大規模言語モデルによる cross-lingual transfer の性能評価

田代雄介[1]　山口流星[1]　鈴木彰人[1]　辻晶弘[1]
[1] 株式会社 三菱 UFJ トラスト投資工学研究所
{tashiro,yamaguchi,suzuki,tsuji}@mtec-institute.co.jp

概要

近年、大規模言語モデルの発展が著しく、自然言語生成の様々な研究が進展している。一方、識別タスクなどの自然言語理解への大規模言語モデルの応用はあまり研究が活発ではない。本研究では、言語識別タスクの中でもモデルの多言語性能が重要である cross-lingual transfer について、サイズや学習データが異なる複数のモデル・設定を用いて、大規模言語モデルの性能を評価する。実験の結果、ターゲット言語のデータがない zero-shot 設定でもモデルの大規模化により転移される知識が増加すること、F1-score の上昇には大規模言語モデルでも few-shot による少量のターゲット言語を必要とすることなどが分かった。

1　はじめに

近年、ChatGPT や Llama2[1] をはじめとする大規模言語モデル (LLM) が急速に発展している。それにともない、自然言語生成 (NLG) の様々なタスクにおいて大規模言語モデルの実用可能性が高まり、活発に研究が行われている。

一方で、大規模言語モデルの自然言語理解（NLI）タスクに対する応用は、従来の比較的小規模なモデルでも実行可能だったこともあり、自然言語生成タスクに比べて注目度が低い。だが、自然言語理解タスクにおいても、モデルや学習データセットの大規模化の恩恵はあると思われる。特に、近年の大規模言語モデルは多言語化が進んでおり、多言語性能が必要なタスクにおいて、従来のモデルを上回る性能を発揮することが期待される。

そこで本研究では、言語識別タスクにおける cross-lingual transfer について、大規模言語モデルの性能を検証する。cross-lingual transfer とは、あるタスクにおいてターゲット言語の学習データが少ないとき、リソースが豊富なソース言語の学習データを用いてモデルを学習し、そのモデルをターゲット言語に適用する手法である。本研究では、日英ニュース記事のタグ付け問題を対象として、ターゲット言語を日本語、ソース言語を英語とした実験を行い、大規模言語モデルを含む複数モデルの識別性能を比較評価する。

2　関連研究

2.1　大規模言語モデル

モデル構造や学習手法の発展にともない、パラメータ数が数十億以上の大規模言語モデルが近年活発に研究・公開されている。それらの多くは英語の単言語モデルだが、LLaMA2[1] をはじめとした多言語の大規模言語モデルも登場している。LLaMA2 は約 2 兆トークンの英語中心の多言語テキストを用いて事前学習されたモデルであり、その性能の高さから後発のモデルにおける継続学習にも利用されている。例えば Elyza[2] は、約 200 億トークンの日本語テキストを用いて LLaMA2 の継続学習を行い、日本語タスクにおいて他の大規模言語モデルを上回る性能のモデルの構築を実現している。

2.2　大規模言語モデルの自然言語理解タスクへの応用

大規模言語モデルを識別などの自然言語理解タスクに用いるには、大きく 2 つの方法が挙げられる。1 つは生成モデルとして利用する方法である。例えば識別タスクを解く場合、「以下の文章が ESG に関連する場合は 1, そうでなければ 0 と答えよ」という入力を与えることで、識別結果が出力として得られる。

もう 1 つは、大規模言語モデルに出力層を追加して識別モデルとし、教師あり学習で自然言語理解タスクを解く方法である。自然言語理解タスクの教師あり学習では、こちらの方がモデルの性能をより引き出せるといわれている。例えば Li et al.[3] は、LLaMA2 を識別モデルとして用いて識別タスクの教師あり学習を行い、大規模言語モデルを生成モデル

として用いる場合や、BERT などの比較的小規模なモデルを用いた教師あり学習と比べて、高い識別精度が得られることを報告している。

2.3 Cross-lingual transfer

cross-lingual transfer は、豊富なソース言語の情報を用いてモデルを学習し、ターゲット言語においてそのモデルを利用する手法である。ターゲット言語の学習データが全くない場合を zero-shot cross-lingual transfer、少量ある場合を few-shot cross-lingual transfer と呼ぶ。多言語の言語モデルは cross-lingual transfer において有効であることが、多くの研究で報告されている。例えば Wu and Mark[4] は、様々なタスクでの multilingual BERT の有効性を示している。

