションの内容は次の通りである．

正規化情報 WRIME 正規化データセットで施された正規化事例・分類カテゴリは，感情分析タスクでの有用性を想定したタスク特化の事例（例：感情記号 “(笑)” → “<8>”）や，後段タスクへの影響が軽微と思われる事例（例：記号の変換 “「悪の教典」” → “『悪の教典』”），原文からの意味の変化が生じ得る，表記の正規化の範疇を超えると思われる事例（例：ネットスラング “ググったら” → “検索すると”）も含んでいる．そこで，本データセットでは，単語レベルの逸脱的表記の揺れの解消や誤記の訂正にあたる正規化事例のみ収録することを目的に，オリジナルの正規化事例の一部カテゴリのみ残し[*15]，さらに一部の事例の追加・削除・編集を行った[*16].

固有名 各投稿に対し，人名・キャラクター名，組織名，地名・施設名，プロダクト名，イベント名に該当する固有名（正式名称およびその他の呼称）に関して，該当するスパンと固有名のフラグの情報を付与した．なお，固有名の崩れた表記に対する標準的な表記は，正規化情報ではなく後述の補足情報として付与した（例：“でーんーまーあーくー” → “デンマーク”）.

補足情報 各投稿に対し，固有名の略称や，非自明な固有名・スラング等を含む場合，該当するスパンと簡潔な説明テキストを「補足情報」として付与した[*17]．略称・略語とされる表現の中には，テキスト中に出現する略語を展開後の表現で単純に置換すると，日本語の文章として不自然になったり，翻訳内容に大きな影響を与えることが懸念されるものがある[*18]．そのため，略称については逸脱的表記の正規化とは分け，補足情報という扱いとした．補足情報の実例として，たとえば，“ハロハピ” に対して “『ハロー、ハッピーワールド!』”，“カブ” に対して “「あつまれどうぶつの森」の仮想的な株” というテキストが付与されている．

3.3 翻訳作業

日本語投稿の英語および中国語への翻訳作業を翻訳会社に委託した．翻訳作業は，日英翻訳者 5 名・日中翻訳者 5 名により，(1) 翻訳担当者による翻訳とセルフチェック，(2) 目標言語母語話者によるバイリンガルチェックの手順で実施された．

翻訳作業の仕様を，「目標言語の自然な表現を用いて，原文の語彙的な意味を伝えることに焦点を当て，逸脱的表記の逸脱的表記への変換および記号的な表現の変換は避ける」という趣旨の下，以下のように定めた．

翻訳単位 翻訳の単位は「投稿」とする．原投稿の文数と翻訳後の文数を同数にする必要はない．

訳出スタイル 原文中に崩れた表現（逸脱的表記）が含まれる場合，翻訳文では目標言語の崩れた表現を用いることはしない．目標言語の正書法や，標準的な記法・punctuation の使用法から逸脱しない範囲で，SNS 投稿として一般的なカジュアルな表現・スタイルを用いる．原文で方言が用いられている場合，翻訳文では目標言語の共通語・標準語を用いて翻訳する．

[*14] WRIME-MT および WRIME 正規化データセットは，利用規約に同意した利用申請者に対して提供される予定である．

[*15] Kondo らの分類体系 [36] のうち，誤字脱字-タイプミス/誤用，異表記-伏字/発音の崩れ/同音異表記，強調表現-音の挿入/繰り返し，を正規化事例として残し，異表記-略語/外来語については後述する補足情報の扱いとした．

[*16] たとえば，準体助詞「ん」が「の」に一律に変換されている事例群については，“-んだ” → “-のだ” など，正規化によって現代日本語の表現としてやや特殊なニュアンスや不自然さが生じている可能性があるものは，削除した．

[*17] オリジナルの正規化情報のうち異表記-略語/外来語にあたるものを補足情報とし，さらに必要に応じて第一著者が追加した．

[*18] たとえば，「ズッ友からのプレゼント」という文章を「ずっと友達からのプレゼント」とすると不自然さが生じる（日本語母語話者がこの文章をはじめから作成する状況は極めて稀と考えられる）．「KY」（「空気が読めない（人）」）なども同様である．

正規化情報　翻訳対象は原文である．崩れた表現を反映しないようにするための参考情報として，原文を整った表現で書き換えた「正規化後文」（表 2 参照）も提示する．

補足情報　原文中の略語・固有名詞・スラング等について，その意味や正式名称等が補足情報欄に記載されている場合がある（表 2 参照）．原文の解釈は，補足情報に記載された内容を前提として行う．

固有名詞タグ　原文中の固有名詞（名詞を超える句や節の表現も含む）は，目標言語の一般的・自然な表記を用いて翻訳する．表 2 の例のように，原文中の固有名詞は “<1>” と “</1>” のようなタグで囲まれている場合があり，その場合には翻訳文中の対応する箇所を同じタグで囲む．固有名詞の訳語として複数の妥当な表現・表記が考えられる場合，“||” 記号で区切って併記する．

ネットスラング・文化依存表現　ネットスラングや，その他日本特有と思われる文化に依存した表現については，目標言語で対応する表現がある場合はそれを使用し（例：“3 密”→“Three Cs”），特にない場合には原言語での意味を反映した目標言語での自然な訳を作成する．説明的な訳は可能な範囲で避ける．

顔文字・アスキーアート　顔文字・アスキーアートは翻訳対象から除外し，翻訳文には含めない．なお，顔文字等が句点の役割を果たしている場合，翻訳文で句点相当の記号を挿入する必要がある．その際，翻訳文において感嘆符・疑問符などを使用し，原文中の顔文字等が表しているニュアンスを反映してもよい．

