생성형 거대 언어 모델에서 일관성 확인 및 사실 검증을 활용한 Hallucination 검출 기법

진 명[1], 김건우[2*]
[1]경상국립대학교 AI 융합공학과 석사과정
[2]경상국립대학교 컴퓨터과학부 교수

wlsaud222@gnu.ac.kr, gunwoo.kim@gnu.ac.kr

Hallucination Detection for Generative Large Language Models Exploiting Consistency and Fact Checking Technique

Myeong Jin[1], Gun-Woo Kim[2]
[1]Dept. of AI Convergence Engineering, Gyeongsang National University
[2]Dept. of Computer Science & Engineering, Gyeongsang National University

요 약

최근 GPT-3 와 LLaMa 같은 생성형 거대 언어모델을 활용한 서비스가 공개되었고, 실제로 많은 사람들이 사용하고 있다. 해당 모델들은 사용자들의 다양한 질문에 대해 유창한 답변을 한다는 이유로 주목받고 있다. 하지만 LLMs 의 답변에는 종종 Inconsistent content 와 non-factual statement 가 존재하며, 이는 사용자들로 하여금 잘못된 정보의 전파 등의 문제를 야기할 수 있다. 이에 논문에서는 동일한 질문에 대한 LLM 의 답변 샘플과 외부 지식을 활용한 Hallucination Detection 방법을 제안한다. 제안한 방법은 동일한 질문에 대한 LLM 의 답변들을 이용해 일관성 점수(Consistency score)를 계산한다. 거기에 외부 지식을 이용한 사실검증을 통해 사실성 점수(Factuality score)를 계산한다. 계산된 일관성 점수와 사실성 점수를 활용하여 문장 수준의 Hallucination Detection 을 가능하게 했다. 실험에는 GPT-3 를 이용하여 WikiBio dataset 에 있는 인물에 대한 passage 를 생성한 데이터셋을 사용하였으며, 우리는 해당 방법을 통해 문장 수준에서의 Hallucination Detection 성능이 baseline 보다 AUC-PR scores 에서 향상됨을 보였다.

1. 서론

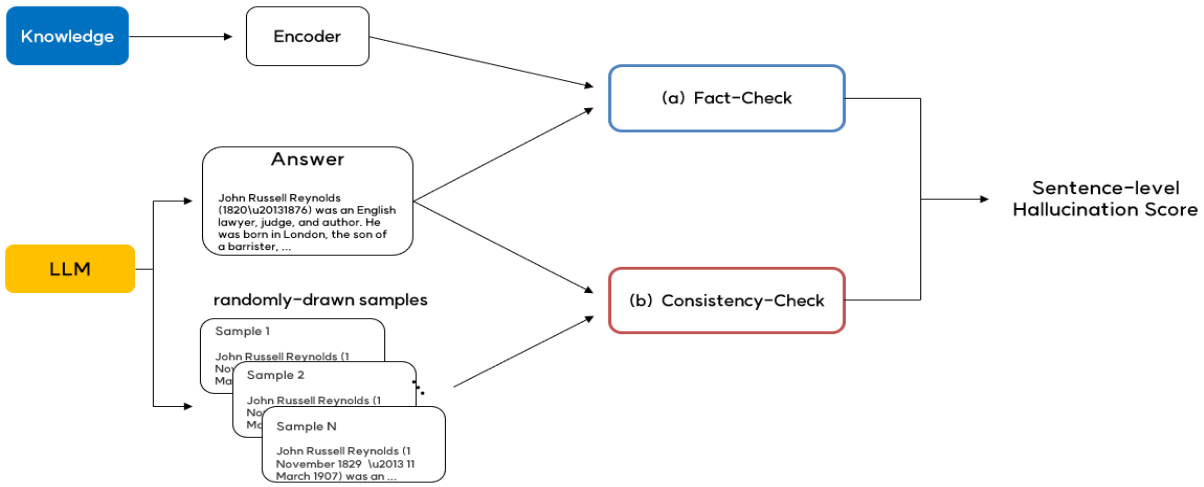
최근 GPT-3[1], LLaMa[2]와 같은 생성형 거대 언어모델(Generative Large Language Model)들이 공개되었고, 해당 모델들은 여러 사용자들의 다양한 질문에 대해 유창하고 사실적인 답변을 생성한다는 이유로 많은 관심을 받고 있다. 또한, 이러한 모델들은 정보 검색, 문서 요약 등 다양한 분야에서 활용되고 있다.

