少量ショットに対する大規模言語モデル（LLM）を用いた人工データ生成による精度向上の試み

山本大輝 [1] 佐々木峻 [1]
[1] アクロクエストテクノロジー株式会社
{yamamoto, sasaki}@acroquest.co.jp

概要

近年，大規模言語モデル（Large Language Model：LLM）の発展により，質問応答や文書生成をはじめとする多様な自然言語処理タスクで高い精度が実現されている．従来の BERT やロジスティック回帰と比較して，LLM はより優れた性能を示しているが，その学習および推論には大量の GPU リソースを必要とするため，業務での安定運用には高額なコストが伴う．

一方で，BERT やロジスティック回帰などの従来手法を利用することで，運用コストを削減しながらも一定の性能を維持する方法が考えられる．しかし，分析に必要な十分なデータを集められない事例も多く，これがモデルの性能向上を妨げる要因となる．

本研究では，LLM を利用して学習データを人工的に生成する手法に着目し，文書分類タスクにおける軽量モデルの精度向上に寄与するかを検討した．実験の結果，参照なし文書生成プロンプトが最も効果的であり，軽量なモデルでも精度向上が可能であることを確認した．

1 はじめに

近年，LLM が発展してきたことにより，ゼロショットや少数ショットのデータを活用できる機会が増加している．しかし，LLM をシステムに組み込んで運用する場合，パラメータ数向上に伴い，計算速度が遅くなるため，複数の GPU を利用することが概ね必須となる．要件次第では LLM の導入はコストが見合わず，かつ，難しいものとなる．そのため，特定のタスクを解かせる BERT[1] や TF-IDF とロジスティック回帰のような軽量モデルでも精度を高くできることは運用コスト面においても大きな意義がある．更に十分なデータを確保できず，データセットの総量が少ないケースが考えられるため，本研究では少量のデータセットによる影響を検討する．

本研究では，少量ショットのデータセットをベースとしたデータ拡張方式を複数提案する．提案手法によって構築したデータセットに対して，ロジスティック回帰や BERT という軽量モデルを用いた分類手法を適用した．複数の提案手法の有効性を調査した結果，少ないデータの場合でも精度向上の有効性が認められる手法が見つかった．

2 関連研究

自然言語処理の分野では，データ拡張（Data Augmentation）を用いてデータを増強し，モデルの性能向上を図る手法が数多く研究されている．関連する研究として，EDA（Easy Data Augmentation）[2] が挙げられる．EDA は，単語の類義語置換，ランダム挿入，ランダム削除，ランダム交換といった手法を用いてデータを拡張し，モデルの性能向上に寄与することを示している．

また，逆翻訳（Back Translation）を利用したデータ拡張手法も広く研究されており，生成される複数のパターンを評価することで，その有効性を示した研究もある [3]．

さらに，LLM を活用したデータ拡張手法として，Ye らの提案 [4] が注目されている．この手法では，タスクとラベルの説明文を入力することで，関連するデータセットを生成する仕組みが提案されている．

日本語の分野でも，LLM で疑似データを作成して少数ショットで学習する JapaGen が提案された [5] が，データ拡張を実施しても，精度に貢献できなかった例も報告されている [6]．