ハッシュタグ　ハッシュタグ（「#」で始まる文字列）は目標言語の表現に翻訳する．文を構成する単語として使われている場合，翻訳文中の適切な位置で訳出する．ハッシュタグと前後の単語との間に半角スペースを挿入する．

アカウント名・URL　原文に含まれているアカウント名（「@」で始まる英数字記号列）と URL は，翻訳文にそのまま残す．（実際には，作業対象の投稿でアカウント名や URL を含むものはなかった．）

4. 実験

WRIME-MT データセットを用いて，ソーシャルメディアテキストの機械翻訳タスク（日英，日中）における，最先端の翻訳特化モデル・汎用言語モデルの精度を評価する．具体的に，次の二つのシナリオの実験を行う．

- 実験 1：翻訳モデルの翻訳精度評価（開発セット）．複数のモデルシリーズ・モデルサイズ（最小 0.5B〜最大 70B）のモデルの精度を評価し，各モデルの精度の違いの傾向を確認する．
- 実験 2：正規化適用時の翻訳モデルの翻訳精度評価（テストセット）．異なる翻訳精度（低/中/高）の翻訳モデルを対象に，正規化モデルによる正規化適用により翻訳精度の向上が見られるかを確認する．

	#Post	#Norm	#Ent	#Supp
All	769	451	707	563
Train	5	4	6	5
Dev	64	28	59	62
Test	700	419	642	496

表 3　WRIME-MT の記述統計．#Post，#Norm，#Ent，#Supp 列はそれぞれ，投稿，正規化，固有名，補足情報の事例数．

4.1 実験設定

データ分割

WRIME-MT の 769 件を，学習セット 5 件，開発セット 64 件，テストセット 700 件となるようランダムに分割した．データセットの記述統計を表 3 に示す．学習セットは，decoder-only モデルの文脈内学習の few-shot 事例用に作成したものである．ただし，各 decoder-only モデルで 0-shot および 5-shot 推論を実施し，開発セットにおける精度を確認したところ，多くのケースで 5-shot により精度が低下したため，以降の実験では 0-shot の結果のみ報告する[*19]．

評価指標

評価指標には，BLEU[*20] [38]，COMET[*21] [39]，Term Success Ratio（TSR） [40] を用いた[*22]．BLEU はシステム出力と参照訳，COMET は原文とシステム出力と参照訳を入力とし，それぞれ n-gram の包含率，深層学習モデル（XLM-RoBERTa）が予測したスコアにより，システム出力の品質を推定する指標である．TSR は，原文中の各用語に対する参照訳中の訳語を，システム出力が含んでいるかを Fuzzy match で評価する指標[*23]であり，本実験では固有表現を評価対象の用語とした．3 指標とも，0–100 の値の範囲で表示する．

翻訳モデル

翻訳精度の評価対象として，多言語および日本語中心モデ

[*19] 5-shot 推論において，“Japanese:\n{ja_text}\n\nEnglish:\n{en_text}\n\n” といった形式の事例を 5 件分並べ，最後に翻訳対象の原文（と “English:\n”）を連結したプロンプトを使用したところ，モデルの生成結果は，翻訳対象文の訳文に続けて独自の原文と訳文をいくつも並べたような出力が多く見られ，この点が各指標のスコアが下がる主な要因となったと考えられる．

[*20] sacreBLEU [37]（`https://github.com/mjpost/sacrebleu`）を使用した．
（日英：“`nrefs:1|case:mixed|eff:no|tok:intl|smooth:exp|version:2.5.1`”，日中：“`nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.5.1`”）

[*21] `https://huggingface.co/Unbabel/wmt22-comet-da`

[*22] 原文中の固有名について，参照訳は妥当な複数の固有名の情報を持つ場合がある（3.3 節）．BLEU と COMET では，1 つの固有名を当てはめた単一の参照訳を用いて計算し，TSR では複数の正解固有名のいずれかに一致しているかという基準で計算した．

[*23] `https://pypi.org/project/fuzzywuzzy/`，`partial_ratio()` を使用．

Normalizer	Twitter			14 domains		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
DeBERTa-L	64.0	52.8	61.4	78.5	56.7	72.9
Sarashina2.2-3b	79.0	67.4	76.3	78.0	66.2	75.3

表 4 JMLN [20] テストセットでの正規化モデルの正規化精度.

ルを用いた（各モデルの正式な Hugging Face ID は付録 A.1 に示す）．具体的には，多言語の encoder-decoder モデルである NLLB-200-3.3B [41]，多言語の翻訳特化 decoder-only モデルである TowerInstruct-13B [42]，X-ALMA-13B [43]，GemmaX2-28-9B [44]，多言語の汎用 decoder-only モデルである Qwen3 [45] の各サイズのモデル，日本語中心の汎用 decoder-only モデルである TinySwallow [46]，Llama-3.1/3.3-Swallow [47]，Sarashina2/2.2 [48, 49] の各サイズのモデルである（公開されている場合は指示学習済みモデルを用いた）．各モデルについて，0-shot での推論または文脈内学習を行い，各 decoder-only モデルでは付録 A.1 に示す英語または日本語の指示テキストを含むプロンプトを使用した．

正規化モデル

正規化モデルとして, Higashiyama ら [20] が構築したモデルを用いた．具体的には，Japanese Multi-Domain Lexical Normalization Dataset（JMLN）[*24]を用いて正規化タスクで fine-tuning された encoder-only および decoder-only モデルで, 正規化精度が高かった日本語 DeBERTa[*25] [50] ベースのモデル（Full-Seg-POS 法を採用，以降 DeBERTa-L-Norm モデルと呼ぶ）と，Sarashina2.2-3b[*26]ベースのモデル（Struct 法を採用，以降 Sarashina2.2-3b-Norm モデルと呼ぶ）を使用した．

