

LLM を利用した文書分類のための Data Augmentation

小野寺優¹ 新納浩幸²

¹ 茨城大学工学部情報工学科 ² 茨城大学大学院理工学研究科情報科学領域
20t4022l@vc.ibaraki.ac.jp hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

概要

大規模言語モデル (Large Language Model:LLM) は様々な自然言語処理タスクに利用され、多くの成果を挙げている。本研究では、文書分類に対する Data Augmentation に LLM を利用することを試みる。LLM を Data Augmentation に利用する場合、少数のラベル付きデータからどのようなプロンプトで文書を生成すればよいか重要となる。ここでは、少数の訓練データからキーワードと要約文を生成し、その2要素で構成されたプロンプトを試みる。実験では、livedoor ニュースコーパスの文書分類タスクを行い、本プロンプトの有効性を調査した。その結果、本プロンプトの有効性を示すことはできなかった。これを踏まえて、LLM を用いて文書分類に対するデータを生成する上での課題を考察する。

1 はじめに

大規模言語モデル (Large Language Model:LLM) は、膨大な量のテキストデータから学習された言語モデルであり、質問応答、文章要約、あるいは機械翻訳など様々な自然言語処理タスクに利用され、多くの成果を挙げている。また同時に LLM の高い文生成能力を生かした様々な利用方法も検討されている。

本研究では、LLM の利用の一つとして文書分類タスクに対する訓練データの自動構築を試みる。具体的には、Data Augmentation のアプローチを取る。つまり、少数の訓練データを元に、LLM を利用して各分類ラベルに対する文書を生成し、そのラベル付き文書を元の少数の訓練データに追加することで分類精度を向上させる。

LLM を用いた Data Augmentation の手法はいくつか提案されている。Yoo ら [1] は、文とラベルの組から構成されたプロンプトを用いて、文とラベルを同時に LLM に自動生成させる手法を提案した。また、Sahu ら [2] は、特定のラベルに属する文から few-shot プロンプトを作成し、Data Augmentation を

行うアプローチを提案した。しかし、これらの手法で扱われるデータは短い文であり、長文のような few-shot プロンプトを構成するのが難しいデータに対する有効性は明らかになっていない。

本研究では、ChatGPT(GPT-3.5) を用いて、キーワードと文書の要約文で構成されたプロンプトにより、長文に対応した Data Augmentation 手法を試みる。キーワードは各分類ラベルの文書から抽出することから獲得し、要約文は訓練データとなる文書を LLM に要約させることで生成する。実験では、livedoor ニュースコーパス¹⁾の文書分類タスクを行い、本プロンプトの有効性を調査した。しかし、モデルの精度向上は見られず、本プロンプトの有効性を示すことはできなかった。ここでは上記実験結果を報告するとともに、LLM を用いて文書分類に対するラベル付きデータを生成する上での課題を考察する。

2 関連研究

自然言語処理において、データ不足を緩和させることを目的とする手法は広く研究されている。He ら [3] は、LLM にアノテーションを行う際の思考過程を説明させ、それを元にしたプロンプトを作成し利用することで、LLM によるラベルなしデータに対するアノテーション作業を行わせる AnnoLLM という手法を提案した。これにより、LLM によってアノテーション作業が代替できる可能性が示された。

少数の訓練データに対してデータを水増しすることでモデルの性能を向上させることを目的とする Data Augmentation というアプローチがある。Wei ら [4] は、訓練データの文書に対して、類義語置換、ランダム挿入、ランダム交換、あるいはランダム削除の4つの方法を用いた EDA(Easy Data Augmentation) という手法を提案した。

近年では、LLM を用いた Data Augmentation の手法も検討されている。Yoo ら [1] は、画像処理の分

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

野に利用される Mix-up というアプローチを参考にし、GPT3Mix という手法を提案した。この手法では、文とラベルの組からなるプロンプトを作成し GPT-3 に与えることで、新たな文とラベルの組を自動生成させる。また、Sahu ら [2] は特定の一つのラベルに属する文のみから構成されるプロンプトを作成し GPT-3 に与えることで、ラベルを考慮した新たな文を生成する Data Augmentation 手法を提案した。更に、中町ら [5] は特定の一つのラベルに属する文に加えて、キーワードとなる単語をプロンプトに与えることで、語彙と品質を考慮した文生成による Data Augmentation 手法を提案している。

3 プロンプトの作成

本研究では、キーワードと文書の要約文を含むプロンプトを LLM に与えることで、データの生成を行う。文生成を行うための LLM は、ChatGPT(GPT-3.5) を使用する。本研究の流れを図 1 に示す。

