対訳構造の指示調整は言語間転移を促進するのか

佐藤 美唯 [1] 西潟 優羽 [2] 秋信 有花 [3] 倉林 利行 [3] 倉光 君郎 [2]
[1] 日本女子大学大学院 理学研究科 [2] 日本女子大学 理学部
[3] NTT ソフトウェアイノベーションセンタ
m1916038sm@ug.jwu.ac.jp kuramitsuk@fc.jwu.ac.jp

概要

多言語 LLM や低資源言語 LLM の開発において，言語間での事前学習データの不均衡は大きな技術課題である．現状は，英語中心の事前学習データが圧倒的に多い一方、低資源言語のデータ収集や作成は容易ではない．言語間転移は英語で学習した知識を活用して低資源言語の性能を高めることが期待されるが，その発生条件や原理は未解明な部分が多い．本研究では，指示調整を通じて言語間転移を促進させることを目指し，指示文を二言語の対訳構造にした対訳指示調整 (Parallel Instruction Fine-Tuning, PIFT) を提案する．PIFT の効果を日本語からのコード生成タスクで調査した結果，単言語の指示文を用いた指示調整と比較して性能が向上することを確認した．

1 はじめに

多言語 LLM や低資源言語 LLM の開発は，非英語圏話者にとって文化や伝統を守るための重要な課題である [1]．英語中心の事前学習データを学習した LLM は，英語では高い性能を発揮する一方で，非英語言語では性能が制限されることが多い [2]．LLM の性能は事前学習データの量が多ければ多いほど向上することが知られている [3]．しかし，英語でさえもデータの枯渇が懸念される中 [4]，言語資源の乏しい低資源言語ではデータを収集することも難しい状況である．

この状況に対応する手段の一つとして，言語間転移の有効性が議論されている．言語間転移とは，ある言語で学習した知識を異なる言語へ転移することを指す [5]．特に，英語などの高資源言語で学習した知識を人為的に転移させることができれば，低資源言語の性能向上に貢献できる [6]．

本研究の目的は，言語間転移を促進するための指示調整手法を探求し，低資源言語の性能を向上させることである．我々は指示調整形式の違いが言語間転移に与える影響を調査するとともに，指示文に二言語の対訳構造を取り入れた対訳指示調整 (Parallel Instruction Fine-Tuning, PIFT) を提案する．PIFT は，指示文に高資源言語と低資源言語の対訳で与えることにより，低資源言語の指示に対して，高資源言語からの知識転移を期待する設計になっている．

我々は，自然言語からのコード生成タスクを用いて PIFT の効果を調査した [7]．コード生成は，事前学習データの多くが英語で構成されている領域であるが [8]，使用する自然言語が異なっても生成されるコード自体は変化しない特性がある．我々は，公開済みモデルに対して，英語-日本語順および日本語-英語順の対訳構造を持つ指示調整形式による PIFT を適用した．その結果，PIFT は，従来の単言語指示調整や二言語指示調整と比べて性能を向上させ，さらに指示文の言語順序が影響を与えることを確認した．

2 多言語 LLM の現状と課題

本節では，多言語 LLM および低資源言語 LLM の開発における現状と課題についてまとめる．

2.1 言語資源

LLM 開発において，自然言語間で事前学習データの言語資源量に大きな不均衡があることは重大な課題の一つである．この不均衡は，事前学習データの構築に利用されるテキストデータの言語分布に起因する．例えば，英語のウェブページは全体の 46.0% を占めており，日本語の約 9 倍に相当する[1)]．このようなテキストデータを事前学習した LLM は，英語での性能が高い一方で，低資源言語の性能は制限され，英語中心の LLM になる [9, 10]．

低資源言語は，言語資源量が極めて限られてお

1) https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

り，高品質なテキストデータやクリーニングのための資源も不足しがちである．そのため，低資源言語 LLM の開発は，技術的および経済的にも困難である．

さらに，言語資源量の不均衡は言語間転移の有効性にも影響を与える．高資源言語で事前学習された LLM は低資源言語への知識転移が期待されるが，低資源言語特有の構文，語彙，文化的背景を十分に反映できない場合がある [11]．これらの課題を解決するために，追加事前学習や指示調整などの手法が提案されているが，その効果はデータの多様性や品質に大きく依存している．