비록 LLMs 가 유창하고 사실적인 답변을 생성한다고 하지만, LLMs 의 답변에는 종종 Inconsistent content 와 non-factual statement 가 포함되는 경우가 있다. 이는 사용자들로 하여금 잘못된 지식을 받아들이고 전파하게 되는 등의 문제를 야기할 수 있다. 이로 인해 발생하는 피해를 예방하기 위해서는 LLMs 의 답변에서의 Hallucination Detection 에 관한 연구가 필요하다.

Inconsistent content 란 생성된 답변에 있는 문장들의 내용이 서로 연관성이 없는 경우를 말한다. 이러한 답변의 Inconsistency 는 동일한 질문을 사용하여 계속해서 답변을 생성하고, 생성된 답변을 샘플로 활용하여 일관성 점수를 계산하고 이를 통해 일관되지 않은 내용의 문장을 검출할 수 있다. 하지만 이러한 일관성 확인 방법은 외부 지식(External Knowledge)을 사용하지 않기 때문에 답변의 사실성에 대해서는 판단하기 어렵다. 따라서 사실이 아닌 내용을 포함하는 non-factual statement 의 검출을 위해서는 외부 지식을 활용한 사실 검증이 필요하다.

이에 본 연구에서는 문장 수준의 Hallucination Detection 을 위해 일관성 확인과 외부 지식을 활용한 사실 검증을

* 교신저자(Corresponding Author)



(그림 1) 제안된 Hallucination Detection 구조

동시에 진행하는 방법을 제안한다.

먼저, 문장 수준의 일관성 확인을 통해 답변에서 각 문장들의 일관성 점수(Consistency score)(그림 1. b)를 계산한다. 그런 다음 외부 지식을 활용해 사실성 점수(Factuality score)(그림 1. a)를 계산하고, 각각의 점수를 가중합한다. 계산된 점수를 문장 수준의 Fact Score 로 사용하여 Hallucination Detection 을 진행한다.

2. 관련 연구

2.1 Hallucination Detection for Large Language Model

Text generation 분야의 다양한 task 에서의 Hallucination 에 관한 연구와 조사는 활발히 진행되고 있다[3]. Manakul et al. (2023) [4]은 zero-resource black-box approach 를 통한 Hallucination Detection 연구를 진행했다. 해당 접근법은 외부 지식없이 동일 질문에 대해 생성된 답변들을 샘플링하고, 해당 답변에서의 일관성을 계산한다. 하지만 이러한 방법은 외부 지식을 사용하지 않기 때문에 생성된 답변이 실제로 사실인지 판단하지 못하는 문제가 있다.

2.2 Fact Verification

외부 지식을 활용하여 생성된 답변의 Hallucination 을 검증하는 것은 사실 검증(Fact Verification task)과 유사하다. Wadden et al. (2022) [5]은 해당 연구에서는 사실 검증을 위해 사용하는 외부 지식 토큰의 개수가 512 를 초과하는 경우가 있었기 때문에 Longformer model [6]을 encoder 로 사용하여 full-context encoding 을 진행하였다.

본 연구에서 외부 지식으로 사용하는 데이터 또한 512 개 이상의 토큰을 가지는 데이터(Section 4.1)가 존재하기 때문에 Longformer model 로 full-context encoding 을 진행하였다.

3. 방법

Notation. R 은 사용자의 질문을 통해 LLM 이 생성한 답변을 의미한다. 앞과 동일한 질문을 사용하여 추가적으로 생성한 N 개의 답변 샘플들은 $\{S^1, S^2, \ldots, S^N\}$으로 표기한다. 또한, i번째 문장에 대한 문장 수준의 hallucination 점수 $S_{\text{HALLUCINATE}}(i)$는 $S_{\text{HALLUCINATE}}(\text{i}) \in [0.0, 1.0]$의 범위를 가지며 i번째 문장이 hallucination 된 문장이라면 $S(i) \rightarrow 1.0$가 되도록 설계되었다.

