E-011

日本語不均衡データに対しての ChatGPT を利用したかさ増し手法の提案
Proposal of a Data Augmentation Method Using ChatGPT for Japanese Imbalanced Data

澤崎 夏希[1)]　遠藤 聡志[2)]
Natsuki Sawasaki　Satoshi Endo

1　はじめに

現在、大規模データの利用が進み、機械学習を用いた手法により多くの問題が解決されている。特に大規模言語モデルの性能向上によって、これまでより飛躍的に幅の広い課題が解決されるようになり、アノテーションなどの人手での作業が前提であったタスクに対しても人間の作業者と同等以上の精度を発揮することが確認されている [1]。一方で、独自のドメインや少量のデータに依存した問題の解決には課題が残っており、リソースの少ない研究者・企業にとっての障壁となっている。

分類問題において、カテゴリを持つデータセットはカテゴリ毎のデータ数が揃っていることが望ましいが、実際には各カテゴリのデータ数が不均衡な場合がある。不均衡なデータセットに対して学習を行った場合、データ量の多いカテゴリの特徴を過学習し、少量データの特徴をうまく学習できないことが知られている [2]。不均衡データへの対策として、各カテゴリのデータ量を揃えるダウンサンプリング手法とかさ増し手法が挙げられる。ダウンサンプリング手法は少量データに合わせてデータを削減し各カテゴリのデータ数を揃える手法で、かさ増し手法は少量データの特徴を加味した生成を行いデータ量を揃える手法である。画像処理分野においては多くのかさ増し手法が用いられている一方で、自然言語処理の分野においては離散的なベクトルを扱うなどの複雑さから汎用的な手法が確立されていない [3]。

少量カテゴリへの対策として、大規模言語モデルを利用したアプローチと、生成モデルを活用した手法が注目されている。事前学習された BERT[4] などの言語モデルで高精度なベクトル表現を獲得するアプローチに対し、ChatGPT[5] などの生成モデルでは学習データそのものを生成するアプローチが検討されている [6]。特に学習データを増やす生成モデルのアプローチは汎用性が高いが、出力の予測と制御が難しく使用する言語によっても大きく性能が変化することが知られ [7]、英語に対しての精度が高いことが知られている一方で、日本語に大しての性能調査はあまり行われておらず、有効性の検証が課題となっている。そこで本研究では ChatGPT にペルソナを用いたデータかさ増し手法を提案し生成データの制御を目指すと共に、分類問題へ利用した際の精度変化を分析することで、日本語における ChatGPT の制御性とかさまし手法としての有効性についての検討を行う。

1)　†琉球大学 理工学研究科 総合知能工学専攻
Department of Integrated Intelligence Engineering, Graduate School of Science and Engineering, University of the Ryukyus.
2)　†琉球大学 工学科 知能情報コース
Intelligent Information Course, Engineering department, University of the Ryukyus

2　先行研究

2.1　係り受けを加味したかさ増し生成手法

著者らの過去の研究に係り受け構造を加味したかさ増し手法生成がある [8]. 分子化合物の生成モデルを元に係り受け構造を加味して生成するモデルを提案し、ニュースタイトルに対して、ルールベース手法と文章生成手法を用いてそれぞれの類似度を比較した.

結果としてどの手法も類似度の獲得が難しく、提案手法によるかさ増しも分類問題の精度精度向上には課題が残っていた. これは人手で設定した特徴量や係り受け構造を加味した特徴では、十分に学習データの特徴を保持できていない事を示しており、より元のデータの特徴を獲得できるかさ増し手法が必要なことを表している.

2.2　Chataug

Dai らは臨床データに対して ChatGPT を用いてかさ増しを行い、従来手法と比較して分類精度の向上を行った [6]。使用されたデータセットはいずれも短文かつ英語で構成されており、文章全体の文脈等を加味する必要がある記事データは対象としていない。

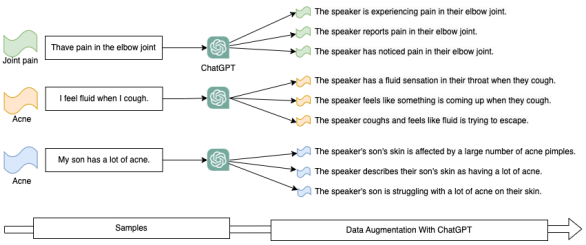

図 1　Chataug

3　提案手法

本研究では、ペルソナを用いた ChatGPT によるデータかさ増し手法を提案する。かさ増し行う際に重要な要素として同じ意味で違う表現の文章を生成することが必要と考えた。そこで、ChatGPT への指示として異なるユーザの特徴として定義したペルソナを加えることで、同じ内容の記事をそれぞれのペルソナに合わせた異なる表現の文章として生成することができると考えた。