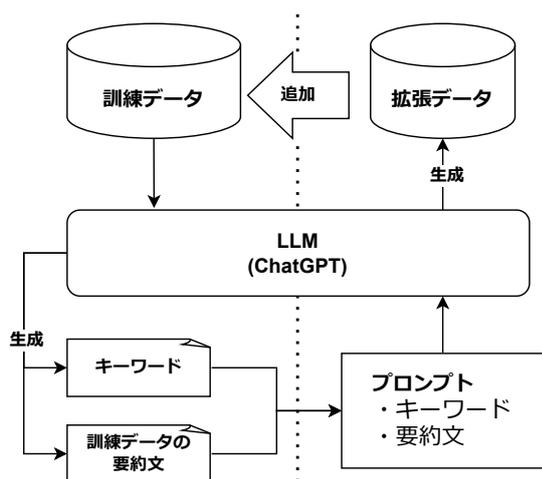


図 1 本研究における Data Augmentation の流れ

3.1 プロンプトに使用するデータ

プロンプトに使用するデータとして、livedoor ニュースコーパスを使用する。これは、NHN Japan 株式会社が運営する「livedoor ニュース」からニュース記事を収集し、可能な限り HTML タグを取り除いたものである。このコーパスには、9つのカテゴリのニュース記事が格納されている。

3.2 プロンプトの構成

本研究では、使用される文書が長文であることを想定し、キーワードと文書の要約文から構成される 3 種類のプロンプトを考案する。

```

あなたは最高のニュース記事ライターです。以下の制約を守ったニュース記事を書くことができます。例文は実際の記事の要約文の例です。制約とキーワード、例文を参考にして、まったく新しいニュース記事を書いてください。

#制約
(1)生成する記事の文字数は1000文字程度とする
(2)記事全体の流れが自然になるようにする

#キーワード
{出力したいカテゴリのキーワード3つ}

#例文
例1: {出力したいカテゴリのテキストの要約文_1}
例2: {出力したいカテゴリのテキストの要約文_2}
...
例10: {出力したいカテゴリのテキストの要約文_3}

#出力
生成文1:
  
```

図 2 プロンプト 1 のテンプレート

```

あなたは最高のニュース記事ライターです。以下の制約を守ったニュース記事を書くことができます。例文は9つにカテゴリ分けされたニュース記事の要約文を1つずつ提示したものです。制約とキーワード、例文を参考にして、まったく新しいニュース記事を書いてください。

#制約
(1)生成する記事の文字数は1000文字程度とする
(2)記事全体の流れが自然になるようにする
(3)カテゴリ「{出力したいカテゴリ名}」の記事を書くこと

#キーワード
{出力したいカテゴリのキーワード3つ}

#例文
「独女通信」の例: {独女通信ラベルに属するテキストの要約文}
「ITライフハック」の例: {ITライフハックラベルに属するテキストの要約文}
...
「トピックニュース」の例: {トピックニュースラベルに属するテキストの要約文}

#出力
生成文1:
  
```

図 3 プロンプト 2 のテンプレート

```

あなたは最高のニュース記事ライターです。以下の制約を守ったニュース記事を書くことができます。例文は実際のニュース記事の要約文です。制約とキーワード、例文を参考にして、まったく新しいニュース記事を書いてください。

#制約
(1)生成する記事の文字数は1000文字程度とする
(2)記事全体の流れが自然になるようにする

#キーワード
{出力したいカテゴリのキーワード5つ}

#例文
例1: {出力したいカテゴリ以外のテキストの要約文_1}
例2: {出力したいカテゴリ以外のテキストの要約文_2}
...
例8: {出力したいカテゴリ以外のテキストの要約文_8}

#出力
生成文1:
  
```

図 4 プロンプト 3 のテンプレート

要約文は、ChatGPT に元となるニュース記事の本文を一つずつ入力することで生成する。各プロンプトにおける要約文の使い方を以下に示す。

プロンプト 1

特定の一つのカテゴリに属するニュース記事本文の要約文 10 個与える (図 2)。

プロンプト 2

各カテゴリに属するニュース記事本文の要約文を 1 個ずつ与える (図 3)。

プロンプト 3

特定の一つのカテゴリに属するニュース記事本文の要約文を除いた、他の 8 カテゴリのニュース記事本文の要約文を 1 個ずつ与える (図 4)。

また、キーワードは ChatGPT に生成した要約文

10 個を与えることでカテゴリごとに生成させる。与えるキーワードの数は、プロンプトの種類により変わり、プロンプト 1 及びプロンプト 2 では 3 つ、プロンプト 3 では 5 つとする。具体的なキーワードは、付録の表 4 に示す。

