日付入り LLM 文書翻訳評価用データセット

岩月憲一　根石将人
株式会社みらい翻訳
{iwatsuki,neishi}@miraitranslate.com

概要

大規模言語モデル（LLM）による翻訳の評価に際し，評価に用いるデータが LLM の事前学習データに含まれている場合，適切な評価ができない。本研究では，この問題を回避するために，頻繁に更新されるニュース系のリソースを利用して，英日の文書翻訳評価用データセットを構築した。本データセットは公開されており，不定期に更新される予定である。また，利用者が自ら更新できるよう，データ構築に用いたソースコードも合わせて公開する。

1 はじめに

大規模言語モデル（LLM）の評価にあたっては，評価用のデータが LLM の事前学習データに含まれていること（リーク）による影響が懸念されている[1]。これは機械翻訳にも同様に当てはまり，実際にリークの影響により正しく評価できないことが指摘されている[2]。公開されている評価用データセットは，それ自体だけでなく，その出典や引用等が事前学習データに含まれる場合がある。この問題を回避するためには，モデルが構築された時点よりも後に作成された文書を評価に用いることが考えられる。

そこで，本研究では，各文書の公開日の情報が付与されている LLM 文書翻訳評価用のデータセットを構築した[i]。これによって，評価したい LLM の事前学習データに含まれていないと考えられるデータのみを用いて適切に翻訳の評価を行うことができる。

また，新たに公開される LLM を評価するためには，データセット自体も更新される必要がある。本データセットは不定期に更新される予定であり，既に 2024 年 5 月, 9 月, 12 月に更新済みである。また，データセットの更新を利用者が自ら行えるよう，構築に必要なソースコードも合わせて公開する。

作成したデータセットを用いて，英日方向の翻訳評価実験を行った。まず，2021～2022 年の文書データを用いて，4 種類の LLM を LoRA[3]によりファインチューニングした。続いて，リークの可能性がない 2024 年 10～11 月のデータと，可能性がある 2023 年 2～4 月のデータ（一部，後述）を用いて翻訳を行い，各 LLM の COMET[4]と BLEU[5]を算出した。結果として，最新でない文書による評価はリークの影響を受ける可能性が示唆され，構築したデータセットの有用性が示された。

2 データセット構築

2.1 文書レベルと段落単位

商用の翻訳システムに入力される翻訳対象の多くは文書である。例えば，論文の PDF ファイルを入力するという使い方がなされる。文書画像を入力とする翻訳評価用データセット[6]も提案されているが，テキスト抽出のエラーが翻訳モデル単独での評価に影響するのを避けるため，本研究ではプレーンテキストとして抽出された文書に着目した。

LLM は従来より長い入力に対応するようになったが，1 つの文書を全て読み込めるモデルは少ない。しかし，文単位まで細かくして処理すると，文を跨いだ文脈の情報が失われてしまい，適切な翻訳が原理的に難しい。そこで，本研究では，入力長と文脈の連続性のバランスをとり，段落を最小単位に設定する。つまり，文書は複数の段落からなると定義され，それより細かくはしない。段落を最小単位とする実験設定は以前から提案されており[7]，WMT24 でも採用された[8]。

2.2 データソース

データソースの選定にあたり，以下の 5 つの条件を設定した。

1. 対訳関係であり，文書間のアラインメントが容易であること
2. 段落の情報があること

[i] 公開先：https://github.com/miraitranslate/mirai-full-document-eval

3. 頻繁に更新されること
4. 内容に新規性があること（過去の文章をそのまま転載するような文書，例えば法律の条文を掲載したり，既存の文学作品を載せたりするようなものでないこと）
5. 著作権その他の権利上の問題がないこと

以上の条件を満たすデータソースとして，日本政府の一部のウェブサイトにある，ニュースリリースが挙げられた。特に，金融庁，首相官邸，経済産業省，財務省のウェブサイトを選定した。

これらのウェブサイトにおいては，英語のページと日本語のページの対応関係がハイパーリンクによって明示されている。また，HTML タグによって段落が明示されている場合が多かった。さらに，政府標準利用規約（第 2.0 版）に則っており，「複製，公衆送信，翻訳・変形等の翻案等，自由に利用できます。商用利用も可能です」といった文言が利用規約に書かれている[ii]。

