한국어 검색 요약 모델 개발 제안서

팀 명	그런데 말입니다.
팀 원	백승주 권민욱 (경상국립대학교 AI융합공학과) / 안유민 (경상국립대학교 AI정보공학과)
제 목	높은 유용성과 신뢰도를 가진 RAG 기법을 활용한 QA 모델 성능 개선

1. Project Introduction

본 프로젝트의 목표는 관련 문서를 기반으로 특정 질문에 대한 간결하고 정확한 요약을 생성하는 대규모 언어 모델(LLM)의 성능을 향상 시키는 것입니다. 질문이 주어지면 관련된 문서를 추출하고, LLM은 추출된 문서를 참고하여 질문에 대한 설명 요약과 정답을 출력합니다. 따라서, 본 프로젝트의 연구 요구 사항(Research requirements)은 다음과 같습니다.

참조 문서 탐색	특정 질문을 해결하는 데 유용하고 신뢰할 수 있는 문서를 효율적으로 식별하고 검색합니다.
필요한 콘텐츠 요약	참조한 문서의 내용을 요약하여 질문에 대한 답변에 필요한 정보를 제공합니다.
정확한 답변 생성	요약된 내용을 바탕으로 질문에 대한 정확한 답변을 생성합니다.

표 1 Research requirements

연구팀은 질의에 대한 답변에 있어 선택한 참조 문서의 유용성과 신뢰성의 중요성을 강조합니다. 참조 문서의 활용도와 신뢰도가 높으면 생성된 답변의 정확도가 크게 향상되고 설명 요약의 품질이 향상될 수 있습니다. 또한, 사용 가능한 대규모의 관련 문서 코퍼스를 보유하면 다양한 도메인에 걸쳐 LLM의 성능을 더 잘 일반화할 수 있습니다.
이를 위해 연구팀은 아래 Figure 1과 같은 아키텍처를 제안합니다.

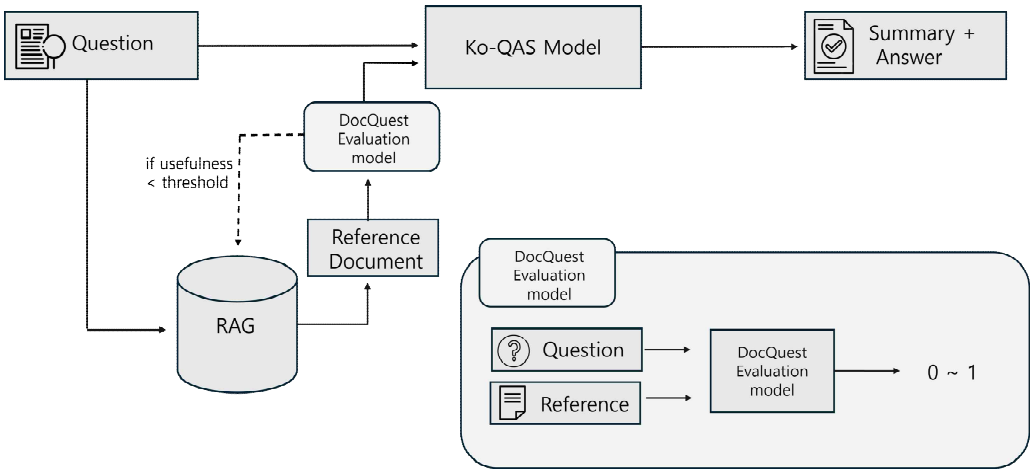

그림 1 한국어 검색 요약 LLM 아키텍처

아키텍처는 크게 세 가지 요소로 구성됩니다:

1. 참조 문서 선별(RAG: Retrieval-Augmented Generation): 대규모 코퍼스가 저장되어 있는 데이터베이스로부터 사용자의 질문에 유용할 수 있는 참조 문서를 추출하여 반환합니다.
2. 유용성 평가 (DocQuest Evaluation LLM, UL): RAG에서 반환한 참조 문서가 질문 해결에 얼마나 도움이 될 수 있는지를 평가하여 RAG가 유용한 참조 문서를 선택하는데 보조합니다.
3. 한국어 답변 요약 생성(Ko-QAS Model): 최종적으로 높은 유용성이 평가된 참조 문서를 기반으로 질문의 답변에 대해 참조 문서의 내용을 요약하고, 이를 바탕으로 정답을 생성합니다.