4 実験

4.1 文書分類に使用するモデル

rinna 社から公開されている日本語 RoBERTa モデルを使用した。これは Hugging Face 社の Transformers ライブラリから、"japanese-roberta-base" という名前で公開されているモデルである²⁾。このモデルは事前学習に日本語 Wikipedia と CC-100 が用いられている。

4.2 データセット

本研究では、提案手法の評価を行うために、livedoor ニュースコーパスを用いた文書分類タスクを行う。各カテゴリに対してラベルを割当て、各ラベルの文書をランダムに同数取り出す。そして、取り出したラベルとテキストの組から、訓練データ、検証データ、テストデータの 3 つを作成する。このときのデータの割当ては、訓練データと検証データが各カテゴリ 10 件ずつの文書を取り出した合計 90 件、テストデータが各カテゴリ 100 件ずつの文書を取り出した合計 900 件となる。

拡張訓練データは、ここで割当てられた 90 件の訓練データを元に作成する。拡張データは、提案するプロンプト一つにつきカテゴリごとに 3 件ずつ、すなわち 27 件のデータを作成する。したがって、拡張後の訓練データの数は元の訓練データ 90 件に拡張データ 27 件を加えた 117 件のデータとなる。

4.3 文書分類モデルの学習

文書分類モデルの学習時の設定を以下に示す。

- 最適化関数：確率的勾配降下法 (SGD)
- 学習率：0.001
- バッチサイズ：3

エポック数の決定には、アーリーストッピングを用いる。検証データの Loss が 3 エポック連続で最小値に更新されなかった場合に学習が中断され、モデルの構築を行う。このとき、学習の中断がなかつ

た場合の最大エポック数は 100 と設定する。

4.4 実験結果

RoBERTa を訓練データで fine-tuning することで、モデルの構築を行う。構築したモデルは、文書分類タスクの正解率により評価する。

実験では、以下の 5 つのパターンにおける訓練データセットを作成し、正解率を比較した。

- 元の訓練データ 90 件 (baseline)
- 拡張訓練データ (プロンプト 1)
- 拡張訓練データ (プロンプト 2)
- 拡張訓練データ (プロンプト 3)
- 元の訓練データ 117 件

この評価結果を以下の表 1 に示す。

訓練データ	正解率
元データ 90 件	0.7655
拡張データ (プロンプト 1)	0.7636
拡張データ (プロンプト 2)	0.7573
拡張データ (プロンプト 3)	0.7573
元データ 117 件	0.7989

結果として、Data Augmentation を行った場合、モデルの性能をわずかに悪化させてしまうことが分かった。

5 考察

5.1 実験結果の考察

livedoor ニュースコーパスの文書分類において本プロンプトによる Data Augmentation を試みた場合、モデルの性能はわずかに低下してしまった。これは、新たに生成したデータの中にノイズとなるデータが含まれていたことが要因だと考えられる。

図 5 は、実験におけるカテゴリごとの正解率を示している。Livedoor HOMME や SMAX、トピックニュースは、全てのプロンプトにおいて正解率が減少していることが読み取れる。これは、Livedoor HOMME やトピックニュースは多様な分野を取り扱うニュースカテゴリのため、LLM が少量の訓練データの中からニュース記事の特徴を学習するのが難しかったのではないかと考える。また、SMAX は似た内容のニュース記事を持つ IT ライフハックや家電チャンネルと分類モデルが混同してしまい正解率が減少してしまったと考えられる。

