# llmMT+1: 非英語言語対 LLM 翻訳の実現法の検討

傅 星儿 [1]　永田昌明 [2]　Chenhui Chu[1]
[1] 京都大学大学院 情報学研究科　[2]NTT コミュニケーション科学基礎研究所
xinger@nlp.ist.i.kyoto-u.ac.jp　masaaki.nagata@ntt.com
chu@i.kyoto-u.ac.jp

## 概要

近年，大規模言語モデルを使った機械翻訳が注目を集めている．しかし，そのモデルの多くは英語中心の言語対のみを対象としており，英語を含まない非英語言語対の翻訳がサポートされておらず，性能が低い問題が残されている．この問題に対して本論文では，対象言語の訓練状況に応じて 3 つに場合分けして，英語中心の翻訳モデルにおいて非英語言語対の翻訳を実現する方法について体系的に検討する．

## 1　はじめに

LLaMA-2[1] をはじめとする大規模言語モデル (以下 LLM という) は，多くの自然言語処理タスクにおいて高い性能を達成している．Xu ら [2] は，LLaMA-2 に対して 2 段階の訓練を行うことで，英語中心の言語対，すなわち英語が目的言語か原言語のとちらかになる言語対の翻訳に対象とした ALMA というモデルを提案した．この ALMA は，翻訳対象である英語以外の言語の単言語データを使って継続事前学習 (Continual Pre-Training: CPT) を行い，翻訳対象となる言語対の高品質な対訳データで Supervised Fine-Tuning(以下 SFT という) を行うことにより， 130 億の LLM を使って GPT-3.5 に匹敵する翻訳精度を達成した [2].

また，Alves ら [3] は，Tower を提案した．Tower は，LLaMA-2 に対してまず単言語データと対訳データ両方を用いた CPT を行った後に，高品質な対訳データを使った SFT を行うことで，ALMA より高く，130 億パラメータで GPT-4[4] に匹敵する翻訳精度を達成した．

しかし，それら手法は，英語中心の言語対のみを対象としており，英語を含まない非英語言語対の翻訳の性能が低い問題が残されている [5].

この問題に対して本論文では，対象言語の訓練状況に応じて，3 つの場合に分けて訓練データを選択する CPT を提案し，非英語言語対の翻訳を実現する方法について体系的に検討する．注目する言語に応じて単言語か対訳データを選択し，CPT を行った後，少量の対訳データで SFT を行う．得られたモデルを Flores200 で評価した．結果，以下の知見を得られた：

- 非英語言語対が両方とも既に事前訓練済みの場合，非英語言語対の対訳データで SFT を行う．
- 非英語言語対の片方が既に事前訓練済みの場合，もう片方の単言語データで CPT を行い，非英語言語対の対訳データで SFT を行う．CPT の際には事前訓練済みの言語の単言語データをリプレイする．
- 非英語言語対が両方とも事前訓練済みではない場合，2 つの言語の単言語データで同時に CPT を行い．非英語言語対の対訳データで SFT を行う．

## 2　関連研究

近藤ら [6] および Guo ら [7] は，ALMA における単言語データの CPT の後に大量の対訳データで CPT を行ってから高品質な対訳データで SFT を行うことにより，ALMA を上回る翻訳精度を達成した．

しかし，これらの手法は英語中心の翻訳対を対象としているため，非英語言語対の翻訳精度を考慮していない．なお，非英語言語対の対訳データは希少であるため，大規模な訓練が困難である．

## 3　提案手法

Tower や ALMA 等の手法は，前述したように英語中心の言語対のみを対象としている．本研究では，Unbabel/TowerBase-13B-v0.1[1)] (以下 TowerBase-13B という) を使って既存の LLM 翻訳モデルに新しい非英