해당 아키텍처를 통해 참조 문서의 유용성과 신뢰성을 보장함으로써, 생성된 답변의 정확도와 설명 요약의 품질을 향상시킬 수 있습니다. 또한, 다양한 도메인에 걸쳐 LLM의 성능을 잘 일반화할 수 있어, 보다 광범위한 질문에 대한 효과적인 대응이 가능합니다.

2. Dataset & Pre-processing

본 연구에서는 Ko-QAS Model Fine Tuning을 위해 “문서-질의-요약-답변”의 형식을 가진 데이터셋을 구축합니다. 본 데이터셋은 한국어를 기반한 오픈 데이터셋과 자체 수집 데이터셋의 조합으로 구성되며, 이 데이터는 여러 파이프 라인의 정제를 거쳐 데이터 품질을 향상 시킵니다.

2.1 Open-Dataset

첫 번째로 ‘문서-질의-답변’으로 이루어진 데이터셋입니다. 대표적으로 AIhub(https://www.aihub.or.kr/)의 '표 정보 질의 응답 데이터'입니다. 요약 데이터는 포함 되어있지 않기에, OpenAI API를 이용하여 답변을 설명할 요약 데이터를 생성합니다.

두 번째로 ‘문서-요약’ 형태로 이루어진 데이터셋입니다. 대표적으로 AIhub 의 ‘문서요약 텍스트’입니다. 질의와 답변 데이터는 포함되어 있지 않기에, OpenAI API를 이용하여 질의와 답변 데이터를 생성합니다. 위 데이터 세트는 건축, 공공행정, 과학기술, 교육, 금융, 의료 등 다양한 도메인으로 이루어져 특정 도메인에 치우치지 않도록 해, 일반화 된 성능을 발휘하고자 합니다. 또한, 한국 지능정보 사회진흥원, 국가 통계 포털, 한국 학술정보에서 원천데이터를 수집하여 총 3단계 정제 및 검증과정을 거쳐 정확도가 검증됐고, 문서 성능 평가 ‘ROUGE’ 로 성능평가를 완료했습니다.

세 번째로 ‘문서-질의-답변-요약’으로 이루어진 데이터셋입니다. 이는 본 테스크에 가장 적합한 데이터 셋으로 MarkAI에서 공개한 ‘Ko-commercial dataset’입니다. 해당 데이터셋의 형태를 기반으로 학습 데이터셋을 구축합니다.

네 번째 형태의 데이터는 최신 정보를 반영하기 위해 2024년부터 작성된 기사, 기고문, 잡지 등을 웹에서 크롤링(Crawling)하여 수집합니다. 수집된 데이터의 신뢰성을 평가하기위해, 질문에 대한 답변에서 거짓된 정보가 있는지 확인하는 과정을 거칩니다. 이러한 과정을 통해 최신 데이터의 품질을 보장하고, 정확하고 신뢰할 수 있는 정보를 제공할 수 있도록 합니다.

2.2 Pre-processing

데이터 전처리는 LLM의 성능과 정확성을 높이기 위해 필수적입니다. 불필요한 데이터와 잡음을 제거하고, 데이터를 일관된 형식으로 변환하여 모델이 효율적으로 학습할 수 있게 합니다. 전처리는 또한 오류를 수정하고 다양한 소스의 데이터를 통합하여 일관성을 유지합니다. 이를 통해 모델이 중요한 패턴을 학습하고, 더 신뢰성 있는 결과를 제공합니다. 전처리를 철저히 수행하면 LLM의 성공적인 학습과 예측 성능이 보장됩니다.

