LLM의 신뢰도 향상을 위한 GraphRAG

방대호, 최현민, 강재모*

경북대학교

dkrlwnstn1@knu.ac.kr, iissaacc@knu.ac.kr *jmkang@knu.ac.kr

Enhancing the Reliability of LLM through GraphRAG

Daeho Bang, Hyunmin Choe, Jaemo Kang

KyungPook National Univ.

요 약

본 논문은 대규모 언어 모델의 단점을 보완할 수 있는 검색 증강 생성(Retrieval-Augmented Generation)의 최근 트렌드인 그래프 구조를 채용한 새로운 분야인 GraphRAG 시스템에 대해 고찰하였다.

Ⅰ. 서 론

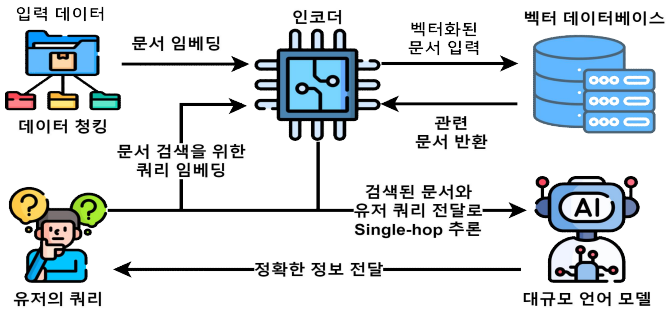
ChatGPT와 같은 대규모 언어 모델(LLM)이 급속히 발전하며, LLM은 사용자의 질문에 대해 신속하고 일관된 답변을 제공할 수 있는 환경이 조성되었다. 그러나 이러한 모델들은 'AI 환각(AI hallucination)' 문제를 가지고 있다. 이는 LLM이 학습하지 않은 정보를 잘못된 방식으로 생성하여 응답하는 현상으로, 특히 신뢰도가 중요한 응용 분야에서 심각한 문제로 대두되고 있다. 예를 들어, 사용자가 실존하지 않는 사건에 대해 질문하면, 모델은 실제로 존재하는 것처럼 허구의 응답을 생성할 수 있다.

이러한 문제를 해결하기 위해 검색 증강 생성(Retrieval-Augumented Generation, RAG) 시스템이 도입되었다[1]. RAG는 외부 데이터 소스를 검색하여 답변의 정확성을 높이려는 접근법이다. 그러나 RAG 역시 특정 질문에 대해 단편적인 정보만을 제공하거나 문맥적인 의미를 충분히 반영하지 못하는 경우가 발생할 수 있다.

이를 보완하기 위해 GraphRAG가 제안되었다[2]. GraphRAG는 RAG의 한계를 극복하기 위해 지식 그래프를 결합한 접근법으로, 문서 내의 개체 간 관계를 분석하여 보다 종합적이고 풍부한 정보를 제공한다. 본 논문에서는 GraphRAG의 개념과 구현 방법을 소개하고, 이를 활용한 실험을 통해 RAG와의 성능 차이를 비교했다. 결과적으로, GraphRAG는 문맥적으로 더 포괄적이고 신뢰성 있는 응답을 생성할 수 있음을 본 연구에서 입증하였다.

Ⅱ. 본론

일반 LLM은 주어진 질문에 대해 학습된 패턴을 바탕으로 답변을 생성한다. 이 과정에서 모델은 방대한 양의 데이터를 활용하여 자연스러운 문장을 생성할 수 있지만, 데이터 내에서 정확한 정보를 선택하지 못하거나, 학습하지 않은 정보에 대해서는 거짓 정보를 생성할 위험이 있다. 이러한 AI 환각 문제는 특히 신뢰도가 중요한 응용 분야에서 큰 문제로 대두되고 있다. RAG는 이러한 문제를 해결하기 위해 제안된 기술들 중 하나다. RAG는 외부의 신뢰할 수 있는 데이터를 대규모 문서 집합에 입력한다. RAG를 통한 질의가 요청되면, RAG 시스템은 대규모 문서 집합에서 해당 정보를 포함한 문서를 검색하고, 이를 기반으로 답변을 생성한다. 이로써 RAG 알고리즘을 사용해 도출한 응답은 더 신뢰할 수 있게 된다[그림 1]. 그러나 RAG는 특정 질문에 국한된 검색 결과(Single-hop)를 활용하기 때문에, 전체 문서 집합에서 중요한 정보를 놓치거나 행간에 존재하는 문맥적 의미를 놓칠 수 있다.

