대형 언어 모델의 질의응답 생성 텍스트 환각 분류 및 텍스트 평가 척도의 한계 탐구

배수현[O], 이동헌
고려대학교 수학과
{baeshstar, holy}@korea.ac.kr

Exploring Hallucination Types in Question-Answering Generation and Limitation of Text Evaluation Metrics

SuHyun Bae[O], Donghun Lee
Department of Mathematics, Korea University

요 약

대형 언어 모델의 생성 텍스트는 환각 현상 때문에 신뢰성 및 안정성 문제가 있다. 질의응답에서의 환각을 판별하기 위해 생성 텍스트 환각을 총 5가지로 분류하는 방식을 제안하고 BLEU, METEOR 그리고 ROUGE 등 기존 자연어 처리 성능 척도가 이를 잘 판별하는지 실험하였다. 실험 결과, 각각이 잘 판별하지 못하는 유형이 명확히 존재하였다. 이를 통하여, 본 논문에서 정의한 5가지 유형의 환각 판별을 잘하는 새로운 척도의 필요성과 그 설계 방법을 시사한다.

1. 서 론

최근 몇 년 동안, 인공지능과 자연어 처리(Natural Language Processing) 기술의 눈부신 발전은 ChatGPT로 대표될 수 있는 대형 언어 모델들이 대거 개발되었다. 이러한 언어 모델들은 번역뿐만 아니라 질의응답, 텍스트 생성 등 NLG(Natural Language Generation) 즉, 다양한 언어 생성 작업을 마치 사람이 한 것처럼 탁월하게 잘 해낸다고 알려져 있다[1].

그러나, 이렇게 높은 수준의 결과물을 낼 수 있음에도 불구하고 모델들은 가끔 환각(Hallucination)이라는 오답을 내곤 한다. 환각이란 모델이 생성한 답이 사실과 다르거나, 맥락에서 벗어난 답인 경우를 말한다[2]. 이러한 환각 때문에 AI 생성물을 의학, 수학 등의 전문 분야에서 사용하기에는 신뢰성과 안정성 문제가 발생한다[3]. 이 때문에 생성된 자연어 텍스트의 환각 유형을 분류하고 처리하는 것은 전문 분야에서 AI 활용성을 높이는 데에 매우 중요하다.

본 논문은 2. 에서 모델 생성 텍스트에서 관찰되는 다양한 환각 유형을 탐색한다. 기존 연구 [2]는 NLG에서의 환각을 크게 Intrinsic, Extrinsic 두 가지 유형으로 분류한다. Intrinsic 유형은 모델 추론에, 출처에서 확인 가능한 오류가 들어있는 것이다. Extrinsic 유형은 모델 추론이 출처에서 확인이 불가능한 경우이다.

하지만 이 방식으로는 질의응답에서 일어나는 모든 경우의 환각 현상을 전부 분류하기엔 역부족이다. 그 이유는 첫 번째, 이 논문에서는 사람도 한눈에 알아보기 힘든 교묘한 환각만을 취급하였기 때문에 완전한 문장이 아닌 단순한 환각들을 분류하지는 않았기 때문이다. 두 번째, 정보 출처와 모델 추론만을 비교하였기 때문이다. “SQuAD”, “WikiQA” 등 다른 QA 데이터셋 같은 경우 정보 출처뿐만 아니라 질문과 그에 따른 모범 답안도 같이 포함돼 있다[4, 5]. 질문과 모범 답안도 같이 고려되어야 모델의 답안이 얼마나 정답과 괴리가 있는지 비교할 수 있을 것이다.

그다음 4. 에서 기존 텍스트를 평가하는 데 사용하는 METEOR 및 BLEU 같은 기존 척도들이 각 환각 유형을 잘 탐지하는지 실험한다. 이때, 세 척도 이외에 “BERTscore” 나 “SelfcheckGPT”같은 언어 모델 및 신경망 기반 수단은 고려하지 않았다[6, 7]. 가능한 적은 연산 자원으로 환각 판별을 수월하게 할 수 있는지 실험하기 위함이다. 이를 통해 기존 척도들의 환각 판별 성능을 비판적으로 검토하고 내재한 한계를 밝히고 개선 방안을 모색하는 것을 목표로 한다.

2. 환각 유형 분류

이 단원에서는 영문으로 된 모델 생성 텍스트에서 나타날 수 있는 환각을 분류하는 절차를 새로 소개한다. 새로운 분류 절차에 의하면 환각 유형은 총 5가지이다. 절차를 진행하며 어떻게 유형을 분류하는지 설명한다.

