大規模言語モデル時代の機械翻訳の展望

東山翔平 *

国立研究開発法人情報通信研究機構　ユニバーサルコミュニケーション研究所
* 責任著者: shohei.higashiyama@nict.go.jp

あらまし　近年の大規模言語モデルの発展は目覚ましく，自然言語処理の諸技術，特に機械翻訳を含むテキスト生成技術は大きな進展を遂げている．本稿では，大規模言語モデルを用いた機械翻訳研究の進展と主な課題を紹介するとともに，今後の展望を述べる．

キーワード　機械翻訳，多言語処理，大規模言語モデル

1. はじめに

近年の大規模言語モデル（Large Language Model; LLM）の発展は目覚ましく，自然言語処理の諸技術，特に機械翻訳を含むテキスト生成技術は大きな進展を遂げている(注1)．本稿では，GPT-3 [2] 以降のデコーダ構造の言語モデルに焦点を当て，LLM を用いた機械翻訳研究の進展と主な課題を紹介するとともに，今後の展望を述べる．

2. 翻訳と機械翻訳

機械翻訳は，処理内容の性質の点でも，プロセスの範囲の点でも，人間による翻訳と異なるものである．翻訳とは，「起点言語文書が担う役割を果たす目標言語文書を産出する操作」[3] であり，そのための翻訳対象文書に関する背景知識や用語等の調査も伴う行為である．対して，機械翻訳が行っていることは，「起点言語の文字列を目標言語の文字列に変換するという統計的な記号処理」[3] である．両者が異なることを端的に表し，機械翻訳が「X 文 Y 訳」と称されることもある [4]．また，商業翻訳サービスが多くのプロセスを含むワークフローにより遂行されているのに対し，機械翻訳が担い得るプロセスはそのごく一部である．たとえば，翻訳・通訳に関する国際標準規格である ISO 17100(注2)では，制作前・制作・制作後プロセスからなる翻訳ワークフローが定義されており，「翻訳」プロセス以外にも，翻訳者自身による「セルフチェック」，翻訳者以外による「バイリンガルチェック」，「最終検品」が必須のプロセスとして定められている [5]．一方，典型的な機械翻訳は，入出力上は ISO 17100 における「翻訳」プロセスに相当する工程を行っているとみなせるのみで，チェックにあたるプロセスを含んでいない．したがって，機械翻訳結果を商業翻訳で利用する際は，人間による全数検査・修正を経た品質保証が不可欠である．具体的には，機械翻訳結果の利用および修正を前提とする翻訳ワークフローはポストエディット（後編集）と呼ばれ，ポストエディットについての国際標準規格 ISO 18587(注3)が規程されている．

3. 大規模言語モデルによる機械翻訳の進展

上述のように，翻訳との間に根本的な違いがあることを前提としつつも，機械翻訳はこの 10 年の間に大きな発展を遂げている．2014〜2016 年頃には，深層学習を用いたニューラル機械翻訳（Neural Machine Translation; NMT）が登場・普及し，大幅な翻訳精度の向上という顕著な発展があった [6]．LLM による機械翻訳（以降，LLM 翻訳と呼ぶ）は，文字列を操作する「統計的な記号処理」である点に大きく変わりはないものの，NMT による発展に続く大きな転換点を生みつつある．LLM 翻訳は，ChatGPT(注4)などの LLM チャットボットサービスを用いて誰もが気軽に利用できるようになり，その翻訳能力の高さや指示に基づいた対応能力の柔軟さは簡単に体感できる．定量的にも，一般的な評価指標に基づく翻訳精度評価（文単位を基本とする翻訳結果の適切さの自動および人手評価）において，LLM 翻訳は従来型 NMT(注5)と同等以上であることが報告されており [7], [8]，本分野での盛んな研究開発により，今後も精度面での継続的な発展が見込まれる．また，LLM 翻訳に期待できる点は，翻訳を模した処理としての機械翻訳の品質向上 [9](注6)の観点のみにとどまらず，従来の機械翻訳からのスコープの拡大も視野に入る [10]．

具体的には，LLM の長所を生かした翻訳シナリオとして，たとえば以下の実現性が高まっている．

(1)　文脈を考慮した長い文章の翻訳
(2)　適切な訳語や訳出スタイルの使用
(3)　インタラクティブな翻訳
(4)　シームレスな言語・タスク横断処理

(1) と (2) は，人間による翻訳では必須の要求事項となる観点である．これらは従来の機械翻訳からの課題であったが [11], [12], [13], [14], GPT-3.5 や GPT-4 [15] を中心とする LLM により，従来以上の性能が達成されている [16], [17](注7)．