1） https://huggingface.co/Unbabel/TowerBase-13B-v0.1

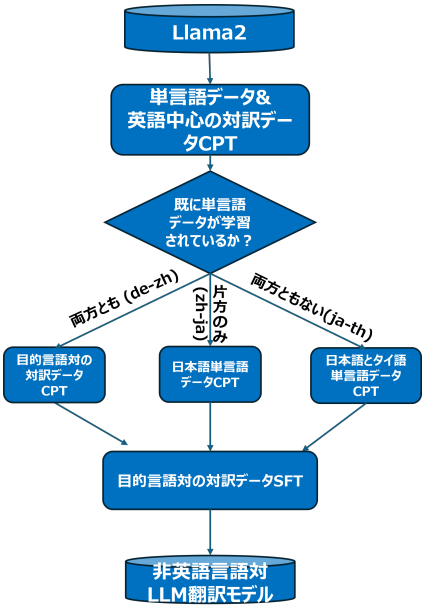


**図 1**　提案手法の概要図

語言語対を 1 つ追加する方法を検討する．

英語中心の LLM 翻訳モデルに新しい非英語対を追加する場合，その非英語対が既に継続事前訓練されているか否かに応じて，3 つの場合が考えられる: (1) 両方とも事前訓練済みの場合，(2) 片方が事前訓練済みの場合，(3) 両方とも事前訓練済みではない場合．Tower は，ドイツ語，スペイン語，フランス語，イタリア語，韓国語，ノルウェー語，ポルトガル語，ロシア語，中国語を対象としているので，本研究では，(a) の例としてドイツ語と中国語，(b) の例として中国語と日本語，(c) の例として日本語とタイ語を選択した．

また，それぞれの言語対に対して，図 1 のような訓練手法を提案する．

- (1) 両方とも事前訓練済みの場合: すでに単言語データを使った訓練が行われたため，目的言語対の対訳データを用いて CPT を行う．
- (2) 片方が事前訓練済みの場合: 事前訓練されていない単言語データで CPT を行う。破壊的忘却を防ぐために，少量の事前訓練済みの単言語データを用いたリプレイも行う．
- (3) 両方とも事前訓練済みではない場合: 二つの言語の単言語データで同時に CPT を行う．

CPT より，新しい言語対の翻訳に必要となる知識を LLM が獲得することを期待する．CPT を行ったモデルに対し，さらに SFT を行うことで非英語言語対の LLM 翻訳モデルを構築する．

# 4　実験設定

## 4.1　データセット

### 4.1.1　CPT

CPT では，ドイツ語・中国語の対訳データとして WikiMatrix[2] から抽出した 10 万文を使用した．日本語の単言語データとして JParaCrawl v3.0[3] からサンプリングした 16 億トークンを使用した．また，破壊的忘却を防ぐためのリプレイデータは WMT24 データセット[4] から 5%[5] 中国語とドイツ語の単言語データをサンプリングした．タイ語の単言語データとして翻訳タスクに適していると評価される CC100[9] から抽出した 10 億トークンを利用した．なお，すべてのトークン数は LLaMA-2 のトークナイザーを用いて計測した．

### 4.1.2　SFT

SFT の訓練データは，前述した 3 つの言語対の対訳データを全てサポートする TED2020[6] を使用した．各言語対において，各翻訳方向 TED2020 の訓練データから 5,000 件をランダムサンプリングし，併せて 10,000 件サンプルを使用した．これらのデータに対し，ALMA を参考に以下のプロンプトを適用した．ただし，本論文では，Tower が英語データを中心に訓練されたモデルであることを考慮し，英文の指示文に対する理解度が最も優れると仮定し，プロンプトの指示文を全部英文とした．なお，評価時にモデルの推論プロンプトも同じものを使用した．