첫 번째 절차는 모델 답안이 제대로 문장 형식을 지니는지 판별한다. 만약 그렇지 않다고 판단되면 유형1로 분류된다. 유형1은 문장 구조가 엉망이거나, 5형식 중 하나가 아닌 유형이다. 예를 들면 동사가 없고 명사만 반복되는 경우, 완전한 문장이 아니므로 유형1로 분류된다. 보통 모델의 추론 성능이 좋지 않을 때 발생하는 오류이다.

두 번째 절차는 모범 답안, 질문과 같은 내용에 관한 얘기를 하고 있는지 판별한다. 그렇지 않다고 판단되면 유형2 동문서답형 환각으로 분류한다. 주로 질문, 모델 추론 간 문장 내용의 연관성이 떨어지는 경우다.

세 번째 절차는 출처와 모범 답안, 모델 추론의 키워드를 비교한다. 그 이전에 답안과 추론 간 키워드의 표제어, 어간, 동의어를 고려해 토큰 alignment를 확인한다. 그 후 모범 답안에 없는 키워드가 모델 추론에 있으면, 유형3인 사족 및 의미 변동형으로 분류한다.

세번째 절차에서 유형3으로 분류되지 않은 것 중 모범 답안에 있는 키워드가 모델 추론에 없는 경우, 유형4 정보 구현 부족형으로 분류된다. 이 경우, 키워드 하나가 없어도 의미가 매우 달라지는 경우도 있다.

마지막 분류 절차는 Cross 존재 여부이다. Cross는 [8]에서 소개된 개념으로 문법적 인과 순서의 변동을 판별 가능하다. 키워드 매칭 했을 때 Cross를 생성하여 인과 관계에 변동이 있는 경우, 유형5로 분류한다.

표 1: 유형별 예시

유형1	질문	Where is the cat?
	정답	The cat is in the house
	오답	The house the house the house
유형2	질문	What is your favorite food?
	정답	My favorite food is pizza
	오답	I am hungry
유형3	질문	How many cats are in there?
	정답	There are seven cats
	오답	There are two cats
유형4	질문	Tell me about Seoul
	정답	Seoul is one of the biggest cities in the world
	오답	Seoul is the biggest city in the world
유형5	질문	What happened to the deer?
	정답	The hunter shot the deer
	오답	The deer shot the hunter

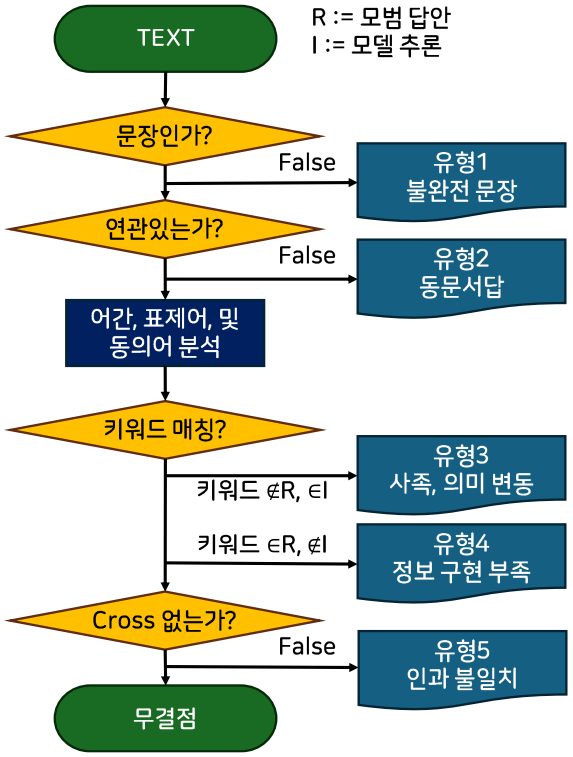

그림 1: 환각 분류 순서도

3. 텍스트 품질 평가 척도

3.1. BLEU

BLEU는 번역 텍스트의 품질을 평가하기 위해 개발된 척도로, 생성된 텍스트와 실제 텍스트 간의 n-gram 일치도를 측정한다[9]. 높은 BLEU 점수는 더 자연스러운 번역을 나타내며, BLEU-n들의 기하 평균을 계산하여 최종 점수를 계산한다. 이 실험에서는 BLEU-1, BLEU-2, 및 BLEU-3을 통해 실험하였다.