本研究に近い先行研究として、Ye et al.[5] が挙げられる。Ye et al. では学習データが異なる複数の大規模言語モデルについて、含意分析や QA タスクでの転移性能の比較を行っている。だが彼らは、大規模言語モデルを生成モデルとして利用する形でタスクを解いており、なおかつ他の識別モデルとの比較も行っていないため、モデルの性能を十分に引き出せていない可能性がある。対して、本研究では大規模モデルを識別モデルとして利用し、いくつかのモデルで比較を行う。さらに、zero-shot や few-shot など複数パターンでの比較を行う。

3 実験設定

3.1 目的

本研究では、表 1 に示す 5 つのベースモデルを用いて、2 つの Research Question(RQ) を検証する[1)]。

RQ1: 多言語モデルにおける事前学習テキスト量やモデルサイズは、cross-lingual transfer での転移性能にどのような影響を与えるか

多言語モデルが大規模になり、また事前学習におけるテキスト量も増えると、転移性能が向上することが期待される。そこで、大規模言語モデル ELYZA と小規模な mBERT の転移性能を比較することで、モデルサイズの影響を確認する。また、事前学習に大量の日本語データを用いている ELYZA と、相対的に日本語データが少ない Llama2 を比較することで、ターゲット言語の事前学習テキスト量が転移性能に与える影響を確認する。

1) 表中の*印は、日本語データ量推定時に 2 バイト ≒ 1$token$ という近似を用いて計算したことを意味する

RQ2: few-shot cross-lingual transfer の有効性は、大規模言語モデルとそれ以外で異なるかどうか

ターゲット言語のデータ量が少量のとき、多言語モデルでの few-shot cross-lingual transfer が単言語モデルの学習に比べて有効であることは、小規模な言語モデルでは示されている。一方で、その有効性が大規模言語モデルにおいてどうなるのかは明らかではない。そこで、多言語大規模モデルである ELYZA の few-shot transfer と単言語大規模モデルである OpenCALM での学習の比較を行い、さらに小規模モデルである mBERT の few-shot transfer と BERT の比較も行うことで、大規模モデルでの few-shot transfer の有効性を確認する。

表 1 ベースモデル一覧。「token 数」の列は学習データにおける日本語 token 数を表す。

言語	サイズ	名称	token 数
多言語	大 (7B)	ELYZA[6]	200 億
多言語	大 (7B)	Llama2[7]	20 億
多言語	小 (700M)	mBERT[8]	約 15 億*
単言語	大 (7B)	OpenCALM[9]	不明
単言語	小 (450M)	BERT[10]	約 20 億*

3.2 タスク設定

多言語の自然言語理解タスクとして、本研究では日英ニュースのタグ付けを考える。タグ付けは、与えられたニュース文章が指定したカテゴリに属するかどうかを判別してタグを付与する二値分類タスクである。本研究では、金融ニュースベンダーである LSEG 社が提供しているロイター・ニュース・アーカイブ (Machine Readable News) における Headline（見出し）データをニュースデータとして用い、金融領域で注目度が高い ESG トピックのタグ付与を対象とした分析を行う。タグの正解データは、提供データにあらかじめ付与されているタグをベースとして人手で補正したものを用いる。

本研究では zero-shot transfer, few-shot transfer 両方の検証のために、表 2 に示す複数のデータパターンを用いる。zero-shot cross-lingual transfer には日本語データが 0 件の「ja0」、few-shot cross-lingual transfer には日本語データが 16 件の「ja16」、104 件の「ja104」がそれぞれ対応しており、各ベースモデルについてデータパターンに応じた 3 種類のモデルを学習することで、zero-shot と few-shot 両方での評価を行う。なお、データ作成においては、学習データは 2021 年 4 月〜2022 年 4 月、テストデータは 2022 年 8 月〜

2022 年 10 月のニュースから、ラベルの正負が 1:1 の均衡データになるようそれぞれサンプリングした。

表 2 実験データの件数パターン。なお、単言語モデルを学習する際は英語データを用いず日本語データのみで行う。

		学習		テスト
パターン名	パターン	英語	日本語	日本語
ja0	英語のみ	1024	0	512
ja16	日本語含 (小)	1024	16	512
ja104	日本語含 (中)	1024	104	512

