자연언어처리 기말 대체 과제 최종 보고서

생명과학전공 2303050 박은서

Abstract

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 소요 된다. 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 반영이 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. BioRAG은 이러한 문제점을 해결하고자, 최신 논문을 토대로 질병-유전자 관계를 검색 및 LLM 기반의 근거 및 검색 결과를 제공하는 패키지이다. 기존 연구와의 주요 차별점은 최신 논문 반영이 가능한 것, 의미 기반 검색(semantic search) 이후 RAG과 LLM 기반의 답변을 제공하는 것, 근거 문장을 제공하는 것 등이다. 생물의학 도메인 고유의 도전과제(다양한 유전자 명명법, 동의어, 약어, 신약물/대상 유전자 네이밍 등)를 극복하기 위해 HGNC 완전 사전+SciSpaCy BC5CDR 모델 조합의 entity normalization을 구현하였으며, 이는 기존 범용 NLP 접근과 차별화되는 핵심 기술이다. 특히 본 프로젝트는 정량적 성능 비교보다, 생물의학 도메인에서 RAG 시스템을 실제로 작동시키기 위해 필요한 데이터 수집, 전처리, 엔티티 정규화, 실시간 검색 파이프라인 설계의 난이도를 실증적으로 탐구하는 데 초점을 두었다. 하드웨어 및 데이터 부족 이슈로 인해 아쉬운 점이 있으나, 이후의 확장성을 고려할 때 이 패키지를 통해 이를 통한 연구 생산성 향상, 답변 신뢰성 확보, 확장성 등의 효과가 기대된다.

Pipeline

data 수집 초기에는 Entrez api로 pubmed에 접근, 데이터(pubmed id, article title, abstract) 수집 → data pre-processing & embedding 파이프라인 구축 후 동일 방식을 이용하나 실시간으로 Pubmed에서 검색하고 이후 과정을 거쳐 결과를 제공하도록 구현

data pre-processing abstract이 비어 있거나 문자열이 아닌 경우 제외하도록 처리 → NLTK의 sent_tokenize로 문장 분리 → SciSpacy의 en_ner_bc5cdr_md 모델, hgnc 레코드 활용해 NER 처리(bc5cdr은 disease에 최적화, gene NER에 도움 되도록 추가 데이터 이용) → entity abel(GENE, DISEASE)에 따라 gene, disease 명으로 normalization

Embedding 수집하고 전처리한 pubmed abstract에 대해 SentenceTransformer all-mpnet-base-v2로 문장 임베딩 (768차원) → faiss로 벡터 인덱스 구축

Semantic search 임베딩 벡터 활용, 활용해 top-k 문장 검색 방식의 semantic search 구현 (사용자 질문 임베딩-faiss 인덱스 간 top-k 유사도 검색)

LLM & RAG pubmed 쿼리 + 사용자 질문 결합 → MedAlpaca-7B 모델을 이용한 최종 RAG 응답 생성 (LLaMA 기반의 MedAlpaca 모델을 사용, 사실 기반의 생물의학 응답을 생성)

주요 변경사항

multi-hop 구현 PAR 방식을 이용해 룰 기반으로 1-hop 검색 결과에서 top-k gene/disease를 선정한 뒤 그

에 대해 중간질의를 만들어 2-hop을 구성하는 방식의 코드 구현을 시도하였다. 하지만 기존 single-hop module보다 평균 3-4배 증가하는 등 시간은 매우 오래 걸리면서도 정확도가 떨어지고 noise가 심하여 성능 비교 결과 최종적으로 single-hop으로 돌아가는 것으로 결론지었다. 이러한 증상의 원인은 2-hop에서 생성된 중간 질의가 원래 의도에서 벗어나는 현상인 query drift, 룰 기반의 질의로 인한 정확한 relation extraction의 어려움 등으로 추정하였다.

검증 테스트

batch test LLM RAG 답변 생성 이후에 검증을 진행하는 것은 현실적으로 어려움이 있어 가능한 수준의 테스트를 진행하였다. LLM RAG 답변 생성시 하드웨어 이슈로 한 개의 질문 당 몇 십분이 소요되어 100개 이상의 데이터셋에서 작동하는 것이 불가능했다. 이에, 이번 프로젝트의 파이프라인을 참고하여

1. 사용자 질문에서 pubmed qeury에 들어갈 단어를 잘 읽어내는지(pubmed는 논문 db로 검색어에 실제 논문과 일치하는 문장을 포함해야 반환하므로, 일반 검색 엔진처럼 전체 질문 쿼리를 삽입할 수 없다. 이에 사용자 질문에서 pubmed용 쿼리(NER 했을 때의 disease 혹은 gene)를 뽑아내고, 이를 통해 Pubmed 논문을 검색한 뒤 전체 질문과 함께 이용해 semantic search하는 방식으로 구현하였다.),

2. pubmed용 쿼리를 토대로 검색했을때 top-k 문장이 타당한지

를 검토하는 테스트를 구성으며, 데이터셋은 주요 질병-유전자 관계를 고려하여 구성하였다. (적합한 데이터셋을 소규모로 구하는 것이 쉽지 않아 사전 지식 기반으로 데이터셋을 구성, 관련 db들을 활용하여 교차검증 하였다.)

real-time test 사용자 입력을 받아 실시간 시스템 결과물(LLM + RAG)이 나올 수 있는지를 검증하였다. 실시간으로 pubmed 검색을 지원하는 만큼, 다소 시간은 소요되나 여러 단어를 돌려보았을 때 꽤 좋은 성능을 보이는 것을 확인했다. 특히 disease->gene 검색보다 gene->disease 검색에서 좋은 성능을 보인다.

Discussion

본 프로젝트는 생물의학 도메인 고유의 기술적 난이도를 극복한 실용적 RAG 시스템을 제시한다. 이는 생물학 연구자의 문헌 조사 생산성 향상을 위한 첫걸음으로서, 지속적인 성능 개선과 임상 적용 가능성을 기대할 수 있다.

주요 기술적 기여 SciSpaCy + HGNC 사전 조합으로 GENE/DISEASE 동시 처리하는 Hybrid biomedical NER을 구현하였고, 실시간 PubMed 연계를 통해 정적 DB 대비 최신성 확보하였으며 모듈화된 RAG로 single-hop 안정성 + multi-hop 확장성을 보장하였다.

도메인 특화 도전과 극복 복잡한 entity normalization을 위해 HGNC 완전 사전 구축하였으며, 하드웨어 제약을 피하고자 CPU fallback + lazy loading 최적화하였다. 또, 실시간성 요구에 맞추어 retriever 단독 평가 체계를 구축하였다.

향후 확장 방향 멀티홉을 재도입하여 LLM agent (LangGraph) + KG (DisGeNET) 연계 방식으로 성능을 끌어올리거나, Cross-encoder 또는 LLM-as-judge로 Reranking 강화하거나, gpu를 확보해 BioASQ 챌린지 데이터셋 정량 평가하는 등 하드웨어와 시간적 한계로 인해 아직 시도해보지 못한 부분들을 방학 중 연구실 환경을 이용해 해결해보고자 한다. 어느정도 속도와 성능을 보장하는 환경이 확보된 후, 여러 LLM을 이용해 그간의 성능을 비교해보는 것도 재미있는 연구가 될 것이라고 생각한다.