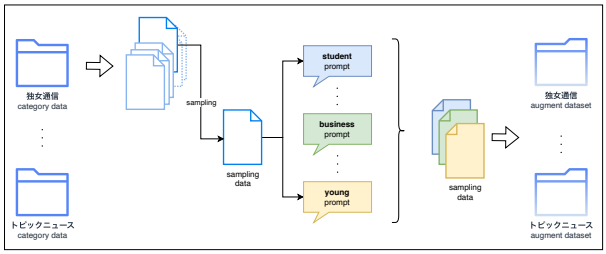

図 2　ChatGPT にペルソナを用いたかさ増し手法

3.1 プロンプトエンジニアリング

ChatGPT などに代表される大規模言語生成モデルは、使用する際に生成についての指示を自然文で行うことで、ある程度ユーザの意図に沿った応答を返すことが知られている [9]。これにより出力の予測が難しい生成モデルを制御し適切な応答が得られる可能性が向上する一方で、不適切なプロンプトを与えた場合は出力の制御が困難になる。一般に適切なプロンプトの予測は難しく経験則を元に作成されることも多いが、今回は ChatGPT4.0 を利用したプロンプトの拡張を行なった。媒体資料等で表現されている読者ターゲットを元にペルソナの雛形を作成し、ChatGPT4.0 を利用しその他のペルソナ候補を生成を指示した。最後に出力された文章に人手を微修正を行なったものを使用している。

3.2 ペルソナ

作成した各ペルソナのプロンプトを表 1 に記す。ペルソナの定義は主に年齢層を中心に行い、専門性の強弱を合わせることで異なる表現が得られるのを期待している。また、本研究のかさ増しで使用されるモデルは「*gpt-3.5-turbo-0301*」を利用している。

ペルソナ	プロンプト
student	高校生程度 10代後半 を想定。社会的な興味関心は低く、身近で話題性の高い記事を好む
engineer	専門的な技術職に従事している想定。時事情報や技術的な記事、家電、IT ガジェットなどを好む
housewife	専業主婦 30代から40代程度 を想定。時間があるため、ファッションや旅行、エンタメやゴシップ記事など揮発性の高いコンテンツを好む
business	20 代から 40 代の企業勤務者を想定。マーケットトレンド、産業ニュース、経済の動向、リーダーシップやキャリア開発に関連する記事を好む。また、ビジネスに影響を及ぼす政治的な出来事についても関心が高い。
senior	60 代以上のリタイアした個人を対象とします。健康とウェルネス、旅行、趣味、芸術と文化、金融計画やリタイアメントに関連する記事に興味がある。
researcher	専門的な知識や研究領域に深い関心があり、その分野に関連する深い洞察や研究成果を探しています。また、研究助成金の動向、学術界のニュース、研究のエチカルな問題についても興味がある。
young	デジタルネイティブとも言えるこの世代は、ソーシャルメディアトレンド、ポップカルチャー、環境問題、社会正義についての記事に興味があります。

表 1 設定したペルソナプロンプト

4 データセット: livedoor news corpus

使用するデータセットは、livedoor news corpu を使用する [10]。このデータセットは NHN Japan 株式会社が運営する「livedoor ニュース」の一部を、可能な限り HTML タグを取り除いて作成したカテゴリ分類に用いられるデータセットであり、それぞれのカテゴリのデータ数は表 2 に示す通り、不均衡なデータセットになっており、記事ごとの文章量にも大きく差がある。また、商品名などそれぞれの記事に固有の語彙が存在するため、テストデータを切り出した際に、学習データ内に存在しない語彙が現れる可能性がある。

そこで、予備実験として、カテゴリ分類を 2 種類のベクトルで行い、データセットとベクトル化手法の相性を調査した。使用したベクトルは文脈を加味する BERT と、単語の頻度情報を重視する TF-IDF とし、文脈と単語情報のどちらを重視するべきか分類精度から判断する。結果として、ほとんどのカテゴリにおいて TF-IDF を使用した方が分類精度が高かった。よってこのデータセットは単語の頻度情報を重視するのが適切と考え、以降のベクトルは TF-IDF を使用する。

category	size	BERT	TFIDF
独女通信	870	82.23%	**85.71%**
IT ライフハック	870	86.79%	**88.11%**
家電チャンネル	864	78.68%	**88.20%**
livedoor HOMME	511	81.44%	**82.41%**
MOVIE ENTER	870	90.19%	**91.00%**
Peachy	842	79.61%	**80.78%**
エスマックス	870	92.07%	**99.50%**
Sports Watch	900	**94.23%**	89.20%
トピックニュース	770	87.62%	**92.53%**