(注1)：本稿の主題からはやや離れるが，国内の自然言語処理研究コミュニティでの LLM に対する受け止め方については，文献 [1] を参照すると，ChatGPT が出現した 2022 年末〜2023 年初頭頃の空気感を感じられるように思われる．
(注2)：https://www.iso.org/standard/59149.html
(注3)：https://www.iso.org/standard/62970.html
(注4)：https://openai.com/index/chatgpt/
(注5)：エンコーダ・デコーダ構造の機械翻訳モデルを指すものとする．
(注6)：人間の翻訳方略を模した LLM 翻訳の研究の一例として，He ら [9] は，Chain-of-Thought プロンプトを用いて，入力文の翻訳に必要な知識を LLM から引き出した上で複数の翻訳文候補を生成させ，品質推定モデル（本節 (8) で後述）により最良の候補を選択する枠組みを提案し，商用 LLM（GPT-3.5）およびオープン LLM（Alpaca，Vicuna）での翻訳精度向上を実現している．
(注7)：Wang ら [16] は，多分野の文書翻訳（中国語 → 英語，英語 → ドイツ語/

(1) **文脈を考慮した長い文章の翻訳**とは，長い文章において生じる談話的現象を適切に扱える必要があることを指す．これには，翻訳結果（目標言語テキスト）の文章全体を通じて文法的・語彙的な一貫性があることや，適切に代名詞や指示語の照応を解決（指示対象を特定）したり，省略されている語句を補完したりした上で翻訳結果に反映することなどが含まれる [11]．LLM 翻訳においては，LLM の長い文脈を捉える能力が役立っているとみられる．たとえば，Pang ら [8] はドイツ語 → 英語翻訳において，従来型 NMT では単語数 90 以上の入力テキスト（翻訳対象テキスト）に対して翻訳精度が顕著に低下していくのに対し，LLM では 500 単語以上の入力に対しても大きな低下が生じないことを報告している．

(2) **適切な訳語や訳出スタイルの使用**とは，対象分野ごとの常識あるいは指定された翻訳仕様に応じて，従うべき用語・言い回し・スタイルなどがあることを意味する．たとえば，野球に関する英語文書の和訳で “make (a/the) start” を「先発登板する」と訳す必要があることや，日本の法令文書の英訳で「(第 1) 項」を “paragraph (1)” と訳す必要がある(注8)といった例が挙げられる．LLM 翻訳では，適切なプロンプト（LLM への入力テキスト）を与えることで，指示学習（Instruction Tuning）により獲得された指示追従能力が効果を発揮しているとみられる．また，入力テキスト外の関連情報を外部文書群から取得して LLM の入力に加える検索拡張生成（Retrieval-Augmented Generation; RAG）[18] により，従うべき訳語やスタイルの具体例を与えることも有効とみられ，Moslem らの研究 [17] はその一例である(注9)．

(3) **インタラクティブな翻訳**とは，システムが出力した翻訳結果に対する対象箇所や修正方針などのユーザのフィードバックに応じて，動的に出力内容が変更・修正される状況を指す(注10)．(4) **シームレスな言語・タスク横断処理**とは，同様に指示を通じて翻訳にとどまらない文章校正，要約，質問応答などの処理を，複数言語にまたがりながら行わせるような状況を指す．(3) と (4) は，特別に設計されたシステム（たとえば，インタラクティブ翻訳システム [19] や言語横断要約手法 [20]）を除き，従来の機械翻訳にとってはスコープ外といえるシナリオである．また，人間の翻訳者に翻訳を依頼する状況と比べると，即時的に回答が得られることに加え，翻訳の範疇を超えるタスクを容易に依頼できることは機械特有の利点になる．

以上のような点を踏まえると，LLM 翻訳システムは，人間の翻訳品質に近づいていく側面と，人間の翻訳とは異なる方向性で進歩していく側面との両面での発展が期待され，多言語テキストの読解・産出（つまり，「読む」「聞く」「書く」「話す」活動）の支援のための汎用多言語処理システムといった役割も担うようになると考えられる．