3.2. METEOR

METEOR는 번역 품질을 평가하는 다른 척도로, unigram 단위의 단어 정렬 및 동의어 사용을 고려하여 번역 품질을 평가한다[8]. precision만 고려하는 BLEU와 달리 precision과 recall을 둘 다 고려한다.

3.3. ROUGE

ROUGE는 문단 요약 품질을 평가하는 데 사용되는 척도로, 생성된 텍스트와 실제 요약 간의 유사성을 측정한다[10]. 이 연구에서는 가장 긴 공통 문장(LCS)를 사용하는 ROUGE-L을 사용해 실험하였다. n-gram의 크기를 사전에 정의하지 않는 장점이 있기 때문이다[11].

4. 환각 분류 성능 평가 실험

4.1. 데이터셋 및 실험 설정

HaluEval에서 제시한 "`QA_data`" 데이터셋을 사용했다[12]. 이 데이터셋은 정보 출처인 knowledge, 질문 question, 모범 답안 right_answer과 환각 답변인 hallucinated_answer 항이 주어져 있다. 하지만 대부분의 정답이 문장이 아닌 단답형 하나의 단어로 되어 있었기 때문에 "`Vicuna`" 모델을 통해 단어로 된 정답을 문장으로 다시 만드는 데이터 보강을 진행하였다[13].

데이터 보강 후, `nltk` 라이브러리에서 BLEU, METEOR 점수를, `rouge_scorer` 라이브러리에서 ROUGE-L 점수를 계산하여 정답과 환각 답안을 비교하였다. 그 후, 점수가 기준점 0.7 이상인 경우를 모아 앞서 정의한 5가지 환각 분류에 부합하는지 확인하였다.

기준점을 0.7로 설정한 이유는 [11]에서 사람의 평가와 척도를 비교할 때 사람의 평가 점수를 1, 2, 3, 4, 5점으로 했을 때, 중간값인 3점을 넘으면 환각이라고 판단했기 때문이다. 그래서 3점과 4점의 중간값 3.5점을 각 척도 점수의 범위인 [0,1]에 맞춰 정규화한 0.7 이상의 점수들을 환각이라고 판단하였다.

4.2. 실험 결과

위 방법에 따라, **4.1.** 의 샘플 총 10,000개 중 무작위 선별한 1,150개에 대해 데이터 보강 및 전처리를 적용하고 기준 점수를 0.7로 설정하여 BLEU, METEOR, ROUGE-L 점수 중 하나라도 그 이상인 경우들만을 골라 분석 대상 81개를 얻을 수 있었다.

분석 대상 81개를 **2.** 에서 소개한 환각 분류 절차에 의거, 유형을 판별하고 유형별 척도 점수의 평균을 계산했다. 이때, 짧은 유형3은 환각 답이 질문에 맞는 키워드만 쓰여 있는 단답형인 경우이다. 이 경우 동사가 없어 완전한 문장은 아니지만 단답으로 취급될 수 있으므로 유형1 대신 짧은 유형3으로 세부 분류하였다. 81개 중 짧은 유형3은 19개, 긴 유형3은 46개, 유형4는 12개, 유형5는 4개로 분류되었다. 유형1, 2가 없는 이유는 두 유형의 샘플은 전부 기준점 이하의 점수가 나왔기 때문이다.

상기 이유로 자료에는 유형3, 4, 5만을 표기하였다. 또한 BLEU-2와 BLEU-3도 81개에서 전부 0점이 나와 BLEU-1만을 표기하였다.

BLEU-1 점수는 짧은 유형3을 평균 점수 0.7421을 기록하며 잘 잡아내지 못하였으며 긴 유형3, 유형4, 유형5는 0.2921, 0.3990, 0.3030으로 안정적으로 잡아내지는 못하였다.

METEOR 점수는 짧은 유형3은 0.0481로 매우 잘 잡아냈지만, 유형5, 긴 유형3, 유형4 순으로 각각 평균 0.7857, 0.7109, 0.5866으로 잘 잡아내지는 못하였다.

ROUGE-L 점수는 짧은 유형3은 0.0563으로 매우 잘 잡아냈지만, 긴 유형3, 유형5, 유형4 순으로 각각 0.7298, 0.6888, 0.6761로 잘 잡아내지 못하였다.

