RAG 기반 LLM을 적용한 공지사항 챗봇 시스템 개발

고상희*, 유승종*, 이석현*, 조현준*, 최기영*, 이동욱**, 정설영***

Development of Announcement Chatbot System with RAG-based LLM

Sanghui Ko*, Seungjong Yu*, Seokhyun Lee*, Hyunjoon Cho*, Kiyeong Choi*, Dongwook Lee**, and Seol-young Jeong***

요 약

RAG(Retrieval Augmented Generation) 기반 시스템은 LLM(Large Language Models,)에서 발생하는 여러 문제를 완화하며, 외부 데이터 소스를 실시간으로 검색하여 더욱 신뢰할 수 있는 답변을 생성한다. 본 논문은 경북대학교 컴퓨터학부의 공지사항을 대상으로 RAG기술을 활용한 질의응답 시스템을 제안한다. 공지사항 데이터를 크롤링하고 이를 벡터로 임베딩하여 검색할 수 있는 QA chatbot 시스템을 구축하였고, 실험 결과 해당 시스템은 기존의 챗봇 시스템에 비해 더 정확한 답변을 제공하는 것으로 나타났다.

Abstract

The RAG-based system alleviates a number of problems arising from Large Language Models (LLM) and generates more reliable answers by searching external data sources in real time. This paper proposes a Q&A system using RAG (Retrieved Augmented Generation) technology for announcements in the Department of Computer Science at Kyungpook National University. We built a QA chatbot system that can crawl announcement data and embed it as a vector to search for it, and the experiment showed that the system provides more accurate answers than the existing chatbot system.

Key words

rag, langchain, llm, qa system, chatbot, web crawling, vectorstore, embedding

* 경북대학교 컴퓨터학부 학사과정, gsh244400@naver.com, bigbell999@knu.ac.kr, tjrgus0703@gmail.com, bluelol5478@gmail.com, rldud1237@knu.ac.kr

** (주)데이터스트림즈 수석연구원, dwlee@datastreams.co.kr

*** 경북대학교 컴퓨터학부 교수(교신저자), snowflower@knu.ac.kr

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (2021-0-01082)

Ⅰ. 서 론

정보의 빠르고 정확한 전달은 공공기관과 교육기관의 중요한 과제이다. 특히 학교의 공지사항은 학생들이 자주 확인하는 필수적인 정보이며, 이를 효율적으로 관리하고 제공하는 시스템의 필요성은 높아지고 있다. 이러한 맥락에서 자연어 처리 기술을 활용한 질문응답 시스템(QA system)은 공지사항에 대한 자동 응답과 정보 검색을 가능하게 한다. 인간과 기계 간의 상호작용이 점차 발전하면서, LLM을 기반으로한 대화형 챗봇 시스템이 우리의 일상 생활에서 더욱 중요한 역할을 하고 있다[1]. 그러나 LLM 기반 시스템은 개인 정보 보안 문제, 정확도, 환각 현상 등으로 인해 여러 문제점이 발생할 수 있다[2].

이러한 문제를 해결하기 위해 RAG 모델이 주목받고 있다. RAG는 검색 기반 모델을 생성형 언어 모델에 결합하여 생성된 텍스트의 질과 정확성을 향상시키는 기술이다[3]. 이를 통해 LLM의 단점을 보완하고, 최신 정보에 대한 접근성을 높일 수 있다. 본 연구에서는 경북대학교 컴퓨터학부의 공지사항을 대상으로 RAG 모델을 적용하여 질문응답 시스템을 구축하고, 그 성능을 평가한다.

Ⅱ. 시스템 구성

이 시스템은 대학 웹사이트에서 공지사항을 스크래핑하여, 각 공지사항의 주요 텍스트 내용과 게시날짜, URL과 같은 메타데이터를 추출한다. 이후 텍스트는 보다 정밀한 검색을 가능하게 하기 위해 작은 청크로 분할된다. 각 텍스트 청크는 두 가지 유형의 임베딩을 생성하여 인덱싱되며, 이는 다음과 같다.

1. 희소 임베딩: TF-IDF 벡터화를 통해 텍스트 청크의 희소 표현을 생성한다.

2. 밀집 임베딩: 임베딩 모델을 사용하여 텍스트의 밀집 벡터 표현을 생성한다.

임베딩 모델에서는 ollama와 upstage 두 가지 방식으로 접근하였다.

Ollama는 모델을 쉽게 배포하고 사용할 수 있는 환경을 제공하여, 사용자의 복잡한 설치나 설정 없이도 쉽게 언어 모델을 활용할 수 있다. 하지만 Ollama는 로컬에서 실행되므로 사용자의 컴퓨터 성능에 직접적인 영향을 받는다. 따라서 로컬 시스템의 자원 한계로 인해 성능이 저하될 수 있다.

반면 Upstage는 한국어 이해와 생성에서 높은 정확도를 가지고 있으며, 한국어의 다양한 뉘앙스를 포착할 수 있는 능력이 뛰어나다. 또한 Upstage는 필요에 따라 경량화된 모델 버전을 제공하여, 낮은 자원에서도 빠르게 동작할 수 있도록 한다.

그림 1은 ollama와 Upstage의 임베딩 모델 성능 평가 결과를 나타낸다. Ollama 실행 시간은 492초, Upstage 실행 시간은 8초로 나타났다.

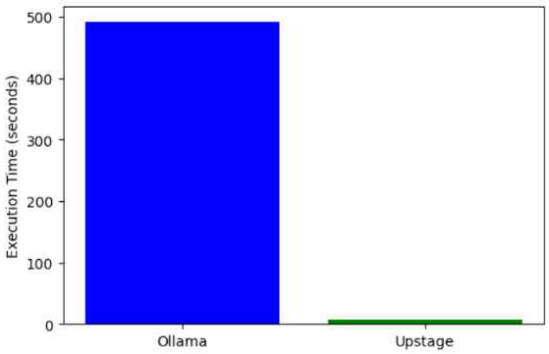

그림 1. ollama와 upstage 성능 비교

두 가지 방법을 비교한 결과, upstage 임베딩 모델이이 본 프로젝트에 적합하다는 것을 알 수 있다.

이렇게 생성된 밀집 임베딩은 관련 메타데이터(텍스트 내용, URL, 날짜)와 함께 효율적인 유사도 검색을 위해 Pinecone 벡터 데이터베이스에 저장된다.

사용자가 쿼리를 제출하면 시스템은 다음 단계를 통해 이를 처리한다. 먼저 쿼리를 TF-IDF 벡터로 변환한 후, 저장된 문서 벡터와의 코사인 유사도를 계산하여 희소 검색을 수행한다. 이후 Upstage 모델을 통해 쿼리를 임베딩하고, Pinecone 인덱스에서 밀집 유사도 검색을 수행한다. 최종적으로 시스템은 희소 검색과 밀집 검색의 결과를 결합하여 가장 관련성 높은 문서를 식별하는 하이브리드 검색 접근 방식을 사용한다. 이를 통해 두 가지 검색 방법의 장점을 모두 활용하여 검색의 정확도를 향상시킨다.

그림 2는 본 시스템의 전체 구성도를 보여주는 다이어그램이다.

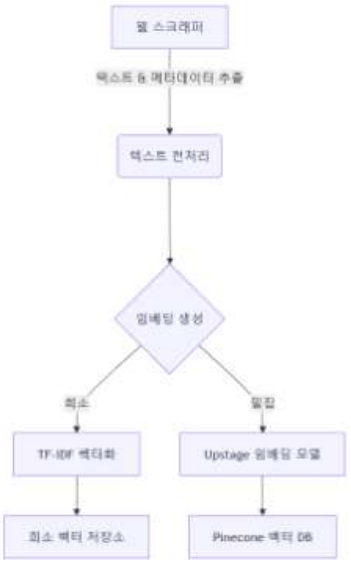

그림 2. 시스템 다이어그램

III. 구현 결과

본 시스템과 기존 시스템의 성능을 비교하기 위해 동일한 질문 세트를 사용하여 다음과 같은 실험을 진행했다.

전통적 질문응답 시스템: 미리 학습된 사전 훈련 모델을 기반으로 질문에 대한 답변을 생성한다. 이 시스템은 추가적인 문서 검색 없이 모델 내에서 답변을 생성한다.

RAG 기반 질문응답 시스템: 질문이 들어오면 관련된 문서를 검색하고, 검색된 문서를 바탕으로 답변을 생성한다.

그림 3과 4는 각각 전통적 질문응답 시스템의 실험 결과와 RAG 기반 질문응답 시스템의 실험 결과를 나타낸다.

그림 3에서 RAG 모델을 적용하지 않은 챗봇은 사전 훈련된 지식에만 의존하기 때문에, 실시간으로 업데이트되는 정보에는 접근하지 못한다. 즉, 경북대학교 TOPCIT 시험 날짜나 튜터 관련 공지사항 같은 질문을 받았을 때, "해당 페이지를 방문해 최신 공지를 확인하세요."와 같은 반응을 보인다. 최신 정보를 다루지 못하는 한계가 있어 사용자가 원하는 구체적인 정보를 제공하지 못하기 때문이다.

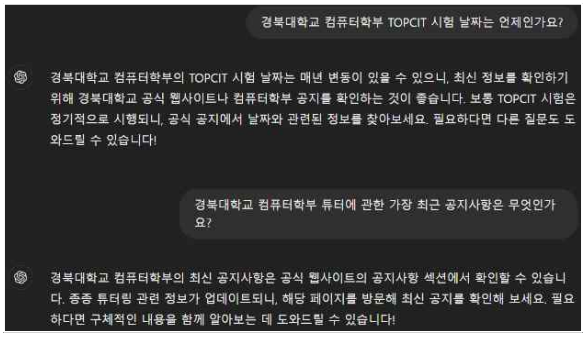

그림 3. RAG 모델을 적용하지 않은 챗봇

그림 4. RAG 모델을 적용한 챗봇

반면, RAG가 적용된 챗봇은 실시간으로 공지사항 데이터베이스에 접근하여 정보를 검색하고 제공할 수 있다. 사용자의 질문이 들어오면, 챗봇은 미리 수집된 경북대학교 컴퓨터학부 공지사항 데이터를 바탕으로 최신 정보를 검색한 후 적절한 답변을 제공한다. 예를 들어, 경북대학교 탑싯 시험 날짜나 수강 변경에 대한 정보에 대해 질문하면 챗봇이 공지사항에서 해당 정보를 찾아 정확한 날짜와 내용을 알려준다.

IV. 결 론

본 연구에서는 경북대학교 공지사항을 대상으로 RAG 기반 질문응답 시스템을 구축하고 그 성능을 평가하였다. RAG 모델은 기존의 LLM 기반 챗봇 시스템에서 발생할 수 있는 환각 문제를 줄이며, 검

색된 공지사항 데이터를 바탕으로 더욱 신뢰성 있는 답변을 생성할 수 있었다. 그러나 성능 향상을 위해 이미지 임베딩의 통합과 결과 결합 방법 개선이 필요하다. 특히, 이미지 임베딩을 통해 공지사항에 포함된 시각적 자료에 대한 이해를 높이고 검색 결과에 이미지 관련 정보를 포함시킬 수 있다[4]. 이를 위해 CLIP(Cross-Lingual Image Pretraining)과 같은 모델을 활용하여 텍스트와 이미지를 동시에 벡터화하는 방법이 유용할 것이다. 또한, 희소 검색과 밀집 검색의 가중치를 조절함으로써 쿼리의 특성에 맞춰 결과의 정확도를 향상시킬 수 있다. 향후 연구에서는 이러한 개선점을 실험하고 실제 성능을 측정하여 챗봇의 응답 정확성 및 속도를 더욱 향상시킬 수 있을 것으로 기대된다.

참 고 문 헌

[1] 신성필. 초거대 AI 의 기반모델 (Foundation Model) 개념 및 표준화 동향. 한국통신학회지 (정보와통신), 2023, 40.6: 12-21.

[2] 정효정(Alice Jung),송주현(Juhyun Song),서상훈(Sanghoon Seo),임진효(Jinhyo Lim),이현상(Hyunsang Lee),and 김동균(Dongkyun Kim). "RAG 기반 LLM 성능 평가 및 검증을 위한 LangChain 활용 RAGA 방법론 연구." 대한전자공학회 학술대회 2024.6 (2024): 2567-2572.

[3] 조찬영(Chanyoung Jo),강성준(Seongjun Kang),and 정현준(Hyunjun Jun). "RAG기반 랭체인을 이용한 생성형 AI 챗봇 구현." Proceedings of KIIT Conference 2023.11 (2023): 460-463.

[4] 전준형(Joon Hyoung Jun),김상철(Sang-Chul Kim),김주철(Joo Chul Kim),and 윤성준(Seong Joon Yoon). "웹 애플리케이션 서버(WAS)에서의 검색 증강 생성(RAG) 기술을 이용한 지식 기반 QA 문제 해결." 한국통신학회 학술대회논문집 2024.1 (2024): 1109-1114.