2) <https://huggingface.co/rinna/japanese-roberta-base>

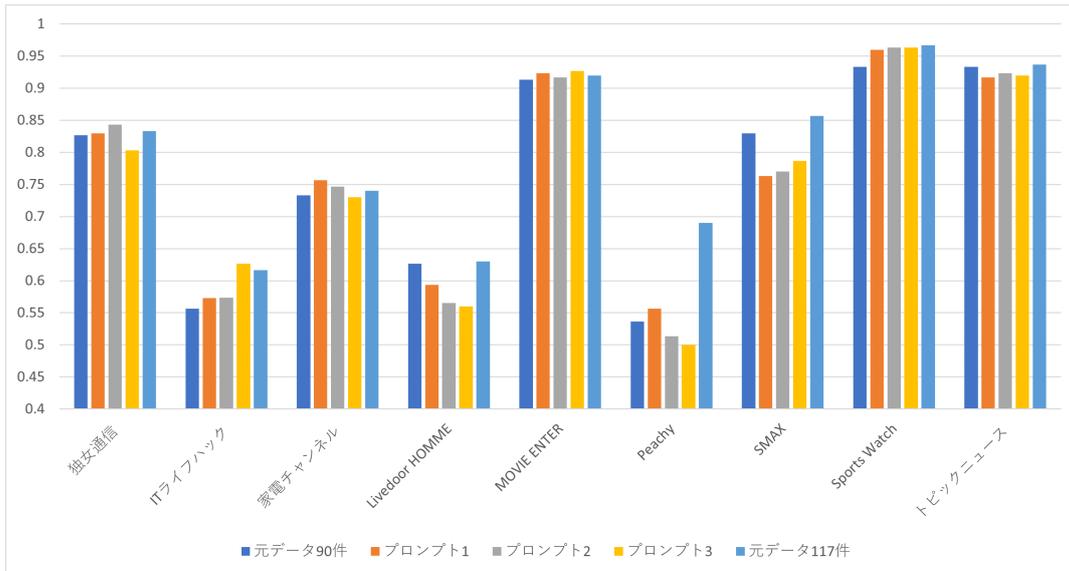


図5 カテゴリごとの正解率

5.2 キーワード変更による実験

3.2節では、ChatGPTにテキストの要約文を10個与えることでキーワードの生成を行っていた。本節では、主観的ではあるが、人手で要約文を読み、カテゴリごとのキーワードを作成した場合で同様の実験を行う。ここで作成した具体的なキーワードは、付録の表5に示す。人手作成のキーワードによる拡張訓練データの評価結果を以下の表2に示す。

表2 キーワード変更時のモデルの評価結果

訓練データ	正解率
拡張データ (プロンプト 1)	0.7570
拡張データ (プロンプト 2)	0.7552
拡張データ (プロンプト 3)	0.76

表1及び表2より、キーワードの変更は、本プロンプトにおいてモデルの性能を向上させることはできないことが分かった。すなわち、キーワードは本プロンプトにおいてモデルの性能に与える影響が小さいと考えられる。

5.3 データ数変更による実験

4.4節では、各カテゴリ3件ずつ、合計27件の拡張による実験を行った。ここで有効性が見られなかった要因の一つとして、拡張するデータ数がカテゴリによっては少なすぎた可能性もある。本節では、拡張するデータ数を各カテゴリ6件ずつの合計54件とし、同様の実験を行い評価する。この評価結果を表3に示す。

表3 データ数変更時のモデルの評価結果

訓練データ	正解率
拡張データ (プロンプト 1)	0.7581
拡張データ (プロンプト 2)	0.7537
拡張データ (プロンプト 3)	0.7474
元データ 144 件	0.7935

表1と表3の結果を比較すると、拡張するデータ数を増やしてもモデルの性能は向上させることができないことが分かった。これは、拡張データの数が増えたことによりノイズとなるデータが更に増えてしまったことや本プロンプトがChatGPTにおいて多くのデータを生成するのに適したものではなかったことが要因として考えられる。

6 おわりに

本研究では、キーワードと文書の要約文により構成されたプロンプトによるLLMを利用したData Augmentationの手法を試みた。キーワードや要約文の与え方を変えた3種類のプロンプトを考案し、日本語RoBERTaによるlivedoorニュースコーパスの文書分類タスクにより評価を行った。その結果、モデルの性能の向上は見られず、試みたプロンプトから生成された文によるData Augmentationの有効性を示すことはできなかった。

今後は、Data Augmentationのみに括らず、LLMによる質の良いデータ生成とその活用方法の可能性の調査に取り組んでいきたい。

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] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. Gpt3mix: Leveraging large-scale language models for text augmentation, 2021.
- [2] 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 Weiyang Shi, editors, **Proceedings of the 4th Workshop on NLP for Conversational AI**, pp. 47–57, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [3] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. Annollm: Making large language models to be better crowdsourced annotators, 2023.
- [4] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.
- [5] 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.

A キーワードの詳細

4 節及び 5.2 節で使用したキーワードを以下に示す。プロンプト 1, 2 では、各表の左から 3 つのキーワードを使用し、プロンプト 3 では、5 つ全てのキーワードを使用する。

