ChatGPT による日本語常識道徳データセットの拡張

大橋 巧 [1]　中川 翼 [2]　彌冨 仁 [1,2]
[1] 法政大学理工学部応用情報工学科　[2] 法政大学理工学研究科応用情報工学専攻
{takumi.ohashi.4g, tsubasa.nakagawa.5p}@stu.hosei.ac.jp　iyatomi@hosei.ac.jp

概要

近年の人工知能の発展や社会への参入により，人工知能に人間の一般的な道徳観を持たせることが望まれるが，現在公開されている学習させるための日本語常識道徳データセットではカバーしている事例が少ない．本研究では ChatGPT による事例のバリエーションを補うデータ拡張手法を提案する．既存の日本語常識道徳データセットを拡張したデータセットでモデルを学習・評価することで提案手法による性能の向上を確認するとともに，GPT-4 Turbo による推定結果と比較することで，拡張データセットで学習したモデルは日本特有の文化や道徳に対する理解が求められる事例に対して，より効果的な推定が可能であることが示唆された．

1　はじめに

近年の人工知能の発展や社会への参入に伴い，人工知能の倫理を巡る議論は活発化している．人工知能をよりよく活用するためには，人間に近い価値観を持つことが必要であり，どのようにして人工知能に倫理を教えるか議論が行われている [1, 2]．その中で現在，人工知能に一般的な道徳観を学習させるためのデータセットが構築されている [3, 4, 5]．

竹下ら [5] は日本語では唯一の常識道徳データセット JCommonsenseMorality（JCM）を構築し，公開した．道徳的判断に関するデータセットはいくつか存在するがそのほとんどは英語圏のデータセットであり，文化差や言語の違いを考えれば日本語の常識道徳データセットは希少である．彼らはこのデータセットを用いて評価実験を行い，Hendrycs ら [3] が構築した英語の常識道徳データセットを用いた実験結果と比較した上で，日本語における道徳的判断には日本特有の文化や慣習も考慮した，より複雑な道徳理解が求められることを示唆した．だが，現状のデータセットでは事例のバリエーションが少ないため，学習データの事例を増やし語彙を補うことが望まれる．

データのバリエーションを補うデータ拡張（Data Augmentation）は学習データの量や質に性能が依存する機械学習分野において効果的であり，自然言語処理（Natural Language Processing; NLP）の分野でも行われている [6, 7]．例えば，Easy Data Augmentation（EDA）は同義語を置換したり，単語を交換，挿入，削除したりなどの単純な操作を行い，効果を示している [7]．また，OpenAI が 2022 年 11 月に公開した対話型の大規模言語モデル（Large Language Models; LLM）である ChatGPT を用いた手法も行われている [8, 9, 10]．Dai ら [10] は Few-shot のテキスト分類タスクのためのデータ拡張手法として，ChatGPT によるデータ拡張を提案し，Amazon レビューの分類や医学領域の NLP タスクといった複数のタスクで既存のデータ拡張手法より高い効果を示した．既存のデータ拡張手法は生成されるデータの精度や多様性に限界があったが，ChatGPT は豊富な知識を持っているかつ，学習時に人間のフィードバックによる強化学習（Reinforcement Learning from Human Feedback; RLHF）[11] を行っていることから，より有益で多様なデータを生成することができる．

本研究では ChatGPT を用いて日本語の常識道徳データセット JCM の拡張を行い，事例のバリエーションを補う手法を提案し，拡張したデータセットで BERT [12] や RoBERTa [13] といった事前学習モデルをファインチューニングして道徳理解に関する性能を評価した．また， JCM を拡張したデータセットを用いて学習した最良のモデルと，主に英語のデータを多く学習している現時点で最先端の LLM である GPT-4 Turbo による道徳的判断の推定結果と比較し，日本特有の文化や慣習を含む，より複雑な道徳理解が求められる事例に対する推定能について検証した．

