JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 7, pp. 81-88, July 2024
https://doi.org/10.9708/jksci.2024.29.07.081

Design of a Question-Answering System based on RAG Model for Domestic Companies

Gwang-Wu Yi*, Soo Kyun Kim*

*Student, Dept. of Computer Engineering, Jeju National University, Jeju, Korea
*Professor, Dept. of Computer Engineering, Jeju National University, Jeju, Korea

[Abstract]

Despite the rapid growth of the generative AI market and significant interest from domestic companies and institutions, concerns about the provision of inaccurate information and potential information leaks have emerged as major factors hindering the adoption of generative AI. To address these issues, this paper designs and implements a question-answering system based on the Retrieval-Augmented Generation (RAG) architecture. The proposed method constructs a knowledge database using Korean sentence embeddings and retrieves information relevant to queries through optimized searches, which is then provided to the generative language model. Additionally, it allows users to directly manage the knowledge database to efficiently update changing business information, and it is designed to operate in a private network to reduce the risk of corporate confidential information leakage. This study aims to serve as a useful reference for domestic companies seeking to adopt and utilize generative AI.

▸ **Key words**: Retrieval-Augmented Generation, Generative AI(Generative Language Model), Question-Answering System, Korean Sentence Semantic Search

[요 약]

생성형 AI 시장의 급속한 성장과 국내 기업과 기관의 큰 관심에도 불구하고, 부정확한 정보제공과 정보유출의 우려가 생성형 AI 도입을 저해하는 주된 요인으로 나타났다. 이를 개선하기 위해 본 논문에서는 검색-증강 생성(Retrieval-Augmented Generation, RAG) 구조 기반의 질의응답시스템을 설계·구현하였다. 제안 방법은 한국어 문장 임베딩을 사용해 지식 데이터베이스를 구축하고, 최적화된 검색으로 질문 관련 정보를 찾아 생성형 언어 모델에게 제공된다. 또한, 이용자가 지식 데이터베이스를 직접 관리하여 변경되는 업무 정보를 효율적으로 업데이트하도록 하고, 시스템이 폐쇄망에서 동작할 수 있도록 설계하여 기업의 기밀 정보의 유출 가능성을 낮추었다. 국내 기업 등 조직에서 생성형 AI를 도입하고 활용하고자 할 때 본 연구가 유용한 참고자료가 되길 기대한다.

▸ **주제어**: 검색-증강 생성, 생성형 AI(생성형 언어 모델), 질의응답시스템, 한국어 문장 의미 검색

• First Author: Gwang-Wu Yi, Corresponding Author: Soo Kyun Kim
*Gwang-Wu Yi (marilynext@naver.com), Dept. of Computer Engineering, Jeju National University
*Soo Kyun Kim (kimsk@jejunu.ac.kr), Dept. of Computer Engineering, Jeju National University
• Received: 2024. 05. 23, Revised: 2024. 07. 15, Accepted: 2024. 07. 16.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

블룸버그 인텔리전스(Bloomberg Intelligence)의 생성형 AI 시장 전망에 따르면 전 세계적으로 생성형 AI 시장이 매년 평균적으로 42% 성장하여 2032년에는 1.3조 달러로 2022년 대비 30배 이상 매우 가파른 성장을 할 것이라 예상한다[1].

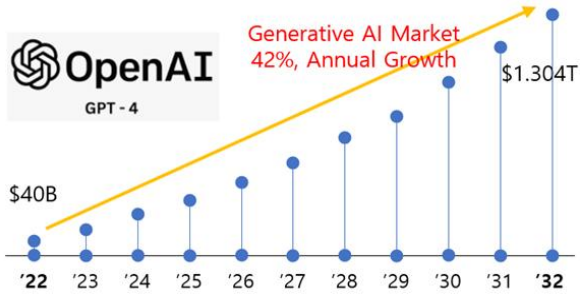

Fig. 1. Generative AI Market Outlook, Reconstructed [1]

최근 ChatGPT의 선풍적인 인기에 힘입어, 국내의 많은 기업과 기관들도 생성형 AI를 도입하는 데 적극적으로 움직이며 관심을 나타내고 있다. 국내의 한 IT 기업이 조사한 '기업 및 기관의 생성형 AI 활용현황'에 따르면, 국내 기업과 기관의 보안, IT 담당자들은 18.6%가 생성형 AI를 도입하여 활용 중이고 57.8%가 도입을 고려하고 있다고 응답하였다[2]. 하지만 그림2와 같이 도입을 주저하게 만드는 우려도 동시에 존재한다. '정부 부문 생성형 AI 챗봇 활용현황 및 전망에 관한 공무원 인식 조사'에서 본인 업무에 생성형 AI를 활용할 경우, 예상되는 문제점으로 부정확한 정보제공(71.3%), 기밀 유출 등 공공정보 보안(51.2%)이 과반으로 높은 응답 비중을 차지하였다[3].

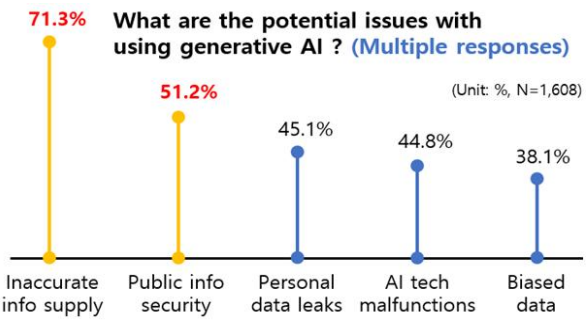

Fig. 2. Survey on Concerns from Generative AI Use, Reconstructed [3]

설문 조사의 응답은 ChatGPT와 같은 범용 생성형 AI의 사용 경험을 기반으로 하는데, 이러한 관점에서 기업과 기관이 우려하는 생성형 AI의 '부정확한 정보제공'의 원인을 다음 세 가지로 분석해 볼 수 있다.

