大規模言語モデルが生成した記事の自動検出

西田 祥人[†] 濱本 惇之介[†] 松浦 哉太[†] 杉原 壮一郎[‡] 梶原 智之[‡] 二宮 崇[‡] 後藤 功雄[‡]
愛媛大学工学部工学科[†] 愛媛大学大学院理工学研究科[‡]

1. はじめに

近年、大規模言語モデル (LLM: Large Language Model) の発展により、自動生成された文章の流暢性は著しく向上しており、人間が書いた文章と見分けることが難しいほどの高い品質を持つようになってきている。そのため、記事作成や議事録作成など、今まで人間が取り組んできた多くの文章作成タスクにおいて、LLM の活用が期待されている。

一方で、LLM によって自動生成された文章には、誤情報やバイアスを含む可能性があり、信憑性が保証されないという問題がある。近年では SNS の普及によって情報へのアクセスが容易になっているため、このような誤情報やバイアスを含む情報が広範囲に拡散し、社会全体へ悪影響を及ぼす可能性が高まっている。特に報道分野では、LLM が生成した誤情報を含む記事がユーザーの不利益につながるリスクがあり、またその参照元として表示される報道機関の信用の低下を招く可能性があるため、対策が求められている。

この問題を解決するために、LLM で生成された文章を検出する研究が英語を中心に進められている [1-2]。しかし、日本語の新聞記事においてはこのような研究に十分に組み込まれていない。

本研究では、このような課題に対処するため、人間が書いた新聞記事と LLM が生成した記事のデータセットを作成し、LLM が生成した記事を自動的に検出する手法を提案する。実際の新聞記事と、3 種類の LLM から生成された文章を用いて BERT [3] と RoBERTa [4] を訓練し、人間が書いた記事か、LLM が生成した記事かを自動分類するモデルを作成した。実験により、作成したデータセットで訓練した分類器は人手による分類よりも高い精度で分類できることを示した。

表 1 データセットの構成 (記事数)

	人間が作成	LLM が生成
訓練	500	500
検証	150	150
評価	150	150

2. 提案手法

2.1 データセット

本研究では、人間が書いた記事として、愛媛新聞の記事を使用する。このとき、512 トークンより長い記事を除外し、800 件の記事を選択した。これをもとに、ChatGPT (gpt-4o)、Qwen2.5 (Qwen2.5-72B-Instruct)¹、Swallow (Llama-3.1-Swallow-70B-Instruct-v0.1)² の 3 つのモデルを用いて、LLM で生成した記事を用意した。ChatGPT では、人間が書いた新聞記事の見出しを与え、その見出しに沿った内容を出力するようにプロンプトを用いて指示することで記事を生成した。Qwen2.5 および Swallow では、人間が書いた新聞記事とその見出しを例として与え、few-shot 学習によって記事を生成した。人間が書いた新聞記事と LLM が生成した記事で同じ見出しを持つ、合計 1,600 件の記事からなるデータセットを作成し、表 1 のように訓練データ 1000 件、検証データ 300 件、評価データ 300 件にランダムに分割した。

2.2 分類器

分類器には、東北大学が提供する BERT³ と京都大学が提供する RoBERTa⁴ の 2 つの事前学習済みモデルを用いた。最適化手法に Adam を使用し、訓練データのバッチサイズを 8、学習率を 1e-5 として訓練した。検証用データにおけるクロスエントロピー損失が 3 回改善されなくなったときに訓練を停止した。TITAN RTX の 140W の GPU を使用し、それぞれのモデルに対して 3 種類の

Automatic Detection of Articles Generated by Large Language Models

Shoto Nishida[†] (nishida@ai.cs.ehime-u.ac.jp)

Junnosuke Hamamoto[†] (hamamoto@ai.cs.ehime-u.ac.jp)

Kanata Matsuura[†] (j520396b@mails.cc.ehime-u.ac.jp)

Soichiro Sugihara[‡] (Sugihara@ai.cs.ehime-u.ac.jp)

Tomoyuki Kajiwara[‡] (kajiwra@cs.ehime-u.ac.jp)

Takashi Ninomiya[‡] (ninomiya.takashi.mk@ehime-u.ac.jp)

Isao Goto[‡] (goto.isao.fn@ehime-u.ac.jp)

Ehime University

¹ <https://huggingface.co/Qwen/Qwen2.5-72B-Instruct>

² <https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.1>

³ <https://huggingface.co/tohoku-nlp/bert-base-japanese-char-v3>

⁴ <https://huggingface.co/ku-nlp/roberta-base-japanese-char-wwm>

表2 BERT の分類結果 (正解率%)

評価データ	訓練データ		
	ChatGPT	Qwen2.5	Swallow
ChatGPT	100.0	92.0	78.7
Qwen2.5	87.7	99.3	96.7
Swallow	66.7	96.3	99.7

表3 RoBERTa の分類結果 (正解率%)

評価データ	訓練データ		
	ChatGPT	Qwen2.5	Swallow
ChatGPT	100.0	99.7	89.7
Qwen2.5	89.3	99.7	97.3
Swallow	75.7	95.0	100.0

LLM で作成したデータセットでそれぞれファインチューニングを行い、6通りのモデルを作成した。

2.3 評価

ファインチューニング後のモデルに対し、3つの LLM で作成した評価データを入力し、分類結果の正解率を求めた。また、各 LLM の評価データから 50 件抽出して人手で分類し、正解率を求め、モデルによる分類結果と比較した。

3. 実験結果

BERT での分類結果の正解率を表 2 に示す。各列はファインチューニングに用いられたデータが ChatGPT, Qwen2.5, Swallow のうちの LLM で作成されたものであるかを表しており、各行は評価データがどの LLM で作成されたものであるかを表している。同様に、RoBERTa での分類結果の正解率を表 3 に示す。

表 2 および表 3 から、多くの組み合わせで高い正解率を達成していることが示され、訓練された言語モデルは LLM が生成した記事か人間が作成した記事かをほとんど見分けられることがわかった。訓練データと評価データの作成に用いられた LLM が同じ場合、99%以上の高い精度で分類できており、訓練データと評価データの作成に用いられた LLM が異なる場合でも多くの組み合わせで高い正解率を達成している。特に、Qwen2.5 で生成されたデータを用いて訓練された BERT および RoBERTa は、異なる LLM から生成された評価データに対しても 90%以上の精度で人間が作成した記事か LLM が生成した記事かを分類できていることがわかる。

3 種類の評価データを人手で分類したときの正解率を表 4 に示す。どの評価データに対しても分類器を用いた正解率は人手評価の正解率よりも高いことが示された。

また、提案手法を用いて自動分類する場合は 100 件あたり数十秒で分類できるが、人手で分類する場合には 100 件あたり 1~2 時間程度の時間を要する。このことから、新聞記事が人間によって作成された記事か LLM が生成した記事かを

表4 人手による分類結果

評価データ	正解率
ChatGPT	52%
Qwen2.5	30%
Swallow	62%

分類する場合は、人手で分類するよりもモデルを使用して分類する方が精度とコストの面の両方で優れていることが分かった。

4. おわりに

本研究では、LLM が生成した記事を分類するためのデータセットとモデルを構築し、評価データに対して人手による分類よりも高い精度を示した。今後は、一部が LLM によって生成された記事の検出にも取り組みたい。

謝辞

株式会社愛媛新聞社からデータを提供いただきました。

参考文献

- [1] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection. In Proc. of SemEval, 2024.
- [2] Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi and Paolo Rosso. Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains. arXiv: 2309.11285, 2023.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of NAACL, 2019.
- [4] 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: 1907.11692, 2019.