HyDE를 활용한 RAG 기반 학사공지 챗봇의 성능 향상

전지수[1], 박준서[1], 최윤창[1], 이윤정[2]
[1]홍익대학교 컴퓨터공학과 학부생
[2]홍익대학교 도시공학과 학부생

wltnjeon01@g.hongik.ac.kr, jsphilip54@g.hongik.ac.kr,
bambini77@g.hongik.ac.kr, dbswjd6654@g.hongik.ac.kr

A Performance Improvement of RAG-based Academic Notice Chatbot Using HyDE

Ji-Su Jeon[1], Jun-Seo Park[1], Yun-Chang Choi[1], Youn-Jeong Lee[2]
[1]Dept. of Computer Engineering, Hongik University
[2]Dept. of Urban Planning and Design, Hongik University

요 약

본 연구는 대학생들의 학사 정보 접근성을 향상시키기 위해 학교 챗봇을 개발하는 것을 목표로 하였다. 기존 RAG 방식의 정확도 한계를 극복하고자, HyDE 기법을 공지사항 데이터셋 특성에 맞게 확장하였다. 각 문서의 핵심 내용을 요약한 제목이 포함된 데이터셋의 구조를 활용해, 사용자의 질문에 대한 가상의 답변과 가상의 제목까지 함께 생성하는 확장된 접근법을 제안했다. 이는 기존의 Raw Query 방식 대비 약 90.45% 검색 정확도 향상을 이루었다. 본 연구의 확장된 HyDE 기법은 학교 챗봇을 비롯한 다양한 시스템에 적용할 수 있어 정보 접근성을 크게 개선할 것으로 기대할 수 있다.

1. 서론

많은 대학에서 기존 대학 홈페이지의 복잡성과 검색 기능의 한계를 극복하기 위해 챗봇 도입을 추진하고 있다[1]. 그러나 대부분의 챗봇은 사전 정의된 질문에만 응답하는 룰베이스 시스템을 사용해 정보 접근성이 떨어진다.

최근 대규모 언어 모델(LLM)의 발전으로 사용자와 자연스러운 대화를 통해 정보를 제공하는 챗봇들이 급속히 등장하고 있다. 그러나 LLM 만으로는 특정 도메인의 최신 정보나 상황별 맥락을 정확히 파악하는 데 한계가 있다. 이를 극복하기 위해 검색 증강 생성(Retrieval-Augmented Generation, RAG) 기법이 주목받고 있다[2].

2. 본론

본 연구에서는 대학 공지사항 DB를 기반으로 RAG 기법을 활용한 학사 정보 챗봇을 구현하였다. RAG 기법으로 공지사항 DB에서 사용자의 질문과 관련된 공지사항 문서를 검색하고 이를 바탕으로 LLM이 구체적이고 정확한 답변을 생성할 수 있었다. 그러나 사용자 질문만으로 관련된 문서를 정확하게 찾는 데 한계가 있었다. 이에 본 연구에서는 검색 성능 향상을 위해 고급 RAG 기법인 HyDE를 도입하였다.

2.1 HyDE (Hypothetical Document Embedding) 적용

HyDE는 사용자의 질문과 타겟 문서 간에 의미론적 간극을 좁히기 위해 가상의 답변을 사용해 문서 검색에 활용하는 기법이다[3].

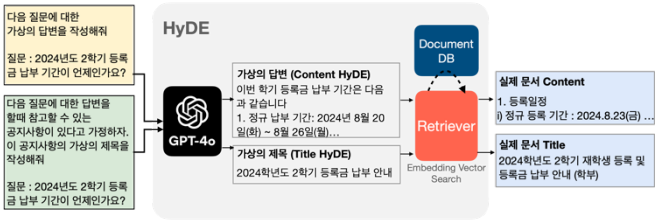

그림 1. 본 연구에 적용한 HyDE Process

본 연구는 공지사항 데이터셋의 특성을 활용한

개선된 검색 방법을 제안한다. 그림 1과 같이, GPT-4o 모델로 생성한 가상의 답변과 가상의 제목을 검색 쿼리로 사용하였다. 이는 기존 HyDE 프레임워크의 가상 답변 활용 방식을 적용할 뿐만 아니라, 가상 제목까지 추가로 생성해 검색에 활용함으로써 검색 정확도를 높이고자 하였다.

