LLM を用いたキーワードに基づく 文書分類ためのデータ拡張の試みと評価

小野寺優 [1]　新納浩幸 [2]
[1] 茨城大学大学院理工学研究科情報工学専攻　[2] 茨城大学大学院理工学研究科情報科学領域
{24nm714r, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

概要

大規模言語モデル (LLM) の高い文生成能力を活かすことで，データ拡張 (Data Augmentation) による分類モデルの性能改善が期待できる．しかし，この手法でニュース記事の分類のような多値分類において精度を向上させた研究はまだ少ない．本研究では，ニュース記事を対象として LLM が生成したデータを訓練データに水増しすることで，分類精度が改善されるかを検証した．カテゴリーのキーワードを与えて新たな文書を生成するアプローチを試みたが，このアプローチでは精度が低下することが確認できた．また，拡張したデータの埋め込み表現を分析することで，元となる訓練データと異なる分布のデータが生成され，分類の際のノイズとなっていることが明らかになった．

1　はじめに

データ拡張 (Data Augmentation) とは，少量のデータから新たなデータを生成し，訓練データに水増しすることでモデルの性能改善を目指す手法である．従来の自然言語処理 (NLP) におけるデータ拡張手法は，元のデータにわずかな変化を与えることで，類似したデータを水増しする手法が主流であった．例えば，逆翻訳 [1] や単語の編集 [2, 3] が挙げられる．近年，大規模言語モデル (LLM) の発展に伴い，より多様で自然な言語表現を持つ疑似データを生成する手法 [4] が提案されている．特に，LLM を活用したデータ拡張は，その生成能力の高さを活かした様々なアプローチ [5, 6, 7, 8, 9, 10] が報告されている．しかし，文書分類タスクにおいて，特に多クラス分類のような複雑なタスクにおいては，LLM を活用したデータ拡張の有効性を示した研究はまだ少ない．

本研究では、日本語のニュース記事を対象とした多クラス分類タスクにおいて、LLM が生成したデータを活用することで、分類モデルの性能向上に繋がるかどうかを検証した。データの生成手法として，LLM に各カテゴリーの説明文とキーワードを与え，そのカテゴリーに属するようなニュース記事を生成させた．

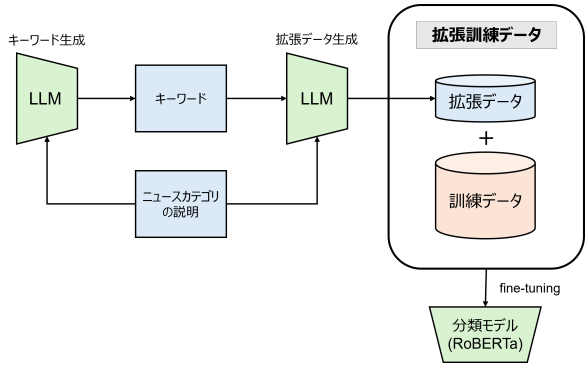

図 1　本研究におけるデータ拡張の流れ

実験では，Livedoor ニュースコーパス[1)] を用いて，拡張前後のデータセットにおいて，分類モデルの精度を比較した．その結果，LLM で生成したデータを拡張したデータセットを用いた場合，分類精度の精度は，元のデータセットの場合と比較して，減少する傾向があることが確認できた．

また，LLM が生成した文書の埋め込み表現を分析した結果，生成されたデータは，元の訓練データとは異なる分布を持つことが明らかになった．この結果は，LLM が多様な文書を生成できる一方で，生成されたデータがノイズとなっている可能性が高いことを示唆している．

2　関連研究

NLP の分野におけるデータ拡張は，従来，元の文書を一度別の言語に翻訳し，その後再び元の言語に翻訳することで拡張データを生成する逆翻訳 [1] や元の文書内の単語に対して類義語置換，ランダム交換，ランダム削除，ランダム挿入といっ