2.3 構築手順

まず，各ウェブサイトより，対訳になっているウェブページの HTML ファイルを取得した。次に，HTML ファイルから，段落ごとに文章を抽出した。一部のウェブページにおいて，段落が HTML タグによらず，改行＋スペースによって表されていたが，これも段落として処理した。見出しや箇条書きも抽出対象としたが，ヘッダやフッタなど，そのウェブサイトに共通する部分については，内容に新規性がなく，文書の中身ではないと判断し除外した。また，表のみの文書も収録しなかった。抽出後，段落数が英日で一致しない文書があったが，これらは英語版のウェブページが，日本語版の翻訳ではなく要約であり，対訳ではないため除外した。

リークに関する情報として，CommonCrawl[iii]の各バージョンに英日それぞれの文書が含まれているかどうかも記録した。多くのオープンな[iv]LLM に事前学習データとして CommonCrawl やこれを加工したデータセットが用いられている（例えば，LLM-jp[9]，Swallow[10]，Sarashina2[11]）ことから，日付だけでなくこの情報もリークの判別に有用だと考えられるためである。これは CommonCrawl より提供されている収録 URI リストと照合して調べた。照合対象は，CC-MAIN-2023-06 以降のバージョンとした。なお，古いバージョンに収録されているウェブページが新しいバージョンにも収録されるという包含関係は必ずしも成り立たないことに注意されたい。

表 1　文書数

公開年	10 段落以下かつ 英語 1,024 語以下	全体
～2020	1,294	1,407
2021	430	460
2022	306	327
2023	452	477
2024	247	260
合計	2,729	2,931

2.4 収録項目

構築したデータセットには，各文書について，以下の項目を収録した。

1. 文書 ID
2. 英文ページの出典 URI
3. 和文ページの出典 URI
4. 英語の段落の配列
5. 日本語の段落の配列
6. 文書の公開年月（整数型）
7. 文書の公開年月日（文字列型）
8. 段落数
9. CommonCrawl への収録の有無

2.1 節で述べた通り，段落を最小単位とするため，文書は段落に分割されている。英語と日本語の段落数は一致することが保証される。故に段落数の情報は冗長であるが，使用時の利便性のために含めた。

文書の公開日の情報は，月単位でのフィルタリングを想定し，年月日と年月の 2 種類を用意した。

2.5 統計

構築したデータセットに含まれる文書数を表 1 に示す。古いデータについては，評価データとしての使用ではなく，訓練データとしての使用や，リークに関する実験での使用を想定して収録した。また，LLM が対応する最大コンテキスト長はモデルによっては小さいため，短い文書のみを抽出して用いることが考えられる。ここでは，10 段落以下かつ英語

[ii] ただし，加工後のデータを政府等が作成したと誤解を招くような様態は認められていない。
[iii] https://commoncrawl.org/
[iv] オープンソースという意味ではない。

表 2　平均トークン数

トークナイザ（huggingface）	10 段落以下かつ英語 1,024 語以下		全体	
	英語	日本語	英語	日本語
llm-jp/llm-jp-3-13b	196.3	158.7	255.4	208.5
sbintuitions/sarashina2-13b	227.5	142.4	295.1	185.9
tokyotech-llm/Swallow-13b-hf	218.9	222.8	283.4	290.1

が 1,024 単語以下のデータに絞り込んだ場合の数値を示した。次章の評価実験ではこの条件によって絞り込んだデータのみを用いた。

平均トークン数については，トークナイザに依存するため，一例を表 2 に示す。トークナイザの種類は，huggingface のモデル名によって表している。文書の言語毎に算出した。

3 評価実験

3.1 手法

構築したデータセットを用いて, LLM による文書翻訳の精度を評価した。言語方向は英日である。

実験設定は，次のとおりである。LLM の公開日を念頭に，リークの可能性の有無によって 2 種類のデータセットを作成した。リークの可能性がないデータには，公開日が 2024 年 10～11 月のデータを使用した。文書数は 35 件であった。リークの可能性があるデータについては，CommonCrawl の CC-MAIN-2023-23 版までに日本語の文書が収録されており，かつ公開日が 2023 年 2 月から 4 月であって，さらに LLM-jp Corpus v3[v]に含まれているデータを使用した。尤も，実際の事前学習データは非公開であるため，確実にリークがあるとは言えない。ただし LLM-jp-3 の LLM については，LLM-jp Corpus v3 が実際の事前学習データであると想定されるため，リークがある確度が高いと考えられる。なお，モデルに対して文書の前半を入力し，その後の内容を予測させることで事前学習データへの混入を検知する方法[12]ではリークの可能性の高い文書を絞り込めなかった。