表 1　JCommonsenseMorality（JCM）の例文

文章	ラベル
赤ちゃんにお酒を飲ませる	1
赤ちゃんに薬を飲ませる	0

2　提案方法

2.1　データセット

本研究では，日本語では現時点で唯一の常識道徳データセット JCommonsenseMorality（JCM）[1)][5] を用いた．データセットに含まれている文章とラベルの例を表 1 に示す．文章は文章の一部，つまり状況や行為が変わることにより道徳的評価が変化する 2 文 1 組で構成されており，道徳的に許容できる（1）か否（0）かの 2 種類のラベルが付与されている．データ数は学習データが 13,975 文，検証用データが 1,996 文，テスト用データが 3,992 文の計 19,963 文である．

このデータセットは量が限られ，一部の単語が変化した 2 文 1 組の構成であるので，目的を実現する学習用データとして不足している．例えば，表 1 の文章の一部を変化させた，「赤ちゃんにコーヒーを飲ませる」や「赤ちゃんにガムを飲ませる」のような場合は適切に分類できるかは不明である．我々が身に着けているような知識や経験を反映するような事例を網羅することは難しいが，それぞれのペアに対応したパターン増やすことで，以前より幅広い内容を含むデータセットの構築が期待できる．

2.2　拡張手法

本研究では JCM に対して，ChatGPT を用いたデータセットの拡張方法を提案する．拡張に用いる ChatGPT のモデルは，GPT-3.5 Turbo[2)]（2023 年 11 月 6 日時点のモデル）である．拡張方法のフレームワークを図 1 に示す．

まず，データセットから文章の一部が変わることにより道徳的評価が変化するペアを取り出す．このペアの事例のバリエーションを増やすため，ペアの一致している部分を抽出する．それぞれの文章に対して，日本語 NLP ライブラリである GiNZA[14] で単語分割して前方から一致する部分と後方から一致

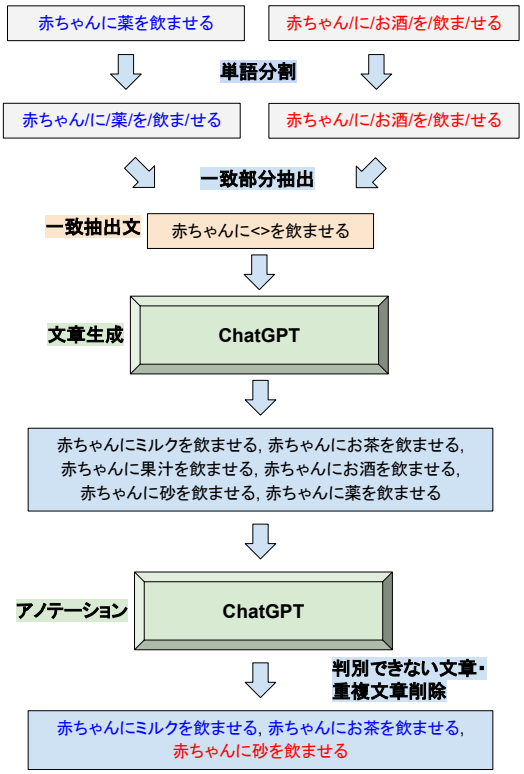


図 1　拡張手法のフレームワーク

する部分を抽出し，その間に <> を挿入した文を作成した．以下この文を一致抽出文と呼ぶ．ただし，一致抽出文が <> を含めて 6 文字以下の場合，適切に検出できていない可能性が高いため，そのペアでは一致抽出文を生成しない．

次に ChatGPT にこの一致抽出文に類似した文章で道徳的に許容できる文章，道徳的に許容できない文章が含まれる 6 文を生成させた．ここで生成された 6 つの文章に対して ChatGPT にこの文章が道徳的に「許容できる」「許容できない」，または「判別できない」か判定を行わせ，その結果をもとにアノテーションを行った．このとき，「判別できない」というラベルを加えたのは，文脈がおかしい文章や道徳的判断しにくい文章を除外するためである．

