E-009

大規模言語モデルを活用したニューラル機械翻訳器の検討
Exploring Neural Machine Translation with Large Language Models

美野 秀弥† 衣川 和尭† 河合 吉彦†
Hideya Mino Kazutaka Kinugawa Yoshihiko Kawai

1. はじめに

近年，大規模言語モデルの出現により要約や対話などの複数の生成タスクが 1 つのモデルで対応できるようになっている．一方で，単一の生成タスクに特化した従来手法のモデルとの比較は十分に行われていない．そこで，本稿では生成タスクの 1 つである機械翻訳に焦点を当て，大規模言語モデルを活用した機械翻訳器を構築してその翻訳精度を検証した．NHK ニュースの日英・英日機械翻訳実験を行い，追加学習データが多い場合は従来手法の機械翻訳器の翻訳精度が高く，追加学習データが少ない場合は大規模言語モデルを活用した機械翻訳器の翻訳精度が高いことを確認した．

2. 実験

2.1 実験概要

本稿では，大規模言語モデル(Large Language Model: LLM)を活用した機械翻訳器として事前学習済み LLM を対訳データで追加学習（ファインチューニング）したモデルを用いる．そして，同じ対訳データを用いて学習した従来手法の機械翻訳器と翻訳精度を比較する．

2.2 実験設定

事前学習済み LLM Meta 社が公開した商用利用可能なデコーダモデルの LLM である Llama2[1]の 7B，13B，70B サイズの Hugging Face モデルを使用した．Llama2 には対話用に指示学習を行った Llama-2-Chat モデルと指示学習前の Llama-2 モデルがある．本稿では，対話を目的としていないため Llama-2 モデルを用いた．

LLM のファインチューニング 量子化した Low Rank Adaptation(QLoRA)[2]を用いた．Low Rank Adaptation は，事前学習済み LLM の重みを固定したまま別途用意した少数の追加パラメータを学習するファインチューニング手法であり，計算コストを抑えつつ事前学習済み LLM の重みを更新するフルファインチューニングと同等の性能を発揮することが報告されている．本稿では，学習データ全体を 10 回学習し，別途用意した開発データで最も精度が高かったチェックポイントのモデルを採用した．図 1 に本稿で用いた学習の概要図を示す．

従来手法の機械翻訳器 比較する機械翻訳器は，12 層のトランスフォーマーベースのエンコーダ-デコーダモデル[3]を Sockeye3[4]で実装した．その他のハイパーパラメータは Sockeye3 のデフォルト値を用いた．

† 日本放送協会 Japan Broadcasting Corporation

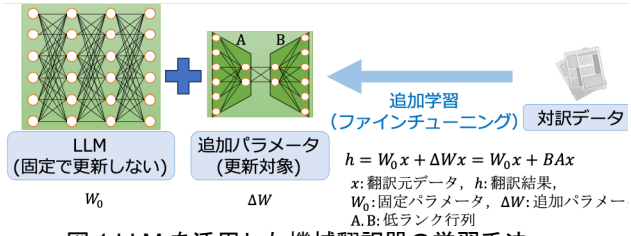

図 1 LLM を活用した機械翻訳器の学習手法

	日英	英日
Llama-2-7B ファインチューニング	14.8	40.0
Llama-2-13B ファインチューニング	15.6	43.5
Llama-2-70B ファインチューニング	16.0	44.0
従来手法	24.3	52.2

表 1 日英・英日機械翻訳実験結果

タスク NHK ニュースの日英・英日翻訳の精度を比較した．翻訳は，1 文ごとに実施した．

データセット NHK の日本語記事を人手で英訳したデータと，NHK の英語記事を人手で日本語訳したデータを用いた．学習データは 95 万文対（日本語記事の英訳データ：60 万文対，英語記事の日本語訳データ：35 万文対），開発データは 500 文対，テストデータは 1000 文対 [1)]用意した．学習時，翻訳時のデータのフォーマットは，日英翻訳は“JA:{日本語文} EN:”とし，英日翻訳は“EN:{英語文} JA:”とした．

評価尺度 機械翻訳タスクで一般的に用いられる評価尺度である BLEU[5]を用いて評価した．

2.3 実験結果

表 1 に実験結果を示す．

Llama-2 のファインチューニングモデル間の比較では，日英・英日翻訳ともに，モデルのパラメータサイズを増やすことで翻訳精度は向上した．パラメータサイズが大きいモデルの方がより多くの知識を学習できると考えられる．ただし，パラメータサイズが 13B と 70B との間の翻訳精度の差（日英：0.8，英日：3.5）は 7B と 13B との差（日英：0.4，英日：0.5）と比べて小さい値となっており，パラメータサイズをより大きくすることによる効果は限定的であると考えられる．

