zero-shot cross-lingual transfer における言語の多様性の効果

佐藤 匠真 [1]　　新納 浩幸 [2]
[1] 茨城大学工学部情報工学部　[2] 茨城大学大学院理工学研究科情報科学領域
19t4035n@vc.ibaraki.ac.jp　hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

概要

mBERT のような多言語の事前学習済みモデルでは，あるタスクに対して言語 A で学習したモデルが別の言語 X の同じタスクに対してそのまま利用できる，いわゆる zero-shot cross-lingual transfer が可能である．本論文では上記の言語 A に当たる部分を複数の言語を組み合わせた場合の効果について調査する．タスクは Amazon レビューの評判分析とし，モデルを学習する言語を英語，ドイツ語，フランス語の各組み合わせとした．そしてテストデータは日本語の Amazon レビューの 2000 文書とし，zero-shot cross-lingual transfer を行った．実験の結果，訓練データに言語の多様性を含める効果があることが判明した．

1　はじめに

近年，mBERT[1]，XLM[2]，XLM-RoBERTa-XL[3] など多言語間の汎用的な特徴表現を学習する多言語の事前学習済みモデルが提案されている．このようなモデルの応用の一つとして zero-shot cross-lingual transfer がある．zero-shot cross-lingual transfer は，あるタスクに対して言語 A で fine-tunining したモデルを別の言語 X の同じタスクに対してそのまま利用する手法である [4]．通常，言語 A を英語などのメジャーな言語，言語 X をリソースが少ない，あるいはないようなマイナーな言語に設定することで，低資源言語の学習の問題解決に利用される．

一方，zero-shot cross-lingual transfer を行う場合，もとのモデルは単言語のモデルであっても可能であることが示されており [5]，多言語の言語モデルにおける言語の多様性が zero-shot cross-lingual transfer に影響を与えているのかどうかがはっきりしていない．本研究では評判分析をタスクに設定して，この点に対する調査を行う．

調査の方法としては言語 A から言語 X への zero-shot cross-lingual transfer を行う際に，言語 A の訓練データの他に，言語 B や言語 C の訓練データも利用することで言語に多様性のある訓練データから学習する場合と，それと同サイズの言語 A (あるいは B や C) の言語に多様性のない訓練データから学習する場合とどちらが zero-shot cross-lingual transfer において効果があるのかを調べる．

実験ではタスクを Amazon レビューの評判分析とし，上記言語 A，B，C を英語，ドイツ語，フランス語，言語 X を日本語とした．つまりテストデータは日本語の Amazon レビュー文書である．実験の結果，訓練データに言語の多様性を含める効果があることが判明した．

2　関連研究

2.1　Multilingual BERT

mBERT(Multilingual BERT) は BERT と同じモデルアーキテクチャと学習方法を持っている多言語の事前学習済みモデルである [1]．mBERT が事前学習に用いる Wikipedia のデータには 104 の言語が含まれている．mBERT では WordPiece モデリングにより，モデルが言語間で埋め込み表現を共有することが出来ている．語彙には様々な言語のキャラクタが入っており，語彙数は約 12 万 token ある．日本語の漢字は 1 文字で 2 文字の漢字単語はない．日本語のの 2 文字以上の token は約 1000 個ほどで，「になっている」，「となっていた」などの長いものもある．

cross-lingual な表現学習をしたモデルとして Goyal らの XLM-R がある [6]．XLM-R は様々なクロスリンガルベンチマークにおいて mBERT より大幅に優れている．また，XLM-R は低リソース言語において特に優れた性能を示している．従来の XLM モデルと比較して XNLI の精度がスワヒリ語やウルドゥー語で向上している．このような，結果を得るためには

- 正の伝達と容量の希薄化

• 高リソース言語と低リソース言語のスケール

のトレードオフが性能に関係している．さらに，XLM-R は GLUE と XLNI において強力な単言語モデルと非常に高い競争力を有している．

2.2 zero-shot cross-lingual transfer

zero-shot cross-lingual transfer は single-source transfer とも呼ばれ，ソース言語 (多くの場合，高リソース言語) でモデルを訓練し，その後ターゲット言語へ直接転送する手法である [4].