1）　https://www.rondhuit.com/download.html

た操作を行うことでデータを生成する EDA(Easy Data Augmentation)[2] のような言い換えベースのアプローチが主流であった．

近年では，LLM の発展により，プロンプトを活用したデータ生成手法が広く研究されるようになっている．Sahu ら [6] は，特定のラベルに属する文書を収集し，Few-shot prompting を用いてデータを生成した．中町ら [7] は，文書の例とキーワードを組み合わせて Few-shot Prompting を実施し，データ生成行った．しかし，これらの手法は，LLM のコンテキストウィンドウの制限を考慮すると，ニュース記事のような長文データの拡張に適用が困難である．

この課題に対処するため，長文要約を用いたデータ拡張手法が提案されている．王ら [8] は，元データの要約文を訓練データに追加し，短い要約文から元の長文へ段階的に学習を行うことで，2 値分類タスクにおいてモデルの精度を向上させた．小野寺ら [9] は，LLM から生成されたニュース記事の要約文とキーワードを組み合わせて Few-shot prompting を行い，データを生成してカテゴリーの 9 値分類においてデータ拡張を行ったが，モデルの精度を向上させることはできなかった．

また，Zero-shot によるデータ生成も検証されている．藤井ら [4] は，タスクとラベルの簡潔な説明のみを用いてデータ生成を行う手法と，それに加えて文書の内容語をランダムに与える Knowledge-Assisted Data Generation(KADG) という 2 つの手法を提案した．Li ら [10] は，LLM がランダムに生成したトピックを基にデータを生成を行う Zero-shot topic generation を提案した．本研究では，ニュース記事のカテゴリーに関する説明文を基に，LLM が連想したキーワードからデータを生成する Zero-shot Prompting を用いてデータ拡張を試みる．

3 拡張訓練データセットの構築

藤井ら [4] は，タスクおよびラベルの簡潔な説明と文書の内容語を LLM に与えることで，疑似データを生成する手法を提案した．また，Li ら [10] は，LLM がランダムに生成したトピックを基にデータを生成・拡張する手法を提案している．本研究では，日本語の文書分類タスクにおいて，これらの手法と類似するが，キーワードの生成手法およびそれらの提示方法において独自性のあるデータ拡張手法を試みた．具体的には，以下の 2 段階の工程で拡張データ生成を行った．

```
Livedoorニュースの「{category}」に関連するキーワードを30個挙げてJSON形式で出力してください。¥n

{category}の特徴：¥n
-{features}¥n

出力フォーマット：¥n
{{
    "category": "{category}",
    "features": "{features}",
    "keywords": [
        "キーワード1",
        "キーワード2",
        ...
    ]
}}
```

図 2 キーワード生成のプロンプトテンプレート．{category} には対象のカテゴリー名，{features} にはカテゴリーの簡潔な説明が挿入される．

```
Livedoorニュースの「{category}」のようなニュース記事を以下の特徴とキーワードを参考にして生成してください。¥n

{category}の特徴：¥n
- {features}¥n

{category}のキーワード：¥n
- {keywords}

出力フォーマット：(以下の形式で出力してください。
タイトルや本文に引用符は使用しないでください。)¥n
{{
    "category": "{category}",
    "label": "{label}",
    "title": "<記事タイトル>",
    "body": "<記事本文>"
}}
```

図 3 拡張データ生成のプロンプトテンプレート．{keyword} には対象カテゴリーのキーワードをランダムに 5 個，{label} には対象カテゴリーのラベル番号が挿入される．{category} と {features} は図 2 と同様．

キーワード生成 LLM にニュースカテゴリーの簡潔な説明文を与え，キーワードをリスト形式で 30 個生成させる．

拡張データ生成 LLM にニュースカテゴリーの簡潔な説明文と，生成されたキーワードの中からランダムに抽出した 5 個を与え，拡張データとなる文書を生成させる．この際，キーワードの組み合わせはシード値を設定することで毎回変更した．