3.3 学習方法

各言語モデルの fine-tuning では、モデルの出力部分に 2 値分類を行う head-layer を付与し、クロスエントロピー損失最小化による学習を行う。表 1 に示すモデルのうち、大規模言語モデルの fine-tuning には、効率的なパラメータ学習手法の一種である LoRA[11] を用いる。一方、表 1 でのサイズが小のモデルについては全パラメータの学習を行う。エポック数は 10 と設定し、学習率は大規模言語モデル 5×10^{-4}、それ以外のモデル 5×10^{-5} を初期値として線形に減衰させた[2)]。また、実験は乱数を変えてそれぞれ 3 回行い、平均値での評価を行った。

4 実験結果

表 3 各モデルの F1-score (括弧内は標準偏差)

	データパターン		
ベースモデル	ja0	ja16	ja104
ELYZA	0.691	0.777	0.857
	(0.088)	(0.082)	(0.007)
Llama2	0.631	0.773	0.839
	(0.053)	(0.036)	(0.006)
mBERT	0.694	0.675	0.740
	(0.014)	(0.004)	(0.029)
OpenCALM	-	0.674	0.781
	-	(0.028)	(0.015)
BERT	-	0.595	0.679
	-	(0.028)	(0.017)

4.1 RQ1: 多言語モデルにおける事前学習テキスト量やモデルサイズの影響

本節では多言語モデルである ELYZA, Llama2, mBERT について、分類精度の評価を行う。評価指標としては、各モデルの絶対的な分類精度を表す F1-score、各モデルのサンプル間の相対的な分類精度を表す AUROC の 2 つを用いる。F1-score での評価結果が表 3、AUROC での評価結果が表 4 である。

表 4 各モデルの AUROC (括弧内は標準偏差)

	データパターン		
ベースモデル	ja0	ja16	ja104
ELYZA	0.915	0.920	0.929
	(0.007)	(0.004)	(0.006)
Llama2	0.923	0.911	0.919
	(0.002)	(0.012)	(0.001)
mBERT	0.836	0.835	0.814
	(0.005)	(0.010)	(0.033)
OpenCALM	-	0.726	0.861
	-	(0.048)	(0.011)
BERT	-	0.635	0.726
	-	(0.018)	(0.010)

まず zero-shot の「ja0」での F1-score を比較すると、ELYZA と mBERT での値はほぼ同等であり、Llama2 は両モデルをやや下回っている。一方で、AUROC の比較では、ELYZA, Llama2 の値が mBERT を大きく上回っていることが見てとれる。このことから、大規模言語モデルでの zero-shot transfer により得られたモデルは、絶対的な分類精度は高くないものの、テストサンプル間の相対評価の性能が優れているといえる。相対評価の性能が優れているということは、モデルが ESG 関連の日本語とそうでない日本語を区別できていることを意味しており、英語データのみを用いた学習でも日本語の意味を捉えたモデルが得られていると考えられる。他方、絶対的な分類精度が低いということは、入力が ESG 関連かどうかの識別境界を適切に決められていない。言い換えると、zero-shot では知識自体の転移はできても識別境界の転移が難しいことを示唆している。

続いて few-shot である「ja16」「ja104」での結果を確認すると、ELYZA, Llama2 の F1-score が mBERT を上回っており、絶対的な分類精度での転移性能が向上している。このことは、少量の日本語データが識別境界を決めるための助けとなったことを意味している。一方で AUROC に注目すると、どのモデルでも「ja16」「ja104」での結果は「ja0」での結果と大差なかった。つまり、日本語データの少量追加は相対評価の性能の向上には繋がらず、zero-shot からの知識自体の上積みはほぼなかったと考えられる。

なお、大規模言語モデルである ELYZA と Llama2

2) ただし学習率 5×10^{-4} だと学習が進まなかった OpenCalm+ja16 の組み合わせのみ、学習率は 7×10^{-3} と設定した。

の間にはスコアの差がほぼなく、事前学習における日本語データ量の転移性能への影響は確認できなかった。今回の ESG 識別タスクは難易度が高くないため、Llama2 の事前学習に用いられた日本語データでも十分だったのであろう。