ただし，上記に限らず，LLM 翻訳において解決・発展が望まれる課題は他にも多く存在する．以下，主なものを挙げる．

(5) **低資源言語への対応**：これまでの機械翻訳の成功は主に高資源言語に対してのものと言える．従来型 NMT では，高精度な翻訳のために大規模な対訳テキスト（数百万〜数千万文程度）からの学習が必要であるし，LLM では，明示的に対応付けられた対訳テキストから学習していなくても高い翻訳能力を持つ場合があるとはいえ，膨大な量の（多言語の）単言語テキストでの事前学習が必要である．実際，従来型 NMT と同様，LLM 翻訳でも，非英語間の翻訳や低資源言語の翻訳において翻訳精度が低くなっていることが示されている [21]．高資源言語に対して高性能な LLM をベースにしつつ，少ないデータで低資源言語に適応可能にするような方法が求められる．

(6) **文化的バイアスの解消**：LLM が有する知識や出力する「意見」は，一部の言語に対応する文化圏のものに偏っていることが確認されており [22], [23]，これは LLM の学習データの大部分を一部の言語，典型的には英語のテキストが占めているという状況に起因する．このことは，資源が少ない言語ほど，対応する文化圏に特有の情報について適切な内容のテキストを生成できないことや，文化に依存する事物や概念（Cultural-Specific Items; CSI）を他言語へ適切に翻訳できないことなどに繋がる [24]．そのため，英語中心の LLM に対し，対象言語や地域ごとの文化差に対する感度を高めたり，多様な文化圏の知識を取り込れたりするための方法が必要となる．

(7) **モデルの小規模化**：巨大な LLM の学習・利用には多大な費用・計算資源・消費電力等がかかるというコスト面の理由，機密情報を含む文書をオンプレミス環境で扱う必要があるというセキュリティ面の理由などから，高い性能を維持しながらモデルサイズ（パラメタ数）を小規模に抑えることが重要となる．この方向性に関連する有望な研究の一つに Xu らの研究 [25] がある．Xu らは，事前学習済みの LLM に対し，5 言語の目標言語の単言語データでの学習と，少量の高品質な対訳データ（5 言語合計 58,000 件）での学習という 2 段階の追加学習（fine-tuning; 微調整）を行うことで，パラメタ数 7B（billion）および 13B のモデルで GPT-3.5(注11)と同等以上の翻訳精度を達成している．

最後に，上述の課題の解決に加えて，(8) LLM による**信頼性の高い自動評価・翻訳品質推定**(注12)の実現も期待される．機械翻訳システム出力の翻訳品質を測る自動評価尺度として，最

ロシア語）において，GPT-3.5/4 が大半の分野で商用機械翻訳サービスと同等以上の自動評価スコアを達成したこと，オープン LLM（対訳データで fine-tuning した Llama 2) でも一部分野でそれらに近いスコアを達成したことを示した．Moslem ら [17] は，GPT-3.5 において，翻訳メモリとみなした対訳テキストから曖昧検索で抽出した対訳事例をプロンプトに加えることで，従来型 NMT よりも高い自動評価スコア（英語 →5 言語）が達成されること，さらに用語集から抽出した用語対訳をプロンプトに加えることで，人手評価スコア（英語 →3 言語）が向上することを示した．

（注8）：『法令用語日英標準対訳辞書（令和 6 年 3 月改訂版）』（https://www.japaneselawtranslation.go.jp/ja/dicts/download）より．

（注9）：(2) に関して，他に，LLM 内部知識の活用の観点も関連する．LLM は広範な分野にまたがる大量・多言語のテキストからの事前学習を通じて分野特有の表現を学習していると想定されるものの，推論時にそれが表出するとは限らないため，学習済みの知識を適切に引き出せるようにすることも重要になる．