2 モデルの JMNL テストセット（Twitter ドメイン，Twitter を含む 14 ドメイン全体）における正規化精度（適合率，再現率，$F_{0.5}$ スコア）を表 4 に示す[*27]．再現率は Sarashina2.2-3b が勝っており，適合率については，14 ドメイン平均では 2 モデルでほぼ同等であるものの，Twitter ドメインでは Sarashina2.2-3b が勝っている．

なお，WRIME 正規化データセットのうち WRIME-MT テストセットと重複していない投稿については，正規化モデルの学習に利用可能である．JMLN とアノテーション基準は異なるものの，2 種類の正規化データセットの両方を用いることで正規化および後段タスクの精度向上に寄与するかについては，今後検証したい．

[*24] JMLN の学習セットは 13k 文, 5.9k 正規化事例からなる．JMLN データセットおよび正規化モデルのソースコードは公開予定となっているが，本稿投稿時点では未公開.

[*25] https://huggingface.co/ku-nlp/deberta-v2-large-japanese-char-wwm

[*26] https://huggingface.co/sbintuitions/sarashina2.2-3b

[*27] 文献 [20] では各モデルについて 2 回の実行結果の平均精度を報告しているが，表 4 には，4.3 節の実験に使用した，2 回のうち 1 回のモデルチェックポイントの結果を示した.

4.2 基本の翻訳精度評価

実験 1 として，表 5 に示すように，WRIME-MT 開発セットにおける各翻訳モデルの翻訳精度を評価した．各 decoder-only モデル・各翻訳方向では，日本語および英語の指示テキストのプロンプトを試し，精度（BLEU と COMET スコアの和）が高かった方の結果を記載している．

参考に，システム出力の代わりに原文（source text）および参照訳（reference text）を使用した場合の各評価指標のスコアも示した．BLEU スコアが極めて低い値になっている一方，COMET スコアは，特に中国語への翻訳において，比較的高い値となった．翻訳結果の品質・妥当性を判断する際は，このような点も考慮しつつ，BLEU，COMET スコアの両方が十分高い値であるかを踏まえる必要がある[*28]．

結果は次のようにまとめられる．(1) 全体として，サイズが大きいモデル，あるいは公開時期が新しいモデルの翻訳精度が高い傾向が見られるが，その限りではないケースもあった（たとえば, Sarashina2 では 70B より 7B モデルが高精度）．(2) 日英方向で相対的に高精度（COMET スコア 70 以上）であったモデルは GemmaX2-28-9B，Sarashina2.2-3B-Instruct，Llama3.3-Swallow-70B-Instruct，Qwen3（8B 以上）であり，日中方向で高精度であったモデルは X-ALMA-13B, GemmaX2-28-9B, Sarashina2.2-3B-Instruct, Qwen3（4B 以上）であった．したがって, GemmaX2-28-9B, Sarashina2.2-3B-Instruct および Qwen3（8B 以上）は両方向で翻訳精度が高いモデルであり，特に Sarashina2.2-3B-Instruct はモデルサイズに比して高精度と言える．(3) TSR については，ほとんどのモデルで日英よりも日中方向で値が低い．これは，中国語において多くの固有名が漢字と英字の複数の妥当な表記（例：“Twitter” と “推特”）を持つ一方，参照訳では，（3.3 節の通り複数の固有名詞の訳語を記述可能な仕様であるものの，実際には,）それらが網羅されていないことが多い点の影響が考えられる．

4.3 正規化適用時の翻訳精度評価

実験 2 として，表 6 および表 7 に示すように，原文（Normalizer=None），正規化モデル（DeBERTa-L-Norm, Sarashina2.2-3B-Norm）による原文の正規化結果，人手付与された正解正規化文（Oracle）の 4 種類の入力を用いて，WRIME-MT テストセットでの各翻訳モデルの翻訳精度を比較した．翻訳モデルについては，日英・日中方向の各翻訳精度が最も低かった NLLB-200-3.3B，翻訳精度が中程度であった TowerInstruct-13B，翻訳精度が高かった

[*28] 実際，中国語への翻訳タスクで日本語のテキストが出力されている事例も見られた．たとえば，TinySwallow-1.5B-Instruct では，“Japanese:\n(中略) 初代以外だとルビサファ大好きだからな〜\n\nChinese:\n” の入力に対し，“初代以外だと、ルビサファが好きなんだね。” と，日本語による会話的応答のようなテキストが出力された事例があった.