3.1. Sentence-level Consistency Score

$\mathcal{B}(.,.)$는 두 문장 사이의 BERTScore 를 의미한다. BERTScore 를 이용하여 추가적으로 생성된 샘플들의 문장 중 가장 점수가 높은 문장들의 평균 점수를 계산하고 이를 일관성 점수($S_{\text{Consistency}}(i)$)로 사용한다(그림 1. b).

$$S_{\text{Consistency}}(i) = \frac{1}{N}\sum_{n=1}^{N} \max_{k}\left(\mathcal{B}(r_i, s_k^n)\right) \quad (1)$$

r_i는 R에 있는 i번째 문장을 나타내고, s_k^n은 n번째 샘플 S^n의 k번째 문장을 나타낸다. (1)은 일관성 확인을 통해 r_i가 포함하는 정보가 여러 번 생성된 답변에서 일관되게 나타난다면 해당 정보는 사실일 것이라 가정한다. 반대로 해당 정보가 생성된 답변에서 일관되지 않고 서로 다르게 나타난다면 해당 문장이 hallucination 되었을 거라 가정한다.

3.2. Sentence-level Factuality Score

$\mathcal{C}(.,.)$는 코사인 유사도를 의미한다. r_i와 full-context encoding 이 진행된 외부지식 벡터 K사이의 코사인 유사도를 계산하고, 이를 사실성 점수($S_{Fact}(i)$)로 사용한다(그

림 1. a).

$$S_{Fact}(i) = \mathcal{C}(K, r_i) \quad (2)$$

3.3. Sentence-level Hallucination Score

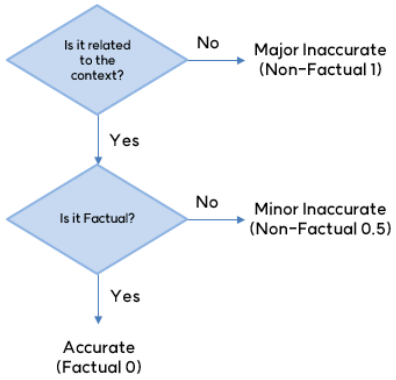
식 (1)을 통해 계산된 일관성 점수($S_{\text{Consistency}}(i)$)에는 외부 지식이 사용되지 않기 때문에 해당 문장이 실제로 사실인지 판단할 수 없다. 때문에 식 (2)를 통해 계산된 문장 수준의 사실성 점수($S_{\text{Fact}}(i)$)를 일관성 점수와 가중합하여 문장 수준의 Hallucination 점수($S_{\text{HALLUCINATION}}(i)$)를 계산한다. 가중치($\beta_1$, β_2)는 각각 0.5, 0.5 로 설정하여 계산하였다.

$$S_{HALLUCINATION}(i) = 1 - \left(\beta_1 \cdot S_{Consistency}(i) + \beta_2 \cdot S_{Fact}(i)\right) \quad (3)$$

식 (3)을 통해 계산된 점수를 통해 최종적으로 문장 수준의 Hallucination Detection 을 진행한다.

4. 실험

4.1. 데이터셋



(그림 2) 데이터 주석처리 과정

실험에는 Manakul et al. (2023) [4]의 연구에서 사용한 데이터셋을 사용하였다. 사용된 데이터는 기본적으로 두 단계를 거쳐 만들었다. 먼저 Wikibio dataset[7]에 있는 인물에 대한 Wikipedia 형식의 글을 GPT-3 를 이용하여 생성한다. 그 다음 일일이 문장 단위로 주석처리를 진행했다 (그림 2). 각 주석의 기준은 아래와 같다.