本研究では，既存研究ではあまり検証されていなかった，データセットが少数データの場合につい

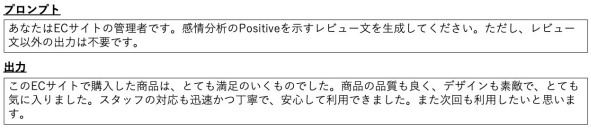

図 1　参照なし文書生成の例

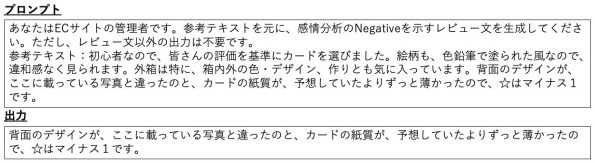

図 2　参照あり文書生成の例

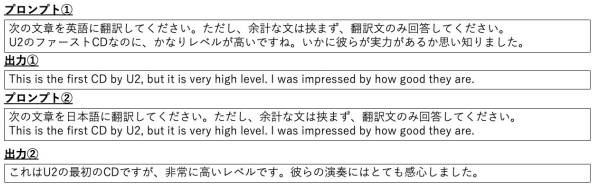

図 3　翻訳拡張の例

て，複数のプロンプト生成パターンを提案し，条件別に評価した．

3　提案手法

本研究では，LLM を利用したデータ拡張手法を提案する．この手法では，LLM を用いて新たなデータを生成し，モデルの性能向上を図ることを目的としている．具体的には，以下の 3 つのデータ拡張手法を提案する．

1. **参照なし文書生成**　文書のコンセプトを入力として与え，同じクラスに属する新たな文書を生成する手法である．LLM は入力されたコンセプトを解釈し，既存データに依存しない新規データを生成することが期待される．具体例は図 1 を参照されたい．
2. **参照あり文書生成**　文書のコンセプトと参考文を入力として与え，同じクラスに属する文書を生成する手法である．この手法では，LLM が入力されたコンセプトに加え，参考文から得られる文脈情報を考慮して，より精緻な文書を生成することが期待される．具体例は図 2 を参照されたい．
3. **翻訳拡張**　日本語の文書を英語に翻訳し，その後，英語から再び日本語に翻訳することで新たな文書を生成する手法である．この手法により，同じ意味を持ちながらも表現の異なる文書を得ることが可能である．具体例は図 3 を参照されたい．

参照なし文書生成では，LLM が与えられたコンセプトを基に，新規性の高い文書を生成することを目指している．一方，参照あり文書生成では，コンセプトに加えて参考文を使用することで，文書生成の方向性をより具体的に示唆することを狙いとしている．最後に，翻訳拡張では，異なる表現を持つ日本語文書を生成することで，多様性の高いデータセットを作成することを期待している．

4　実験

本研究では，提案手法およびモデルの性能を評価するために，MARC-ja データセットを用い，ロジスティック回帰と，BERT の派生モデルである DeBERTa[7] の 2 つのモデルを比較した．また，比較においては，データセットのサンプル数，複数のモデル，および提案手法の違いが性能に与える影響を検討した．

4.1　モデリング

本実験では，学習させるベースモデルとして，ロジスティック回帰および DeBERTa（ku-nlp/deberta-v3-base-japanese）を使用した．DeBERTa の最適化には Adam[8] を用い，エポック数はデータセットの規模に応じて調整した．

ロジスティック回帰　文字列の tri-gram を用いた TF-IDF 特徴量を入力として適用する．

DeBERTa　クラス分類用のヘッドを追加したモデルを使用し，分類タスクに適用する．

4.2　データセット

実験には，MARC[9] を日本語化した MARC-ja データセットを利用した．データセットの構成は以下の通りである．

- **データサンプリング**　MARC-ja データセットの学習データから 10 件，100 件，1000 件をランダムに 5 回抽出し，これらのデータセットをベースとして分析を行った．結果は 5 回の抽出による平均値を用いて評価した．また，比較として MARC-ja のデータセット全体である 187528 件を利用した結果も算出する．

- **データ拡張**　本論文で掲載している 3 つのデータ拡張手法である参照なし文書生成，参照あり文書生成，翻訳拡張を実施した．本提案のデータ拡張により生成した数はデータの件数と同数を作成している．
- **データ拡張に利用した LLM**　抽出したデータセットを基に，LLM を利用して新しいデータ拡張データセットを作成した．本実験では，以下の 2 種類の LLM を利用した
  – llm-jp-3.7b-instruct
  – llm-jp-13b-instruct

