E-001

大規模言語モデル生成文の識別と特徴量の分析

Analyzing Feature Representations for Discrimination of LLM-Generated Text

若狭 春輝[1)]　神野 健哉[1)]
Haruki Wakasa　Kenya Jin'no

概要

本研究では，人間が執筆した文章と，大規模言語モデル (LLM) によって機械生成された文章を，BERT に代表されるエンコーダーモデルを用いて分類する際，文章の識別に用いられる特徴量が LLM 特有の文章生成確率に依拠しているのか，あるいはその他の文章的特徴に基づいているのかを検討した．生成方法を制御した文章に対して分類を行い，エンコーダーモデルから抽出される特徴量に与える影響を分析した．その結果，分類は生成確率の違いによらず高精度に行えることを確認し，特徴量が潜在的な文章特徴を捉えていることが示唆された．本研究は，機械生成された文章の識別における LLM 特徴空間に内在する情報の理解に貢献する．

1　はじめに

近年，GPT-3[1] などに代表される大規模言語モデル (LLM) の発展により，自然な文章を生成することが可能になっている．この技術は文章生成の効率化や対話システムの精度向上など，さまざまな応用も期待される一方で，フェイクニュースの拡散や教育現場での不正利用といった社会問題も引き起こす可能性がある．こうした背景のもと，人間が書いた文章か機械生成された文章かを人間が識別することはできない [2] と言われているため，自動的に識別する方法が注目されている．しかし，多くの既存手法は，生成モデルのトークン生成確率に依存しており，未知のモデルや内部情報にアクセスできない場合は性能が低下しやすいという課題がある [3].

本研究では，事前学習済みの RoBERTa[4] から抽出される CLS トークンを用い，線形分類器によって，人間文と LLM 文を判別する手法を提案する．提案手法は，生成モデルの内部構造や生成確率に一切依存しないため，異なる生成モデル，生成条件に対して頑健であることを示す．また，異なるドメインに対しての汎用性を検証する．さらに，RoBERTa の CLS ベクトルに主成分分析 (PCA) を適用することで，人間文と LLM 文の RoBERTa 特徴量の分離傾向を可視化し，線形分類器の有効性を示す．

1)　東京都市大学 総合理工学研究科 情報専攻 Informatics, Graduate School of Integrative Science and Engineering, Tokyo City University

2　文章の識別手法

本研究では，事前学習済みの RoBERTa から抽出される CLS トークンを使用し，線形分類器で人間文と LLM 文を判別する手法を提案する．図 1 に示すように，入力文章を事前学習済みの RoBERTa に入力し，得られた出力から CLS トークンに対応する 768 次元のベクトルを抽出する．分類器は線形 SVM を使用し，2 クラス分類を行う．RoBERTa 自体のパラメータは学習時には更新せず，線形分類器のみの学習を行う．本手法の大きな特徴は，生成モデルや文章の生成確率に依存しない点にある．生成モデルにアクセスできない場合や，未知の LLM によって生成された文章であっても，文章があれば識別可能であり，汎用性の高い手法である．

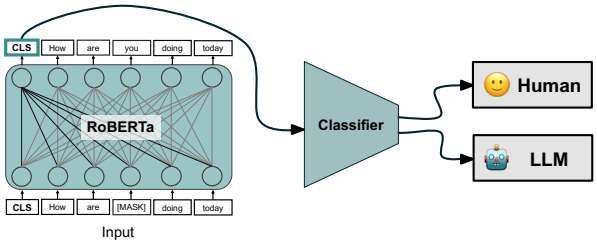

図 1　RoBERTa と線形分類器による識別モデル

3　実験

異なるドメイン，生成モデル，文の生成確率において RoBERTa と線形 SVM を用いた識別手法の有効性を評価するために実験を行った．人間文のデータセットとしてニュース記事 xlsum[7] と物語文 WritingPrompts(WP)[8] の 2 種類を使用する．LLM 文は見出しまたはタイトルをもとに Gemma[5] または Llama3[6] で生成した．生成時の温度パラメータ (temperature) はデフォルトの値として一般的に使われる 0.8 を基準とした．この温度パラメータはトークン生成における確率分布の滑らかさを調整するものであり，文の生成確率 [9] に大きく影響を与える．

さらに，ドメイン，生成モデル，及び文章の生成確率に対する頑健性を評価するために，次の 3 つの実験条件を設定した．1 つ目はあるデータセットで学習し，異なるデータセットで評価を行うクロスドメイン設定である．2 つ目は，ある生成モデルで学習し，別の生成モデルで生成された文を識別するクロスモデル設定である．3 つ目は，評価に使用する LLM 文生成時の温度パ

ラメータの値を高く設定し，通常の温度パラメータで生成された文で訓練した分類器で識別する設定である．表 1 に，通常の温度パラメータで生成した文を訓練および検証に使用した際の，クロスドメインおよびクロスモデルの分類精度を示す．同じデータセット，生成モデルの場合には 0.99 以上の非常に高い精度が得られた．また，異なるドメインや生成モデルであっても 0.89 から 0.98 程度の高い精度を維持しており，汎用性，頑健性に優れていることを確認した．