以上のように，言語資源量の不均衡は LLM の公平性や汎用性を損なう要因である．本研究分野において，言語間転移を促進する手法や低資源言語のデータ構築手法の確立が喫緊の課題である．

2.2 課題設定

言語資源量の不均衡はプログラミング分野においてますます顕著である．最新の技術情報の多くは英語を中心に発信され，その後もしばし英語で議論されている [12]．その結果，英語と非英語言語間で言語資源量に大きな差が生じている．ソースコードデータセット The Stack に含まれる自然言語比率は，英語が 94%を占める一方で，日本語はわずか 1%未満である [8]．

また，言語資源量の不均衡は，コード生成タスクにおける言語間の性能差にも表れている．我々は複数のモデルの英語と日本語からのコード生成性能を評価してきたが，英語 LLM や多言語 LLM は言語間性能差があることを確認した [13]．さらに，Qiwei らは 23 種の自然言語からのコード生成性能を評価し，LLM における言語間性能差があり，言語によってもその差が異なることを示している [14]．これらの結果は，LLM が異なる自然言語で表現された同等の意味を正確に理解し，回答することが難しい点を示唆している．したがって，プログラミング分野では，英語で学習した知識をいかに低資源言語へ転移させるか，その手法が重要になる．

3 対訳指示調整

本研究では，言語間転移を促進させる指示調整手法として対訳指示調整を提案する．本節では，2.2 節の課題設定を踏まえて，コード生成タスクを例に説明する．

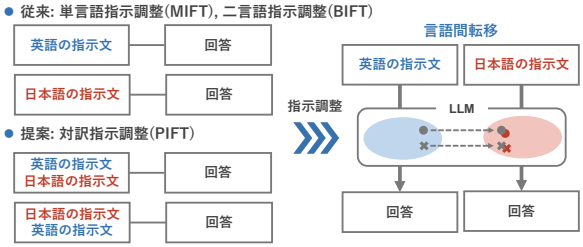

図 1 従来手法 MIFT・BIFT と提案手法 PIFT の概要

3.1 指示調整

指示調整は，指示文とそれに対応する回答がペアになった指示調整データを用いて LLM を微調整し，指示に追従するよう調整する手法である [15]．以下に，コード生成タスクにおける指示調整データの例を示す．

指示 (英語)

Write a function to calculate the sum of two numbers.

回答

```
def calculate_sum(a, b):
    return a + b
```

単言語指示調整 (Monolingual Instruction Fine-Tuning, MIFT) は，単一言語の指示調整データを用いて微調整する手法である．また，二言語指示調整 (Bilingual Instruction Fine-Tuning, BIFT) や多言語指示調整は，単一言語の指示調整データを翻訳した多言語版を作成し，それらのデータを併用して微調整する手法である [?]．コード生成タスクの場合は，回答がコードで共通であるが，自然言語の場合は指示文と同様に翻訳したデータを使用する．

3.2 対訳指示調整

対訳指示調整 (Parallel Instruction Fine-Tuning, PIFT) のアイディアはシンプルである．PIFT は指示文に高資源言語と低資源言語の対訳構造を取り入れることにより，言語間転移を促進する手法である．本研究では，図 1 に示す通り，高資源言語として英語，低資源言語として日本語を使用し，次の二種類の対訳構造を採用した．

指示（英語-日本語順）

Write a function to calculate the sum of two numbers.
2 つの数の合計を計算する関数をかきなさい．

指示（日本語-英語順）

2 つの数の合計を計算する関数をかきなさい.
Write a function to calculate the sum of two numbers.

対訳構造の言語順序は英語を先に配置する形式（英語-日本語順）と、日本語を先に配置する形式（日本語-英語順）の 2 種類ある．本研究では日本語を英訳する順序の日本語-英語順形式をメインに採用した.

