Retrieval 을 활용한 Finetuning-free LLM-generated text detection

한진모, 강주연, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실
{jmhan, jykang}@hi.snu.ac.kr, nkim@snu.ac.kr

Finetuning-free Retrieval-based LLM-generated text detection

Jinmo Han, Ju Yeon Kang, Nam Soo Kim
Human Interface Laboratory,
Department of Electrical and Computer Engineering and INMC,
Seoul National University

요 약

본 연구는 대규모 언어 모델(LLM)이 생성한 텍스트를 효율적으로 판별하기 위한 경량 탐지 기법을 제안한다. 상당수의 탐지 기법은 사전학습 언어 모델에 대한 추가적인 파인튜닝으로 인한 높은 계산 비용을 수반한다. 본 연구는 이러한 한계를 극복하기 위해, 별도의 파인튜닝 과정 없이 학습 데이터의 임베딩을 사전에 인코딩하고, 이를 통해 정의한 Local LLM-ness vector 만을 활용하여 텍스트의 생성 주체를 판별하는 방법을 제시한다. 제안 기법은 별도 파인 튜닝 없이 사전 학습된 언어 모델의 표현력만으로 기존의 통계 기반 기법들의 성능을 능가하는 실용적인 성능을 달성한다.

Ⅰ. 서 론

대규모 언어 모델의 발전으로 인해, 인간이 작성한 텍스트와 기계가 생성한 텍스트를 구별하는 문제는 콘텐츠 신뢰성 및 정보 무결성 측면에서 중요한 연구 과제로 부상하고 있다. 기존의 LLM 생성 텍스트 탐지 연구는 주로 지도 학습 기반의 분류 모델을 파인튜닝하거나, 통계적 특성 또는 언어 모델 확률을 활용하는 방식에 의존해왔다[1] [2]. 그러나 이러한 접근은 새로운 모델이나 도메인에 대한 일반화 성능이 제한적이거나, 반복적인 언어 모델 호출로 인한 계산 부담이 크다는 문제점을 가진다. 본 논문에서는 이러한 문제의식을 바탕으로, 파인튜닝을 전혀 사용하지 않으면서도 실용적인 탐지 성능을 달성할 수 있는 간단한 임베딩 기반 판별 기법을 제안한다. 핵심 아이디어는 인간 텍스트와 기계 텍스트가 임베딩 공간에서 형성하는 국소적 구조 차이를 활용하는 것으로, 소수의 샘플만으로도 의미 있는 판별 축을 구성할 수 있다는 점에 있다.

Ⅱ. 본론

1. 제안하는 방법

본 연구에서는 대규모 언어 모델(LLM)이 생성한 텍스트를 판별하기 위해, 파인튜닝 없이 임베딩 공간에서 정의되는 국소적 판별 벡터를 활용하는 방법을 제안한다. 구체적으로, 인간 텍스트와 기계 텍스트 각각에서 코사인 유사도가 가까운 순서로 20 개의 샘플을 선택한다. 각 텍스트는 BERT-base 모델을 통해 벡터로 변환된다[3]. 토큰 임베딩에 대한 mean pooling 을 적용하여 문장 단위 임베딩을 구성한다. 이후 인간 텍스트 임베딩의 평균을 Local Human Centroid 로, 기계 텍스트 임베딩의 평균을 Local Machine Centroid 로 정의한다. 두 중심 벡터의 차이를 Local LLM-ness vector 로 정의하며, 이는 해당 국소 영역에서 인간 텍스트에서 기계 텍스트로 향하는 판별 방향을 나타낸다. 판별 대상 텍스트는 동일한 방식으로 임베딩된 후 Local LLM-ness vector 에 대해 직교 투영되며, 이 투영 값이 클수록 기계 생성 텍스트일 가능성이 높다고 판단한다.

2. 실험 설계

제안하는 방법의 성능을 평가하기 위해, 본 연구에서는 NeurIPS 에서 공개된 DetectRL 데이터셋을 사용하였다[4]. DetectRL 은 다양한 대규모 언어 모델로 생성된 텍스트와 인간이 작성한 텍스트를 포함하는 공개 벤치마크로, 실제 생성 환경을 반영한 다양한 데이터 분포를 제공한다. 구체적으로, DetectRL 데이터셋의 Direct Prompt(직접 프롬프팅), Prompt Attack(프롬프팅을 통한 분류기 공격), Paraphrase Attack(생성된 문장을 paraphrase 하는 프레임워크를 통한 분류기 공격), Perturbation Attack(일부 단어나 철자 등을 수정함으로써 분류기 공격) 시나리오에서의 성능을 평가하였다. 비교 기준은 LLM 의 추론을 통한 통계량을 요구하는 Entropy[2], DetectGPT[3] 기법을 선정했다.

3. 실험 결과

제안 기법의 각 시나리오별 AUROC / F1 Score 는 순서대로 아래 표와 같다.

Method	Direct Prompt	Prompt Attacks	Para. Attacks	Pert. Attacks
Entropy	26.5 / 0.0	26.2 / 0.0	64.6 / 57.6	68.6 / 69.0
DetectGPT	52.8 / 40.9	51.8 / 38.0	31.8 / 16.9	18.2 / 0.0
Proposed	**97.2 / 97.2**	**92.6 / 92.8**	**91.1 / 89.0**	**88.5 / 78.1**

실험 결과를 통해, 제안한 방법이 의미 있는 판별 성능을 달성함을 확인하였다. 특히 소수의 샘플을 활용한 국소적 중심 구성만으로도 안정적인 판별 신호를 얻을 수 있음을 보였으며, 이는 임베딩 공간에서의 구조적 차이가 LLM 생성 텍스트 탐지에 효과적으로 활용될 수 있음을 시사한다.

Ⅲ. 결론

본 논문에서는 파인튜닝 없이 LLM 생성 텍스트를 탐지할 수 있는 간단하고 효율적인 임베딩 기반 기법을 제안하였다. 제안 방법은 소수의 인간 및 기계 텍스트 샘플만을 이용해 국소적 판별 축을 구성하고, 이를 통해 새로운 텍스트의 생성 주체를 판단한다는 점에서 실용성이 높다. 특히 Local LLM-ness vector 라는 개념을 통해, 임베딩 공간에서의 구조적 차이를 직접적으로 활용할 수 있음을 보였으며, 이는 복잡한 분류기나 추가 학습 없이도 경쟁력 있는 성능을 달성할 수 있음을 시사한다. 향후 연구에서는 국소 영역 선택 방식이나 K 값 변화에 따른 안정성 분석, 그리고 다양한 임베딩 모델에 대한 확장 가능성을 검토할 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 2026 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의해 지원되었음

참 고 문 헌

[1] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn." Detectgpt: Zero-shot machine-generated text detection using probability curvature," International Conference on Machine Learning, ICML 2023, Honolulu, Hawaii, USA, July 23- 29, 2023, Proceedings of Machine Learning Research, 2023.

[2] Thomas Lavergne, Tanguy Urvoy, and François Yvon." Detecting fake content with relative entropy scoring," Proceedings of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, Patras, Greece, July 22, 2008, CEUR Workshop Proceedings, 2008.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 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), 2019.

[4] Wu, Junchao, et al." Detectrl: Benchmarking llm-generated text detection in real-world scenarios," Advances in Neural Information Processing Systems 37, pp. 100369-100401, 2024.