**例: 中国語・日本語**

**日中翻訳のプロンプト**

Translate this from Japanese to Chinese:
**Japanese:** { 原言語文 }
**Chinese:**

**中日翻訳のプロンプト**

Translate this from Chinese to Japanese:
**Chinese:** { 原言語文 }
**Japanese:**

2)　https://opus.nlpl.eu/WikiMatrix/corpus/version/WikiMatrix
3)　https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/
4)　https://www2.statmt.org/wmt24/mtdata/
5)　5%という割合は，既存研究 [8] を参考にした．なお，言語対設定 1 ではドイツ語・中国語であるため，リプレイデータにも両方を含めた．
6)　https://opus.nlpl.eu/TED2020/zhen/v1/TED2020

**表 1** ドイツ・中国語語翻訳評価結果

| | de-zh | | zh-de | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| TowerBase-13B | 29.9 | 82.6 | 10.1 | 65.9 |
| TowerInstruct-13B | **34.6** | **86.8** | 17.0 | 73.1 |
| TowerBase-13B-SFT | 32.8 | 85.5 | **18.5** | **83.2** |
| TowerBase-13B-CPT-SFT | 33.0 | 85.7 | 18.3 | **83.2** |

**表 2** 中国語・日本語翻訳評価結果

| | zh-ja | | ja-zh | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| TowerBase-13B | 9.9 | N/A[8] | 25.1 | 84.2 |
| TowerInstruct-13B | 16.8 | 74.3 | 29.6 | 74.3 |
| TowerBase-13B-SFT | 19.2 | 86.8 | 27.3 | 85.9 |
| TowerBase-13B-CPT-SFT | **20.6** | **87.2** | **30.5** | **86.4** |

## 4.2 比較モデル

本研究では，翻訳に特化した既存のモデルに新しい言語対を導入しているため，その言語対においての翻訳性能をまず評価した．対象モデルは，Tower の 1 段階目の CPT を行った TowerBase-13B と Tower の 2 段階目の SFT を行った Unbabel/TowerInstruct-13B-v0.1[7](以下 TowerInstruct-13B という) とした．

評価実験では，対象とする言語に応じた CPT を TowerBase-13B に行った後，SFT を行った後の翻訳精度を評価した．公平な評価を図るため，TowerBase-13B に対しても SFT を行って評価した．

以下に実験に使用したモデルをまとめる．

- TowerBase-13B: ベースモデル
- TowerInstruct-13B: ベースモデルに対し英語中心の対訳データ等で SFT したモデル
- TowerBase-13B-SFT: 提案手法の SFT を適用したベースモデル
- TowerBase-13B-CPT-SFT: 提案手法モデル

## 4.3 モデル訓練ハイパーパラメータ

以下に評価実験で使用したハイパーパラメータを示す．CPT では，最適化手法として AdamW[10] を使用した．学習率は Tower と統一して，最大で $3.0 \times 10^{-5}$，最小で $3.0 \times 10^{-6}$ の cosine scheduler を使用した．通常の事前訓練と同様に次単語を予測するよう訓練を行った．CPT は全て Full-Parameter で行った．

SFT は CPT と同様に最適化手法として AdamW を用いた．訓練に使用したエポック数を 1，学習率を $2.0 \times 10^{-4}$ とした．SFT は LoRA チューニング [11] で行い，学習可能なパラメータは約 3100 万となり，元のモデルのパラメータ数の約 0.24%となった．

7） https://huggingface.co/Unbabel/TowerInstruct-13B-v0.1

8） COMET は語彙的な類似度のみを評価しているので，翻訳の出力が目的言語になっていない場合は N/A とした．

**表 3** 日本語・タイ語翻訳評価結果

| | ja-th | | th-ja | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| TowerBase-13B | 0.8 | N/A | 5.1 | 67.4 |
| TowerInstruct-13B | 15.8 | 49.1 | 8.9 | 62.0 |
| TowerBase-13B-SFT | 5.3 | N/A | 8.2 | 72.3 |
| TowerBase-13B-CPT-SFT | **27.6** | **68.0** | **14.4** | **81.7** |

