한글 대형 언어 모델에 워터마킹 기술 적용 및 분석

김성식, 이재구*
국민대학교
*jaekoo@kookmin.ac.kr

Applying Watermarking Technique to a Korean Large Language Model

Sungsik Kim, Jaekoo Lee*
College of Computer Science, Kookmin University.

요 약

본 논문은 한글 대형 언어 모델에 워터마킹 기술의 적용 가능성을 평가한다. 최근 대형 언어 모델은 급격한 발전을 이루었지만, 동시에 윤리적 문제와 악용 가능성도 부각되고 있다. 언어 모델의 워터마킹 기술 접목은 모델 출력에 고유한 정보를 추가하여 언어 모델 사용 여부를 검증하는 방식으로 제안되었다. 본 연구에서는 한글 언어 모델에 워터마킹 기술을 적용하여 실험하였다. 한글 언어 모델의 출력 문장 품질을 워터마킹 적용 강도에 따라 분석하였으며, 영어 기반 언어 모델에 비해 성능이 크게 저하됨을 시각적으로 확인하였다. 또한, 선형 회귀 기울기를 이용한 수치적 분석을 통해 한글 언어 모델이 워터마킹에 따른 성능 하락에 강건하지 못함을 보였다. 우리는 한글 언어 모델에 맞는 워터마킹 기술의 발전을 위해서 한글의 언어학적 특성에 맞는 기술 연구의 필요성을 시사한다.

Ⅰ. 서론

최근 대형 언어 모델(Large Language Model)은 급격한 발전을 이루고 있다. 특히 ChatGPT 와 같은 모델은 여러 산업에 많은 영향을 미치고 있다. 그러나 거대한 활용 가능성과 함께 악용 위험성도 증가하고 있다[1].

위와 같은 문제점들을 해결하기 위해 언어 모델 출력에 워터마킹(Watermarking) 기술을 추가하는 방법이 등장하고 있다. 언어 모델에 워터마킹 기술을 접목하는 것은 언어 모델의 출력에 사람은 인식할 수 없지만 알고리즘적으로 기계는 인식할 수 있는 패턴을 추가하여, 해당 문구가 언어 모델에 의해 출력된 것임을 검증할 수 있도록 하는 것을 의미한다[2]. 해당 기술이 실제 언어 모델에 적용된다면, 언어 모델 사용 여부를 쉽게 파악하여, 다양한 악용 가능성의 해결책이 될 수 있다.

본 논문에서는 한글 대형 언어 모델에 워터마킹 기술을 적용하여 결과를 분석하고자 한다. 워터마킹 기술 적용에 따른 한글 모델 성능의 하락을 분석하고, 시각적인 예시를 통해 분석을 수행한다.

Ⅱ. 본론

본 논문에서는 워터마킹 기술 적용을 위해 Kirchenbauer 등[2]이 제안한 방식을 사용하였다. 우선, 언어 모델이 프롬프트(Prompt)와 $t-1$번째까지 생성한 토큰(Token)들을 입력으로 받으면, t번째 토큰을 생성하기 위해 모든 토큰들에 대한 로짓(Logit) l_t을 출력하게 된다. 워터마크를 추가하기 위해 $t-1$번째 토큰을 정해진 해시 함수(Hash Function)에 입력하고 나온 값을 난수 생성기(Random Number Generator)에 초기값(Seed)으로 입력하게 된다. 다음으로, 초기값이 지정된 난수 생성기에 의해 전체 토큰 집합 $\mathcal{V}$를 각각 비율이 γ와 $(1-\gamma)$인 그린 집합(Green List) $\mathcal{G}$와 레드 집합(Red List) $\mathcal{R}$로 구분하게 된다. 최종적으로 $\mathcal{G}$에 속하는 토큰의 로짓에 미리 정의한 δ를 더해 해당 토큰들이 뽑히게 될 가능성을 매우 높인다.

문장에 워터마크가 있는지 검증하기 위해 Z-검정(Z-test)를 사용한다. 길이가 T인 문장이 워터마크에 대한 정보 없이 작성되어 토큰을 $\mathcal{V}$에서 임의로 뽑았다면, 문장 안에 있는 그린 집합 $\mathcal{G}$에 속하는 토큰의 개수인 $\mathrm{N}_{\mathcal{G}}$는 횟수가 T, 확률이 γ인 이항 분포를 따르고, T가 충분히 크다면 평균이 $\gamma\mathrm{T}$, 분산이 $\gamma(1-\gamma)\mathrm{T}$인 정규 분포를 따른다. 따라서 귀무 가설(Null Hypothesis)을 "문장이 워터마크에 대해 알지 못한 채로 생성되었다." 라고 설정한다. 이를 검정하기 위한 Z-검정통계량(Z-test Statistic)은 다음과 같다.

$$z = \frac{N_{\mathcal{G}} - \gamma\mathrm{T}}{\sqrt{\gamma(1-\gamma)\mathrm{T}}} \quad (1)$$

문장에 대해 해당 Z-검정통계량을 계산하고, 미리 정의된 기각역(Reject Region) $z > \alpha$에 의해 귀무 가설의 기각 여부를 결정하게 된다.