3.3 設計

LLM における言語間転移の原理は未解明な部分が多くあるが，いくつかの仮説がある．PIFT の設計合理性は以下の仮説に基づいている.

LLM が多言語で高性能を発揮するためには，言語間の意味的整合性が重要な要因である仮説がある [16]．高資源言語と低資源言語間で語彙や文法構造を共有し，その整合性を学習させることで効率的な言語間転移が可能になる．一方で，意味的整合性が欠如すると英語以外の言語で性能が著しく低下することが報告されている．よって，PIFT では指示と回答で対訳関係を持たせるのではなく，指示文に二言語の対訳構造を取り入れる設計とし，より言語間の意味的整合性を学習できるよう工夫した.

次に，英語を中心に事前学習した LLM は非英語の指示文が与えられた際に，一度英語に翻訳し処理している内部翻訳仮説である [17]．日本語の指示文が与えられた際は，一度英語の指示文に内部で翻訳され，コードを生成する順番になる．よって，PIFT は日本語を英訳する順序の日本語-英語順形式の方が効果が高くなると考えられる．この順序は，日本語の指示文に対して，英語を通じて適切な応答を引き出すための重要な設計要素である.

4 実験

本節では．PIFT の効果を自然言語からのコード生成タスクで調査した結果をまとめる．PIFT は日本語からのコード生成性能を向上させ，英語からのコード生成性能を低下させないことを確認する.

4.1 コード生成タスク

コード生成タスクは，自然言語記述からコードを生成するタスクである．このタスクは，英語と非英語間の言語資源量に大きな不均衡が存在する領域であり，英語で学習した知識を低資源言語に転移させる手法は有望である．評価には以下の 3 種類のベンチマークを使用した.

- HumanEval[7]/JHumanEval[13]: 標準的なベンチマークとその日本語版.
- CL-HumanEval[18]：HumanEval の自然言語記述のみに着目したベンチマーク.
- SakuraEval: HumanEval に含まれない日本文化や日本語処理に関する問題中心のベンチマーク.

これらのベンチマークは関数定義とドキュメンテーション文字列を含むプロンプトから関数の続きとなるコード生成性能を評価する．評価指標には，生成されたコードがユニットテストに合格した割合を示す pass@k($k = 1$) を採用した [19]．本実験では 3 種類のベンチマークスコアの平均値を算出する.

4.2 指示調整データの作成

指示調整データセットには，OpenCoder[20] プロジェクトが公開する指示調整データの一部を使用した. 我々は，特に少量の指示調整データであっても効果的に言語間転移を引き起こす効果に着目していたため，educational_instruct の 110K 件のうち，指示文に含まれる英単語数が多い順に 2000 件, 4000 件, 6000 件と 8000 件を抽出した．日本語の指示文は，英語の指示文を DeepL API を用いて機械翻訳し作成した．なお，DeepL はいくつかの機械翻訳ツールを試した結果，忠実な和訳，品質，多言語 LLM 系の実験で採用されている点を考慮して採用した.

4.3 指示調整

我々は，公開されたモデルに MIFT の英語 (MIFT-En) と日本語 (MIFT-Ja)，BIFT, PIFT の英語-日本語順 (PIFT-EnJa) と日本語-英語順 (PIFT-JaEn) の計 5 種類の手法による指示調整を行った.

使用した公開済みモデルは Qwen2.5-0.5B である．このモデルは，英語，中国語，日本語を含む 29 言語のテキストデータ (18T トークン) で事前学習されている．事前に HumanEval と JHumanEval を用いてコード生成性能を評価した結果，言語間性能差が確認されたため使用した.

指示調整時には，Qwen2.5[21] のチャットテンプレートを使用し，それぞれの手法に適した形式で行った．パラメータ設定は OpenCoder[20] プロジェクトが公開する設定を参考にした.