「判別できない」というラベルが付けられた文章や元データと重複する文章，生成した文章同士で重複した文章を除外し，必要のないデータを取り除いた．最後に偏りが発生しないように，1 つの組から生成された文章の中で道徳的に許容できる，できない文章それぞれ 3 つを上限に拡張データセットに採用した．

この手順に従い，各組で文章の生成とアノテー

1）https://github.com/Language-Media-Lab/commonsense-moral-ja
2）https://platform.openai.com/docs/models/gpt-3-5

表 2　拡張前と拡張後のデータ数

	許容できる (0)	許容できない (1)	計
元	7,515	6,460	13,975
拡張後	19,535 (+12,020)	11,649 (+5,189)	31,184 (+17,209)

ションを行い，データセットを拡張した．拡張後のデータ数は表 2 に示すように，拡張前の約 2.2 倍となった．

3　実験

3.1　データ拡張の評価

我々は，拡張前と拡張後のデータセットでそれぞれ事前学習モデルの BERT[12] と RoBERTa[13] をファインチューニングし，JCM のテストデータに対して道徳的に許容できるか否かの評価を行い，拡張の効果を確認した．

事前学習モデルには，日本語版 Wikipedia で事前学習した BERT[3] と 日本語版 Wikipedia，日本語版 CC-100 で事前学習した RoBERTa[4] を用いた．損失関数に交差エントロピー誤差，最適化手法に AdamW[15] を使用し，エポック数は 20 エポックの early-stopping を適用した．学習率は BERT $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 4 \times 10^{-5}, 5 \times 10^{-5}\}$，RoBERTa $\{1 \times 10^{-6}, 2 \times 10^{-6}, 3 \times 10^{-6}, 4 \times 10^{-6}, 5 \times 10^{-6}\}$，バッチサイズはどちらのモデルも {8，16，24，32} を使用した．検証用データでパラメーターチューニングを行い，F1 スコアで最も高い精度を示したハイパーパラメータでテスト用データを評価した．このときシード値を変えて 5 回ずつ行い，それぞれのスコアの平均を求めた．また，比較のため ChatGPT（GPT-3.5 Turbo）と GPT-4 Turbo[5]（どちらも 2023 年 11 月 6 日時点のモデル）で同様に推定を 1 回ずつ行い，評価した．

3.2　日本特有の文化を含む文章に対する評価

本実験で用いる JCM には日本語特有の表現や日本独自の文化を含む文章がいくつか見られるため，提案手法で JCM を拡張したデータセットを学習したモデルと主に英語のデータを多く学習している GPT-4 Turbo では推定結果に違いが出ると考えられる．そこで我々は，道徳ラベルの推定結果からそれぞれのモデルが日本特有の文化や慣習を含む文章に対して正しく推定できているのか分析した．拡張データセットでファインチューニングした RoBERTa と GPT-4 Turbo による推定結果から，片方が誤りもう一方が正解する文章をそれぞれ抽出した．RoBERTa は実験で得たシード値を変えた 5 回の結果を用いた．抽出した文章に対して，以下の 2 つの手法で日本特有の文章が含まれている割合をそれぞれ求めた．

日本文化いろは事典による手法　日本文化に関するキーワードを取り上げ，基本事項を紹介している日本文化いろは事典[6] にある 253 個のキーワードのいずれかが含まれている文章の割合を求めた．

人手評価による手法　10 名の日本人を対象に道徳判別の推定結果を伏せた形で日本特有の単語や言い回しが含まれているか判断をするブラインドテストを行い，その割合を求めた．テストに用いるデータは，RoBERTa と GPT-4 Turbo の片方が誤り，もう一方が正解する文章に対して 50 文ずつランダムサンプリングを行い，RoBERTa の 1 回の結果ごとに 2 セットずつ，合計 10 セット作成し，1 名あたり 1 セット評価をした．