4.2 RQ2: few-shot cross-lingual transfer における多言語モデルの優位性

本節では多言語モデルである ELYZA, mBERT の few-shot tranfer での性能と、単言語モデルである OpenCALM, BERT の少量日本語データで学習したときの性能を比較する。表 3 の「ja16」「ja104」のデータパターンでの結果を見ると、ELYZA が OpenCALM を 0.103, 0.076 だけそれぞれ上回っている。また mBERT が BERT を 0.080, 0.061 それぞれ上回っている。つまり、多言語モデルでの few-shot transfer の優位性は、モデルが大規模かどうかを問わずにはっきりと見てとれる。単言語モデル同士の比較では OpenCALM の精度が BERT を上回っており、大規模化の効果を受けているが、同時に ELYZA の転移性能も大規模化により mBERT 対比で向上したため、few-shot transfer の優位性はあまり変わらなかったのだといえる。なお、表 4 の AUROC での比較結果も同様であり、few-shot での優位性は評価基準を問わないことが見てとれる。

4.3 解釈性

前節の評価で、zero-shot での学習でも大規模言語モデルは日本語知識を学習できていそうなこと、few-shot での追加日本語データは知識に寄与をあまりもたらしていないことが示唆された。本節ではこのことの定性的な裏付けのために、解釈性ツールの一種である LIT[12] を用いて LIME[13] による寄与度評価を行い、モデルがどのような根拠で識別を行っていたかを調査する。LIME は個々の予測結果に関する局所的な解釈を与える手法であり、各テストデータの識別における各 token の寄与度を出力する。この寄与度を全テストデータについて計算した後に、token 毎の寄与度平均を算出する。この寄与度平均が高いほど、識別に影響している token であることを意味する。cross-lingual transfer の性能が良いモデルは、ほとんど英語のみの学習データを用いていても、ESG に関連する日本語 token の寄与が高いことが期待される。

分析対象のモデルとしては ELYZA を利用し、表 2 の各データパターンについて平均寄与度上位の token を比較する。比較結果を示したのが表 5 である。まず英語データのみを用いた「ja0」の列を見ると、エネルギー、環境、炭といった ESG 関連と思われる日本語 token が寄与度上位にきている。このことから、zero-shot でもモデルが日本語の ESG に関する知識を学習できているという前節での主張が、定性的にも裏付けられる。また、3 つのデータパターンを比較すると、寄与度上位の token にほぼ差が見受けられない。このことは、日本語データの少量追加による識別性能への影響が小さいことを意味しており、やはり前節での結果と整合的である。

一方で、「ja104」の 8〜10 位には「削」「油」「債」といった他 2 つのモデルでは上位に出てこなかった token が並んでおり、日本語データを加えた学習により、モデルが捉える語句が一部変化している。このことから、若干ではあるが、英語データからは捉えられない知識を日本語データが有していた可能性が示唆される。

表 5 学習データパターン別の平均寄与度上位 token 比較

順位	ja0	ja16	ja104
1	エネルギー	エネルギー	エネルギー
2	環境	炭	環境
3	EV	燃	炭
4	炭	環境	燃
5	燃	EV	EV
6	電	気	気
7	力	電	電
8	気	候	削
9	候	排	油
10	排	力	債

5 おわりに

本研究では、ニュースタグ付けタスクにおける cross-lingual transfer について大規模言語モデルの性能評価を行った。実験の結果、モデルの大規模化により zero-shot で転移される知識が増加することが確認された。一方で、絶対的な識別精度の上昇のためには大規模言語モデルにおいても少量のターゲット言語を必要とするという結果も得られ、大規模言語モデルが常に転移性能を向上させるわけではないことも分かった。今後の課題としては、別の様々なタスクについて同様の結果が得られるかの検証が挙げられる。

免責事項

本稿の見解は著者個人のものであり，筆頭著者の所属する組織の公式見解を示すものではない。

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] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-japanese-llama-2-7b, 2023.

[3] Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised llama finetuning. **arXiv preprint arXiv:2310.01208**, 2023.

[4] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 833–844, 2019.

[5] Jiacheng Ye, Xijia Tao, and Lingpeng Kong. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. **arXiv preprint arXiv:2306.06688**, 2023.

[6] ELYZA-japanese-Llama-2-7b-fast-instruct. https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b-fast-instruct.

[7] Llama-2-7b. https://huggingface.co/meta-llama/Llama-2-7b.

[8] mBERT (bert-base-multilingual-cased). https://huggingface.co/bert-base-multilingual-cased.

[9] OpenCALM-7B. https://huggingface.co/cyberagent/open-calm-7b.

[10] BERT (bert-base-japanese-v2). https://huggingface.co/cl-tohoku/bert-base-japanese-v2.

[11] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2021.

[12] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models, 2020.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016**, pp. 1135–1144, 2016.