近年の BERT 等の事前学習済みモデルは，言語をまたいだ転移学習が可能であることが知られている．例えば英語やフランス語などの間には，似た語彙も多く，英語で学習した知識がフランス語のタスクに有用なのは当然のことである．さらに，この転移学習は事前学習用の言語 (L1) と微調整用の言語 (L2) との間に，共通の語彙が全く無くても可能である．この結果から，言語をまたいだ転移学習には人間の言語の何らかの構造的特徴が関連していると考えられている．

Li らの事前学習された言語モデルにおける新たな cross-lingual 構造 [7] によると，多言語エンコーダの最上層に共有パラメータが存在するため，単言語コーパス間で共有する語彙がない場合やテキストが非常に異なる領域のものであっても zero-shot cross-lingual transfer が可能であると示されている．

Artexte らの単言語表現の言語伝達可能性について [5] では，言語 L1 で事前学習された単言語モデルを言語 L2 コーパスを利用し新しいトークンの埋め込みを学習することにより新しい言語 L2 に転送した (共有サブワードの概念がない) モデルは最先端の多言語モデルと zero-shot cross-lingual transfer ベンチマークで同等に機能した．これは，多言語モデルでは語彙の共有も共同事前トレーニングも必要ないことを示している．

3 訓練データの言語の多様性

図 1のように英語のデータ 3300 文から訓練データ 1000 文，検証データ 100 文，計 1100 文の組を 3 つ作りそれぞれ E1，E2，E3 とした．ドイツ語も同様に訓練データ 1000 文，検証データ 100 文，計 1100 文の組を 3 つ作りそれぞれ D1，D2，D3 とした．フランス後も同様に訓練データ 1000 文，検証データ 100 文，計 1100 文の組を 3 つ作りそれぞれ F1，F2，F3 とした．

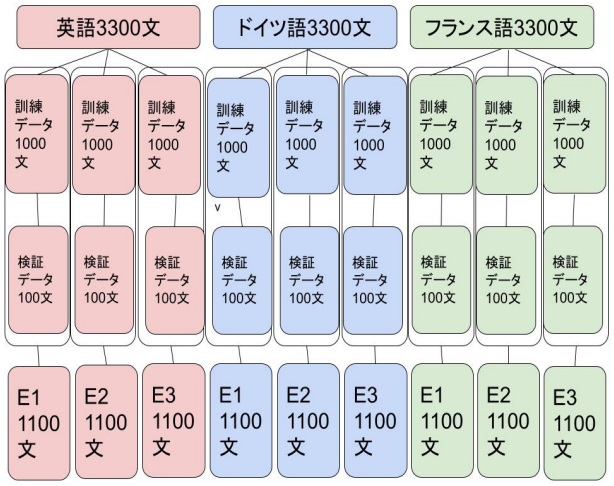

図 1 1100 文の訓練データ

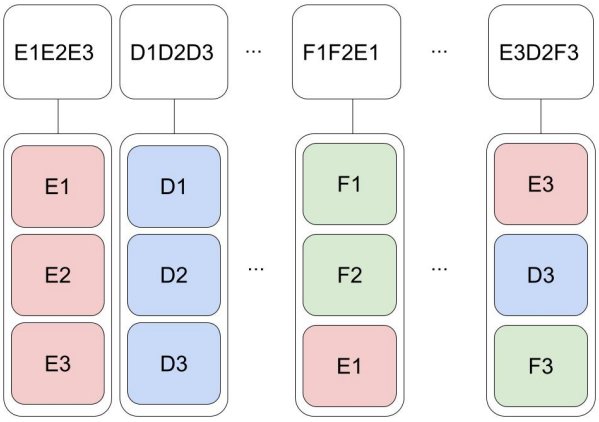

図 2 言語の多様性を持たせた 3300 文の訓練データ

さらに，訓練データに言語の多様性の違いを持たせるために図 2のように，図 1で作成したデータを 3 つ合わせて訓練データ 3000 文，検証データ 300 文の新たな組を作った．1 つの言語から構成される言語の多様性を持っていない E1E2E3，D1D2D3，F1F2F3 の 3 組，2 つの言語から構成される言語の多様性を持っている E1E2F1，E1E2D1，D1D2E1，D1D2F1，F1F2E1，F1F2D1 の 6 組，3 つの言語から構成される最も言語の多様性を持っている E1D1F1，E2D2F2，E3D3F3 の 3 組，計 12 組を作成した．