본 논문에서는 워터마킹 기술을 한글 기반 언어 모델에 적용하여 실험하고 분석한다. 한글 언어 모델로 생성한 문장들과 워터마킹을 적용한 예시는 [그림 1]에서 확인할

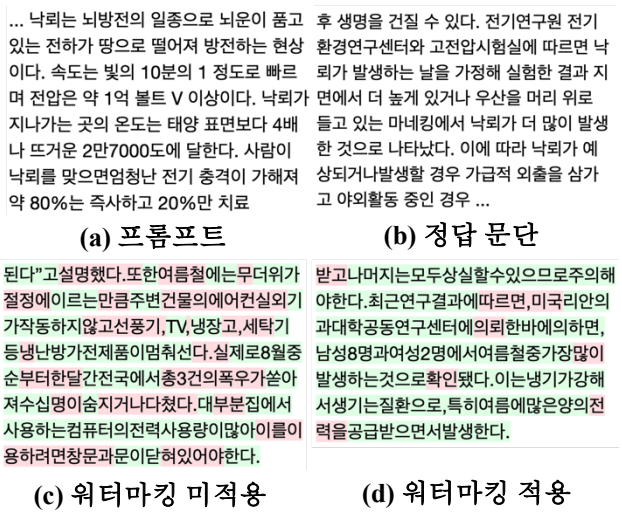
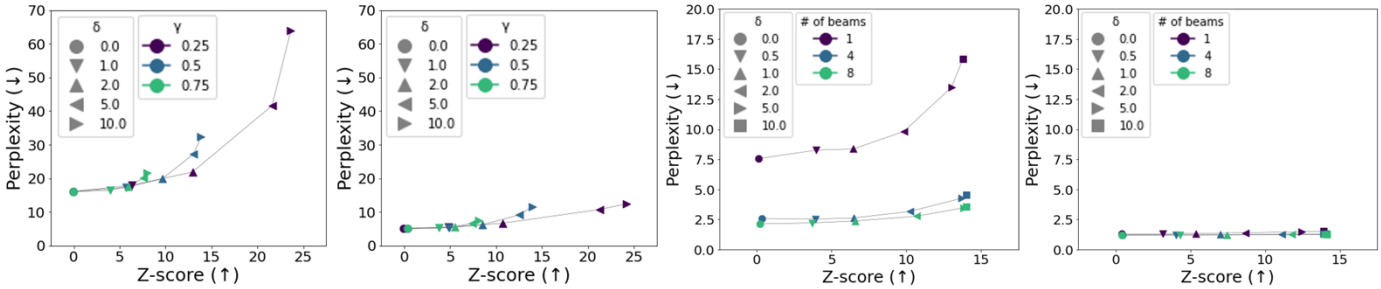


그림 1. 한글 언어 모델 워터마킹 (δ = 2.0, γ = 0.5)



(a) 한글 모델 (샘플링) **(b) 영어 모델 (샘플링)** **(c) 한글 모델 (빔 탐색)** **(d) 영어 모델 (빔 탐색)**

그림 2. 한글과 영어 모델의 Perplexity 와 Z-score 상관관계

수 있다. 토큰 배경색은 각각 녹색은 $\mathcal{G}$, 적색은 $\mathcal{R}$에 속하는 토큰임을 의미한다.

Ⅲ. 실험

한글 언어 모델 실험을 위해 모델로는 KoGPT-2[3]를 사용하였고, 데이터 집합으로는 2022 년 7 월 1 일부터 2022 년 7 월 10 일까지의 뉴스를 모은 네이버 뉴스 데이터 집합[4]의 테스트 집합을 사용하였다. 영어 언어 모델의 실험 환경은 Kirchenbauer 등[2]이 제안한 방식과 동일하다.

평가 지표로서 모델이 출력한 언어의 품질을 평가하는 Perplexity 와, 문장이 워터마킹이 잘 되었는지 판단하는 평균 Z-score(수식 1)이다. Perplexity 가 낮을 수록 생성된 문장의 품질이 좋으며, Z-score 는 높을 수록 문장이 워터마킹이 잘 되었다는 것을 의미한다.