A.1 ChatGPT が生成したキーワード

ChatGPT に生成させたカテゴリごとのキーワードを以下の表 4 に示す。

表 4 ChatGPT が生成したキーワード

カテゴリ名	キーワード
独女通信	ダイエット, 女性, 映画, 女優, ヨーグルト
IT ライフハック	Ultrabook, Android, テクノロジー, アプリ, セキュリティ
家電チャンネル	スマートフォン, テクノロジー, プロジェクター, マッサージャー, メディア
Livedoor HOMME	借金, ファッション, ゴルフ, ドライブ, 転職
MOVIE ENTER	映画, 公開, 日本, 主題歌, 俳優
Peachy	女性, ファッション, iPhone, イベント, チャリティ
SMAX	ソフトウェア更新, Android, スマートフォン, 不具合修正, NTTドコモ
Sports Watch	W 杯, 日本代表, 選手, インタビュー, マッチ・試合
トピックニュース	ネット掲示板, ニュース記事, メディア, 反応・コメント, イベント・出来事

A.2 人手で作成したキーワード

ニュース記事の要約文を読み、独自に作成したカテゴリごとのキーワードを以下の表 5 に示す。

表 5 人手で作成したキーワード

カテゴリ名	キーワード
独女通信	女性, 恋愛, 結婚, 健康, 美容
IT ライフハック	スマートフォン, タブレット, アプリ, セキュリティ, ネットワーク
家電チャンネル	家電, 携帯電話, 電子, 機器, 映像
Livedoor HOMME	仕事, ブランド, キャリア, 転職, 金
MOVIE ENTER	映画, 主人公, 監督, 俳優, 演技
Peachy	女性, ファッション, かわいい, 手帳, イベント
SMAX	スマートフォン, モバイル, 発表, 機能, アップデート
Sports Watch	選手, 試合, スポーツ, 監督, 記者
トピックニュース	芸能人, 事件, 政治, 話題, 意見

B プロンプトと生成文の例

3.2 節では、各プロンプトにおけるテンプレートを作成した。本節では、テンプレートに従って作成したプロンプト 1 とその生成文の一例を以下の図 6 及び図 7 に示す。

<p>あなたは最高のニュース記事ライターです。以下の制約を守ったニュース記事を書くことができます。例文は実際の記事の要約文の例です。制約とキーワード、例文を参考にして、まったく新しいニュース記事を書いてください。</p> <p>#制約 (1)生成する記事の文字数は1000文字程度とする (2)記事全体の流れが自然になるようにする</p> <p>#キーワード 「ダイエット」、「女性」、「映画」</p> <p>#例文 例1：多くの人が様々なダイエットを試みても、その継続が難しいことを経験している...(省略) 例2：最近、女性専用車両での通勤中に見かける女性のマナーの悪さについての記事です。...(省略) ... 例10：近年、アイドルがメディアで注目を浴びており、特にAKB48がその代表格となって...(省略)</p> <p>#出力 生成文1：</p>
--

図 6 独女通信におけるプロンプト 1 の例

<p>「映画ヨガ」が女性たちのダイエットに革命をもたらす</p> <p>多くの女性が悩む「ダイエット」。その継続が難しいことはよく知られています。しかし、今回、新たな方法が登場し、女性たちに大きな魅力を持って受け入れられています。ヨガインストラクターの城所恵美先生が、「続けやすい」ダイエット方法として、家で映画を観ながら行える「映画ヨガ」を紹介し、多くの女性たちの注目を浴びています。</p> <p>この方法の魅力は、家庭で手軽に実践できる点です。必要なのはインターネット接続とXbox 360で利用できるZuneビデオのみ。このアプローチは、映画好きな女性にとって理想的で、ダイエットを楽しみながら続けることができます。</p> <p>「映画ヨガ」では、異なるヨガポーズがさまざまな映画ジャンルに合わせて紹介されています。たとえば、アクション映画を鑑賞する際には「ウォーリア」のポーズをとることで、力強く勇敢な気分を味わえます。恋愛映画には「鳩」のポーズが適しており、心が開放され感情豊かになる手助けになります。サスペンス・ホラー映画を楽しむ際には「舟」のポーズがおすすめで、身体が不安定なバランスポーズをとることで、恐怖感と緊張感をより感じやすくなります。そして、SF映画に対して「飛行機」のポーズが紹介されており、キープが難しいため高いダイエット効果が期待されます。</p> <p>映画ヨガは、ダイエットだけでなく、映画鑑賞の楽しみも提供します。紹介された映画はすべてZuneビデオで配信されており、映画ヨガを試しながら観賞することができます。これはXbox 360、Zuneビデオ、およびZuneファンに関連する情報で、多くの女性たちにとって魅力的なダイエット方法として注目されています。映画ヨガを通じて、ダイエットを続けつつ映画の楽しみを味わう新たなアプローチが女性たちに提供されています。</p>

図 7 上記のプロンプト 1 による生成例

図 7 のように、生成文にはタイトルが同時に生成されるときがある。本研究では、このようなタイトル部分は取り除き、本文のみを使用した。