zero-shot cross-lingual transfer を行うため訓練データには日本語を含んでいない．

4 実験

Google が公開している mBERT を用いた文書分類タスクで zero-shot cross-lingual transfer を行うときに，訓練データに言語の多様性を持たせた場合の識別精度と持たせなかった場合の識別精度を比較し，

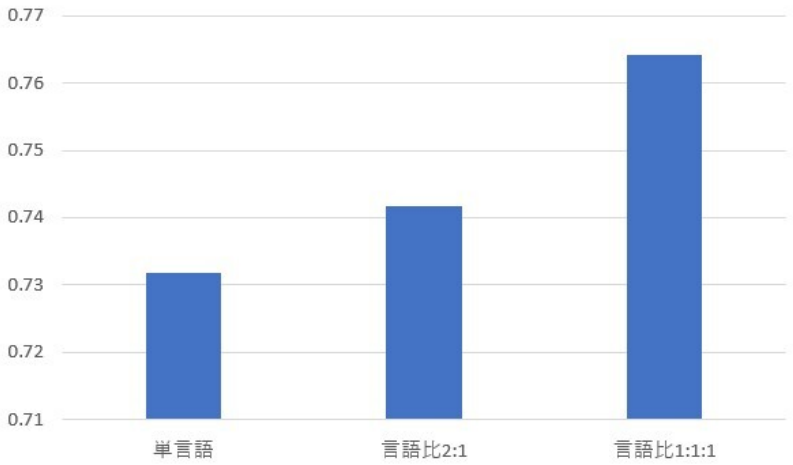

図 3　実験結果の平均値のグラフ

訓練データに言語の多様性を持たせる効果を確認した．

4.1　事前学習済みモデル

Google で公開されているモデル (BERT-Base, Multilingual Cased) を使用した．これは Hugging Face 社の Transformers ライブラリから，モデル名'bert-base-multilingual-cased' で利用できるモデルである．The lagest Wikipedias 内でトップ 104 の言語のデータが事前学習コーパスに利用されている．

4.2　実験用データセット

実験には Webis-CLS-10 データセット[1)]を用いた．このデータセットには日本語，英語，ドイツ語，フランス語が収録されている．ラベルはレビューの星の数であり，1 から 5 までの 5 段階評価である．ただし，ラベルが 3 のデータは存在しない．本実験ではラベルが 1，2 のデータを negative，4，5 のデータを positive として評判分析 (2 値分類) を行った．

このデータセットには日本語，英語，ドイツ語，フランス語それぞれに books，dvd，music の 3 つの領域がある．各言語の各領域に訓練データ 2000 文，テストデータ 2000 文がある．本実験ではすべて music の領域のデータを用いた．

4.3　文書分類器の作成

作成した訓練データ 3000 文，検証データ 300 文の 12 組で BERT-Base， Multilingual Cased に対して 10epoch の fine-tuning を行った．

4.4　実験結果

訓練データの言語の多様性で作成した訓練データを用いて fine-tuning を行ったモデルで評判分析を解いた．すべて日本語のテストデータ 2000 で評価した．単言語の訓練データで fine-tuning を行い評価を行った場合の結果を表 1，訓練データの言語比 2:1 で fine-tuning を行い評価を行った場合の結果を表 2に，訓練データの言語比 1:1:1 で fine-tuning を日本語で評価を行った結果を表3に示す．また，実験結果の平均値のグラフを図3に示す．