データ生成に使用した LLM は，Anthropic 社が提供している Claude 3.5 Sonnet モデル[2] の `claude-3-5-sonnet-20241022` と OpenAI 社が提供している GPT4o モデル[3] の `gpt-4o-2024-11-20` の 2

2） https://docs.anthropic.com/ja/docs/about-claude/models
3） https://platform.openai.com/docs/models/gp

表 1　各分類モデルにおける 5 回の平均正解率と標準偏差

	Baseline	20 × 9 拡張			40 × 9 拡張		
		+Claude	+GPT	Baseline*	+Claude	+GPT	Baseline*
正解率	92.17(0.32)	90.56(0.72)	91.69(0.15)	92.07(0.95)	90.80(0.42)	91.40(0.42)	92.85(0.83)

種類である．図 2 と図 3 に，プロンプトのテンプレートを示す．生成における LLM のハイパーパラメータ設定は付録の A.2 節に示す．

4　実験

本研究では，長い文書に対する深い文脈理解が必要とされるニュース記事カテゴリーの多値分類タスクにおいて，本手法が有効性を検証した．

4.1　データセット

本手法の評価には，Livedoor ニュースコーパスを使用した．Livedoor ニュースコーパスには 9 つのカテゴリーのニュース記事が含まれており，記事はタイトルと本文から構成されている．本実験では本文のみを用い，分類モデルによる 9 値分類を実施した．具体的には，各カテゴリーに 0〜8 までのラベル番号を割り当て (表 3)，分類モデルがニュース記事のラベル番号を予測するタスクを設定した．

Livedoor ニュースコーパスには合計 7367 件のデータが含まれるが，カテゴリー毎のデータ数は不均衡である．このため，モデルの fine-tuning で用いる訓練データおよび検証データは均衡を保つように調整した．具体的には，以下のようにデータを分割し，データセットを構築した．

訓練データ　各カテゴリーから 200 件 (合計 1800 件) を抽出し，そのうち 100 件 (合計 900 件)，120 件 (合計 1080 件)，140 件 (合計 1260 件) の 3 種類のデータセットを構築．

検証データ　各カテゴリーから 50 件 (合計 450 件) を抽出し，データセットを構築．

テストデータ　残りのデータ (合計 5117 件) を使用してデータセットを構築．テストデータのカテゴリーごとのデータ数は不均衡である．

4.2　文書分類の設定

分類モデルのベースとして，rinna が Hugging-Face にて公開している日本語 RoBERTa[11] モデルの `rinna/japanese-roberta-base`[4] を使用した．本実験で fine-tuning によって構築される分類モデルは以下の通りである．

Baseline　合計 900 件の元の訓練データセットで学習させた分類モデル

Baseline+Claude　元データに Claude 3.5 Sonnet が生成した拡張データを追加して学習した分類モデル

Baseline+GPT　元データに GPT4o が生成した拡張データを追加して学習した分類モデル

Baseline*　Baseline+Claude や Baseline+GPT のデータ数と同じ数の元データで学習した分類モデル

分類モデル学習時のハイパーパラメータの詳細は付録の A.2 に示す．

本研究では，データ拡張の規模を以下の 2 種類に設定した．カテゴリーごとのデータ数は常に均衡になるように調整した．

- 各カテゴリー 20 件 (合計 180 件)
- 各カテゴリー 40 件 (合計 360 件)

これに対応して，Baseline*モデルは 4.1 節で構築した訓練データセット (各カテゴリー 120 件および 140 件の抽出データ) を使用して学習を行った．

4.3　文書分類の結果

構築した分類モデルを用いた文書分類を 5 回行い，その平均正解率を算出して本手法の有効性を評価した．その結果を表 1 に示す．

結果として，Baseline モデルに比べて本手法の正解率が低下する傾向があることが確認できた．また，拡張データの量を増やした場合，正解率がさらに低下する傾向が見られた．