（注10）：自然言語でシステムと「対話」する方法は，インタラクティブ（対話的）な手法の一形態といえる．

（注11）：GPT-3（パラメタ数 175B）[2] 以降，OpenAI の後継モデルのパラメタ数は公表されていない．

（注12）：参照訳（人間により予め作成された翻訳結果の正解例）を用いず，原文と機械翻訳出力のみから翻訳品質を自動評価する方法は品質推定と呼ばれる．

近のニューラル評価尺度よりも GPT-4 やオープン LLM を用いた評価尺度の方が，システムレベル評価(注13)では人手評価スコアに基づく方法との相関が高いと報告されている [26], [27]．また，より実用性の高い評価シナリオでの品質評価に取り組んだ研究として，Treviso ら [28] は，入力された起点言語テキストおよび目標言語テキスト（何らかの機械翻訳システムによる翻訳結果）に対し，誤り内容の解説文と，訂正後の翻訳結果を出力する多言語モデル xTower を提案した(注14)．解説文は約 6 割のケースで有益な解説であったと分析されており，訂正後の翻訳結果により全体としての翻訳精度が向上することが示されている．こうした研究がさらに発展し，実用性と信頼性の高い自動評価が可能になれば，チェックの観点で人間の翻訳とのギャップを縮めることや，ポストエディットなどの翻訳ワークフローの効率性を高めることに繋がると考えられる．

4. おわりに

本稿では，翻訳と機械翻訳の違いを踏まえつつ，LLM を用いた機械翻訳研究について，八つの観点 (1)〜(8) の翻訳シナリオや課題に焦点を当てた現状と展望の議論を行った．挙げた観点は網羅的なものではないため，他の話題については文献 [8], [10] なども参照されたい．本分野での盛んな研究開発により，継続的な発展，特に人間の翻訳品質に近づいていく側面と，人間の翻訳とは異なる方向性で進歩していく側面との両面での発展が期待される．

謝　辞

本稿の内容について有益なコメントをくださった内山将夫氏，田中英輝氏，藤田篤氏に感謝いたします．

文　献

[1] 乾健太郎，“ChatGPT の出現は自然言語処理の専門家に何を問いかけているか，” 自然言語処理，vol.30，no.2，pp.273–274，2023.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” Advances in Neural Information Processing Systems, vol.33, pp.1877–1901, Curran Associates, Inc., 2020.

[3] 藤田　篤，山田　優，影浦　峡，“翻訳と機械翻訳：年次大会のテーマセッションを通じての知見，” 自然言語処理，vol.27，no.4，pp.975–981，2020.

[4] 影浦　峡，“改めて、翻訳とは何か：Google NMT が使える時代に，” 言語処理学会 第 23 回年次大会 発表論文集，pp.931–934，2017.

[5] 田嶌奈々，“特集「言語サービスの国際規格」『何でも教えてキカク』今さら聞けない基本知識 − おさらい編，” JTF ジャーナル，no.280，pp.6–11，日本翻訳連盟，2015.

[6] 中澤敏明，“機械翻訳の新しいパラダイム：ニューラル機械翻訳の原理，” 情報管理，vol.60，no.5，pp.299–306，2017.

[7] T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, P. Koehn, B. Marie, C. Monz, M. Morishita, K. Murray, M. Nagata, T. Nakazawa, M. Popel, M. Popović, and M. Shmatova, “Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet,” Proceedings of the Eighth Conference on Machine Translation, pp.1–42, Association for Computational Linguistics, Singapore, Dec. 2023.

[8] J. Pang, F. Ye, L. Wang, D. Yu, D.F. Wong, S. Shi, and Z. Tu, “Salute the classic: Revisiting challenges of machine translation in the age of large language models,” 2024. https://arxiv.org/abs/2401.08350

[9] Z. He, T. Liang, W. Jiao, Z. Zhang, Y. Yang, R. Wang, Z. Tu, S. Shi, and X. Wang, “Exploring Human-Like Translation Strategy with Large Language Models,” Transactions of the Association for Computational Linguistics, vol.12, pp.229–246, 03 2024.

[10] C. Lyu, Z. Du, J. Xu, Y. Duan, M. Wu, T. Lynn, A.F. Aji, D.F. Wong, and L. Wang, “A paradigm shift: The future of machine translation lies with large language models,” Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp.1339–1352, ELRA and ICCL, Torino, Italia, May 2024.

[11] S. Maruf, F. Saleh, and G. Haffari, “A survey on document-level neural machine translation: Methods and evaluation,” ACM Computing Surveys, vol.54, no.2, pp.1–36, March 2021.

[12] A. Fujita, “Attainable text-to-text machine translation vs. translation: Issues beyond linguistic processing,” Proceedings of Machine Translation Summit XVIII: Research Track, pp.215–230, Association for Machine Translation in the Americas, Virtual, Aug. 2021.

[13] K. Semenov, V. Zouhar, T. Kocmi, D. Zhang, W. Zhou, and Y.E. Jiang, “Findings of the WMT 2023 shared task on machine translation with terminologies,” Proceedings of the Eighth Conference on Machine Translation, pp.663–671, Association for Computational Linguistics, Singapore, Dec. 2023.