表 3　各モデルの分類性能

	Accuracy	Precision	Recall	F1
BERT（元データセット）	0.789	0.782	0.762	0.771
BERT（拡張データセット）	0.788	0.763	0.797	**0.778**
RoBERTa（元データセット）	0.848	0.849	0.822	0.835
RoBERTa（拡張データセット）	0.864	0.853	0.858	**0.855**
ChatGPT (GPT-3.5 Turbo)	0.838	0.778	0.916	0.841
GPT-4 Turbo	**0.938**	**0.936**	**0.931**	**0.934**

4　結果と考察

4.1　データ拡張の評価結果

道徳的に許容できるか否かを推定する各モデルの性能を表 3 に示す．BERT，RoBERTa のどちらでも拡張データセットでファインチューニングした方が，元データセットでファインチューニングした場合より高い道徳判別能を示した．また，拡張データセットでファインチューニングした RoBERTa では，

3）https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
4）https://huggingface.co/nlp-waseda/roberta-large-japanese-with-auto-jumanpp
5）https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
6）http://iroha-japan.net/

表 4　『GPT-4 Turbo が推定を誤り, RoBERTa が正解した文章』の例文

文章	正解ラベル
喪中の間に七五三をする	0
正月の御飾を 12/30 にする	0
手水舎を利用しない	1
野球の試合に負けたので、砂を持ち帰る	0

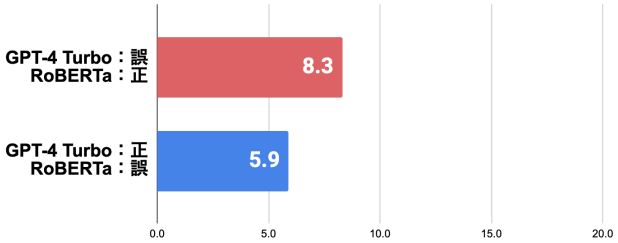
ChatGPT による推定結果よりスコアが上回った．

ChatGPT で類似文章を複数生成したことで元データセットより事例のバリエーションが広がり，拡張したラベルを学習することで ChatGPT が持つ知識範囲をカバーすることができたため，性能向上に繋がったと考えられる．しかし，GPT-4 Turbo による推定結果を上回ることはできなかった．

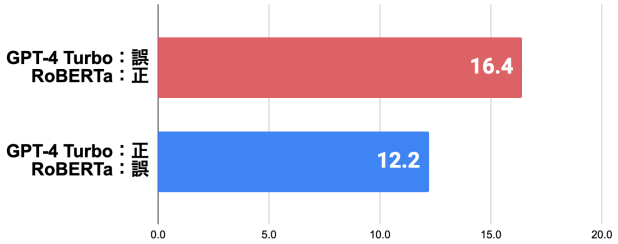
4.2　日本特有の文化を含む文章に対する評価結果

『GPT-4 Turbo が推定を誤り，RoBERTa が正解した文章』の中で日本特有の文化を含むと判断した例文と対応する正解ラベル（道徳的に許容できる：0，許容できない：1）を表 4 に示す．上 2 文は日本文化いろは事典による手法で判断した例文であり，事典に登録されている「七五三」,「正月」という単語がそれぞれ含まれている．下 2 文は人手評価による手法で判断した例文であり，日本文化いろは事典に登録されている単語は含まれていないが，評価を行った全員が日本特有の文化を含んでいると評価した．「野球の試合に負けたので、砂を持ち帰る」という例文では日本特有の高校野球の文化を示しており，この文章を正しく判別するには日本特有の文化や道徳に対する理解が必要であると言える．

