자연언어처리 기말 대체 과제 제안서

생명과학전공 2303050 박은서

제안의 개요

질병-유전자 관계는 신약 개발, 치료 방법 선정 등에 필요하나 현재는 연구자들이 알고 있는 지식을 활용하거나, 매번 논문 db에서 수작업으로 문헌을 찾는 등 많은 시간과 노력이 소요된다. 기존에 disnet 등의 질병-유전자 db는 존재하나 정적이며 최신 논문은 반영되지 않고, 사용자 친화적이지 않은 단점이 있다. 이는 키워드 기반 검색은 가능하나 문장의 의미나 관계 반영 불가하므로 유관 논문 누락 가능성이 존재한다. 또한 검색 결과 해석 및 요약 기능 없어 연구자의 추가 분석 (관련 논문을 찾아 읽어보는 등)을 필요로 한다. 이번 과제에서는 이러한 문제점을 해결하고자, 최신 논문을 토대로 질병-유전자 관계를 검색 및 LLM 기반의 근거 및 요약을 제공하는 서비스를 개발하고자 한다. 기존 연구와의 주요 차별점은 최신 논문 반영이 가능한 것, 의미 기반 검색(semantic search) 이후 RAG과 LLM 기반의 답변을 제공하는 것, 근거 문장 및 논문을 제공하는 것 등이다. 이를 통한 연구 생산성 향상, 답변 신뢰성 확보, 확장성 등의 효과가 기대된다.

작업 범위

PubMed(bio, medical 분야의 논문, 정보가 집약된 사이트) API를 활용해 소규모 도메인 별 논문 abstract을 수집, 문장 단위로 분리 및 biomedical entity tagging (NER) 등으로 pre-processing한 후 disease-gene 관계를 임베딩 및 벡터 인덱스를 구축한다. 이후 사용자 질문을 받으면 임베딩해 semantic search 기법을 통해 Top-k 문장을 검색, RAG와 LLM으로 질문에 부합하는 답변을 생성하는 서비스를 만든다.

문제 해결의 내용

Data & pre-processing PubMed(bio, medical 분야의 논문, 정보가 집약된 사이트) API를 활용해 소규모 도메인 별 논문 abstract을 수집하고, 문장 단위로 분리한 뒤 biomedical entity tagging (NER)을 거쳐 유전자 이름, 질병으로 normalization할 것이다. 이를 통해 json 형식의 산출물이 제공된다.

Semantic search 문장 의미를 반영한 질병-유전자 관계 검색을 위해 임베딩하고 벡터 인덱스를 구축한 뒤 사용자 질문을 임베딩해 Top-k 문장을 검색하고 metadata 반환하는 semantic search 기법을 이용해 유관 논문을 가능한 많이 반영하고 연구자의 질문에 정확히 부합하는 답변을 구성하도록 한다.

RAG & LLM 검색 결과인 Top-k 문장을 프롬프트에 포함, RAG와 LLM을 활용해 답변을 생성한다. 이후 자연어 요약, 근거 문장, PMID 로 구성된 최종 답변을 연구자에게 제공한다. RAG의 환각 현상을 방지하기 위해서 Multi-hop query 방식을 이용하고자 한다. 이는 단일 문서가 아닌 여러 문서를 토대로 RAG 답변을 생성하는 방식으로, 기존 방식에 비해 환각 현상을 완화시켜줄 수 있다. 다만 이 방법은 추론 경로 오류 등의 문제를 일으킬 수 있어, PAR RAG(Plan-then-Act-and-Reviuew RAG) 프레임 워크를 활용하는 전략을 취하고자 한다.

참고 사이트나 참고 논문

1 Credible Plan-Driven RAG Method for Multi-Hop Question Answering (link)

2 MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries (link)

3 RAG Hallucination: What is It and How to Avoid It (link)