5　実験結果

主要な結果を表 1 に示す．本研究では，文書分類タスクにおける正答率を評価指標として採用し，データ拡張手法の有効性を検証した．結果として，文書分類タスクにおいて，精度が向上するデータ拡張手法と，逆に精度が向上しない手法が存在することを確認した．データ数が少ない（10, 100）の条件において，参照なし文書生成を用いた場合には精度が向上した．反面，1000 件のデータを利用した場合，更に，参照ありデータ生成や翻訳拡張を用いた条件では精度向上が見られなかった．また，ロジスティック回帰，DeBERTa では，データが少ない場合でも精度向上率が向上したが，特にロジスティック回帰では，100 件のデータの場合にも精度向上率が DeBERTa よりも高く見られた．

5.1　モデリングの比較

本研究では，データ量の違いがロジスティック回帰，および DeBERTa モデルの性能に与える影響を検討した．その結果，以下の知見が得られた．まず，ロジスティック回帰においては，データ量が少ない場合（例: 10 件，100 件）に精度が顕著に向上することが確認された．しかし，データ量が約 1000 件を超えると，精度の向上は限定的であり，性能の改善がほとんど見られなかった．この理由として，ロジスティック回帰の特徴量として利用している TF-IDF が単語の出現頻度に基づいて特徴量を生成する性質が挙げられる．データが増加することで特徴量の多様性が高まり，これがモデル性能の向上に寄与したと考えられる．さらに，DeBERTa モデルについても，少量データ（例: 10 件，100 件）の場合に大幅な精度向上が見られたが，データ量が 1000 件に達すると精度の変化がほとんど観測されなかった．これは，事前学習済みのモデルを初期値として利用していることにより，既存の言語特性をモデルが学習されていることから，実験データの学習の収束が早かったことが主な要因と考えられる．

5.2　データ拡張モデルの比較

表 2 に主要なモデル比較に必要な結果を掲載した．本実験では llm-jp-3.7b-instruct と llm-jp-13b-instruct を利用したがデータ拡張する上で大きな差分は発生しなかった．今回のモデルでは類似した文章が生成されていたため，差分が発生していない可能性が高い．しかし，より問題生成難易度が高いデータにおいては差分が出る可能性がある．

5.3　データ拡張の効果

本研究では，データ拡張手法の有効性を検証し，以下の結果を得た．

参照なし文書作成　参照なし文書作成を用いた場合，最も高い精度上昇率が得られた．この結果は，参照を用いないことで，参照あり文書作成と比較して生成された文書に多様性が生まれ，分類の性能向上につながったことに起因すると考えられる．

参照あり文書作成　参照あり文書作成のデータを用いた場合，精度が低下する結果となった．多様性の低い生成結果により，モデル性能に悪影響を及ぼした可能性が示唆される．

翻訳拡張　翻訳を用いたデータ拡張でも，精度が低下する結果が得られた．これにより，翻訳されたデータの使用が必ずしもモデルの性能向上に寄与するわけではなく，むしろノイズとして作用し，精度が悪化する場合があることが示唆される．

5.4　拡張方式による比較

拡張方式では，参照なし文書作成の精度が最も高く，他 2 手法は精度が下がることもあった．参照なし文書作成のみ原文を利用していないことから，入力として与えた参照文章とは異なる文書が作成されたことにより精度が大きく向上したと考えられる．

6　おわりに

本研究では，少量ショットデータセットに対する LLM を用いたデータ拡張手法の有効性を検証した．MARC-ja データセットを利用し，ロジスティック回帰および DeBERTa モデルの性能を比較する中で，データ量や拡張方式がモデル精度に与える影響を詳

