検出器の判断に基づく大規模言語モデルの生成テキストの特徴分析

三浦 東子 [1]　谷口 雅弥 [2]　坂口 慶祐 [1,2]　乾 健太郎 [1,2,3]
[1] 東北大学　[2] 理化学研究所　[3]Mohamed bin Zayed University of Artificial Intelligence
miura.toko.r8@dc.tohoku.ac.jp, masaya.taniguchi@riken.jp
{keisuke.sakaguchi, kentaro.inui}@tohoku.ac.jp

概要

本研究では，大規模言語モデル (LLM) が生成するテキスト (LLM-text) の特徴をヒトが書いたテキスト (human-text) との比較から明らかにする．エッセイを対象に 2 つの実験を行った. 1 つ目は human-text と LLM-text から LLM-text を分類する実験，2 つ目は，LLM-text に改変を加えることで LLM-text の検出を妨害する実験である．1 つ目の実験から LLM-text に固有の特徴があることを示した．さらに 2 つ目の実験から，その特徴が「単語の連なりの滑らかさ」であることを示した．そして，この他の特徴を LLM-text が備え得ることや， LLM-text の特徴の軽減と human-text の特徴の付加について今後の課題として議論する．

1　はじめに

近年の大規模言語モデル（LLM）の性能が向上したことで，日常の多くの場面で LLM が利用されるようになった．一方で LLM を制限すべき状況での LLM の利用や，ユーザーが生成テキストに対して感じる違和感の解消が課題となっている．

教育現場は LLM を制限すべき状況の一つである．試験問題や課題に対する正確な回答や流暢なエッセイ，および正常に動作するコードを簡単に生成できるため，回答にかかる時間を大幅に削減することが可能になった．しかしこの利点が学生の学習過程の簡略化に繋がり，得られる経験や知識が減少することが懸念された．この例をはじめとした LLM 乱用を抑制するために，生成テキスト検出器の開発が進められた．一方，攻撃に対する検出器の脆弱性も確認されており，今後の攻撃手法や LLM の発展に合わせて検出器の頑健性向上に努める必要がある．

LLM は対話モデルとしても利用される．企業の自動応答やロボットの発話内容をより自然なものにするためには，生成テキストから LLM らしさを除去する必要がある．

以上で挙げた，検出器の頑健性向上および生成テキストの違和感の軽減は，互いに対立する課題である．なぜなら，検出器は LLM らしさを捉えることで生成テキストを検出するため，生成テキストから LLM らしさを除去することは検出器に対する攻撃となるからである．ヒトが書くテキスト (human-text) と LLM が生成するテキスト (LLM-text) の違いを明らかにすることは，この対立する 2 つの課題の両者への貢献となる．

したがって本研究では human-text と LLM-text の違いの解明を最終的な目的とし，この達成のために，human-text と比較して得られる LLM の特徴の解明に取り組む．

2　関連研究

現在，LLM-text と human-text の区別は人にとって難しいタスクであることが報告されている [1]. 人の感覚に頼らずに LLM-text を判別することの必要性から，様々な検出手法が研究・開発された．代表的な検出手法として，LLM の出力に埋め込んだ透かし (Watermark) を検出する *Watermarking Technology* [2]，LLM の統計的特徴をもとに検出する *Zero-shot Detectors* [3]，human-text と LLM-text を学習する *Fine-tuned LM Detectors* [4]，human-text, LLM-text，および検出器の評価をもとに Attacker が生成したテキストを学習する *Adversarial Learning methods* [5]，誘導するためのプロンプトを使い，LLM を検出器として利用する *LLMs as Detectors*，人と機械のそれぞれによる分類結果を統合して検出する *Human-assisted methods* [6] が挙げられる．

また LLM-text と human-text の統計的な特徴の違いについての研究も行われた．その結果，特定の品詞を含む割合の違いや，LLM-text と比べた human-text