1) 開発データとテストデータは，英翻訳タスクでは日本語記事の英訳データを，英日翻訳タスクでは英語記事の日本語訳データを，それぞれ用いた．

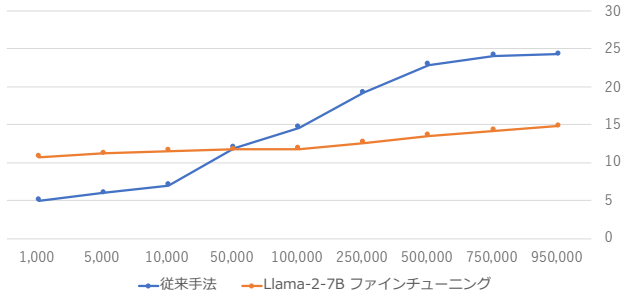

図 2 学習データ量の違いによる日英機械翻訳精度の変化（横軸：学習データ量，縦軸：BLEU スコア）

従来手法との比較では，日英・英日翻訳ともに，Llama-2 のファインチューニングモデルの翻訳精度が従来手法よりも低い結果となった． 本稿で用いた QLoRA によるファインチューニング手法では，従来手法の翻訳精度を上回ることは難しいと考えられる．

従来手法のモデルのパラメータサイズは 1B 程度であり，翻訳タスクのみを扱うのであれば LLM のような大きなパラメータサイズは必要ではない可能性がある．

また，英日機械翻訳結果からランダムに抽出した 50 例を人手で分析した結果，ファインチューニングモデルの翻訳結果の半数以上のデータに翻訳元文にはない情報が過剰訳として生成されていることを確認した．従来手法の翻訳結果では 1 割程度にとどまっており，ファインチューニングモデルは過剰訳の問題が顕著になっていることを確認した．表 2 に，著者が創作した英文の英日機械翻訳結果の例を示す．従来手法の翻訳結果には過剰訳は見られないが，Llama-2-7B のファインチューニングモデルの翻訳結果には複数の過剰訳が確認できる．

3. 分析

2 節の実験では，学習データとして 95 万文対の人手翻訳のデータを用いた．しかし，一般的にはそれだけの分量のファインチューニング用の学習データを用意できる状況は少ない．そこで，日英翻訳タスクにおいて学習データを変化させた場合の翻訳精度を調査した．1000 文対から 95 万文対まで 9 つのデータセットを構築し，それぞれのデータセットを用いて Llama-2-7B のファインチューニングを実施した．図 2 に結果を示す．

Llama-2-7B のファインチューニングのモデルはデータの分量が少ない場合と多い場合との間で翻訳精度の BLEU スコアの差が 5 程度であるのに対して，従来手法のモデルは 20 以上あり，ファインチューニングのモデルは学習データを増やすことによる効果が小さいことを確認した．また，学習データが 5 万文対よりも少ない場合は Llama-2-7B のファインチューニングのモデルの翻訳精度が高くなった．学習データを多く用意できない場合は，LLM のファインチューニングモデルを用いた方が精度が高くなる可能性がある．

表 2 英日機械翻訳例

翻訳元英語	Due to an accident resulting in injury or death at Mitaka Station, the JR Chuo line suspended operation for both directions.
正解日本語訳	ＪＲ中央線の各駅停車は、三鷹駅で起きた人身事故の影響で、上下線で運転を見合わせています。
Llama-2-7B	ＪＲ中央線では、下り新宿発の列車が上り線の三鷹駅で事故によりけがをした人がいることが分かり、両方向の運転を見合わせています。
従来手法	このうちJR中央線の電車は、三鷹駅で起きた人身事故のため、上下線で運転を見合わせています。

4. 結論

本研究では大規模言語モデルを活用したニューラル機械翻訳の検討を行った．既存の LLM を対訳データで追加学習したモデルで日英・英日機械翻訳実験を行い，ファインチューニング用の学習データが多い場合は従来手法の機械翻訳器の翻訳精度が高く，少ない場合は大規模言語モデルを活用した機械翻訳器の翻訳精度が高いことを確認した．今後の研究課題としては，過剰訳の課題解決が挙げられる．また，LLM のファインチューニング手法の違いによる翻訳精度の変化を調査する必要がある．さらに，翻訳作業の効率化を考えると，校正作業などの補助タスクが扱えた方がよい場合があると考えられる．必要な補助タスクを調査し，それらを扱える単一モデルの構築を検討していきたい．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究（課題 225）により得られたものです．

参考文献

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314 (2023).

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. (2017).

[4] Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. Sockeye 3: Fast neural machine translation with pytorch. arXiv preprint arXiv:2207.05851 (2022).

[5] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing 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, Pennsylvania, USA, July (2002).