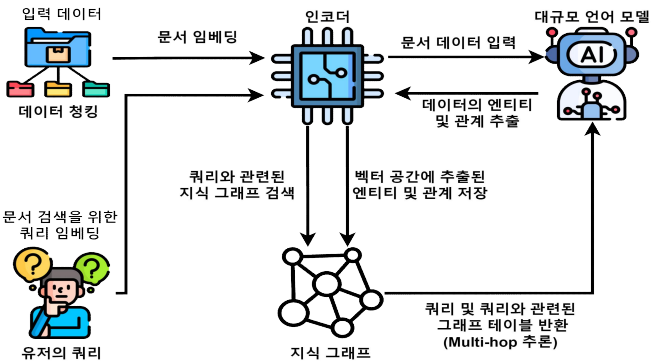


[그림 1]

GraphRAG는 RAG의 이러한 단점을 보완하기 위해 탄생했다. 지식 그래프를 활용하여, 문서 집합 내의 개체 간 관계를 파악(Multi-hop)하고 이를 바탕으로 답변을 생성한다. 이를 통해 RAG보다 더 종합적이고 다양성이 풍부한 응답을 생성할 수 있다. Gra phRAG의 핵심 강점은 지식 그래프를 활용함으로써 단순한 정보 검색에 그치지 않고, 문서 내의 여러 개체와 그들 간의 관계를 파악하여 응답을 생성할 수 있다는 점이다.

이 과정에서 GraphRAG는 문서 내의 개체 간 관계를 기반으로 더욱 풍부하고 깊이 있는 정보를 제공할 수 있다. 예를 들어, 단순히 "대한민국 경제 상황" 에 대한 질문이 주어졌을 때, GraphRAG의 Multi-hop 추론으로 경제 관련 주요 사건뿐만 아니라 정치적 결정, 사회적 반응, 국제적 맥락 등을 모두 통합하여 보다 포괄적이고 다층적인 응답을 생성할 수 있다. 이와 같은 접근법은 특히 복잡한 주제나 다차원적인 문제를 다룰 때 그 진가를 발휘한다. 요약하자면, 일반 LLM 및 RAG 시스템은 개별 문장이나 단편적인 정보에 초점을 맞추어 응답하는 반면에 GraphRAG는 이 정보들

을 연결하고 상호 의존성을 고려하여 문맥적으로 일관된 응답을 제공할 수 있다. 이는 단순한 정보 제공을 넘어서 사용자가 질문의 배경에 대해 더 깊이 이해할 수 있도록 돕는 기능을 수행할 수 있다.[그림 2]



[그림 2]

본 연구에서는 RAG를 위한 벡터 저장소와 GraphRAG를 위한 지식 그래프를 생성하여 실험적으로 그들의 성능을 직접 비교하고자 한다. 중점적으로 확인할 부분은 환각 증상의 여부, 응답의 정확성, 문맥의 포괄적인 이해도이다. 또한 두 방법론 간의 장단점에 대해 비교하겠다.

Ⅲ. 실험 및 결론

본 연구에서는 RAG를 위한 벡터 저장소와 GraphRAG를 위한 지식 그래프를 생성하여, 두 방법론의 성능을 비교했다. 주요 비교 요소는 환각 증상의 여부, 응답의 정확성, 그리고 문맥의 포괄적인 이해도였다. 본 실험을 진행하기 위해 다음과 같은 실험 환경을 구성했다. RAG 시스템은 LangChain 모듈을, GraphRAG 시스템은 Microsoft의 GraphRAG 파이프라인을 이용했다. 언어 모델은 120억 개의 파라미터로 학습된 'Mistral NeMo' 모델을, 임베딩 모델은 'BGE-M3' 모델을 채택했다. 전반적인 과정에서 요구된 Video RAM은 10GB 이하였다. 비교 과정을 위한 데이터 세트는 2021년 발행된 대한민국 종합 일간지 및 지역신문의 뉴스 기사들로 이루어진 데이터 세트를 채택했다[3]. 해당 데이터 세트에서 1000편의 기사를 무작위로 선정하여 모은 다음, 각 기사당 하나의 텍스트 파일로 저장하여 총 1000개의 텍스트 파일로 나누어 사용했다.