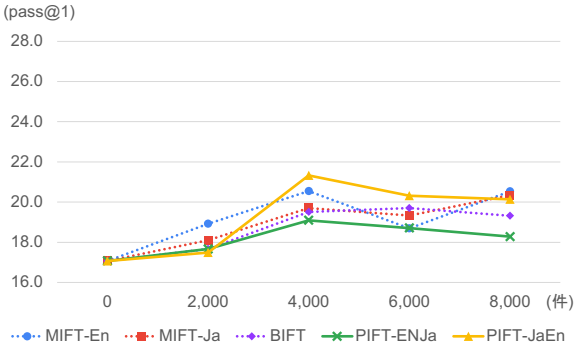

図 2　各指示調整手法における日本語からのコード生成ベンチマークの平均 pass@1 スコアの変化

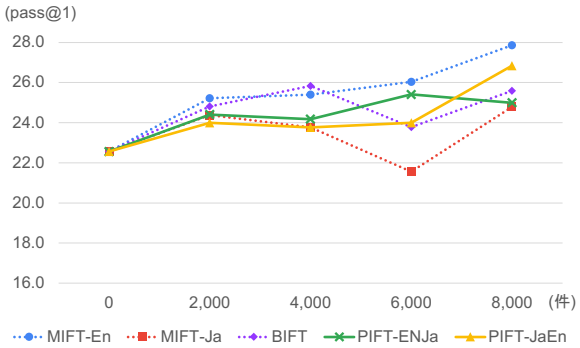

図 3　各指示調整手法における英語からのコード生成ベンチマークの平均 pass@1 スコアの変化

4.4　評価結果

まず，PIFT が日本語からのコード生成性能を向上させるのか，どちらの語順がより向上させるのかを確認する．図 2 にデータ件数別の日本語からのコード生成ベンチマークにおけるスコアの変化を示す．PIFT は，指示調整前 (データ件数 0 件) が 17.1%であるのに対してスコアが向上している．

MIFT と BIFT と比較した結果，PIFT-JaEn は最も高いスコアを示し，データ件数 4000 件時点で 21.3%へと 4.2%向上した．PIFT-JaEn は従来手法よりも相対的にスコアが低くなる結果になった．

そして，PIFT が英語からのコード生成性能を低下しないことを確認する．図 3 に英語からのコード生成ベンチマークにおけるスコアの変化を示す．PIFT は，指示調整前が 22.6%であるのに対して，スコアを向上させていることが確認された．特に，PIFT-EnJa はデータ件数 8,000 件時点で 26.8%のスコアを示し，MIFT-En の 27.9%に次いで向上した．

以上より，PIFT は英語からのコード生成性能を低下させずに，従来手法よりも日本語からのコード生成性能を向上させる手法であることが確認された．また，言語の語順は日本語-英語順の方が英語-日本語順に比べて効果であることも確認された．ただし，データ件数の違いにより変動があることから本実験結果は限定的なものであるため，指示調整データの量や品質，モデルの違いによる変化についてさらなる調査が必要である．

5　関連研究

言語資源量が限られている状況下で，言語間転移を促進することは注目すべき研究課題である [22]．従来手法として取り上げた MIFT でも，ある程度の言語間転移が期待できると示唆している [23, 24, 25]．また，多言語指示調整は MIFT よりも高い効果を示すことが報告されており，言語間の類似性 [26]，データ量 [24]，および言語数 [27] などの要素が影響する．

他にも機械翻訳等の補助タスクを用いて指示調整する翻訳支援調整が提案されている [16]．Zhu らは機械翻訳タスクで指示調整した後に，英語のタスクの指示調整を実施する二段階学習を提案している [28]．我々の提案する PIFT は，一段階で指示調整を行う点が異なっている．

さらに，言語間調整は翻訳タスクに明示的に頼らずに，多言語指示調整を言語横断的に再構築する．Chai らは，非英語の指示と英語の回答を組み合わせて指示調整する [29]. Zhang らは，多言語 LLM が非英語の指示に対してまず英語で考え，その後非英語で応答するよう多段階設計を導入している [30]. PIFT は，指示文に高資源言語と低資源言語の対訳構造を直接取り入れている点が異なっている．

6　むすびに

本研究では，言語間転移を促進させる指示調整形式として対訳二言語指示調整 (PIFT) を提案した．PIFT の効果を日本語からのコード生成タスクで検証した結果，単言語指示調整や二言語指示調整と比較して，性能の向上が確認された．特に，指示文の言語順序が影響を与えることを確認し，日本語英語の順が最も高い効果が得られることがわかった．この手法は，自然言語間での事前学習データ量の不均衡が多い分野においても低資源言語の性能を向上させる手法として期待される．今後は，英語と日本語以外の言語でも同様の傾向が確認されるのかを検証する，

謝辞

本研究は JSPS 科研費 JP23K11374 の助成を一部受けたものです.

参考文献

[1] Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, et al. A survey on large language models with multilingualism: Recent advances and new frontiers. **arXiv preprint arXiv:2405.10936**, 2024.

[2] Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond english: An empirical study on language capability transfer. **arXiv preprint arXiv:2401.01055**, 2024.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361**, 2020.

[4] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. **arXiv preprint arXiv:2211.04325**, Vol. 1, , 2022.

[5] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. **arXiv preprint arXiv:1904.09077**, 2019.

[6] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: Corpora, alignment, and bias. **arXiv preprint arXiv:2404.00929**, 2024.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. **arXiv preprint arXiv:2107.03374**, 2021.

[8] Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. The stack: 3 tb of permissively licensed source code. **arXiv preprint arXiv:2211.15533**, 2022.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[10] 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.

[11] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Association for Computational Linguistics, 2020.

[12] Philip J Guo. Non-native english speakers learning computer programming: Barriers, desires, and design opportunities. In **Proceedings of the 2018 CHI conference on human factors in computing systems**, pp. 1–14, 2018.

[13] 佐藤美唯, 高野志歩, 梶浦照乃, 倉光君郎. Llm は日本語追加学習により言語間知識転移を起こすのか？ 言語処理学会第 30 回年次大会発表論文集, pp. 2897–2900, 東京, 日本, 2024. 言語処理学会. 予稿集.

[14] Qiwei Peng, Yekun Chai, and Xuhong Li. Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. **arXiv preprint arXiv:2402.16694**, 2024.

[15] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. **arXiv preprint arXiv:2109.01652**, 2021.

[16] Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In **Findings of the Association for Computational Linguistics ACL 2024**, p. 7961–7973. Association for Computational Linguistics, 2024.

[17] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. **arXiv preprint arXiv:2402.10588**, 2024.

[18] Miyu Sato, Yui Obera, Nao Souma, and Kimio Kuramitsu. Cl-humaneval: A benchmark for evaluating cross-lingual transfer through code generation. In **Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation**, 2024.

[19] Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy S Liang. Spoc: Search-based pseudocode to code. **Advances in Neural Information Processing Systems**, Vol. 32, , 2019.

[20] Siming Huang, Tianhao Cheng, Jason Klein Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J Yang, JH Liu, Chenchen Zhang, Linzheng Chai, et al. Opencoder: The open cookbook for top-tier code large language models. **arXiv preprint arXiv:2411.04905**, 2024.

[21] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. **arXiv preprint arXiv:2412.15115**, 2024.

[22] Shaoyang Xu, Junzhuo Li, and Deyi Xiong. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models?, 2023.

[23] Niklas Muennighoff, Thomas Wang, Adam Roberts, Stella Biderman, Teven Le Scao, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics.

[24] Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality, 2024.

[25] Nadezhda Chirkova and Vassilina Nikoulina. Zero-shot cross-lingual transfer in instruction tuning of large language models, 2024.

[26] Shaoxiong Ji and Pinzhen Chen. How many languages make good multilingual instruction tuning? a case study on bloom, 2024.

[27] Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions?, 2024.

[28] Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning, 2024.

[29] Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. **arXiv preprint arXiv:2401.07037**, 2024.

[30] Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. Multilingual large language models are not (yet) code-switchers. **arXiv preprint arXiv:2305.14235**, 2023.