검색 성능 평가를 위해 최근 6개월 동안의 학교 공지사항을 바탕으로 GPT-4o Mini 모델을 사용해 질문과 답변으로 이루어진 평가 데이터 셋을 생성하였다. 또한 사람이 직접 검토한 후 최종적으로 426개의 고품질 평가 데이터 셋을 구축하여 실제 사용 시나리오에 더 가깝게 모델의 성능을 평가했다.

2.2 PCA 시각화 및 가설 설정

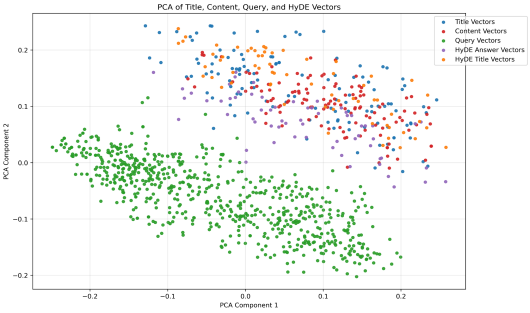

그림 2. 임베딩 벡터 PCA 시각화

그림 2는 다양한 임베딩 벡터들을 PCA를 통해 2차원으로 시각화한 결과를 보여준다. 이 임베딩 벡터들은 문서 제목(Title Vectors), 문서 내용(Content Vectors), 사용자 질문(Query Vectors), 그리고 HyDE로 생성된 가상 답변(HyDE Answer Vectors)과 가상 제목(HyDE Title Vectors)으로 구성되어 있다. 벡터 공간 분석 결과, HyDE로 생성된 가상 답변과 가상 제목의 벡터들이 실제 문서 내용 및 제목의 벡터들과 근접한 영역에 분포하는 것으로 나타났다. 반면, 원래의 사용자 질문 벡터들은 실제 문서 내용 및 제목 벡터들과 상대적으로 멀리 떨어져 있었다. 따라서 HyDE를 통해 원래 사용자 질문의 의도를 잘 파악하면서도, 효과적인 검색을 제공할 수 있을 것이라고 분석하였다.

이러한 분석을 바탕으로, HyDE를 적용한 네 가지 검색 방식을 설계하고 비교 실험을 진행하여 평가하였다. 각 방식에서 임베딩 벡터 간 유사도가 가장 높은 문서의 id와 정답 문서의 id와 비교해 정확도를 계산하였다.

a) Combined HyDE: 가상 답변-실제 문서 내용 유사도, 가상 제목-실제 문서 제목 유사도의 평균
b) Title HyDE: 가상 제목-실제 문서 제목 유사도
c) Content HyDE: 가상 답변-실제 문서 내용 유사도
d) Raw Query: 사용자 질문-실제 문서 제목 유사도와 사용자 질문-실제 문서 내용 유사도의 평균

2.3 실험결과

실험 결과, HyDE를 적용한 검색 방식이 Raw Query 검색 방식보다 우수한 성능을 보였다. 특히, 제목과 내용을 모두 고려한 Combined HyDE 방식이 가장 높은 성능을 달성했다. Combined HyDE 방식은 0.5498의 성능 점수를 기록하여, 기존의 Raw Query 검색 방식(0.2887)에 비해 약 90.45% 향상된 결과를 보였다. 이는 HyDE의 확장 적용을 통해 관련성 높은 문서를 효과적으로 검색할 수 있음을 보여준다.

Title HyDE 방식(0.486762)이 Content HyDE 방식(0.352342)보다 더 높은 성능을 보였다. 이는 공지사항의 제목이 문서 내용을 효과적으로 대표하며, 검색 정확도 향상에 중요한 역할을 한다는 것을 시사한다.

검색 방식	성능 점수
Combined HyDE	0.549898
Title HyDE	0.486762
Content HyDE	0.352342
Raw Query	0.288732

<표 1> HyDE 적용 및 비교

3. 결론

HyDE의 적용과 확장을 통해, 본 연구에서 개발한 학교 챗봇은 기존의 룰베이스 시스템이나 단순 RAG 기반 챗봇 모델보다 관련성 높은 정보를 정확하게 제공할 수 있게 되었다. 이는 학교 챗봇을 비롯한 다양한 시스템에 적용할 수 있어 정보 접근성을 크게 개선할 것으로 기대할 수 있다.

※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] 박한길 외, “인공지능 기반 대학교 정보 알림 챗봇 시스템 개발.” Proceedings of KIIT Conference, 2023.6 (2023), 920-923.
[2] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.
[3] Gao, Luyu, et al. "Precise zero-shot dense retrieval without relevance labels." *arXiv preprint arXiv:2212.10496* (2022).