실험 결과, RAG는 세부적인 정보에 대한 응답에서 우수한 성능을 보였다. 특히, 매우 구체적인 질문에 대해서도 정확하고 상세한 답변을 제공하는 데 강점을 보였다. 반면, GraphRAG는 아주 세부적인 정보를 검색해내기 쉽지 않았다. 이는 GraphRAG가 보다 포괄적이고 문맥적으로 중요한 정보를 중심으로 응답을 생성하기 때문으로 보인다. 환각 증상의 측면에서는 RAG가 상대적으로 더 다루기 쉬웠다. GraphRAG 역시 환각 증상을 걸러낼 수 있었지만, 모델의 크기가 작은 환경에서는 프롬프트에 매우 민감하게 반응하여 적절한 프롬프트는 찾는 데 시간이 오래 걸렸다. 문맥의 포괄적인 이해도에서는 GraphRAG가 RAG보다 훨씬 더 우수한 성능을 보였다. GraphRAG는 문맥적인 의미를 잘 파악하고, 여러 정보 간의 관계를 종합하여 응답을 생성하는 데 뛰어난 능력을 보였다. 반면, RAG는 포괄적인 질문에 대해 환각 증상을 일으키는 경우가 많았고, 질문의 맥락을 충분히 이해하지 못한 채 부분적인 정보만을 제공하는 경향이 있었다.

실험 결과를 종합해보면, RAG는 구체적인 질문에 대해 빠르고 정확한 응답을 제공하는 데 강점을 보였으며, 특히 세부 정보에 대한 응답에서 우수한 성능을 나타냈다. 반면 GraphRAG는 문맥적인 이해도를 기반으로 보다 종합적이고 다차원적인 응답을 생성하는 데 유리했으며, 포괄적인 질문에 대해 더 정확하고 신뢰할 수 있는 응답을 제공했다. 두 방법론의 활용에 소요되는 시간적 자원은 RAG가 압도적으로 효율이 좋았다. GraphRAG와 같은 경우 데이터를 정제하여 그래프로 변환하는 과정은, RAG가 데이터를 벡터화하여 벡터 저장소에 저장하는 시간에 비해 수십 배의 시간을 요구했다.

향후 연구에서는 GraphRAG의 성능을 개선하고 그 적용 범위를 넓히기 위한 여러 가지 접근이 고려될 수 있다. 첫째, 지식 그래프 생성 과정의 효율성을 높이기 위한 알고리즘 개발이 필요하다. 이를 통해 데이터 처리 속도를 향상시키고, 더 큰 규모의 데이터 세트를 실시간으로 처리할 수 있을 것이다. 둘째, GraphRAG의 지식 그래프가 지속적으로 최신 정보로 업데이트될 수 있도록 실시간 데이터 수집 및 통합 시스템을 연구할 수 있다. 이를 통해 변동성이 큰 도메인에서도 최신 정보를 반영한 정확한 응답을 제공할 수 있을 것이다. 셋째, 다양한 도메인에서 GraphRAG의 성능을 테스트하여 그 범용성을 검증하고, 각 도메인에 특화된 지식 그래프를 생성하는 방법을 연구할 필요가 있다. 마지막으로, GraphRAG의 단점을 극복하기 위해 하이브리드 접근법을 고려할 수 있다. 예를 들어, 단순한 질문에는 RAG의 속도와 간결함을 활용하고, 복잡한 질문에는 시간이 걸리더라도 GraphRAG의 깊이 있는 분석을 적용하는 방식으로 두 시스템의 장점을 결합할 수 있을 것이다. 이러한 접근은 실제 응용에서 더 나은 사용자 경험을 제공할 수 있을 것이다.

본 연구를 통해 GraphRAG의 잠재력과 한계를 명확히 분석할 수 있었다. 이 분석을 바탕으로 더 나은 LLM 기반 응답 시스템을 설계하기 위한 기초를 마련했다. GraphRAG는 현재의 기술적 한계를 극복하고, 미래의 AI 응용에서 더욱 신뢰성 있는 정보를 제공하는 데 중요한 역할을 할 수 있을 것으로 기대된다. 앞으로도 여러 분야에서의 지속적인 연구와 개발을 통해 더욱 향상된 성능을 갖춘 지식 기반 LLM 응답 시스템이 등장할 것으로 전망된다.

본 연구는 Microsoft의 공개된 GraphRAG 레포지토리를 기반으로 수행되었습니다. 연구 과정에서 상기 레포지토리를 한국어 데이터에 적합하도록 프롬프트를 수정하였으며, 로컬 환경에서의 테스트를 위해 코드를 추가적으로 조정했습니다. 수정된 코드와 관련 자료는 저희 레포지토리 (https://github.com/richbang/Ko-GraphRAG)에서 확인할 수 있습니다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음" (IITP-2024-2020-0-01808*)

참 고 문 헌

[1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[2] Edge, Darren, et al. "From local to global: A graph rag approach to query-focused summarization." arXiv preprint arXiv:2404.16130 (2024).

[3] AI 허브 뉴스기사 독해 데이터, (https://www.aihub.or.kr/).