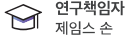


언어 모델의 지식을 활용한 추론 기술 개발



연구책임자
제임스 손



소속학과
김재철시대학원



참여연구원

안나민, Max Glockner, Iryna Gurevych,
이노아, 오필훈, 이예준, Ieva Staliūnaitė,
Gisela Vallejo, Andreas Vlachos



연구실 홈페이지
https://xfact.net

대형 언어 모델(LLM, Large Language Models)이 일상 생활에서 점점 더 많이 사용됨에 따라, 이들 언어 모델의 추론 일관성과 외부 지식 및 사실들이 제시될 때의 추론 방식을 검증하는 것은 매우 중요하다. 본 연구에서는 언어 모델 내의 사실 기반 추론에 대해 조사하였으며, 그 중에서도 질의 응답과 사실 검증과 같은 작업들에서 대해 대형 언어 모델이 다양한 형태의 지식을 어떻게 활용하는지에 대해 실험을 진행하였다. 연구팀은 대형 언어 모델이 사실 검증 작업 중 상충하는 증거나 지식을 바탕으로 모호성을 어떻게 처리하는지에 대해 실험을 진행하였다. 이를 위해, 불확실한 정보에서 맥락적 이해와 추론을 요구하는 새로운 사실 검증 데이터셋, AmbiFC를 제작하였다. 또한 불필요한 정보를 필터링하고 추출된 정보를 요약하는 등의 보다 효율적인 정보 통합 방법을 제안함으로써, 더 적은 검색 결과를 기반으로 모델의 사실 정확성과 실행 시간의 효율을 향상시켰다.

연구배경

대규모 언어 모델(LLM, Large Language Model)의 정보 저장 및 추출 능력은 일관되고 유창한 답변 생성을 가능케 한다. 그러나 한번 매개변수에 학습되어 내재된 상태의 정적 정보는 검토 및 수정이 불가능하며, 이는 오래되거나 변경된 사실을 영구화하여 '환각'과 같은 비사실적 생성으로 이어지기도 한다. 사실 확인 또는 질의응답과 같이 높은 수준의 지식 엄밀성이 요구되는 작업의 경우에는 응답이 사실인지 확인하는 데 필요한 증거를 제공할 수 있는 지식 코퍼스에서 정보를 검색을 필히 해야한다. 모델에 관련 정보를 제공하면 답변의 정확도를 향상시킬 수 있지만 질문과 무관하거나 과도한 정보제공은 오히려 모델 내에서 혼란을 초래해 잘못된 답변을 유도할 수 있다. 그렇기에, 사실 확인과 같은 도전적인 과제에 이를 적용하려 한다면, 검색 결과를 추출된 정보의 충돌이 있는 경우에도 효과적으로 추론할 수 있어야 한다.

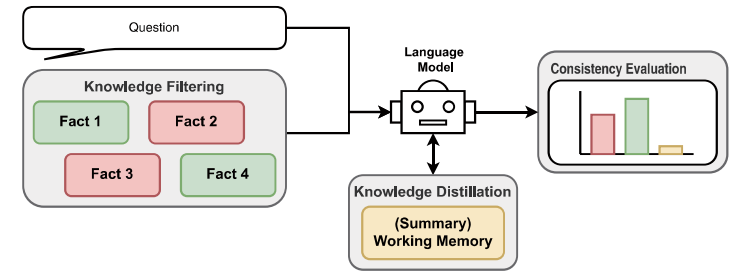
연구내용

본 연구에서는 증거에 대한 주장의 사실 검증을 위해, 모델을 훈련하고 검증하기 위해 새로운 데이터셋인 AmbiFC를 제작하였다. 해당 데이터셋은 Wikipedia 에서 추출한 50,000 여개의 단락들을 활용하여, 총 10,000개의 검증할 수 있는 텍스트 기반의 주장들로 구성되며, 상황과 증거 선택에 따라 사실 검증의 결과가 바뀌는 정황적 추론을 요구하는 차별화된 특징을 지니고 있다. 또한 연구팀에서는 사실 추론 작업을 수행할 때 언어 모델의 안정성과 자체 일관성을 연구했으며, 모호하지 않은 경우에도 언어 모델이 인간과는 다른 다양한 추론을 보여준다는 것을 밝혔다. 마지막으로 질의 응답, 사실 검증과 같은 지식 집약적 추론 작업들을 위해 언

기대효과

검색 증강 생성(RAG)과 같은 방법은 변화하는 세계 지식을 기반으로 적용 가능한 LLM을 구축하는 데 인기를 얻고 있지만, 검색된 정보가 사실적 답변을 생성하는 데 효과적으로 사용되는 것이 중요하다. 본 연구에서는 검색 정보와 대규모 언어 모델에서 생성된 응답 간의 상호 작용이 복잡하게 이루어짐을 보인다. 또한, 모델이 일반적으로 사용되는 증거의 일부만으로도 효율적이고 정확한 질문 답변을 달성할 수 있음을 보여주며, 양에 따른 정밀도의 효율성 향상을 보여준다. AmbiFC 벤치마크는 미묘한 이해와 해석이 가능한 보다 정교한 모델 개발을 장려할 것으로 기대되며, 사실 추론 작업에 대한 LLM 일관성 및 지식 통합에 대한 평가와 함해져 대규모 언어 모델의 신뢰성을 향상시킬 것으로 기대된다.

Diagram



연구성과

M Glockner, I Staliūnaitė, J Thorne, G Vallejo, A Vlachos, I Gurevych: AmbiFC: Fact-Checking Ambiguous Claims with Evidence, Transactions of the Association for Computational Linguistics, 2024
N Lee, NM An, J Thorne, Can Large Language Models Capture Dissenting Human Voices?, Empirical Methods for Natural Language Processing 2023.
P Oh, James Thorne, Detrimental Contexts in Open-Domain Question Answering, Findings of Empirical Methods for Natural Language Processing 2023
Y. Lee, P Oh, James Thorne, Knowledge Corpus Error in Question Answering, Findings of Empirical Methods for Natural Language Processing 2023

연구비 지원

KAIST New Faculty Startup Finding (G04220030)
The AmbiFC dataset was collected with the support of Google