の語彙の豊富さ，単語同士の依存関係の距離が長さの違い，などの知見が集まった [7][8]．このようなコーパスおよび長文を対象とした分析が進む一方で，比較的短い文章に現れる特徴の分析は未だ不十分である．語彙や品詞の統計的な特徴の計算には大量の文書が必要となるため，検出器が検出時に利用することが難しい．したがって，本研究では比較的短い文章から得られる特徴を対象に研究を行う．

3 実験

3.1 仮説: 単語の連なりの滑らかさ

LLM-text の特徴の候補として，「単語の連なりの滑らかさ」を挙げる．この特徴を挙げる理由は 2 つある．1 つは，DetectGPT [3] の提案手法が有効なことである．DetectGPT は生成テキストを検出する手法の一つであり，human-text を書き換えた場合，元の文に比べて対数確率が増加することも減少することもあるが，ほとんどの LLM-text では，書き換えによって対数確率の低下のみが起こるという現象を利用している．つまり，LLM-text が対数確率が最も高くなるような単語の連なりであることを利用した手法であると言える．DetectGPT は高精度で LLM-text を検出できることから，連続する単語の並び方の滑らかさが LLM-text の特徴となっていることが示唆される．

もう 1 つは，単語レベルの言い換えに対する検出器の脆弱性である．ここで，単語レベルの言い換え (word-level paraphrase) は文中の一部の単語を同義語で置換することを，文レベルの言い換え (sentence-level paraphrase) は文章中の一部の文を意味の似た文で置換することを指す．検出器は word-level paraphrase に対して脆弱であることが確認されている [9]．これは，単語が置換された部分で，単語の滑らかな連なりが切断されたことが原因で検出が困難になったものと考えられる．

以上の推測を元に，「単語の連なりの滑らかさ」が LLM-text の特徴の 1 つであるという仮説を立てた．

3.2 実験方針: LLM-text の特徴の同定

「単語の連なりの滑らかさ」が LLM-text の特徴の 1 つであるという仮説を確かめる．

はじめに，実験 1 において非敵対的テキストに対して高精度な検出が可能なモデルを構築する．その後，実験 2 において，実験 1 で構築したモデルに対して下記の検証を行う．

- word-level paraphrase に対する脆弱性を調べ，同義語置換により単語の滑らかな連なりを崩すことが検出の妨げになることを示す．
- 文脈内学習により生成した敵対的テキストに対する頑健性を調べる．攻撃による精度低下が見られない場合，LLM は不自然な単語の連なりを生成できないことが言える．
- sentence-level paraphrase に対する脆弱性を調べ，複数の LLM が生成した文が混ざった文章が検出器に与える影響を明らかにする．検出精度が低下しない場合，検出器は文の連続の滑らかさ，および意味的な滑らかさを考慮しないことが言える．

3.3 実験対象: エッセイコーパス

イギリスの大学生 (学士課程及び修士課程）が書いた 4 分野（AH: 人文科学，LS: 生命科学，PS: 物理科学，SS: 社会科学）のエッセイを集めたコーパス: BAWE (British Academic Written English Corpus [10]) を用いる．検出器は与えられたエッセイを，人が書いたもの (human) か LLM が生成したもの (LLM) かの 2 値に分類した．以降，LLM に分類した場合を「検出した」と言う．学習したドメイン (In-domain) と未知のドメイン (OoD: Out-of domain) を比較するため，学習には AH を使い，その他の分野 (LS，PS，SS) を OoD として検出器のテストに用いる．データの分割は，学習 : 検証 : テスト = 8 : 1 : 1 とし，さらに，文脈内学習のため，AH のテストデータから 10 エッセイ (human:5，LLM:5) を除く．したがってデータの構成は表 1となる．

表 1 データセットの構成

ドメイン	学習	検証	テスト
In-domain	1130	138	132
OoD	-	-	418

LLM-text の生成には GPT-3 (gpt-3.5-turbo-1106 [11]，temperature=0.7) を利用し，以下の手法で生成する．

まず，BAWE 内のエッセイの冒頭 250 単語を元に題目を生成する（プロンプトは付録 A.1）．その後，生成した題目を元にエッセイを生成する．これにより，BAWE 内の 1 エッセイ (*human-essay*) に対して 1 つの生成エッセイ (*LLM-essay*) が存在することにな

る（プロンプトは付録 A.2）.