表 1　単言語での学習

訓練データ	日本語で評価
E1E2E3	0.7310
D1D2D3	0.7450
F1F2F3	0.7190
平均値	0.7317

表 2　言語比 2:1 での学習

訓練データ	日本語で評価
E1E2D1	0.7395
E1E2F1	0.7190
D1D2E1	0.7615
D1D2F1	0.7535
F1F2E1	0.7070
F1F2D1	0.7705
平均値	0.7418

1）　https://webis.de/data/webis-cls-10.html

表 3 言語比 1:1:1 での学習

訓練データ	日本語で評価
E1D1F1	0.7630
E2D2F2	0.7610
E3D3F3	0.7685
平均値	0.7642

実験結果より，mBERT を用いて zero-shot cross-lingual transfer を行う場合の訓練データは言語の多様性を含ませた場合の結果が最も良く，単言語の場合の結果が最も悪いことがわかった．

5 考察

3節で作成した訓練データ 1000 文，検証データ 100 文の組を用いて fine-tuning を行ったモデルで 4節同様に日本語のテストデータを用いた zero-shot cross-lingual transfer を行った．結果を表 4に示す．

表 4 単言語の訓練データ 1000 文

訓練データ	日本語で評価
E1	0.6940
E2	0.7315
E3	0.7000
D1	0.7660
D2	0.7005
D3	0.7480
F1	0.6625
F2	0.7080
F3	0.7130
平均値	0.7137

英語，ドイツ語，フランス語の中で最も精度が良かった E2，D1，F3 を用いて新たな組を作り，その組で fine-tuning を行ったモデルで実験した．精度は **0.7490** だった．これは，精度が良い訓練データを合わせても精度が向上するわけではないことを示している．

さらに，zero-shot cross-lingual transfer ではない日本語の訓練データを用いた評判分析を解く場合に言語の多様性を持たせる効果を確認するために実験を行った．日本語の訓練データ 1000 文，検証データ 100 文の組 J1 を作成した．そして，J1 を用いて 4節で使用したモデルに対して fine-tuning を行い日本語のテストデータ 2000 文で評価した．実験結果を表 5に示す．

表 5 日本語 1000 文で fine-tuning

J1 で fine-tuning したモデル	日本語で評価
E1E2E3	0.7905
D1D2D3	0.7450
F1F2F3	0.8175
単言語の平均値	0.7843
E1E2D1	0.8035
E1E2F1	0.8230
D1D2E1	0.8105
D1D2F1	0.7985
F1F2E1	0.7855
F1F2D1	0.8185
言語比 2:1 の平均値	0.8066
E1D1F1	0.8015
E2D2F2	0.8030
E3D3F3	0.8095
言語比 1:1:1 の平均値	0.8047

表5より日本語を含む訓練データで fine-tuning する場合も zero-shot cross-lingual transfer の場合と同様に訓練データに言語の多様性を含ませた結果が多様性を含ませなかった場合より良くなった．しかし，J1 の 1000 文だけで fine-tuning した結果は **0.8265** だった．この結果は訓練データに目的言語データが含まれている場合 (つまり zero-shot cross-lingual transfer ではない場合) は，訓練データに別言語のデータを含ませない方ががよいことを示している．この点についてはさらに調査をする必要がある．

6 おわりに

本研究では mBERT を用いた zero-shot cross-lingual transfer を行うときに，訓練データとして 3 種類の言語を含むデータ 3 組と 2 種類の言語を含むデータ 6 組と単言語のデータ 3 組を作成し，それぞれで fine-tuning を行い評判分析で精度を比較した．言語の多様性を最も持たせた場合の精度が最もよく，単言語の場合の精度が最も悪かった．

また，zero-shot cross-lingual transfer ではない場合の訓練データに言語の多様性を含めることの効果について調べることを今後の課題とする．

謝辞

本研究は 2022 年度国立情報学研究所公募型共同研究（22FC04）の助成を受けています.

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.

[3] Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling, 2021.

[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, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] 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**, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[7] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics.