RAG 모델의 테이블 데이터 처리 성능 분석 연구

이은빈[1], 배호[2,3]
[1]이화여자대학교 인공지능융합전공 석사과정
[2]이화여자대학교 사이버보안학과 교수
[3]이화여자대학교 인공지능융합전공 교수
eunbinlee@ewha.ac.kr, hobae@ewha.ac.kr

A Study on Improving RAG Model Performance for Tabular Data Analysis

Eun-Bin Lee[1], Ho Bae[2,3]
[1]Dept. of Artificial Intelligence Convergence, Ewha Womans University
[2]Dept. of Cyber Security, Ewha Womans University
[3]Dept. of Artificial Intelligence Convergence, Ewha Womans University

요 약

Retrieval Augmented Generation(RAG)는 Large Language Models(LLM)에서 발생하는 환각 현상, 민감 정보 노출과 같은 한계들을 극복하기 위해 제안된 기술이다. 새로운 정보가 지속적으로 업데이트되는 외부 데이터베이스로 인해 RAG는 최신 정보에 기반한 정교한 답변 생성을 할 수 있어 질의응답이나 대화문과 같은 텍스트 데이터를 다루는 작업에서 강점을 발휘한다. 반면, 테이블 데이터는 행과 열로 구성된 정형화된 데이터로 각 속성 간에 복잡한 관계를 가지고 있어 충분한 문맥 정보를 추출하는데 어려움을 겪으므로 성능 저하 문제가 발생한다. 본 논문에서는 이러한 문제를 해결하기 위해 테이블 데이터를 분석하기 위한 질의 유형을 정보 검색, 특성 분석, 상관관계 분석, 결과 예측, 추세 분석과 같이 5가지로 분류하고, 각 유형에서 모델이 보이는 성능을 평가한다. 특히, 다양한 질의 유형 데이터를 통해 실험을 진행하고 분석하여 질의 유형의 특징과 RAG 성능 간의 관계성을 연구하고자 하며, 이를 통해 테이블 데이터에서의 RAG 성능 개선 방향을 제시하고자 한다.

1. 서론

Large Language Models(LLM)은 AI 기술의 핵심으로 주목받고 있으며, 특히 자연어 처리 분야 중 텍스트 생성, 요약, 번역 등의 작업에서 탁월한 성능을 보이고 있다. 그러나 종종 LLM이 잘못된 정보를 생성하는 환각 현상(Hallucination), 학습 데이터에서 민감한 정보 노출 등 여러 한계가 여전히 존재한다. 이러한 한계로 인해 특정 산업에 최적화되거나 전문적인 지식이 필요한 작업에는 적극적으로 활용되기 어렵다[1].

Retrieval Augmented Generation(RAG, 검색 증강 생성)[2]은 LLM의 여러 문제들을 극복하기 위해 등장한 기술이다. RAG는 인덱싱(Indexing), 검색(Retrieval), 응답 생성(Generation)의 3가지 핵심 구성 요소로 나뉜다. 인덱싱 단계에서 텍스트로 표현할 수 있는 원시 데이터를 검색이 가능한 형태로 구조화하며, 임베딩 과정을 통해 벡터 형식으로 변환해 벡터 데이터베이스에 저장한다. 검색 단계에서는 사용자로부터 입력받은 쿼리를 벡터로 변환한 후, 벡터 데이터베이스에서 유사도를 계산했을 때 가장 높은 유사도를 가진 K개의 문서를 선택한다. 응답 생성 단계에서는 검색된 K개의 문서와 입력 쿼리를 LLM에 모두 전달하여 텍스트 형식의 응답을 생성한다.

RAG는 최신 정보에 기반한 정교한 답변 생성을 가능케 하며, 질의응답(Question-Answering, QA)이나 대화문(Dialog)과 같은 작업에서 강점을 발휘한다[3]. 특히 텍스트 데이터에 대해 탁월한 성능을 보이며, LLM을 단독으로 사용하였을 때 발생할 수 있는 한계점을 보완하고 있다.

그러나 RAG는 테이블과 같은 구조화된 데이터를 다룰 때 성능이 저하된다는 문제가 존재한다. 테이블은 행과 열로 구성된 정형화된 데이터로, 각 속성 간의 복잡한 관계를 가지고 있다. 테이블 데이터는 매우 간결한 정보만 포함하고 있기 때문에 텍스트 데이터에 비해 풍부한 문맥 정보를 제공하지 않아 RAG 모델이 충분한 정보를 추출하는 데 어려움을 겪는다. 테이블 데이터는 다양한 산업에서 많이 활용되는 중요한 데이터 유형이기 때문에 이를 정확

<표 1> 테이블 데이터를 분석하기 위한 질의 유형 개요

질의 유형	질의 목표	질의 예시	성능	
			ROUGE-L	Answer Relevancy
정보 검색	사용자 질의에 맞는 테이블 데이터의 특정 항목 검색	가장 나이가 많은 사람의 학력은 무엇인가?	0.5	0.92
특성 분석	테이블 내의 개별 속성의 통계적 특성을 분석	인종에 따른 소득 차이는 어떻게 되는가?	0.32	0.91
상관관계 분석	테이블 내의 여러 속성 간의 상관관계 분석	연령이 증가함에 따라 소득이 증가하는가?	0.35	0
결과 예측	주어진 테이블 데이터를 바탕으로 미래 경향이나 결과 예측	교육 수준, 직업을 바탕으로 소득을 예측할 수 있는가?	0.62	0.97
추세 분석	시간에 따른 데이터 변화를 분석하여 추세 도출	성별에 따른 소득 증가 패턴이 어떻게 되는가?	0.26	0.96

히 분석하고 해석하는 것은 매우 중요하다.

테이블 데이터를 분석하기 위한 질의는 다양한 유형으로 분류될 수 있으며 각각의 유형에 따라 RAG 모델이 다른 성능을 보일 수 있다. 따라서 질의 유형을 분류하는 작업은 테이블 데이터를 분석하는 과정에서 모델의 성능을 세밀하게 평가하고 개선하기 위해 필수적이다.

본 논문의 2장에서는 RAG를 활용하여 테이블 데이터를 분석하기 위한 질의 유형을 체계적으로 분류한다. 또한, 각 질의 유형에서 모델이 보이는 성능을 분석한다. 이를 통해 테이블 데이터 분석 작업에 대한 RAG 모델의 성능 양상을 분석하고, 3장에서 성능을 개선할 수 있는 연구 방향을 제시하고자 한다.

2. 본론

RAG를 활용한 테이블 데이터 분석을 위한 질의 유형은 <표 1>과 같이 분류할 수 있다. 본 논문에서는 테이블 데이터를 분석하기 위한 질의 유형을 정보 검색, 특성 분석, 상관관계 분석, 결과 예측, 추세 분석과 같이 크게 5가지로 분류하였다. 테이블 데이터에서 어떤 내용을 얻고 싶은지 목적에 따라 다른 질의를 선택할 수 있다.

본 연구에서는 성능 평가를 위해 성인 인구 조사 소득 데이터셋[4]를 활용하였다. 이 데이터셋은 성인 인구의 소득, 직업, 교육 수준 등 다양한 특성을 포함하고 있다. 본 논문에서는 데이터셋을 분석하기 전, 각 데이터의 속성을 정규화 및 전처리 과정을 거쳐 적절한 구조로 변환하였다.

이후, 테이블 데이터를 RAG 모델의 검색에 사용하기 위해 각 특성별로 벡터화하여 검색이 가능하도록 설정하였다. 구체적으로는 HuggingFace의 임베딩 모델을 사용하여 각 데이터 항목을 벡터로 표현하였으며, 벡터 스토어는 FAISS를 사용하였다. 벡터 스토어는 로컬 환경에서 불러와 사용되었으며, 이를 기반으로 데이터 간의 유사도를 측정하고 효과적인 정보 검색을 수행할 수 있도록 설계하였다. 벡터 스토어의 검색 기능은 retriever로 변환되어 RAG 모델의 질의 처리에 활용되었다. 또한 RAG 모델에 사용된 LLM은 gpt-3.5-turbo 모델이다.

본 실험에서는 <표 1>의 질의 예시를 사용하여 각 질의 유형별 테이블 데이터 분석 쿼리에 대한 RAG 모델의 성능을 평가하였다. 평가 지표는 생성된 답변의 정확도를 평가하기 위해 ROUGE-L[5], 생성된 답변이 질문과 관련성 있는지 평가하기 위해 RAGAS 벤치마크[6]의 Answer Relevancy를 사용하였다. 이 지표들은 모두 1에 가까울수록 좋은 성능을 나타낸다.

<표 1>의 성능 결과를 확인했을 때, RAG 모델의 성능은 각 질의 유형에 따라 다양하게 나타나고 있다. ROUGE-L 지표의 경우, 결과 예측 질의 유형에서 가장 높은 값인 0.62를 기록하여 RAG 모델이 결과 예측 질문에 대해 비교적 정확한 답변을 생성할 수 있음을 보여준다. 반면, 상관관계 분석에서는 0.35로 가장 낮은 성능을 보였다. 이를 통해 RAG 모델이 테이블 데이터에서 서로 다른 속성 간의 복잡한 관계를 추론하는 데 어려움을 겪을 수 있음을 시사한다.

Answer Relevancy 지표에서는 대부분의 질의 유형에서 높은 성능을 보였으나, ROUGE-L에서와 마찬가지로 상관관계 분석에서의 점수가 0으로 나타나 모델이 해당 유형의 질문에 대해 적절한 관련성을 가지는 답변을 생성하지 못했음을 알 수 있다. 이와 반대로 결과 예측이나 추세 분석에서는 각각 0.97과 0.96의 높은 관련성 점수를 기록하여 해당 유형의 질문에 대해 RAG 모델이 유의미한 답변을 생

성할 수 있음을 확인하였다.

본 실험을 통해 전반적으로 상관관계 분석과 같은 복잡한 유형의 질문에 대해 RAG 모델의 성능이 낮게 나타났다. 이러한 유형의 질문은 단순한 데이터 검색이나 통계적 특성 분석을 하는 유형에 비해 여러 속성 간의 관계를 파악하고, 이를 바탕으로 추론해야 하므로 RAG 모델이 정확하고 관련성 있는 답변을 생성하는 데 어려움이 있었다. 상관관계 분석의 경우, 테이블 데이터에 존재하는 다양한 변수 사이의 복잡한 상호 작용을 이해해야 하기 때문에 단순한 정보 검색보다 높은 수준의 추론 능력을 요구한다. 따라서, 이러한 복잡한 질의 유형에 대한 RAG 모델의 성능을 향상시키기 위해서는 테이블 데이터를 적절하게 처리할 수 있는 전처리 방법과 더욱 정교한 RAG 아키텍처 모델링이 필요할 것으로 보인다.

결과적으로, <표 1>과 같은 결과는 RAG 모델이 테이블 데이터에서 다양한 유형의 질의에 대해 상이한 성능을 보일 수 있으며, 특히 상관관계 분석과 같은 복잡한 질의 유형에 대해서는 추가적인 모델 개선이 필요함을 시사한다.

3. 결론

본 논문은 RAG 모델의 테이블 데이터에 대한 분석 성능을 향상시키기 위해 테이블 데이터를 다루는 다양한 질의 유형을 분류하고, 각 유형에 대한 RAG 모델의 성능을 분석하였다. 이를 통해 테이블 데이터에 대한 RAG 모델의 성능을 더 깊이 이해할 수 있는 통찰력을 제공하였으며, 특히 테이블 데이터 특유의 구조와 관계성을 반영한 분석의 중요성을 강조하고 있다.

그러나 RAG 모델에 활용 가능한 테이블 데이터의 다양성이 부족하다는 점은 여전히 문제로 남아있다. 금융, 의료 분야와 같이 실제 산업에 기반한 테이블 데이터는 고유한 구조적 특성과 의미를 가지고 있기 때문에, 이러한 데이터셋을 충분히 확보하여 효과적으로 다룰 수 있는 방법이 필요하다. 따라서, 이러한 특성을 고려한 데이터셋을 구축하고, 이를 위한 적절한 전처리 방법에 대해 추가 연구가 필요하다. 이러한 데이터 부족에 대한 한계는 합성 데이터를 활용하여 일부 해소할 수 있다. 합성 데이터는 실제 데이터와 통계적으로 유사한 속성을 가지며, 합성 데이터 생성 과정에서 데이터 스키마와 관계성을 유지하면서 다양한 상황을 반영할 수 있으므로, RAG 모델은 더욱 다양한 테이블 구조 및 관계성을 학습할 수 있을 것이다.

이와 같이 향후 연구에서는 다양한 도메인의 테이블 데이터와 이에 적합한 전처리 방법론에 대한 연구가 필요하다. 이러한 후속 연구를 통해 RAG 모델의 활용 범위를 더욱 넓힐 수 있을 것으로 기대한다.

4. 사사

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2021-II212068, 인공지능 혁신 허브 연구 개발)

참고문헌

[1] Mallen, Alex, et al. “When not to trust language models: Investigating effectiveness of parametric and non-parametric memories.” arXiv preprint arXiv:2212.10511, 2022.

[2] Lewis, Patrick, et al. “Retrieval-augmented generation for knowledge-intensive NLP tasks.” Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459-9474.

[3] Gao, Yunfan, et al. “Retrieval-augmented generation for large language models: A survey.” arXiv preprint arXiv:2312.10997, 2023.

[4] Becker, B. & Kohavi, R. “Adult [Dataset].” UCI Machine Learning Repository, 1996. https://doi.org/10.24432/C5XW20.

[5] Lin, Chin-Yew. “ROUGE: A package for automatic evaluation of summaries.” Text Summarization Branches Out, 2004.

[6] Es, Shahul, et al. “RAGAS: Automated evaluation of retrieval augmented generation.” arXiv preprint arXiv:2309.15217, 2023.