LLM への入力は文書とした。文書は 1 つまたは複数の段落からなる。段落は改行によって示した。

Zero-shot や few-shot では，複数段落の文書の段落数が維持されない問題が頻発したため，LoRA チューニングによって LLM を文書翻訳に特化させた。具体的には，2023 年 1 月のデータでエポック数などのハイパーパラメータ調整を行い，2021～2022 年のデータを用いて LoRA チューニングを行った。

使用した LLM は，Llama-2[13]，LLM-jp-3[9]，Sarashina2[11]，Swallow[14]であり，いずれも base モデルで，パラメータサイズは 13B である。モデルの公開時期はそれぞれ，2023 年 7 月，2024 年 9 月，2024 年 6 月, 2023 年 12 月である[vi]。Llama-2 は, 2022 年 9 月までのデータが事前学習に使用されており[vii]，本実験で用意した 2 種類の評価データどちらにおいてもリークの影響がないモデルとみなせる。2 種類の評価データは異なる文書から構成されるため，翻訳の難易度に差があることが想定され，単純なスコアの比較だけではリークの影響の有無を判断できない。そこで，どちらの評価データについてもリークの影響がない LLM について 2 種類の評価データ間のスコア差分を算出し，これと各モデルのスコア差分を比較することでリークの影響を分析する。

評価指標には, COMET（COMET-22[4]）と BLEU[5]を用いた。2.1 節で述べた通り，最小単位を段落として評価するため，COMET の入力は段落とし，BLEU も段落を文とみなして[15]算出した。BLEU は LLM の翻訳を評価する指標としては適さないとされている[16]が，文章の表層を記憶している可能性を確認する上では有用であると考え，採用した。

3.2 結果と考察

得られた自動評価指標のスコアを表 3 に示す。Llama-2 の COMET（百分率）の値と，それ以外のモデルの値を比較すると，リークの可能性がないデータでは差分が小さいが，リークの可能性があるデータでは差分が大きい。リークの影響により，スコアが不適切に高くなっている可能性がある。

[v] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

[vi] https://about.fb.com/news/2023/07/llama-2/
https://llmc.nii.ac.jp/topics/post-707/
https://www.sbintuitions.co.jp/news/press/20240614_01/
https://www.titech.ac.jp/news/2023/068089

[vii] https://huggingface.co/meta-llama/Llama-2-13b-hf

表 3　自動評価指標のスコア

公開時期*	モデル	リークなし（2024 年 10～11 月）		リークあり（2023 年 2～4 月）	
		COMET	BLEU	COMET	BLEU
2022 年 9 月	Llama-2	88.18	38.57	83.14	13.48
2024 年 9 月	LLM-jp-3	89.65	41.93	86.06	35.68
2024 年 6 月	Sarashina2	89.84	22.07	89.48	32.18
2023 年 12 月	Swallow	89.06	41.80	88.85	30.04

*Llama-2 については，公開時期ではなく knowledge cut-off date である。

表 4　翻訳例（一部段落抜粋・段落は改行区切り）

原文[17]	参照訳[18]	Llama-2	LLM-jp-3	Sarashina2	Swallow
On February 20, 2023, Prime Minister Kishida attended the fourth Tokyo Global Dialogue held in Tokyo. Prime Minister Kishida said in his address, “I am KISHIDA Fumio, Prime Minister of Japan.	令和５年２月２０日、岸田総理は、都内で開催された第４回東京グローバル・ダイアログに出席しました。 総理は、挨拶で次のように述べました。 「日本の総理大臣の岸田文雄です。	令和５年２月２０日、岸田総理は、都内で開催された第４回東京グローバル・ダイアローグに出席しました。 総理は、挨拶で次のように述べました。 「私は、岸田文雄（ふみお）日本国総理大臣であります。	令和５年２月２０日、岸田総理は、都内で開催された第４回東京グローバル・ダイアログに出席しました。 総理は、挨拶で次のように述べました。 「皆さん、こんにちは。岸田文雄です。	令和５年２月２０日、岸田総理は、都内で開催された第４回東京グローバル・ダイアログに出席しました。 総理は、挨拶で次のように述べました。 「内閣総理大臣の岸田文雄です。	令和４年２月２０日、岸田総理は､都内で開催された第４回東京グローバル･ダイアログに出席しました。 総理は挨拶で次のように述べました。 「皆様こんにちは。岸田文雄です。