表 1 実験結果

ベースモデル	データ数	データ拡張モデル	元データ	参照なし文書作成	参照あり文書作成	翻訳	正答率	精度上昇率
ロジスティック回帰	10		○				0.545	
ロジスティック回帰	**10**	**llm-jp-13b**	○	○			**0.615**	**12.84%**
ロジスティック回帰	10	llm-jp-13b	○		○		0.571	4.77%
ロジスティック回帰	10	llm-jp-13b	○			○	0.503	-7.71%
DeBERTa	10		○				0.744	
DeBERTa	**10**	**llm-jp-13b**	○	○			**0.883**	**18.68%**
DeBERTa	10	llm-jp-13b	○		○		0.722	-2.96%
DeBERTa	10	llm-jp-13b	○			○	0.732	-1.61%
ロジスティック回帰	100		○				0.664	
ロジスティック回帰	**100**	**llm-jp-13b**	○	○			**0.729**	**9.79%**
ロジスティック回帰	100	llm-jp-13b	○		○		0.656	-1.20%
ロジスティック回帰	100	llm-jp-13b	○			○	0.645	-2.86%
DeBERTa	100		○				0.915	
DeBERTa	**100**	**llm-jp-13b**	○	○			**0.925**	**1.09%**
DeBERTa	100	llm-jp-13b	○		○		0.888	-2.95%
DeBERTa	100	llm-jp-13b	○			○	0.901	-1.53%
ロジスティック回帰	**1000**		○				**0.809**	
ロジスティック回帰	1000	llm-jp-13b	○	○			0.799	-1.24%
ロジスティック回帰	1000	llm-jp-13b	○		○		0.788	-2.60%
ロジスティック回帰	1000	llm-jp-13b	○			○	0.788	-2.60%
DeBERTa	1000		○				0.946	
DeBERTa	**1000**	**llm-jp-13b**	○	○			**0.946**	**0.00%**
DeBERTa	1000	llm-jp-13b	○		○		0.916	-3.17%
DeBERTa	1000	llm-jp-13b	○			○	0.944	-0.21%
ロジスティック回帰	187528		○				0.906	
DeBERTa	187528		○				0.949	

表 2 データ拡張モデル別実験結果

モデル	データ数	データ拡張モデル	正答率
ロジスティック回帰	10	llm-jp-3.7b	0.751
ロジスティック回帰	10	llm-jp-13b	0.615
ロジスティック回帰	100	llm-jp-3.7b	0.805
ロジスティック回帰	100	llm-jp-13b	0.729
ロジスティック回帰	1000	llm-jp-3.7b	0.814
ロジスティック回帰	1000	llm-jp-13b	0.799
DeBERTa	10	llm-jp-3.7b	0.895
DeBERTa	10	llm-jp-13b	0.883
DeBERTa	100	llm-jp-3.7b	0.926
DeBERTa	100	llm-jp-13b	0.925
DeBERTa	1000	llm-jp-3.7b	0.945
DeBERTa	1000	llm-jp-13b	0.946

細に分析した．

実験結果から，参照なし文書作成が他のデータ拡張手法に比べて最も高い精度を示すことが明らかになった．一方で，参照あり文書作成や翻訳拡張では，生成元のデータと類似していることから，モデルの精度低下を招く場合があることが確認された．

さらに，データ量の増加がモデル性能に与える影響については，ロジスティック回帰や DeBERTa が少量データにおいて顕著な精度向上を示す一方，データ量が増えるにつれて性能向上の限界が見られる結果となった．このことは，データ量とモデルの性能向上の関係が，モデルの特性や学習方法に依存していることを示唆している．

本研究の成果は，少量ショットにおけるデータセットにおいて，参照なしデータ生成を用いて精度向上ができたことを確認し，コスト効率の良い自然言語処理システムを構築できる可能性が見えた．今後は，より多様なデータセットやタスク，生成データの傾向において提案手法の汎用性を検証し，さらなる改良を進めることが課題である．

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 Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. **CoRR**, Vol. abs/1901.11196, , 2019.

[3] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. **CoRR**, Vol. abs/1808.09381, , 2018.

[4] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. 2022.

[5] 藤井巧朗, 勝又智. 日本語タスクにおける LLM を用いた疑似学習データ生成の検討. 言語処理学会第 30 回年次大会, 2024.

[6] 小野寺優, 新納浩幸. LLM を利用した文書分類のための Data Augmentation. 言語処理学会第 30 回年次大会, 2024.

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. **CoRR**, Vol. abs/2006.03654, , 2020.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **ICLR (Poster)**, 2015.

[9] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, 2020.