3.4 実験設計: LLM-text の検出と攻撃

検出器として BERT (bert-base-uncased) [12] を利用し，以下の実験を行う.

実験 1 では，検出器を In-domain (AH) で訓練し，OoD (LS，PS，SS) でテストを行う (ハイパーパラメータは付録 B).

実験 2 では，訓練した検出器に対して各種攻撃を行い，攻撃による検出精度の変化を観察する．攻撃手法は，(1) Parrot [13] による LLM-text の sentence-level paraphrase，(2) NLTK [14] による LLM-text の word-level paraphrase，(3) GPT-3 による文脈内学習，の 3 パターンである. Parrot は T5 ベースのモデルであるため，(1) は文章に異なる LLM が生成した文を混ぜること同義である．言い換えによる攻撃は，各エッセイに対し，(1) は語彙の 40%を，(2) は文の 50%を対象に行った．(3) では，非敵対的生成時に利用した題目と，検出器が human と予測したエッセイを利用する（プロンプトは付録 A.3）.

4 結果・考察

4.1 実験 1: LLM-text の検出

結果を表 2に示す．正解率 (accuracy) は In-domain, OoD ともに 90%を超え，OoD では In-domain に比べて 0.0518 point 低い値となった. 予測の誤りは全て偽陽性 (FP) であり，偽陰性 (FN) は 0 であったことから，検出器は人が書いたエッセイを誤検出する傾向があること，そして OoD ではこの傾向が顕著になることが予想される.

表 2 In-domain と OoD に対する検出精度

test domain	accuracy	FP	FN
In-domain	0.9848	2	0
OoD	0.9330	8	0

4.2 実験 2: LLM-text 検出器への攻撃

結果を表 3に示す．word-level paraphrase は偽陰性の数を大きく増加させたのに対し，sentence-level paraphrase の検出精度への影響は僅かであった. これは，同義語置換により単語の滑らかな連なりを崩すことが検出の妨げになること，そして検出器は文の連続の滑らかさ，および文同士の意味的な繋がりをあまり考慮しないことを意味する.

文脈内学習による検出性能の低下は見られなかった. この結果から，LLM は敵対的なテキストを生成するように誘導するプロンプトを用いた場合であっても不自然な単語の並びを生成することはできないということが言える．つまり単語の並びの滑らかさは，現状における LLM が自ら変えることのできない特徴であることが考えられる. さらに，生成モデルが認識する「LLM らしさ」と，本研究において BERT が学習した「LLM らしさ」の間に違いが存在することも示唆された．ここで用いた手法は *LLMs as Detectors* の 1 例の OUTFOX [15] を参考にしたものであり， OUTFOX は攻撃者と検出器の両者をプロンプトにより実現する．論文中ではこの攻撃が，プロンプトベースの検出器に対して十分な効果を発揮したことから，プロンプトベースの検出器とファインチューニングによる検出器の間で LLM-text の特徴の認識に違いがあると捉えることができる.

5 分析

5.1 語彙の影響と文章の構成の影響

実験により，OoD では検出精度が In-domain に比べて低いことがわかった．OoD での性能低下には，分野特有の語彙や，特定の単語の並び・文の構成が影響した可能性がある．この影響について論じるために，エッセイ中の単語・文をランダムに並べ替えた場合の検出精度を調べた.

表 4より，単語を並べ替えた場合には，全てのサンプルが LLM ではないテキストであると予測された. In-domain と OoD で共に正しい検出が不可能になったことから，検出器は特定の単語ではなく単語の並びを手がかりに判断していると考えられる.

文を並べ替えた場合には FP (誤検出) が減少したが，FN (見逃し) の数に変化はなかった．並び替えが検出精度を下げることは無かったものの，精度への影響自体は存在したことから，OoD における検出精度低下には LLM-text における分野特有の文の構成が存在することが考えられる．しかし単語の場合に比べて誤検出・見逃しの大きな増減が無かったことから，文の並びの滑らかさが与える検出への影響は比較的小さいことがわかった.