Technique	Description
중복 제거 (Deduplication)	텍스트 중복 제거 작업은 모델 학습 시 답변의 다양성을 유지하고 스토리지 효율을 높이기 위해 필요합니다. 이를 위해 TF-IDF를 이용해 문장의 중요도를 계산하고, MINIHASH-LSH를 활용해 유사한 문장을 효율적으로 탐지합니다. 또한, 의미적 유사성 측정 기법 Paraphrase Identification 모델을 사용해 다른 단어로 표현되었지만 동일한 의미를 가진 중복 문장을 식별하여 제거합니다.
Privacy Reduction	사전학습에 사용되는 웹 텍스트 데이터는 민감하거나 개인정보를 포함할 수 있으므로 개인정보 침해의 위험을 방지하기 위해 PII를 제거해야 합니다. 이를 위해 NER(Named Entity Recognition)과 키워드 스포팅을 사용해 RawData에서 인물 개인정보 등 민감한 엔티티를 식별하고 제거합니다. 또한, 민감한 부분은 마스킹하여 대체할 수 있습니다.
데이터 필터링	텍스트 데이터에서 노이즈를 제거하여 품질을 높이고, 유해한 콘텐츠를 식별하여 제거합니다. 저품질 데이터를 식별하여 제거함으로써 데이터의 신뢰성과 유용성을 유지합니다. 또한, 웹 스크래핑 과정에서 발생할 수 있는 HTML, CSS, 자바스크립트 마크업 등을 제거하여 텍스트의 질을 향상시킵니다.
데이터 정형화	텍스트 데이터의 품질을 높이기 위해, 특수문자와 불필요한 공백을 제거합니다. 또한, 텍스트 내 오타를 수정하고, 분석에 도움이 되지 않는 불용어를 제거합니다. 웹 스크래핑에서 얻은 데이터를 파싱하여 필요한 정보를 추출하고 이를 구조화된 형태로 변환합니다. 결측치가 있는 경우 이를 처리하여 데이터의 완전성을 유지합니다.

3. RAG (Retrieval-Augmentation Generation)

한국어 정보 검색 및 요약 기술은 방대한 정보 속에서 정확하고 신속하게 답변을 제공하는 데 매우 중요하다. RAG 시스템은 대규모 문서 데이터에서 필요한 정보를 효율적으로 검색하고, 이를 입력 프롬프트에 포함 함으로써 높은 품질의 응답을 생성하도록 합니다. RAG는 크게 두 가지 단계로 나뉩니다.

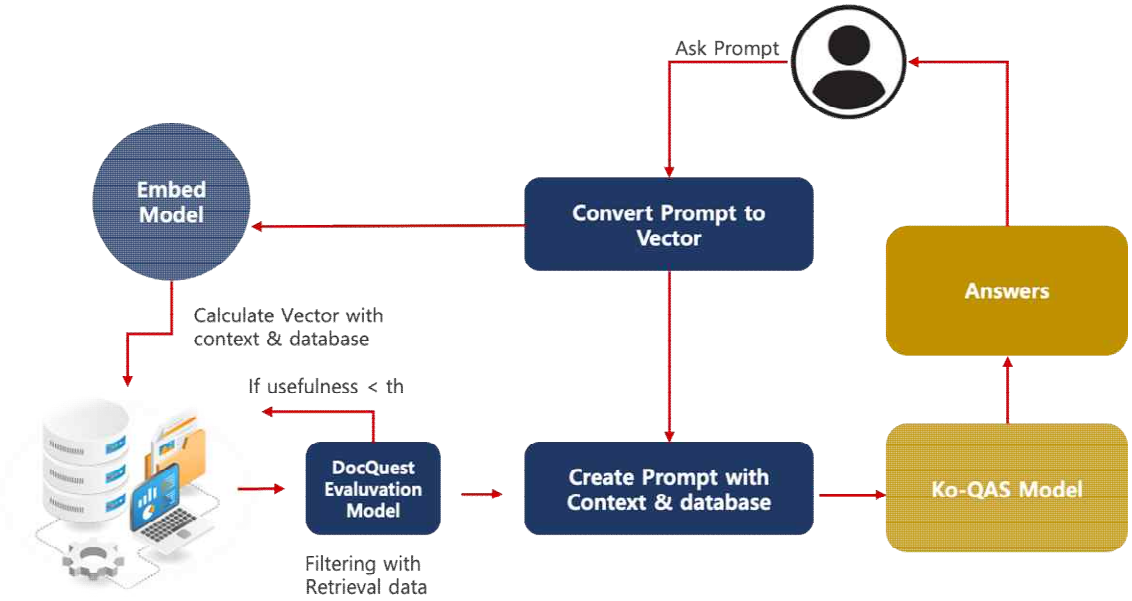

그림 2 RAG Architecture

<Indexing>

Indexing 단계는 문서 데이터를 LLM의 입력으로 사용할 수 있는 작은 조각들로 나누는 과정입니다. 이 단계에서 대규모 문서 데이터를 전처리하여 필요한 정보를 추출하고 정리합니다. 이렇게 전처리된 텍스트 조각들을 임베딩하여 벡터로 변환하고, 이를 데이터베이스에 저장합니다. 이를 통해 대규모 문서 데이터를 효율적으로 저장하고 관리하며, 나아가 Retrieval을 가능하게 합니다. 이때, 이 데이터 베이스에는 최대한 신뢰도가 높은 문서만을 포함하여 모델이 잘못된 정보를 기반으로 응답을 생성하지 않도록 해야합니다.