表 2 に通常の温度パラメータで生成した文を訓練に用い，温度パラメータを高く設定して生成した文を検証に用いた際のクロスドメイン，クロスモデルの分類精度を示す．生成確率を意図的に変化させた文においても通常の文と同等程度に分類可能であることを確認した．

表 1　分類精度 データセット-生成モデル (G: Gemma, L: LLaMA3)

train \ eval	xlsum-G	xlsum-L	WP-G	WP-L
xlsum-G	0.998	0.985	0.897	0.917
xlsum-L	0.995	0.992	0.965	0.948
WP-G	0.892	0.892	1.000	0.960
WP-L	0.973	0.980	0.995	0.998

表 2　分類精度 (検証データ: temperature=1.4)
データセット-生成モデル (G: Gemma, L: LLaMA3)

train \ eval	xlsum-G	xlsum-L	WP-G	WP-L
	(temperature = 1.4)			
xlsum-G	0.998	0.988	0.907	0.887
xlsum-L	0.995	0.998	0.950	0.950
WP-G	0.897	0.907	1.000	0.953
WP-L	0.973	0.980	0.995	0.998

4　RoBERTa 特徴量の PCA

RoBERTa の CLS トークンが文章の意味的な特徴を圧縮して表現していると言われることから，本研究ではその内部表現の分布を可視化することを目的として，PCA を適用した．具体的には，各文章の CLS ベクトル（768 次元）を 2 次元の主成分空間に射影し，人間文と生成文の分布の違いを分析した．図 2 に示すように，人間文と LLM 文は主成分空間上で異なる分布を示しており，とくにデータセット，生成時の温度パラメータの違いにかかわらず，RoBERTa の特徴空間が人間文と LLM 文を分離可能な構造を持っていることが確認された．この結果は，線形分類器が高精度に識別できた背景として，CLS ベクトルが文の生成特性に関する有効な情報を含んでいることを示唆している．

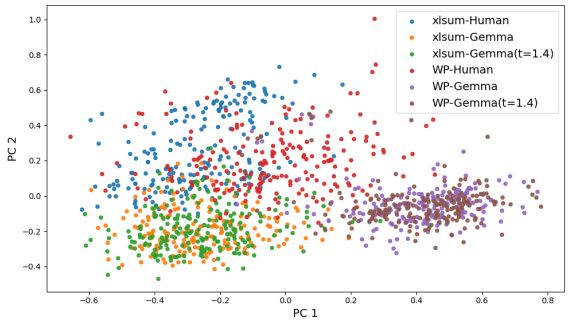

図 2　PC の可視化

5　まとめ

本研究では，人間文と LLM 文を識別する際にドメイン，生成モデル，文の生成確率に依存しない手法について検証を行った．その結果，全ての条件である程度高い精度が得られ，RoBERTa と線形分類器を用いた手法がクロスドメイン，クロスモデル，生成条件に関して頑健であることを確認した．さらに，CLS ベクトルに対して PCA を適用することにより，RoBERTa の特徴量空間が人間文と LLM 文の違いを明確に捉え，線形分類の有効性を示すことができた．今後は，本研究の課題である識別に寄与する特徴の詳細な解釈や，LLM 文を得ることができないなどの悪条件下における識別精度の向上に取り組む予定である．

謝辞

本研究は，科研費 JP23K11266，JP23K28077，JP24K15115，東北大学電気通信研究所共同プロジェクト研究【R06/B14】「深層学習における表現学習に関する研究」の助成によるものです．

参考文献

[1] T. Brown *et al.*, ”Language models are few-shot learners,” in *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[2] S. Gehrmann *et al.*, ”GLTR: Statistical detection and visualization of generated text,” in *Proc. ACL*, pp. 111–116, 2019.

[3] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, ”DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in *Proc. ICML*, 2023.

[4] Y. Liu *et al.*,“RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv:1907.11692*, 2019.

[5] A. Chowdhery *et al.*, “Gemma: Open models based on Gemini research and technology,” *arXiv:2403.03269*, Mar. 2024.

[6] H. Touvron *et al.*,“LLaMA: Open and Efficient Foundation Language Models,” *arXiv:2302.13971*, 2023.

[7] S. Hasan *et al.*, “XL-Sum: Large-scale multilingual abstractive summarization for 44 languages,” in *Proc. EMNLP*, Punta Cana, Dominican Republic, Nov. 2021, pp. 9406–9419.

[8] A. Fan, M. Lewis, and Y. Dauphin, “Hierarchical neural story generation,” in *Proc. ACL*, Melbourne, Australia, Jul. 2018, pp. 889–898.

[9] I. Kuribayashi *et al.*, ”Lower Perplexity is Not Always Human-Like,” in *Proc. ACL*, pp. 5203–5213, 2021.