**表 4** ドイツ語・中国語 SFT の効果

| | de-zh | | zh-de | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| TowerBase-13B-CPT | 29.8 | 82.7 | 10.0 | 66.0 |
| TowerBase-13B-CPT-SFT | **33.0** | **85.7** | **18.3** | **83.2** |

## 4.4 評価方法

予備評価実験及び評価実験でモデルの翻訳性能を評価するために，200 言語の翻訳を評価するデータセット Flores200 を使用した．

評価指標として，BLEU[12] 及び COMET[13] を使用した．BLEU は sacreBLEU[14] を使用した．COMET のモデルは wmt22-comet-da を選択した．

# 5 評価結果

## 5.1 対訳データの CPT の効果

まず，表 1 にドイツ語・中国語において実験の翻訳評価の BLEU 及び COMET スコアを示す．

既に単言語データが Tower の第 1 段階目で学習されたドイツ語・中国語については，対訳データを追加して学習しても，有意な上昇は見られなかった．既に単言語データの CPT がされた言語対に対しては，少量の対訳データで CPT を追加しては効果がみられず，SFT のみで翻訳性能が上昇するとわかる．

## 5.2 単言語データの CPT の効果

表 2，表 3 それぞれ，中国語・日本語，日本語・タイ語の実験の翻訳評価の BLEU 及び COMET スコアを示す．まず，日本語のみ単言語データが学習されていない中国語・日本語においては，日本語の単言語データを学習した結果，両方向での翻訳性能が上昇した．さらに，両方とも単言語データが学習されていない日本語・タイ語においても，両方向ともベースラインより大きく上回っていることが確認できる．以上のことから，単言語データの CPT がされ

**表 5**　独中　LoRA・Full-Parameter の比較

| | de-zh | | zh-de | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| TowerBase-13B-CPT(L)-SFT | 28.9 | 83.1 | 9.5 | 72.4 |
| TowerBase-13B-CPT(F)-SFT | **33.0** | **85.7** | **18.3** | **83.2** |

注記: (L) は LoRA を，(F) は Full-Parameter を意味する．

**表 6**　日中　LoRA・Full-Parameter の比較

| | zh-ja | | ja-zh | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| TowerBase-13B-CPT(L)-SFT | 11.1 | 83.1 | 24.4 | 84.3 |
| TowerBase-13B-CPT(F)-SFT | **20.6** | **87.2** | **30.5** | **86.4** |

注記: (L) は LoRA を，(F) は Full-Parameter を意味する．

ていない日本語及びタイ語に対して，単言語データのみを使用した CPT は，翻訳方向に関係なく性能を向上させることがわかる．

### 5.3　SFT の必要性

表 4 に，ドイツ語・中国語において CPT のみを実行した場合と，その後 SFT を実行した場合の評価を示す．結果からわかるように，CPT のみ実行した場合，評価の精度が変わらかった．一方で，SFT を追加すると，精度の上昇が観察された．モデルを翻訳タスクに特化させるために SFT が必要であることを再確認した．

### 5.4　CPT 手法の選択

表 5，表 6 は LoRA と Full-Parameter を選択した場合，ドイツ語・中国語の対訳データ，日本語の単言語データで CPT を行ったモデルの評価結果をそれぞれ示している．結果からわかるように，CPT を LoRA で行った場合，TowerBase-13B-SFT の翻訳評価よりも下回っている．これは，CPT を LoRA で実行した場合，元のモデルの性能を悪化させる恐れがあると示唆している．

### 5.5　破壊的忘却の防止

表 7 に，破壊的忘却を防ぐための手法として使われた 5%の中国語・ドイツ語単言語データと日本語の単言語データを含めた「ja+5%zh-de」での CPT，日本語の単言語データのみを使用した CPT「ja」，および日本語で追加 CPT されていない TowerBase-13B-SFT の結果を示す．表からわかるように，中国語から日本語まで翻訳する方向においては，「ja」が最も優れた性能を示した．一方で，逆方向では，「ja」は TowerBase-13B-SFT よりも BLEU・COMET スコア両方に下回ることがわかる．CPT