Date	Translator	Inst	ja-en			ja-zh		
			BLEU	COMET	TSR	BLEU	COMET	TSR
	Source text		0.1	51.1	10.2	1.9	63.0	11.9
	Reference text		100.0	93.5	100.0	100.0	95.4	100.0
2022/07	NLLB-200-3.3B		9.9	58.8	22.0	8.7	56.6	1.7
2024/02	TowerInstruct-13B	(ja/ja)	9.3	65.3	30.5	15.4	63.9	15.3
2024/10	X-ALMA-13B-Group6	(en/ja)	18.8	69.4	30.5	15.9	70.9	16.9
2025/02	GemmaX2-28-9B	(en/en)	19.1	71.9	28.8	20.5	73.5	25.4
2025/01	TinySwallow-1.5B-Instruct	(ja/en)	14.4	64.8	22.0	7.8	62.6	8.5
2025/03	Sarashina2.2-0.5B-Instruct	(en/ja)	8.5	66.4	22.0	6.5	62.0	10.2
2025/03	Sarashina2.2-1B-Instruct	(en/ja)	12.5	69.5	32.2	10.4	65.2	8.5
2025/03	Sarashina2.2-3B-Instruct	(en/ja)	19.6	**74.4**	**42.4**	17.6	74.3	16.9
2024/06	Sarashina2-7B	(en/ja)	15.4	67.8	44.1	8.2	64.4	13.6
2024/08	Sarashina2-70B	(en/ja)	14.8	61.6	32.2	8.6	60.7	23.7
2024/11	Llama3.1-Swallow-8B-Instruct	(en/en)	14.8	69.7	37.3	15.3	65.1	15.3
2025/03	Llama3.3-Swallow-70B-Instruct	(en/en)	18.6	72.1	35.6	21.4	65.2	25.4
2025/04	Qwen3-1.7B	(en/en)	8.5	64.2	18.6	8.2	65.2	16.9
2025/04	Qwen3-4B	(ja/en)	17.1	67.7	25.4	23.2	72.5	22.0
2025/04	Qwen3-8B	(en/ja)	21.2	70.6	30.5	24.7	73.9	32.2
2025/04	Qwen3-14B	(en/en)	22.6	71.2	37.3	**26.5**	**75.3**	32.2
2025/04	Qwen3-32B	(en/ja)	**23.7**	72.8	40.7	22.0	74.1	**33.9**

表 5 WRIME-MT 開発セットにおける各翻訳モデルの精度．「Date」列はモデル公開時期，「Inst」列は各モデルで採用したプロンプトの指示言語 (ja-en/ja-zh) を示す．

Translator	Normalizer	All		Standard		Non-standard	
		BLEU	COMET	BLEU	COMET	BLEU	COMET
NLLB-200-3.3B	None	10.5	61.0	11.8	61.3	8.7	60.6
	DeBERTa-L-Norm	9.0	<u>61.4</u>	11.6	61.3	6.7	<u>61.5</u>
	Sarashina2.2-3B-Norm	8.9	<u>62.1</u>	<u>11.9</u>	<u>61.8</u>	6.4	<u>62.4</u>
	Oracle	9.9	62.6	11.9	61.3	7.9	64.2
TowerInstruct-13B	None	11.1	67.1	14.4	68.2	8.5	65.7
	DeBERTa-L-Norm	<u>14.2</u>	<u>68.1</u>	<u>15.6</u>	68.2	<u>12.8</u>	<u>67.9</u>
	Sarashina2.2-3B-Norm	<u>14.0</u>	<u>68.9</u>	14.4	<u>68.7</u>	<u>13.6</u>	<u>69.1</u>
	Oracle	15.2	69.2	14.3	68.1	15.9	70.7
Sarashina2.2-3B-Instruct	None	19.2	74.4	19.1	73.9	19.3	75.0
	DeBERTa-L-Norm	18.8	73.9	18.7	73.8	18.9	73.9
	Sarashina2.2-3B-Norm	18.7	<u>74.5</u>	18.7	<u>74.1</u>	18.7	75.0
	Oracle	18.9	75.1	19.1	74.0	18.6	76.6
Qwen3-32B	None	24.4	74.2	24.1	74.0	24.8	74.5
	DeBERTa-L-Norm	24.3	74.1	23.7	73.9	25.0	74.3
	Sarashina2.2-3B-Norm	<u>24.6</u>	<u>74.9</u>	24.1	<u>74.3</u>	<u>25.2</u>	<u>75.7</u>
	Oracle	25.0	75.4	24.0	74.0	26.0	77.2

表 6 WRIME-MT テストセットにおける各正規化法適用時の各翻訳モデルの精度 (日 → 英)．正規化なし（Normalizer=None）に対し，正規化モデル適用でスコアが向上した場合に<u>下線</u>で表示．

Translator	Normalizer	All		Standard		Non-standard	
		BLEU	COMET	BLEU	COMET	BLEU	COMET
NLLB-200-3.3B	None	8.9	56.1	9.4	56.1	8.3	56.1
	DeBERTa-L-Norm	9.8	56.9	9.4	56.4	10.4	57.6
	Sarashina2.2-3B-Norm	9.3	57.1	9.9	56.6	8.7	57.7
	Oracle	8.8	57.1	9.4	56.2	8.1	58.3
TowerInstruct-13B	None	12.1	65.5	15.5	67.0	9.4	63.6
	DeBERTa-L-Norm	12.9	65.7	15.3	66.5	10.2	64.7
	Sarashina2.2-3B-Norm	13.2	66.6	15.5	67.4	9.9	65.6
	Oracle	14.2	67.0	15.4	66.9	13.0	67.2
Sarashina2.2-3B-Instruct	None	16.5	73.8	16.8	73.0	16.1	74.9
	DeBERTa-L-Norm	16.5	73.5	16.7	72.8	16.2	74.4
	Sarashina2.2-3B-Norm	16.8	73.7	17.0	72.7	16.6	75.0
	Oracle	17.0	74.4	17.0	73.0	17.0	76.2
Qwen3-14B	None	28.2	77.5	28.7	77.6	27.5	77.3
	DeBERTa-L-Norm	27.9	77.5	28.4	77.4	27.2	77.6
	Sarashina2.2-3B-Norm	28.1	78.5	28.5	78.1	27.7	78.9
	Oracle	28.7	79.0	28.7	77.6	28.7	80.9

表 7 WRIME-MT テストセットにおける各正規化法適用時の各翻訳モデルの精度（日 → 中）．正規化なし（Normalizer=None）に対し，正規化モデル適用でスコアが向上した場合に下線で表示．