RoBERTa と GPT-4 Turbo の片方が誤り，もう一方が正解する文章に対する日本特有の表現の割合についての評価を図 2 に示す．『GPT-4 Turbo が推定を誤り，RoBERTa が正解した文章』の方が日本文化いろは事典を用いた手法では 2.4%，人手評価による手法では 4.2%ほど，日本特有の文化や慣習を含む文章の割合が高いことが確認できた．GPT-4 Turbo が推定を誤る文章に日本特有の文化や慣習を含む割合が高い理由として，GPT-4 Turbo の持つバイアスが関係していると考えられる．GPT-4 Turbo は学習時にインターネットからの膨大な量のデータで学習をするとともに RLHF を行っている [11]．この学習データの偏りや人間のアノテーターの価値観や社会文化に影響されることにより，特定の文化や社会，言語



(a) 日本文化いろは事典による手法



(b) 人手評価による手法

図 2　日本特有の表現が含まれる文章の割合の比較

に偏っている可能性がある [16, 17]．一方で，提案手法で拡張したデータセットでファインチューニングしたモデルは，JCM に含まれている日本語特有の表現や日本独自の文化も学習データに含まれているため，GPT-4 Turbo でも推定が難しい日本特有のより複雑な道徳理解が求められる事例に対して，日本特有の文化や道徳を考慮した推定が可能になっていると考えられる．

5　おわりに

我々が提案した拡張手法により，既存の常識道徳データセットを ChatGPT で事例のバリエーションを補ったデータセットに拡張することができた．拡張データを用いてファインチューニングした BERT や RoBERTa は, 拡張前のデータを用いるより高い性能であることを確認したとともに，RoBERTa は ChatGPT を超える性能を達成した．また，この RoBERTa と GPT-4 Turbo による推定結果の違いを分析すると，拡張データセットで学習したモデルは日本特有の文化や道徳に対する理解が求められる事例に対して，より効果的な推定が可能であることが示唆された．

今後は日本語以外の他の言語や文化圏の道徳に関するデータセットに対して，同様に ChatGPT を用いる拡張を行い効果を確かめるとともに，GPT-4 Turbo による推定結果との違いにどのような傾向が見られるか調査を行う．

参考文献

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The Moral Machine Experiment. **Nature**, Vol. 563, No. 7729, pp. 59–64, 2018.

[2] Liwei Jiang, Chandra Bhagavatula, Jenny T Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Regina A. Rini, and Yejin Choi. Can Machines Learn Morality? The Delphi Experiment. **arXiv preprint arXiv:2110.07574**, 2021.

[3] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with Shared Human Values. **Proceedings of the International Conference on Learning Representations (ICLR)**, 2021.

[4] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-life Anecdotes. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 13470–13479, 2021.

[5] 竹下昌志, ジェプカラファウ, 荒木健治. JCommonsenseMorality: 常識道徳の理解度評価用日本語データセット. 言語処理学会第 29 回年次大会, pp. 357–362, 2023. in Japanese.

[6] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. A Survey of Data Augmentation Approaches for NLP. In **Findings**, 2021.

[7] Jason Wei and Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In **Conference on Empirical Methods in Natural Language Processing**, 2019.

[8] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. LLM-powered Data Augmentation for Enhanced Crosslingual Performance. **arXiv preprint arXiv:2305.14288**, 2023.

[9] Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT. **arXiv preprint arXiv:2304.14334**, 2023.

[10] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. ChatAug: Leveraging ChatGPT for Text Data Augmentation. **arXiv preprint arXiv:2302.13007**, 2023.

[11] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training Language Models to Follow Instructions with Human Feedback. **arXiv preprint arXiv:2203.02155**, 2022.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **North American Chapter of the Association for Computational Linguistics**, 2019.

[13] 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. **arXiv preprint arXiv:1907.11692**, 2019.

[14] 松田寛. GiNZA-Universal Dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.

[15] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. **arXiv preprint arXiv:1711.05101**, 2017.

[16] Han Rao. Ethical and Legal Considerations behind the Prevalence of ChatGPT: Risks and Regulations. **Frontiers in Computing and Intelligent Systems**, Vol. 4, No. 1, pp. 23–29, 2023.

[17] Partha Pratim Ray. ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope. **Internet of Things and Cyber-Physical Systems**, 2023.