5　考察

実験では，LLM を用いて生成したデータを拡張した場合，分類の正解率が低下する傾向が確認された．これは，生成データが分類モデルの学習において悪影響を及ぼした可能性が示唆される．

5.1　元の文書と生成文書の距離

生成文書と元の文書との間に埋め込み表現の分布に差が存在する可能性がある．これを確認するた

4)　https://huggingface.co/rinna/japanese-roberta-base

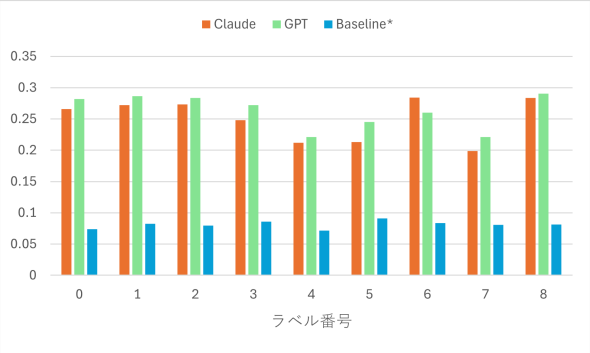

図 4 Baseline で用いた元の訓練データとのユークリッド距離

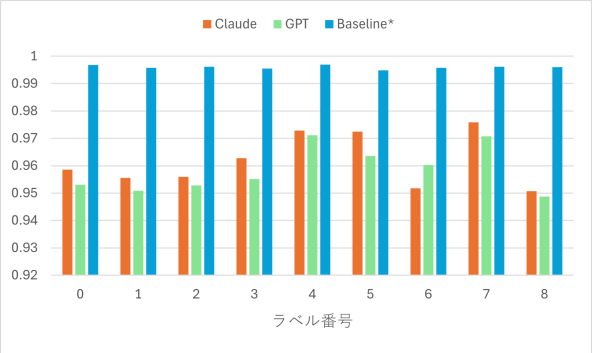

図 5 Baseline で用いた元の訓練データとのコサイン類似度

め，以下の手順で分析を行った．

1. 各カテゴリーの文書に対してクラス平均埋め込み表現を取得する．
2. 取得した埋め込み表現の間でユークリッド距離とコサイン類似度を測定する．

埋め込み表現の取得には，Microsoft 社が Hugging-Face にて公開している多言語埋め込みモデル [12] の `intfloat/multilingual-e5-large`[5] を使用した．文書間の距離の測定結果を図 4 と図 5 に示す．

結果として，LLM が生成した文書は元の文書に比べてユークリッド距離が大きく，コサイン類似度が低いことが確認された．これにより，生成文書は元の文書とは異なる分布に属していることが分かり，分類モデルの学習時に拡張データがノイズとして作用してしまう可能性が考えられる．

5.2 訓練データ数による影響

従来の研究では，データ拡張が特に訓練データが少量の場合に有効であり，データ量が増加するにつれてその効果は減少することが報告されている [2, 3]．この点を踏まえ，Baseline モデルの正解率 (表 1) を見ると，既に高水準の精度が達成されていることから十分な訓練が行われている可能性があると推測される．そこで，Baseline の訓練データの量を各カテゴリー 50 件ずつ (合計 450 件) に減らした場合において，本手法によるデータ拡張を試みた．その結果を表 2 に示す．

結果として，訓練データ量を減らした場合でも，データ拡張を行った方が正解率が Baseline を下回ることが確認された．これにより，単にデータ量が多すぎるために改善が見られなかったのではなく，本手法で生成したデータが分類タスクにおいて悪影響を与えていることが示唆される．

表 2 Baseline のデータ量を 450 件としたときの 5 回の平均正解率と標準偏差

	Baseline	20 × 9 拡張	
		+Claude	+GPT
正解率	89.64(0.09)	88.64(0.52)	88.05(0.39)