表 2 livedoor news corpus

5 かさ増し実験

カテゴリ毎に 80 件のデータをサンプリングし、7 つのペルソナを使いかさ増しを行うことでかさ増しを行った。その後、100 文字以下の極端に短い文章を削除し、学習データとしてある程度の文章長を持つデータにフィルタリングしている。また、比較としてペルソナを用いないかさ増しプロンプトを作成し、比較によりペルソナによるかさ増しの効果を分析する。学習データ、テストデータ、かさ増しデータ、ペルソナ無しかさ増しデータを散布図として t-SNE を用いて可視化したものを図 3 として示す。

これらの図から、かさ増しデータの傾向を分析すると、どのカテゴリもかさ増しデータがある程度集中して分布していることがわかる。これは、1 つのデータからペルソナの数だけ類似データが生成されることに加えて、ChatGPT の出力パターンに類似性があることが示されている。また、学習データの外周部分に位置していることが多いことから、かさ増しによって元のデータとは類似しつつも異なったデータとして表現されていると考えられる。一方でペルソナを用いないかさ増しデータは元のデータから分離した位置に集中しており、元のデータと類似しない異なったデータであることがわかる。ペルソナの有無が学習データの特徴維持に寄与するか類似度実験、分類実験によって明らかにする。

5.1 類似度計測実験

かさ増しデータが、カテゴリの近傍に位置していることが確認できたため、より定量的な元データとの類似性を cos 類似度、BLUE スコアから評価する。cos 類似度はデータ間の意味の類似性を、BLUE スコアは表現の類似性を評価でき、それぞれの平均値を取り、散布図にしたものを図 4 に示す。