Sarashina2.2-3B-Instruct および Qwen3（日英では 32B，日中では 14B）の計 4 モデルを対象とした．また，翻訳精度は，全投稿（All，700 件），人手付与された正規化事例を含まない投稿（Standard，391 件），人手付与された正規化事例を含む投稿（Non-standard，309 件）の 3 つのサブセットに対してそれぞれ算出した．原文に施された正規化が適切である場合，Standard サブセットにおける翻訳精度は同等となり，Non-standard サブセットにおける翻訳精度は向上すると期待される．

表 6 の日英翻訳における結果は次のようにまとめられる．(1) 4 翻訳モデルとも，原文の翻訳精度よりも正解正規化文の翻訳精度の方が高くなり，Non-standard サブセットでは，COMET スコアが 1.6〜5.0 ポイント高くなった（ただし BLEU スコアは下がるケースもあった）．(2) NLLB-200-3.3B，TowerInstruct-13B，Qwen3-32B の 3 翻訳モデルでは，正規化非適用（None）と比べた適用時の精度は，Sarashina2.2-3B-Norm 正規化モデルを用いた場合を中心に向上した．具体的には，Non-standard サブセットでは COMET スコア +1.2〜+3.4，All では +0.7〜+1.8 であった（BLEU スコアは低下・向上の両方のケースが見られた）．(3) 一方，翻訳モデル Sarashina2.2-3B-Instrct では，Sarashina2.2-3B-Norm 正規化モデルを適用した場合，BLEU スコアがやや低下し，COMET スコアは僅かな変化のみであった．(4) 各翻訳モデルについて，DeBERTa-L-Norm モデルによる正規化適用の効果は，Sarashina2.2-3B-Norm モデルに比べて限定的であり，翻訳モデル NLLB-200-3.3B，TowerInstruct-13B においてのみ COMET スコアの向上（Non-standard サブセットで +1 ポイント以上）が見られ，翻訳モデル Sarashina2.2-3B-Instrct では COMET スコアが低下した（Non-standard サブセットで −1.1 ポイント）．(5) 各翻訳モデルについて，Standard サブセットでは，正規化モデル適用による COMET スコアの低下はほぼ見られなかった（−0.1〜+0.5 ポイントの変化）．

表 7 の日中翻訳の結果も，ほぼ同様の傾向であった．つまり，Sarashina2.2-3B-Norm モデルで正規化した場合，NLLB-200-3.3B，TowerInstruct-13B，Qwen3-14B の 3 翻訳モデルで正規化適用の有効性が確認でき，Non-standard サブセットでは COMET スコア +1.6〜+2.0，Standard サブセットでは +0.4〜+0.5，All では +1.0〜+1.1 であった．

議論

2 言語方向についての実験結果から，正規化の有効性について次のようにまとめられる．逸脱的表記を含む原文の意味内容を反映した機械翻訳結果を生成する目的において，人手正規化には劣るものの，正規化モデルによる正規化は一定の有効性が見られた．

具体的には，翻訳モデル NLLB-200-3.3B，TowerInstruct-13B，Qwen3-14B/32B においては，正規化モデル Sarashina2.2-3B-Norm を適用した場合に COMET スコ

アを中心に翻訳精度が向上した．ただし，正規化モデル DeBERTa-L-Norm による正規化の有効性は限定的であり，この違いは，両正規化モデルの Twitter ドメインにおける正規化精度（表 4 の適合率・再現率参照）の違いによる結果と判断できる．

一方，翻訳モデル Sarashina2.2-3B-Instruct においては，正規化モデル Sarashina2.2-3B-Norm 適用時も COMET スコアの変化は僅かで，明確な有効性は確認できなかった．理由として，翻訳モデルが，逸脱的表記と（原言語または目標言語の）標準的表記の対応関係についての「知識」を有している場合，正規化の有無にかかわらず誤訳は生じず，翻訳結果に対して重大な影響を与えないということは十分考えられる．実際，4.4 節で示す事例では，Sarashina2.2-3B-Instruct は，逸脱的表記を含む原文に対して最も頑健な翻訳結果を示した．ただし，同モデルにおいても人手正規化の適用時は COMET スコアの向上が見られたことから，同モデルが有する「知識」は十分ではないこと，正規化モデルが有効となるには正規化精度の向上—より非自明な逸脱的表記に対する正確な正規化—が必要であると言える．

4.4 事例分析

固有名・スラングに対する翻訳結果例

開発セットの事例における，正規化適用なしの状況での翻訳モデル（4.3 節と同じ 4 モデル）の出力事例を表 8 に示す．

例 (a) は，「どうぶつの森シリーズ」の略称 “どう森” を含むテキストである．4 翻訳モデルのうち，Sarashina2.2-3B-Instruct のみ適切な英語名称 “Animal Crossing” を出力しており，他の翻訳モデルは適切な生成に失敗している．

例 (b) は，「YouTube」の投げ銭機能「スーパーチャット」の略称 “スパチャ” を含むテキストである．Qwen3-32B のみ，妥当な英訳 “super chat” を出力している．ただし，サービスの名称のため “Super Chat” という表記がより適切である．他の翻訳モデルは適切な訳の生成に失敗している．