表 5　入力の続きの出力結果

入力[18]	Llama-2	LLM-jp-3	Sarashina2	Swallow
令和５年２月２０日、岸田総理は、都内で開催された第４回東京	オリンピック・パラリンピック競技大会推進本部会合に出席しました。後略	グローバル･ダイアログに出席しました。	栄養サミット２０２３関係閣僚等会合に出席しました。後略	栄養サミットに出席しました。後略

リークの可能性があるデータの翻訳結果の一部を表 4 に示す。第 4 段落以下は省略した。第 1･2 段落は Llama-2 を含めほぼ参照訳通りであるが，第 3 段落は全モデルが参照訳と異なる訳を出力した。さらに, この文書の第 1 段落の一部を各 LLM に入力し，その続きを出力させた結果を表 5 に示す[viii]。LLM-jp-3 のモデルは参照訳の第 1 段落を再現できていることから，当該文書が学習されている蓋然性は高い。しかし，第 2 段落以降は生成されず，翻訳結果においても第 3 段落以降は参照訳と大きく異なる文章が出力された。このことから，使用した LLM の事前学習したデータを完全に再現する能力は高くなく，生成結果の分析ではリークの影響の判断は難しいと言える。適切な評価のためには，リークの可能性がない評価データを使用することが求められる。

4 おわりに

本研究では，日付入り LLM 文書翻訳評価用データセットを構築し, 評価実験を行った。実験の結果，最新でない文書を用いた評価ではリークの悪影響があることが示唆された。LLM は今後も更新され続けると見込まれるが，その評価には，同様に更新され続けるデータセットを使用する必要がある。

[viii] LoRA チューニング前の（zero-shot の）LLM でも同様の傾向であった。

参考文献

[1] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 10776–10787, Singapore, 2023.

[2] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2765–2781, Mexico City, Mexico, 2024.

[3] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In **the Tenth International Conference on Learning Representations**, Virtual, 2022.

[4] 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), 2022.

[5] 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.

[6] Benjamin Hsu, Xiaoyu Liu, Huayang Li, Yoshinari Fujinuma, Maria Nadejde, Xing Niu, Ron Litman, Yair Kittenplon, and Raghavendra Pappagari. M3T: A New Benchmark Dataset for Multi-Modal Document-Level Machine Translation. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)**, pp. 499–507, Mexico City, Mexico, 2024.

[7] Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. Challenges in Context-Aware Neural Machine Translation. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 15246–15263, Singapore, 2023.

[8] 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, et al.. 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, USA, 2024.

[9] LLM-jp. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv**, 2407.03963, 2024.

[10] 岡崎直観, 服部翔, 平井翔太, 飯田大貴, 大井聖也, 藤井一喜, 中村泰士, Mengsay Loem, 横田理央, 水木栄. Swallow コーパス: 日本語大規模ウェブコーパス. In **言語処理学会第 30 回年次大会発表論文集**, pp. 1498–1503, 神戸, 2024.

[11] 清野舜, 李凌寒, 高瀬翔. 大規模な日本語の事前学習言語モデル Sarashina1・2 の公開. https://www.sbintuitions.co.jp/blog/entry/2024/06/26/115641

[12] Shahriar Golchin and Mihai Surdeanu. Time Travel in LLMs: Tracing Data Contamination in Large Language Models. In **the Twelfth International Conference on Learning Representations**, Vienna, Austria, 2024.

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. **arXiv**, 2307.09288, 2023.

[14] 藤井一喜, 中村泰士, Mengsay Loem, 飯田大貴, 大井聖也, 服部翔, 平井翔太, 水木栄, 横田理央, 岡崎直観. 継続事前学習による日本語に強い大規模言語モデルの構築. In **言語処理学会第 30 回年次大会発表論文集**, pp. 2102–2107, 神戸, 2024.

[15] Jindřich Libovický, Thomas Brovelli, and Bruno Cartoni. Machine Translation Evaluation beyond the Sentence Level. In **Proceedings of the 21st Annual Conference of the European Association for Machine Translation**, pp. 199–208, Alicante, Spain, 2018.

[16] Xianfeng Zeng, Yijin Liu, Fandong Meng, and Jie Zhou. Towards Multiple References Era – Addressing Data Leakage and Limited Reference Diversity in Machine Translation Evaluation. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 11939–11951, Bangkok, Thailand, 2024.

[17] https://japan.kantei.go.jp/101_kishida/actions/202302/_00027.html

[18] https://www.kantei.go.jp/jp/101_kishida/actions/202302/20tgd.html