ペルソナを用いたかさ増しデータは、分布が右辺下部に集中していることがわかる。これは cos 類似度が高

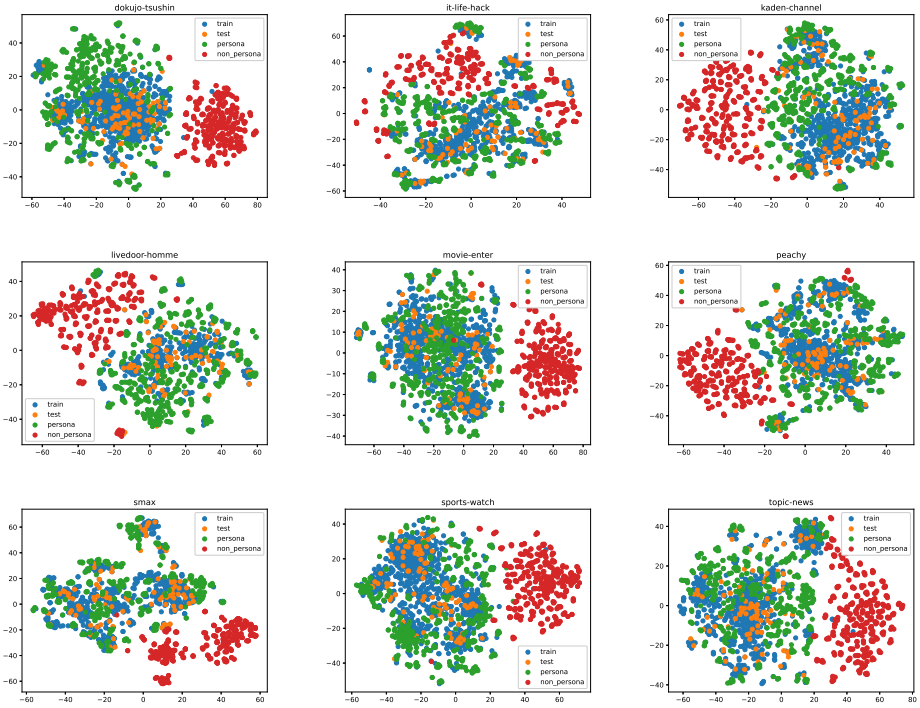

図 3　生成データ散布図

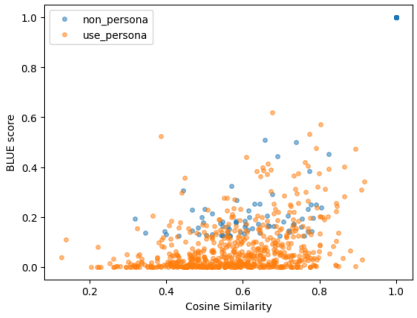

図 4　生成データ類似度実験結果

category	orginal	use persona	non persona
独女通信	80.21%	62.21%	13.23%
IT ライフハック	88.38%	67.38%	23.05%
家電チャンネル	80.45%	52.62%	13.98%
livedoor HOMME	79.54%	59.54%	09.71%
MOVIE ENTER	92.13%	79.13%	05.62%
Peachy	79.46%	55.46%	10.97%
エスマックス	94.20%	73.20%	10.97%
Sports Watch	92.90%	79.74%	03.78%
トピックニュース	93.50%	85.50%	06.58%

表 3　生成データセット精度評価

く、BLUE スコアが低いことを示しており、ベクトルで評価した類似度は高いが、単語レベルの類似性が低いことがわかる。今回のかさ増し手法は、同じ意味で違う表現のデータを獲得することが望ましいため、ある程度意図通りのかさ増しが行えていることがわかる。一方でペルソナを用いないかさ増しデータは、比較的 cos 類似度が低く BLUE スコアが高いところに集中していることから、ペルソナにより意図通りのかさ増しを行えていると考えられる。

5.2　生成分類実験

かさ増しデータがどの程度元のデータの特徴を保持しているか、分類実験により計測した。学習データからかさ増し元となったデータを除き、多層パーセプトロン (MLP) を用いて学習したモデルを使用し、推定結果が元データのカテゴリと一致していれば正解とした。元データと同じカテゴリ分類されていれば、かさ増し元のデータの特徴を保持していると判断できる。かさ増し元データ、ペルソナを使用したかさ増し、ペルソナを使用しないかさ増し 3 種類のデータを用いて精度を評価した。

まずかさ増し元データの分類精度に注目すると、カテゴリ毎に分類精度に差があることがわかる。これはカテゴリ毎の多様性を表しておりかさ増し元データの分類精度が高いものほど、共通した表現や要素が含まれていると考えられる。次にペルソナを使用したかさ増しの分類精度を見ると、ある程度は元のデータの特徴を保持しているものの、分類に必要な重要な特徴を再現できていない可能性が示唆されている。ペルソナを使用していないかさ増しでは、必要な特徴をほとんど保持できておらず著しく低い精度になっていることがわかる。そこで、どのような誤分類が発生しているか確認するため、混同行列を用いてそれぞれの比較を行った。

ペルソナを使用したかさ増しデータの混同行列に注目すると、peachy カテゴリへの誤分類により精度が低下していることがわかる。このカテゴリは含まれるジャンルが広く、10〜20 代向けの特徴を持つカテゴリだが、かさ増しデータに共通としてこのカテゴリの特徴が含まれてしまっている可能性がある。元データはでは peachy カテゴリとしての表現を加味して分類しているが、かさ増しによってこれらのカテゴリの特徴となっている文体等が弱まってしまったことが考えられる。

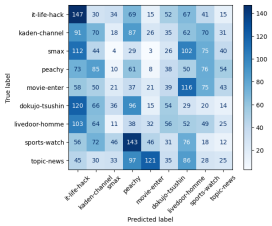

図 5　ペルソナ無し
かさ増しデータ分類結果

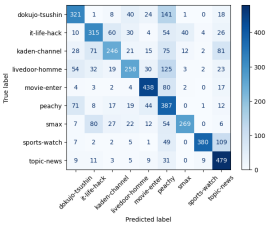

図 6　ペルソナ有り
かさ増しデータ分類結果

ペルソナを使用していないかさ増しでは学習データが全体に分散しており、ランダムに近い形のデータになっていることが確認できる。このことから、ペルソナによって多様性だけなく分類に有効な特徴も保持されていることが示唆された。

6　分類実験

これらの実験からペルソナを使用したかさ増しではデータセットの特徴をある程度獲得したかさ増しが行えていると判断したため、かさ増しデータが分類精度にどの程度寄与するか確認する。かさ増し手法には元のデータが少ない状態でも精度の低下を防ぐことが求められるため、学習データとかさ増しデータの割合を変化させながら分類精度の推移を観察する。データセットは少量カテゴリに合わせた 400 件を上限とし、かさ増し手法では削減した分のデータ量を合計 400 件になるように補填を行い学習データとする (図 7)。分類には MLP を用い、パラメータは表 4 のものを使用し、学習データを 10 件ずつ増加させ、かさ増しデータを 10 件ずつ減少させた結果を図 8 に示す。