[14] Y. Wang, Z. Sun, S. Cheng, W. Zheng, and M. Wang, “Controlling styles in neural machine translation with activation prompt,” Findings of the Association for Computational Linguistics: ACL 2023, pp.2606–2620, Association for Computational Linguistics, Toronto, Canada, July 2023.

[15] OpenAI, “GPT-4 technical report,” 2024. https://arxiv.org/abs/2303.08774

[16] L. Wang, Z. Du, W. Jiao, C. Lyu, J. Pang, L. Cui, K. Song, D. Wong, S. Shi, and Z. Tu, “Benchmarking and improving long-text translation with large language models,” Findings of the Association for Computational Linguistics ACL 2024, pp.7175–7187, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, Aug. 2024.

[17] Y. Moslem, R. Haque, J.D. Kelleher, and A. Way, “Adaptive machine translation with large language models,” Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pp.227–237, European Association for Machine Translation, Tampere, Finland, June 2023.

[18] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” 2024. https://arxiv.org/abs/2312.10997

[19] S. Santy, S. Dandapat, M. Choudhury, and K. Bali, “INMT: Interactive neural machine translation prediction,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pp.103–108, Association for Computational Linguistics, Hong Kong, China, Nov. 2019.

[20] J. Zhu, Y. Zhou, J. Zhang, and C. Zong, “Attend, translate and summarize: An efficient method for neural cross-lingual summarization,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.1309–1321, Association for Computational Linguistics, Online, July 2020.

（注13）：2 つの機械翻訳システムの出力群を比較し，どちらのシステムがより高い翻訳品質であるかを判定する評価方法．

（注14）：xTower には，入力翻訳結果中のエラー箇所とそのエラーレベル（major, minor）も入力として与える必要がある．Treviso らは，人間により付与されたエラー情報と，自動エラー検出モデル xCOMET [29] により検出されたエラー情報の 2 種類を使用・比較し，前者よりも後者のエラー情報を用いた方が解説文の品質は低い一方，訂正後翻訳結果の品質には大きな差がないことを示している．

[21] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Multilingual machine translation with large language models: Empirical results and analysis," Findings of the Association for Computational Linguistics: NAACL 2024, pp.2765–2781, Association for Computational Linguistics, Mexico City, Mexico, June 2024.

[22] S. Shen, L. Logeswaran, M. Lee, H. Lee, S. Poria, and R. Mihalcea, "Understanding the capabilities and limitations of large language models for cultural commonsense," Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp.5668–5680, Association for Computational Linguistics, Mexico City, Mexico, June 2024.

[23] W. Wang, W. Jiao, J. Huang, R. Dai, J.-t. Huang, Z. Tu, and M. Lyu, "Not all countries celebrate thanksgiving: On the cultural dominance in large language models," Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.6349–6384, Association for Computational Linguistics, Bangkok, Thailand, Aug. 2024.

[24] B. Yao, M. Jiang, D. Yang, and J. Hu, "Benchmarking LLM-based machine translation on cultural awareness," 2024. https://arxiv.org/abs/2305.14328

[25] H. Xu, Y.J. Kim, A. Sharaf, and H.H. Awadalla, "A paradigm shift in machine translation: Boosting translation performance of large language models," The Twelfth International Conference on Learning Representations, pp.1–21, May 2024.

[26] T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pp.193–203, European Association for Machine Translation, Tampere, Finland, June 2023.

[27] Q. Lu, B. Qiu, L. Ding, K. Zhang, T. Kocmi, and D. Tao, "Error analysis prompting enables human-like translation evaluation in large language models," Findings of the Association for Computational Linguistics ACL 2024, pp.8801–8816, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, Aug. 2024.

[28] M. Treviso, N.M. Guerreiro, S. Agrawal, R. Rei, J. Pombal, T. Vaz, H. Wu, B. Silva, D. vanStigt, and A.F.T. Martins, "xTower: A multilingual LLM for explaining and correcting translation errors," 2024. https://arxiv.org/abs/2406.19482

[29] N.M. Guerreiro, R. Rei, D. vanStigt, L. Coheur, P. Colombo, and A.F.T. Martins, "xCOMET: Transparent machine translation evaluation through fine-grained error detection," 2023. https://arxiv.org/abs/2310.10482