6 おわりに

本研究では，LLM を用いた 2 段階の Zero-shot 学習に基づくデータ生成手法を提案し，文書分類タスクにおけるデータ拡張の有効性を検証した．具体的には，(1) 簡潔なカテゴリーの説明からキーワードを生成し，(2) そのキーワードを基に記事を生成するという流れでデータ拡張を試みた．しかし，実験の結果，Baseline モデルと比較して分類精度が低下する傾向が確認された．

精度低下の主な原因として，生成データが元のデータと異なる分布に属しており，分類モデルの学習においてノイズとして作用している可能性が示唆された．特に，生成データが元のデータとの埋め込み表現の分布において距離が大きいことが確認され，この分布の差異が分類精度に悪影響を与えたと考えられる．

本研究では，簡潔なカテゴリーの説明文をプロンプトで使用したが，より具体的かつ詳細なカテゴリーの特徴をプロンプトに含めることで，生成データの質を向上させる可能性がある．今後の課題としては，生成データの分布を元のデータへ近づけるための方法の検討することや異なる LLM を使用した際のデータ拡張効果を評価し，モデル選択の影響を調査することが挙げられる．

5） https://huggingface.co/intfloat/multilingual-e5-large

謝辞

本研究は国立国語研究所の共同研究プロジェクト「テキスト読み上げのための読みの曖昧性の分類と読み推定タスクのデータセットの構築」及び JSPS 科研費 23K11212 の助成を受けています.

参考文献

[1] Sam Shleifer. Low resource text classification with ulmfit and backtranslation, 2019.

[2] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.

[3] 高萩恭介, 新納浩幸. BERT を用いた Data Augmentation 手法の改善と JGLUE による評価. 言語処理学会第 29 回年次大会発表論文集, pp. 305–310, 2023.

[4] 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 回年次大会発表論文集, pp. 2284–2289, 2024.

[5] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. Gpt3mix: Leveraging large-scale language models for text augmentation, 2021.

[6] Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. Data augmentation for intent classification with off-the-shelf large language models. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, **Proceedings of the 4th Workshop on NLP for Conversational AI**, pp. 47–57, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] 中町礼文, 西内沙恵, 浅原正幸, 佐藤敏紀. 語彙と品質を考慮したデータ水増しの言語教育支援への適用. 言語処理学会第 29 回年次大会発表論文集, pp. 1924–1929, 2023.

[8] 王悦綸, 吉永直樹. 文書分類のための要約に基づくデータ拡張. 言語処理学会第 30 回年次大会発表論文集, pp. 2449–2454, 2024.

[9] 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 回年次大会発表論文集, pp. 3265–3270, 2024.

[10] Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. Data generation using large language models for text classification: An empirical case study, 2024.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024.

A　付録

A.1　Livedoor ニュースコーパス

Livedoor ニュースの記事が 9 個のカテゴリーに分けられたコーパス．カテゴリーに応じて表 3 のようにラベル番号を付与する．

表 3　Livedoor ニュースコーパスのカテゴリー

番号	カテゴリー名
0	独女通信
1	IT ライフハック
2	家電チャンネル
3	LivedoorHOMME
4	MOVIE ENTER
5	Peachy
6	S-MAX
7	Sports Watch
8	トピックニュース

A.2　実装時の詳細設定

実装の環境は，NVIDIA GeForce RTX 4070(12GB) である．

拡張データ生成時の設定

データ生成には，`claude-3-5-sonnet-20241022` と `gpt-4o-2024-11-20` を使用した．`temperature` は，キーワード生成時は 0.0，拡張データ生成時は多様性のある記事を得るため，0.5 に設定した．`max-tokens` は，1024 とした．

文書分類時の設定

RoBERTa の fine-tuning により，分類モデルを構築した際のハイパーパラメータを以下の表 4 に示す．

表 4　文書分類モデル学習のパラメータ

最適化関数	AdamW
学習率	2e-5
バッチサイズ	8

エポック数は，検証データの Loss が 3 エポック連続で最小値にならなかったときに学習を停止するように `Early Stopping` を設定した．学習が止まらなかったときの最大エポック数は 100 とした．