逸脱的表記に対する翻訳結果例

開発セットの事例における，正規化適用なしまたは正規化モデル適用の状況での翻訳モデル（4.3 節と同じ 4 モデル）の出力事例を表 9 に示す．

例 (c) は，“かほり” という歴史的仮名遣いの表現を含むテキストである．TowerInstruct-13B，Sarashina2.2-3B-Instruct，Qwen3-32B の 3 翻訳モデルは，原文に対しても，Sarashina2.2-3B-Norm モデルによる正規化後文に対しても，妥当な訳を出力した．翻訳モデル NLLB-200-3.3B は，原文の ‘かほりがする” に対して不適切な訳を出力したが，正規化後の “香りがする” に対しては妥当な訳を出力した（ただし，正規化前後とも “初夏” に対する誤訳を含む）．

例 (d) は，童謡の替え歌を歌っている様子が表されたテキストであり，全体が逸脱的表記となっている[*29]．原文に対する翻訳結果では，Sarashina2.2-3B-Instruct のみ妥当な訳文を出力した（ただし，不要な注釈も出力している）．DeBERTa-L-Norm モデルによる正規化後文は，不適切な正規化により解釈不能な文となっており，各翻訳モデルの翻訳結果もそれに応じて元の原文の意味内容を表さないものになっている．Sarashina2.2-3B-Norm モデルによる正規化後文は，一部（“でーん” の部分）の漏れを除き適切な正規化であり，各翻訳モデルとも “もういくつ寝ると” の部分を反映した訳を生成した．“でーんマーク” の部分に対しては，Sarashina2.2-3B-Instruct のみ “Denmark” と出力できている（ただし，不要な “’s flag” も出力している）．

例 (e) は，「サブスク（サブスクリプション）」を表す逸脱的表記 “さぶすく” を含むテキストである．原文に対する翻訳結果では，Sarashina2.2-3B-Instruct のみ，“さぶすく” に当たる部分も含めて妥当な訳文を出力している．正規化後文に対する翻訳結果では，NLLB-200-3B を除き，各

	(a)	
Source	俺もどう森してみたなったど。	
Reference	I wanted to trying playing Animal Crossing too.	
NLLB	I’ve been trying to do that for years.	✗
Tower	I tried to be like him.	✗
Sarashina	I’ve been wanting to try Animal Crossing.	✓
Qwen	I also tried doing something about the forest.	✗
	(b)	
Source	... なんて事ない雑談の生放送でもスパチャが飛んでいく。	
Reference	... casual live chats can still attract tons of super chats.	
NLLB	... even if it’s a live broadcast of a small talk, it’s a spatch.	✗
Tower	... live broadcasts of casual chats are popular. It seems that the target of what people value is changing.	✗
Sarashina	... live streams of casual chats can still receive スパチャ.	✗
Qwen	... even live broadcasts of casual chatter can receive super chat donations.	△

表 8 WRIME-MT 開発セット事例に対する翻訳モデル（NLLB-200-3.3B，TowerInstruct-13B，Sarashina2.2-3B-Intstruct，Qwen3-32B）の出力の例．“...” は原文・参照訳・出力結果の一部を省略して表示していることを表す．“✓”，△，および “✗” は，原文の下線部の表現に対する訳がそれぞれ適切であること，一部不適切な点があること，適切でないことを示す（翻訳結果全体が適切であることを必ずしも意味しない）．

[*29] 参照訳は，逸脱的表記を反映した “Deeeenmaaaark” を含んでいるが，翻訳仕様に基づくと，本来 “Denmark” が適切である．

	Normalizer & Input text	Translator	Output text	Adequacy
(c)		Reference	It smells like early summer in the air!	
	[None] 初夏のかほりがするぜ	NLLB	I'm gonna miss the first summer.	✗
		Tower	The fragrance of early summer	✓
		Sarashina	The scent of early summer is palpable.	✓
		Qwen	The scent of early summer.	✓
	[Sarashina2.2-3b-Norm] 初夏の香りがするぜ	NLLB	Smells like the first summer.	✗
		Tower	The scent of early summer	✓
		Sarashina	The scent of early summer is here.	✓
		Qwen	It smells like early summer.	✓
(d)		Reference	How many more sleeps until Deeeenmaaaark	
	[None/DeBERTa-L-Norm] もーいーくつねーるとー、 でーんーまーあーくー♪	NLLB	♪ I'm not gonna let you down ♪	✗
		Tower	Moikutsubenaru to, donmai aku ♪	✗
		Sarashina	How many more days until Denmark? (Note: This is a ...	△
		Qwen	I'm going to sleep now, goodnight ♪	✗
	[DeBERTa-L-Norm] もーいつねる、ままでく♪	NLLB	♪ ♪ I'll be there, I'll be there ♪	✗
		Tower	Mooitsuneru, Mamadeku ♪	✗
		Sarashina	I don't know when I'll go to sleep yet.	✗
		Qwen	I'm going to sleep now, goodnight ♪	✗
	[Sarashina2.2-3b-Norm] もういくつ寝ると、 でーんマーク♪	NLLB	How much more sleep do you need, Mark?	✗
		Tower	How many more times can I sleep, don't Mark ♪	✗
		Sarashina	How many sleeps until Denmark's flag?	△
		Qwen	How many more naps until Daimark? ♪	✗
(e)		Reference	It looks like I can catch up on it with a subscription service, so I'll watch it all together eventually.	
	[None/DeBERTa-L-Norm] さぶすくでみれそうだから いずれまとめて。	NLLB	I'm sure I'll miss it so I'll have to put it together.	✗
		Tower	I'll catch up on it sometime.	✗
		Sarashina	I think I can catch up through subscription so I'll watch it all at once later.	✓
		Qwen	I think I can watch it on subsoku, so I'll watch it all later.	✗
	[Sarashina2.2-3b-Norm] サブスクでみれそうだから いずれまとめて。	NLLB	I'm sure I'll see it on the subscript so I'll put it together.	✗
		Tower	I'll probably watch it later with a subscription.	△
		Sarashina	It seems like it's available on subscription, so I'll catch up later when I have time.	△
		Qwen	I think I can watch it through a subscription service, so I'll catch up later.	△