また両結果における誤検出の減少と見逃しの増加から，単語や文の並びが不自然になった文章を，LLM が生成したものでは無いものと判断する傾向にあることが考えられる．特定の単語の並びについ

表 3　攻撃に対する検出精度

test domain	attack method	accuracy		FP	FN
In-domain	5-shot learning	0.9848	(–)	2	0
OoD	5-shot learning	0.9330	(–)	28	0
In-domain	sentence-level paraphrase	0.9848	(–)	2	0
OoD	sentence-level paraphrase	0.9430	(↑ 0.0100)	27	0
In-domain	word-level paraphrase	0.7535	(↓ 0.2313)	2	33
OoD	word-level paraphrase	0.7799	(↓ 0.1531)	28	64

ては今後調査する．

表 4　word/sentence level のシャッフルに対する検出精度

level	test domain	accuracy		FP	FN
word	In-domain	0.5000		0	66
word	OoD	0.5000		0	209
sentence	In-domain	1.000	(↑ 0.0152)	0	0
sentence	OoD	0.9833	(↑ 0.0503)	6	0

5.2　入力単語数と検出精度

図 1に入力単語数と正解率の関係を示す．正解率は 1〜20 単語にかけて急激に上昇したが，20 単語以降は伸びが緩やかになった．In-domain と OoD において共に 90%を超えるためには 120 単語の入力を必要とした．

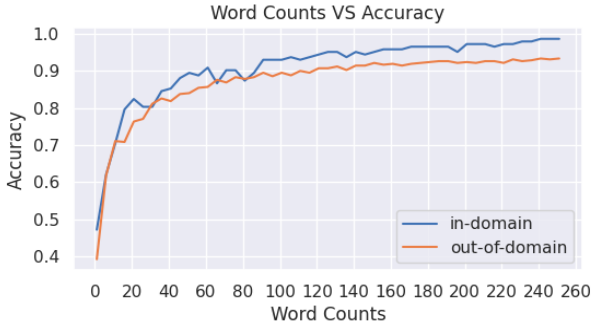

図 1　入力単語数と検出精度の関係

5.3　単語レベルの言い換えの影響

単語レベルの言い換えの影響について調べた．同義語置換する語彙の割合と検出精度の関係を図 2に示す．In-domain，OoD 共に 20%までは影響が小さいが，30%を超えると精度が徐々に低下した．

5.4　議論

実験を通じて，滑らかに連続する単語の並びを崩すことが，生成テキストから LLM らしさを除去するための方法の一つであることを示した．しかしこの方法の中には，文章中の単語をただランダムに並べ替えるなど，同時に人らしさをも失う崩し方が含まれる．また並びを崩すということは文章の完成度・正確性を下げることであり，LLM の進歩に逆行する．LLM らしい特徴の解明は，検出器が生成テキストを見逃す可能性を減らすことや，生成テキストの不気味さの軽減に寄与する．一方で人らしさについては調べられていない．本研究の長期的な目標は，ヒトが書いたテキストと LLM が生成するテキストの違いの明確化を通じて，検出器の頑健性の向上や，生成テキストの違和感の軽減に貢献することであった．人が書いたテキストの誤検出を抑制し，生成テキストに対する人らしさの導入のためには，人らしさに関する知見を得ることも必要である．したがって今後は，人らしさを残した崩し方や，単語の並び以外に注目して LLM らしさの除去または人らしさを増す方法を明らかにする．