6.1　実験設定

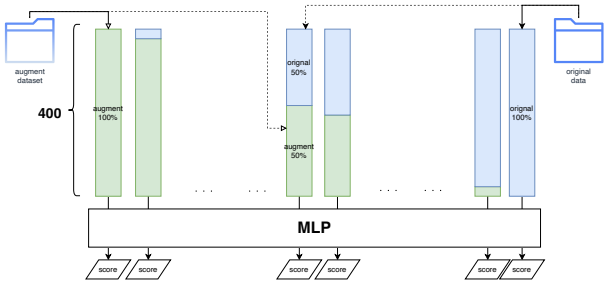

図 7　分類実験手順

parameter	value
hidden layer sizes	200, 150, 50
solver	adam
alpha	0.0001
max iter	300

表 4　ハイパーパラメータ

6.2　実験結果

実験の結果から、ペルソナを使用することで学習データが少ない状態でもある程度分類精度を保持出来ることがわかる。これまでの実験で確認した、元のデータの特徴を保持していることが示された。ペルソナを使用しないかさ増しでは、学習データ少ない場合にモデルの性能を悪化させていることがわかり、プロンプトによる出力結果の制御が重要であることがわかる。

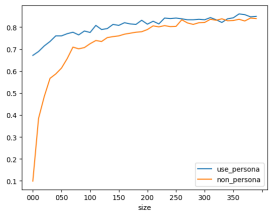

図 8　かさ増し量と
精度推移

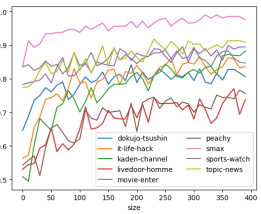

図 9　カテゴリ毎の
分類精度

カテゴリ毎にみると、かさ増しの効果が弱いカテゴリが存在することがわかる。このことから、カテゴリ毎にプロンプトを設計し適用するような手法が効果的な可能性がある。

7　考察・まとめ

ChatGPT に対してペルソナを用いたかさ増しにより、学習データの不均衡さの解消を目指した。ペルソナを使用したかさ増しデータはある程度、元のデータとの類似性を獲得できることが確認できた。これは ChatGPT などの大規模文章生成モデルに対しては、プロンプトによる出力の制御が重要なことが確認できた一方で、出力の予測が困難なことも表している。

今回のかさ増し手法では分類精度の低下を軽減することは出来たが、向上させるまでは至らなかったため、カテゴリ分類に必要な特徴を得るためのプロンプトを新しく作成することで、分類精度を向上させるかさ増しが行える可能性がある。例えば、文章の抽象度を制御し語彙を広げるような手法を用いることで、学習データには不足している語彙を補うといった手法が考えられる。さらに、マルチモーダルな生成モデルを活用した大規模モデルの研究も進められているため、より抽象的なかさ増し手法も検討する。

参考文献

[1] Törnberg, Petter. ”Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.” arXiv preprint arXiv:2304.06588 (2023).
[2] He, Haibo, and Edwardo A. Garcia. ”Learning from imbalanced data.” IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.
[3] Bayer, Markus, Marc-André Kaufhold, and Christian Reuter. ”A survey on data augmentation for text classification.” ACM Computing Surveys 55.7 (2022): 1-39.
[4] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).
[5] ChaaGPT.https://openai.com/blog/chatgpt/
[6] Dai, Haixing, et al. ”Chataug: Leveraging chatgpt for text data augmentation.” arXiv preprint arXiv:2302.13007 (2023).
[7] Lai, Viet Dac, et al. ”Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.” arXiv preprint arXiv:2304.05613 (2023).
[8] SAWASAKI, Natsuki, et al. ”Sentence Generation Method by Extension of MolGAN Using Sentence Graph.” Journal of Japan Society for Fuzzy Theory and Intelligent Informatics 32.2 (2020): 668-677.
[9] White, Jules, et al. ”A prompt pattern catalog to enhance prompt engineering with chatgpt.” arXiv preprint arXiv:2302.11382 (2023).
[10] livedoor ニュースコーパス http://www.rondhuit.com/download.html#ldcc