表 9 WRIME-MT 開発セット事例に対する翻訳モデル（NLLB-200-3.3B，TowerInstruct-13B, Sarashina2.2-3B-Intstruct，Qwen3-32B）の出力の例．Adequacy 列の "✓"，△，および "✗" は，Adequacy の観点でそれぞれ翻訳結果が妥当であること，一部不適切な点があること，不適切であることを意味する（Fluency については不問）．不適切という判断の根拠に当たる個所を赤字で示した（判断根拠が訳抜けの場合，特に修飾なし）．

翻訳モデルは "subscription (service)" と適切な訳語を出力した（ただし，3 モデルとも "まとめて" に当たる訳が欠落している）．

5. おわりに

本稿では，日本語ソーシャルメディア投稿を用いた日英・日中機械翻訳評価のための対訳データセット WRIME-MT の構築と評価について報告した．WRIME-MT を用いて，多言語および日本語中心のオープンモデルの翻訳精度を評価したところ，GemmaX2-28-9B，Sarashina-2.2-3B-Instruct および Qwen3（8B 以上）が両言語方向で高い翻訳精度を示した．また，機械翻訳の前段タスクとして逸脱的表記の正規化について検証したところ，Sarashina-2.2-3B に基づく高精度な正規化モデルを用いた場合，翻訳精度が低程度～高程度であった複数の翻訳モデルで，翻訳精度が向上することを示した．ただし，逸脱的表記に頑健な

Sarashina-2.3-3B-Instruct では，自動正規化の有効性は確認できず，正規化精度に改善の余地があることがわかった．また，逸脱的表記以外の現象では，事例分析から，固有名や略語・スラングについて翻訳エラーが生じることを確認した．

今後の方向性として，逸脱的表記やその他の逸脱的現象に対する言語モデルの頑健性・多言語処理能力を向上させる研究が考えられる．たとえば，原言語同士または原言語・目標言語の疑似対訳や，人手作成された正規化データセットを用いたモデルの事後学習が有効である可能性がある．また，固有名や略称・スラングについては，新出の表現への対応が重要となるため，エンティティリンキングや検索拡張生成 [51] などの戦略が有望と考えられ，今後検証を行いたい．

研究の限界　本データセットに収録された投稿は著者 1 名が選択したものであり（3.1 節），選択が恣意的であることは否定できない．今後，ランダムまたは別の作業者によって選択された投稿を追加し，今回の実験結果（4 節）と，翻訳モデル評価結果の傾向が同様となるかを確認したい．また，本データセットにおける言語情報は著者 1 名が付与したものである（3.2 節）．今後，データセットを拡張し，複数のアノテータがアノテーションを行った際の一致率を計測することを検討している．

謝辞　翻訳作業の仕様定義に関して助言・ご協力いただいた藤田篤氏，Benjamin Marie 氏に感謝いたします．

参考文献

[1] Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 91–102, Florence, Italy, August 2019. Association for Computational Linguistics.

[2] Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 76–91, Online, November 2020. Association for Computational Linguistics.

[3] Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[4] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics.

[5] 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.

[6] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *The 6th International Conference on Learning Representations*, 2018.

[7] Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1916–1920, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Hamzat Olanrewaju Aliyu Amina Gogo Tafida Bashar Umar Kangiwa Nasiru Muhammad Dankolo Abubakar Ahmad Aliero, Bashir Sulaimon Adebayo. Systematic review on text normalization techniques and its approach to non-standard words. *International Journal of Computer Applications*, Vol. 185, No. 33, pp. 44–55, Sep 2023.

[9] Pidong Wang and Hwee Tou Ng. A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 471–481, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[10] José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. Phonetic normalization for machine translation of user generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 407–416, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] Sina Ahmadi and Antonios Anastasopoulos. Script normalization for unconventional writing of under-resourced languages in bilingual communities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14466–14487, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] Alexandre Berard, Ioan Calapodescu, and Claude Roux. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 526–532, Florence, Italy, August 2019. Association for Computational Linguistics.

[13] Benjamin Marie and Atsushi Fujita. Synthesizing parallel data of user-generated texts with zero-shot neural machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 710–725, 2020.

[14] Paul Michel and Graham Neubig. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 543–553, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[15] Ryo Fujii, Masato Mita, Kaori Abe, Kazuaki Hanawa, Makoto Morishita, Jun Suzuki, and Kentaro Inui. PheMT: A phenomenon-wise dataset for machine translation robustness on user-generated contents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5929–5943, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[16] José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. Understanding the impact of UGC specificities on translation quality. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 189–198, Online, November 2021. Association for Computational Linguistics.

[17] Paul McNamee and Kevin Duh. The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 910–918, Marseille, France, June 2022. European Language Resources Association.

[18] Rachel Bawden and Benoît Sagot. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 198–216, Singapore, December 2023. Association for Computational Linguistics.

[19] 笠原要, 斉藤いつみ, 浅野久子, 片山太一, 松尾義博. テキスト正規化技術を用いた CGM 日本語テキスト翻訳. 言語処理学会 第 21 回年次大会 発表論文集, pp. 804–807, March 2015.

[20] Shohei Higashiyama and Masao Utiyama. Comprehensive evaluation on lexical normalization: Boundary-aware approaches for unsegmented languages. arXiv:2505.22273, 2025.

[21] Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 22–64, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[23] OpenAI and Others. GPT-4 technical report. arXiv:2303.08774, 2023.