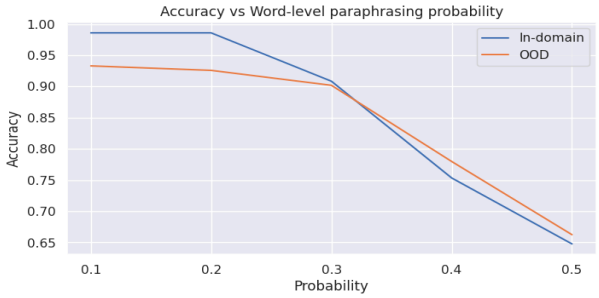

図 2　単語レベルの言い換えの割合と検出精度の関係

本研究では human-text と比較して得られる LLM の特徴を調べた. 実験の結果,「単語の連なりの滑らかさ」が，現在の生成モデルが生成するテキストの持つ特徴の一つであるという知見が得られた．今後の課題として，単語の並び以外に着目した LLM らしさや，人らしさの解明が残されている．

謝辞

本研究は理化学研究所の基礎科学特別研究員制度，JSPS 科研費 JP21K21343 の支援を受けています。

参考文献

[1] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 7282–7296, Online, August 2021. Association for Computational Linguistics.

[2] Evan Lucas and Timothy Havens. GPTs don't keep secrets: Searching for backdoor watermark triggers in autoregressive language models. In Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta, editors, **Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)**, pp. 242–248, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In **Proceedings of the 40th International Conference on Machine Learning**, ICML'23. JMLR.org, 2023.

[4] Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. Deepfake text detection in the wild, 2023.

[5] Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning, 2023.

[6] Luoxuan Weng, Minfeng Zhu, Kam Kwai Wong, Shi Liu, Jiashun Sun, Hang Zhu, Dongming Han, and Wei Chen. Towards an understanding and explanation for mixed-initiative artificial scientific text detection, 2023.

[7] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Tianming Liu, and Xiang Li. Differentiating chatgpt-generated and human-written medical texts: Quantitative study. **JMIR Med Educ**, Vol. 9, p. e48904, Dec 2023.

[8] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

[9] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2023.

[10] Sian Alsop and Hilary Nesi. Issues in the development of the british academic written english (BAWE) corpus. **Corpora**, Vol. 4, No. 1, pp. 71–83, August 2009. Correspondence to: Hilary Nesi, email: h.nesi@coventry.ac.uk.

[11] OpenAI. gpt-3.5-turbo-1106. url:https://platform.openai.com/, accessed: 2023-12-31.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Prithiviraj Damodaran. Parrot: Paraphrase generation for nlu., 2021.

[14] Steven Bird, Ewan Klein, and Edward Loper. **Natural language processing with Python: analyzing text with the natural language toolkit**. " O'Reilly Media, Inc.", 2009.

[15] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In **Proceedings of the 38th AAAI Conference on Artificial Intelligence**, Vancouver, Canada, February 2024.

A　生成プロンプト

A.1　エッセイ題目の生成

エッセイの冒頭 250 単語 [*essay*] を元にした題目の生成には以下のプロンプトを用いた.

Please write the most appropriate problem statement for an essay assignment that would have you write the following essay.
Essay: [essey]
Problem Statement:

A.2　非敵対的 LLM-essay の生成

生成した題目 [*problem statement*] を元にした LLM-essay の生成には以下のプロンプトを用いた.
Given the following problem statement, please write an essay expressing a clear opinion with a minimum of 250 words.
Problem statement: [problem statement]
Essay:

A.3　敵対的 LLM-essay の生成

文脈内学習による敵対的テキスト生成には以下のプロンプトを用いた. 但し [*human problem statement*] は検出器が human と予測したエッセイの題目を, [*human pred essay*] はそのエッセイの冒頭 250 単語を, [*target problem statement*] は敵対的エッセイ生成対象の題目を指す.
Here are the results of detecting whether each essay from each problem statement is generated by a Human or a Language Model(LM).
Problem statement: [human problem statement]
Answer: Human
Essay: [human pred essay]
Problem statement: [human problem statement]
Answer: Human
Essay: [human pred essay]
...
Problem statement: [target problem statement]
Answer: Human
Essay:

B　実験設定詳細

B.1　ハイパーパラメータ

表 5　ハイパーパラメータの設定

パラメータ	設定した値
エポック数	80
バッチサイズ	8
学習率	5e-5