표 2: 유형별 척도 점수 상세표

	짧은 유형3	긴 유형3	유형4	유형5
BLEU-1	0.7421	0.2921	0.3990	0.3030
METEOR	0.0481	0.7109	0.5866	0.7857
ROUGE-L	0.0563	0.7298	0.6761	0.6888

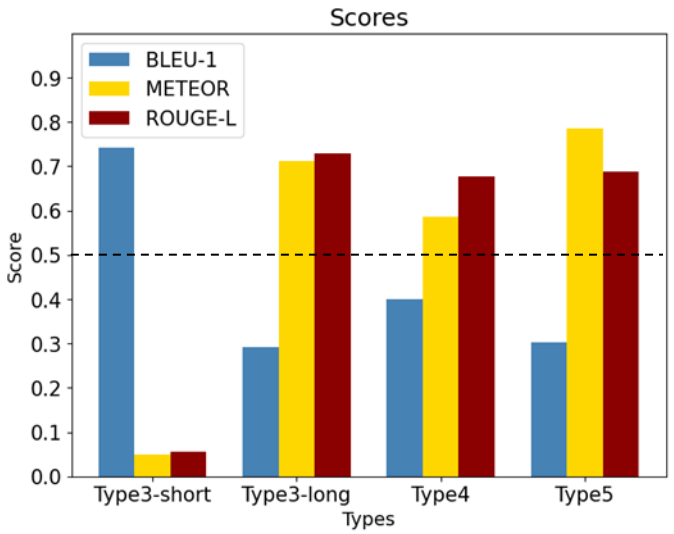

그림 2: 유형별 척도 점수 그래프

5. 결론 및 향후연구

4.2.을 통하여, 각각의 척도들은 위에서 정의한 환각 유형들을 전부 안정적으로 잡아내지는 못한다는 것을 알 수 있었다. BLEU는 상대적으로 긴 유형3, 유형4, 유형5를 잘 판별하고, METEOR와 ROUGE-L은 짧은 유형3을 매우 잘 판단하였다.

위와 같이 각 척도가 잘 판별하는 유형이 명확히 존재한다. 그러므로 세 가지 척도를 잘 조합하고 개선하여 모든 유형을 잘 판별하고, 연산 자원도 가능한 한 적게 드는 새로운 척도를 추후 설계하고자 한다.

나아가, 텍스트뿐만 아니라 수학 문제 풀이 과정 생성 모델 같이 특정 형식을 내놓는 모델에도 이런 척도를 적용하여 환각을 판별할 수 있을지 연구하고자 한다.

6. 참 고 문 헌

[1] J. Li, T. Tang, W. X. Zhao, J. -Y. Nie, & J. -R. Wen, "Pre-trained Language Models for Text Generation: A Survey". *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 4492-4499, 2022.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, & P. Fung, "Survey of Hallucination in Natural Language Generation". *ACM Computing Surveys*, vol 55, no. 12, pp. 1-38, 2023.

[3] M. Sadat, Z. Zhou, L. Lange, J. Araki, A. Gundroo, B. Wang, R. R. Menon, M. R. Parvez, & Z. Feng, "DelucionQA: Detecting Hallucinations in Domain-specific Question Answering". *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 822-835, 2023.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, & P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text". *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383-2392, 2016.

[5] Y. Yang, W. -T. Yih, & C. Meek, "WikiQA: A Challenge Dataset for Open-Domain Question Answering". *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013-2018, 2015.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, & Y. Artzi, "BERTScore: Evaluating Text Generation with BERT". *International Conference on Learning Representations*, 2020.

[7] P. Manakul, A. Liusue, M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models". *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004-9017 2023.

[8] S. Banerjee, & A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65-72, 2005.

[9] K. Papineni, S. Roukos, T. Ward, & W. -J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation". *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318, 2002.

[10] C. -Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries". *Text Summarization Branches Out*, pp. 74-81, 2004.

[11] A. Chen, G. Stanovsky, S. Singh, & M. Gardner, "Evaluating Question Answering Evaluation". *Proceedings of the Second Workshop on Machine Reading for Question Answering*, pp. 119-124, 2019.

[12] J. Li, X. Cheng, X. Zhao, J. -Y. Nie, & J. -R. Wen, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models". *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449-6464, 2023.

[13] W. -L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, & E. P. Xing, (2023, Mar. 30). Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna (downloaded 2024, Apr. 17)