[24] Henny Sluyter-Gäthje, Pintu Lohar, Haithem Afli, and Andy Way. FooTweets: A bilingual parallel corpus of World Cup tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[25] Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 168–176, Hong Kong, November 2019. Association for Computational Linguistics.

[26] Yo Ehara. To what extent does lexical normalization help English-as-a-second language learners to read noisy English texts? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 451–456, Online, November 2021. Association for Computational Linguistics.

[27] Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pp. 126–135, Beijing, China, July 2015. Association for Computational Linguistics.

[28] Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 493–509, Online, November 2021. Association for Computational Linguistics.

[29] Johann Roturier and Anthony Bensadoun. Evaluation of MT systems to translate user generated content. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September 19-23 2011.

[30] Marlies van der Wees, Arianna Bisazza, and Christof Monz. Five shades of noise: Analyzing machine translation errors in user-generated text. In *Proceedings of the Workshop on Noisy User-generated Text*, pp. 28–37, Beijing, China, July 2015. Association for Computational Linguistics.

[31] Maja Popovic, Ekaterina Lapshinova-Koltunski, and Maarit Koponen. Effects of different types of noise in user-generated reviews on human and machine translations including ChatGPT. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pp. 17–30, San Ġiljan, Malta, March 2024. Association for Computational Linguistics.

[32] Leiyu Pan, Yongqi Leng, and Deyi Xiong. Can large language models learn translation robustness from noisy-source in-context demonstrations? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2798–2808, Torino, Italia, May 2024. ELRA and ICCL.

[33] OpenAI. Introducing ChatGPT, 2022. `https://openai.com/index/chatgpt/`.

[34] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2095–2104, Online, June

2021. Association for Computational Linguistics.

[35] Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. A Japanese dataset for subjective and objective sentiment polarity classification in micro blog domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7022–7028, Marseille, France, June 2022. European Language Resources Association.

[36] Risa Kondo, Ayu Teramen, Reon Kajikawa, Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Hideaki Hayashi, Yuta Nakashima, and Hajime Nagahara. Text normalization for Japanese sentiment analysis. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pp. 149–157, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics.

[37] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191. Association for Computational Linguistics, October 2018.

[38] 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. Association for Computational Linguistics, July 2002.

[39] 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. Association for Computational Linguistics, December 2022.

[40] Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 663–671. Association for Computational Linguistics, December 2023.

[41] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. arXiv:2207.04672, 2022.

[42] Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks, 2024.

[43] Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. X-ALMA: Plug play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*, 2025.

[44] Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5420–5443, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[45] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. arXiv:2505.09388, 2025.

[46] Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. arXiv:2501.16937, 2025.

[47] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA, October 2024.

[48] SB Intuitions. Sarashina2, 2024. `https://www.sbintuitions.co.jp/blog/entry/2024/06/26/115641`.

[49] SB Intuitions. Sarashina2.2, 2024. `https://www.sbintuitions.co.jp/blog/entry/2025/03/06/112144`.

[50] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

[51] Bryan Zhang, Taichi Nakatani, and Stephan Walter. Enhancing E-commerce product title translation with retrieval-augmented generation and large language models. In *The 4th International Workshop on Data-Centric AI*, 2024. arXiv:2409.12880, 2024.

付　　録

A.1 実験設定の詳細

実験に用いた事前学習モデルを表 A·1 に示す．翻訳モデルの推論では，decoder-only モデルについて表 A·2 に示すプロンプトを使用した．“{tgt_lang}” は目標言語の文字列（日本語プロンプトでは “英語” または “中国語”, 英語プロンプトでは “English” または “Chinese”），“{src_text}” は翻訳対象の原文を挿入するプレースホルダを表す．encoder-decoder および decoder-only モデル共通で，ハイパーパラメタは num_beams=1 で greedy search（do_sample=False）を用いた．Qwen3 モデルでは，non-thinking モードを適用した．また，モデルの入力や各評価指標のスコア計算時に用いたテキストには，入力前に NFKC 正規化を適用した．

Hugging Face ID
ku-nlp/deberta-v2-large-japanese-char-wwm
sbintuitions/sarashina2.2-3b
facebook/nllb-200-3.3B
Unbabel/TowerInstruct-13B-v0.1
haoranxu/X-ALMA-13B-Group6
ModelSpace/GemmaX2-28-9B-v0.1
SakanaAI/TinySwallow-1.5B-Instruct
sbintuitions/sarashina2.2-0.5b-instruct-v0.1
sbintuitions/sarashina2.2-1b-instruct-v0.1
sbintuitions/sarashina2.2-3b-instruct-v0.1
sbintuitions/sarashina2-7b
sbintuitions/sarashina2-70b
tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2
tokyotech-llm/Llama-3.3-Swallow-70B-Instruct-v0.4
Qwen/Qwen3-1.7B
Qwen/Qwen3-4B
Qwen/Qwen3-8B
Qwen/Qwen3-14B
Qwen/Qwen3-32B

表 A·1 実験において正規化（上段）・翻訳（下段）モデルとして利用した事前学習モデル．

Lang	Prompt
ja	次の日本語のテキストを {tgt_lang} に翻訳してください。最後の日本語テキストに対する翻訳のみ出力し、改行文字（”\n”）は出力しないでください。\n\nJapanese:\n{src_text}\n{tgt_lang}\n
en	Translate the following Japanese text into {tgt_lang}. Output only the translation of the final Japanese text without including any newline characters (”\n”). \n\nJapanese:\n{src_text}\n{tgt_lang}\n

表 A·2 Decoder-only モデルで使用した日本語および英語プロンプト．