로짓에 더해주는 δ가 클수록 모델이 $\mathcal{G}$에서 더 많은 토큰을 고르게 된다, 이는 곧 $N_{\mathcal{G}}$와 Z-score 의 증가를 의미한다. 또한, 이는 문장의 출력 과정에 가해지는 제한이 강해져 Perplexity 의 상승으로 이끌 수 있다.

[그림 2]는 Z-score 와 Perplexity 의 상관관계에 대한 실험 결과를 그래프로 표현한 것이다. 모델이 토큰을 고르는 방식은 모델이 출력한 확률분포에서 토큰을 뽑는 샘플링(Sampling)과 확률이 높은 문장을 1 개가 아닌 n 개를 고르면서 가장 좋은 조합을 찾는 빔 탐색(Beam Search)를 사용하였다. [그림 2.a]와 [그림 2.b]는 δ와 γ에 따른 샘플링 방식 실험 결과이다. 해당 실험 결과를 통해 한글 모델은 워터마킹 적용 강도가 강해질 수록 생성된 문장의 품질이 영어 모델에 비해 더 크게 떨어짐을 확인할 수 있다. [그림 2.c]와 [그림 2.d]는 빔 탐색 방식을 사용하여 γ를 0.5 로 고정하고 실험하였다. 워터마킹 적용에 따라 거의 영향을 받지 않는 영어 모델과 달리 한글 모델은 문장 품질이 떨어지는 것을 확인할 수 있다.

더 나아가, 해당 관계를 수치적으로 검증하기 위해 선형 회귀 방식을 사용하는 것을 제안한다. δ에 따라 그래프를 그리고 선형 회귀한다면, 기울기를 구할 수 있다. 기울기가 작다면 모델이 워터마킹 적용에 따른 문장 품질 손실에 강건하다는 것을 의미한다. [표 1]에서 한글 모델의 경우 기울기가 영어 모델에 비해 훨씬 큰 것을 확인할 수 있다.

임준호와 김현기[5]에 따르면 한글 어절에서 각 토큰마다 성능 기여도를 측정한 결과 첫 번째 토큰보다 마지막 토큰의 기여도가 높았다. 일반적으로 교착어인 한글 어절에서 마지막 토큰은 조사나 어미에 해당하며, 한글 언어 모델은 조사와 어미에 대한 의존성이 높다고 해석할 수 있다. 본 논문에서 사용한 워터마킹 기술은 임의로 토큰 집합을 제한하기 때문에 의존성이 높은 단어들 또한 제한될 수 있다. 따라서, 한글이 교착어라는 특성이 워터마킹 기술을 한글 언어 모델에 적용하는 것에 부정적인 영향을 끼쳤음을 추측할 수 있다.

표 1. 선형 회귀 모델의 기울기

γ	기울기	
	한글	영어
0.25	1.804	0.310
0.5	1.069	0.453
0.75	0.628	0.299
평균	1.167	0.354

(a) 샘플링

빔의 수	기울기	
	한글	영어
1	0.562	0.018
4	0.147	0.006
8	0.106	0.005
평균	0.272	0.009

(b) 빔 탐색

Ⅳ. 결론

본 논문에서는 한글 대형 언어 모델에 워터마킹 기법을 적용하고 분석하였다. 실험 결과, 한글 언어 모델이 영어 언어 모델에 비해 워터마킹에 따른 성능 하락에 강건하지 못했다. 대부분의 언어 모델 연구의 기반 언어는 영어이기 때문에 한글의 교착어라는 언어학적 특성을 반영하지 못한다. 추후, 한글 언어 모델의 범용적인 사용성과 악용의 방지를 위해서 한글 언어적 특성에 맞는 워터마킹 방식의 연구가 필요하다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00516,국가통계데이터에 적용 가능한 차등정보보호 개념을 도출하고 통계분석의 유용성을 보장해야 하는 문제 해결)

참 고 문 헌

[1] Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., ... & Biggio, B. (2022). The threat of offensive ai to organizations. Computers & Security, 103006.

[2] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models. Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research 202:17061-17084

[3] SKT-AI. (2020). KoGPT2. Retrieved from https://github.com/SKT-AI/KoGPT2

[4] Daekeun Kim. (2022). naver-news-summarization-ko. Retrieved from https://huggingface.co/datasets/daekeun-ml/naver-news-summarization-ko

[5] 임준호, 김현기. "사전학습 언어모델의 토큰 단위 문맥 표현을 이용한 한국어 의존 구문분석." 정보과학회논문지 48.1 (2021): 27-34.