- Major Inaccurate (Non-Factual, 1) : 문장이 포함하는 내용이 주제와 관련이 없는 경우
- Minor Inaccurate (Non-Factual, 0.5) : 문장이 포함하는 내용 중 사실이 아닌 내용이 있지만, 전체적으로 주제와 관련된 내용을 포함하는 경우
- Accurate (Factual, 0) : 문장이 포함하는 내용이 사실인 경우

실험에 사용한 외부 지식에는 <표 1>에서 볼 수 있듯이 512 개 이상의 토큰을 가지는 데이터가 약 15.6%비율로 존재하기 때문에 일반적인 transformer 기반의 언어 모델인 BERT 나 RoBERTa 가 아니라 Longformer model 을 통해 full-context encoding 을 진행했다.

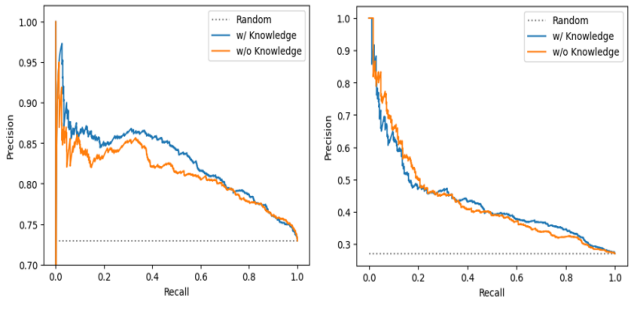
<표 1> Wikipedia text 데이터 통계

#Passages	Avg of tokens	> 512 tokens
238	376.6	15.6%

4.3. 결과

실험 결과에 앞서 먼저 본 연구에서 제안하는 방법이 문장의 사실성을 파악하는 능력을 평가하기 위해 레이블을 두 개의 클래스로 재 분류했다. Major-inaccurate 레이블과 minor-inaccurate 레이블을 전부 *non-factual* 클래스로 묶어주었고, accurate 레이블을 *factual* 클래스로 변경해주었다.

(그림 3)과 <표 2>에서 볼 수 있듯이 외부 지식없이 일관성 점수만을 활용하여 문장 수준의 hallucination detection 을 진행한 경우 보다 제안하는 방법을 통해 문장 수준의 hallucination detection 을 진행할 때 성능이 향상됨을 알 수 있다. 특히, *non-factual* 클래스 검출에서의 성능 향상이 *factual* 클래스 검출에서의 성능 향상보다 큰 것을 보아 보다 정확한 hallucination detection 을 위해 외부 지식 정보가 필요함을 알 수 있다.



(a) Non-Factual Sentences (b) Factual Sentences

(그림 3) PR-Curve of detection *non-factual* and *factual* sentences

<표 2> AUC-PR for sentence-level detection tasks.

Method	Sentence-level (AUC-PR)	
	NonFact	Factual
Random	72.96	27.04
w/o Knowledge	81.20	43.41
w/ Knowledge	**82.66**	**43.42**

5. 결론

본 연구에서는 LLMs 가 사용자의 질문을 받아 생성하는 답변에서의 Hallucination 검출을 위한 방법을 제안한다. 일관성 확인(Consistency-Check)방식을 통해 생성된 답변들이 일관된 정보를 포함하는지에 관한 점수를 계산하고, 추가로 생성된 답변이 가지고 있는 정보가 실제 지식과 비교하여 얼마나 사실적인지를 나타내는 점수를 답변과 외부 지식과의 유사도를 바탕으로 계산하여 이용함으로써 일관성 확인 방식만을 사용한 Baseline 보다 AUC-PR score 에서 성능이 향상됨을 보였다.

6. Acknowledgement

본 논문은 2023 년도 정부(교육부)의 제원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2021R1G1A1006381)

참고문헌

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020. 33:1877–1901.

[2] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023)

[3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv*., 2023. 55(12)

[4] Manakul, Potsawee, Adian Liusie, and Mark JF Gales. "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models." *arXiv preprint arXiv:2303.08896* (2023).

[5] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States. Association for Computational Linguistics. 2022. pages 61–76,

[6] Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." *arXiv preprint arXiv:2004.05150* (2020).

[7] Rémi Lebret, David Grangier, and Michael Auli. Neural Text Generation from Structured Data with Application to the Biography Domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. Association for Computational Linguistics. 2016. pages 1203–1213