<Retrieval>

Retrieval 단계에서는 질문이 들어오면 벡터 데이터베이스와 유사도를 비교하여 가장 비슷한 상위 n개의 문서를 추출합니다. 이 문서들은 질문과 함께 LLM의 입력으로 사용됩니다. 이를 통해 LLM은 보다 신뢰성 높고 정확한 답변을 생성할 수 있습니다. 다만, 유사도만으로는 충분하지 않은 경우, 유용하지 않은 정보를 걸러내기 위해 LLM을 활용한 필터를 구축합니다.

<DocQuest Evaluation model>

DocQuest Evaluation model은 Retrieval을 통해 출력된 문서와 질문을 입력으로 받아, 해당 질문에 대해 문서가 얼마나 문맥상 일치하는지를 0에서 1 사이의 수치로 나타낸다. 1에 가까울수록 해당 질문에 대한 답변에 도움이 된다고 판단하며, 이를 기준으로 임계 값 이상의 문서만을 LLM의 최종 입력으로 사용하게끔 모델링하는 것이 목표이다. 이를 통해 질문에 정확히 부합하는 정보를 가진 문서만을 사용하여 답변의 정확도와 신뢰성을 더욱 높일 수 있을 것이다.

4. Fine Tuning

Foundation model을 fine-tuning하기 위해 이전 유명한 연구들에서 적용한 기법을 사용합니다.

4.1 QLoRA

QLoRA는 Model Quantization와 LoRA(Low-Rank Adaptation)를 결합한 기법으로, 두 가지 주요 방법론을 통해 효율적으로 모델을 미세 조정합니다.

- Model Quantization: Model Quantization는 모델 파라미터의 데이터 타입을 변경하여 모델의 메모리 용량을 줄이는 기술입니다. 일반적으로 32비트 부동 소수점 수치를 사용하는 모델 파라미터를 8비트 또는 16비트 정수로 변환하여 메모리 사용량과 계산 비용을 크게 줄입니다. 이 방법은 모델의 추론 속도를 향상시키고, 메모리 비용을 절감하며, 대규모 모델을 더 작은 하드웨어에서 실행할 수 있도록 합니다.
- Low-Rank Adaptation (LoRA): LoRA는 모델의 전체 파라미터를 조정하지 않고, 모델의 특정 부분에 저차원 행렬을 추가하는 방식으로 미세 조정을 수행합니다. 이는 Full Fine-tuning에 비해 훨씬 적은 양의 추가 파라미터만을 학습시키기 때문에 계산 비용이 크게 줄어듭니다. LoRA는 주로 트랜스포머 아키텍처에서 어텐션 헤드와 같은 특정 모듈에 적용되며, 이는 모델의 성능 저하 없이 효율적인 미세 조정을 가능하게 합니다.

4.2 Fine-tuning with human feedback

사람 피드백을 통한 미세 조정은 모델이 주어진 프롬프트에 대해 여러 응답을 생성하고, 사람이 선호하는 응답을 선택하여 이를 학습하는 반복 과정입니다. 각 단계에서 모델은 새로운 응답을 생성하고, 사람 평가자가 더 선호하는 응답을 선택합니다. 이렇게 수집된 선호도 데이터를 사용하여 모델을 미세 조정합니다. 반복을 통해 모델은 점점 사람의 선호도에 더 가까운 응답을 생성하게 됩니다. 이 방식은 모델의 출력이 사람의 가치와 선호도에 부합하도록 하여 응답의 품질과 관련성을 향상시키고, 변화하는 선호도에 동적으로 적응할 수 있게 합니다. 결과적으로 모델은 사람과 유사한, 선호되는 응답을 생성하여 실제 활용도와 신뢰성을 높일 수 있습니다.

비고*2	