**表 7**　中国語・日本語破壊的忘却を防ぐ効果

| | zh-ja | | ja-zh | |
|---|---|---|---|---|
| CPT 方法 | BLEU | COMET | BLEU | COMET |
| TowerBase-13B-SFT | 19.2 | 86.8 | 27.3 | 85.9 |
| ja | **20.9** | **87.5** | 21.7 | 84.2 |
| ja+5%zh-de | 20.6 | 87.2 | **30.5** | **86.4** |

**表 8**　日本語・タイ語 CPT の順番

| | ja-th | | th-ja | |
|---|---|---|---|---|
| CPT 方法 | BLEU | COMET | BLEU | COMET |
| TowerBase-13B-SFT | 5.3 | N/A | 8.2 | 72.3 |
| ja&th | **27.6** | **68.0** | **14.4** | **81.7** |
| ja+th | 24.1 | 64.6 | 8.4 | 76.2 |

データにリプレイデータを追加することで，両方向においてベースラインの TowerBase-13B-SFT より上回る翻訳性能を達成した．以上のことから，既存の翻訳モデルに新しい言語対を導入する際は，既に存在している言語対のリプレイデータを少量な割合で加えることが必要であると言える．

### 5.6　新しい言語を導入する際の訓練順番

表 8 は，2 つ目の新しい言語，タイ語を追加する際に違う CPT 手法を行った結果を示している．「ja&th」は，日本語とタイ語の単言語データ同時に使用し CPT を行った場合を示し，「ja+th」は，日本語の単言語データで CPT を行った後にタイ語の単言語データを使い CPT を行った場合を示す．表から明らかに，新しい言語を 2 つ追加する場合は，同時に CPT を行う手法が最もいい性能を達成した．「ja+th」は，タイ語までの翻訳性能がベースラインより上昇したが，逆方向ではベースラインよりも劣っていることがわかる．

## 6　おわりに

本論文では，Tower のような英語中心の多言語 LLM 翻訳において，非英語言語対の翻訳の実現方法を 3 つの場合に分けて訓練データを選択する CPT を提案し，非英語言語対の翻訳を実現する方法について体系的に検討した．また，英語中心の多言語 LLM に非英語言語対を 1 つ導入する際に，対訳データより単言語データの効果が顕著であることを確認できた．そして，破壊的忘却を防ぐために，既に存在している言語対のリプレイデータを少量な割合で加えることが必要であることも確認した．

今後の課題として，非英語言語対を追加したことによる英語中心の言語対の翻訳精度への影響を調べることが考えられる．

## 謝辞

本研究は NTT コミュニケーション科学基礎研究所および JSPS 科研費 JP23K28144 の助成を受けたものです.

## 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] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. **arXiv preprint arXiv:2309.11674**, 2023.

[3] Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. Tower: An open multilingual large language model for translation-related tasks. **arXiv preprint arXiv:2402.17733**, 2024.

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[5] 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 Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2765–2781, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] 森下睦 永田昌明近藤海夏斗. 対訳データを用いた継続事前訓練による大規模言語モデルの翻訳精度評価. 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[7] Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. A novel paradigm boosting translation capabilities of large language models. **arXiv preprint arXiv:2403.11430**, 2024.

[8] Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. **Transactions on Machine Learning Research**, 2024.

[9] Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. Sailor: Open language models for south-east asia. **arXiv preprint arXiv:2404.03608**, 2024.

[10] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. **arXiv preprint arXiv:1711.05101**, Vol. 5, , 2017.

[11] 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. **arXiv preprint arXiv:2106.09685**, 2021.

[12] 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, 2002.

[13] Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT 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, 2022.

[14] Matt Post. A call for clarity in reporting bleu scores. **arXiv preprint arXiv:1804.08771**, 2018.