첫 번째, 고유 업무영역에 대한 불충분한 학습이다. 기업과 기관의 업무는 매우 개별적이고 구체적인데, 인터넷에 공개된 정보를 기반으로 학습한 범용 생성형 AI에게 고유 업무에 관한 질문하였을 때 정확한 답변이 기대하기 어렵다. 따라서 고유 업무영역에 대한 추가적인 파인튜닝을 진행하거나 고유 업무영역의 정보를 우선 고려하여 답변하는 기술 구조가 필요하다.

두 번째, 생성형 AI의 최종 학습 시점에 기인하는 부정확한 정보제공 문제이다. 기업과 기관의 업무는 다양한 요인에 따라 수시로 변경된다. 하지만 생성형 AI는 태생적으로 최종 학습 이후에 변경되거나 새로 발생한 정보에 대해서는 부정확한 정보를 제공하거나 의미 없는 답변을 하는 이른바 환각(hallucination) 현상이 발생 될 수 있다. 생성형 AI의 지속 가능한 사용을 위해서는 변경되는 정보를 적기에 효과적으로 반영할 수 있어야 한다.

세 번째, 일반적으로 생성형 AI는 제공정보의 출처를 제시하지 않는다. 정보의 정확성에 관한 판단은 그 내용뿐만 아니라 정보의 출처 제공 여부도 고려된다. 하지만 다양한 지식을 학습하는 과정에서 언어 모델 내부에 파라미터화(parameterized) 되는 기술적 특성 때문에 생성형 AI의 대답을 보고서에 활용하고자 하였을 때 해당 내용의 출처를 명확히 제시할 수 없다. 대답에 대한 출처나 근거를 함께 제공한다면 생성형 AI의 신뢰성을 크게 높일 수 있을 것이다.

또한, 설문 결과에 따르면 생성형 AI의 '부정확한 정보제공'과 더불어 '정보 보안' 문제도 우려하고 있다. 이는 기업과 기관의 이용자가 국내·외 민간이 운영하는 생성형 AI 시스템과 문답을 주고받으면서 기밀 유출되는 우려를 의미한다. 이러한 우려 때문에 일부 대기업에서는 직원들의 ChatGPT 사용을 차단하고[4] 공공부문에서는 내부 업무용 생성형 AI 도입 시 인터넷 등 외부망과 분리해서 구축·운영하도록 하는 지침을 수립하기도 하였다[5].

본 논문은 이러한 생성형 AI의 한계점을 개선하여 국내 기업의 업무 환경에 적합한 질의응답시스템을 설계하고 구현하는 것을 목표로 한다. 본 연구에서는 RAG 구조를 적용하여 지식 데이터베이스 내에서 질문과 관련된 유용한 정보를 검색하고, 이를 바탕으로 생성형 언어 모델이 답변을 생성하도록 한다. 이를 실제로 구현하기 위해 적합한 한국어 문장 임베딩 모델과 생성형 언어 모델을 선정하여 구성한다. 또한, 효과적인 지식 데이터베이스의 구조를 설계하고 지식의 검색, 업데이트와 같은 정보 관리 기법을 최적화한다.

II. Preliminaries

1. Related works

1.1. Retrieval-Augmented Generation

RAG(Retrieval-Augmented Generation) 구조는 자연어 처리 분야에서 실시간 정보 검색과 생성형 언어 모델을 통합한 방식이다. 이 방식은 지식 데이터베이스에서 질문과 관련된 정보를 검색하여 생성형 언어 모델에 제공함으로써 이를 기반으로 더 정확하고 관련성 높은 응답을 생성하는 데 도움 준다. 특히, 생성형 언어 모델이 사전에 학습하지 못한 다양한 도메인의 구체적인 지식이 요구되는 오픈-도메인 질의(Open-domain Question)에 대해서 우수한 응답을 보인다[6].

또한, 기업 등 조직에서 질의응답시스템의 구축하고 유지보수하는 측면에 있어 RAG 구조는 비용을 절감시키고 효율적인 방법을 제공한다. 조직은 고유한 업무 정보를 규정, 지침, 보고서 등 문서의 형태로 보유하고 있다. 이러한 문서들을 단순히 쪼개어 의미 검색이 가능한 벡터 형태로 변환·저장하는 작업을 통하여 손쉽게 지식 데이터베이스를 구축할 수 있다. 이후 정보를 최신 상태로 유지하기 위해 지식 데이터베이스를 수시로 업데이트하면 된다.

반면에, 사전 학습된 언어 모델(pre-trained LM)을 기반으로 특정 업무 분야에 맞게 파인튜닝하는 전략을 선택한다면, 해당 업무 지식에 대한 데이터셋을 준비하고 언어 모델을 훈련 시켜야 한다. 또한, 업무 정보에 작은 변경이 있을 때마다 반복적으로 파인튜닝을 해야 한다. 이는 특히 인공지능 전문가나 고성능 서버 자원이 부족한 중·소규모 조직에 상당한 부담이 될 수 있다.

1.2. Sentence-BERT

센텐스-BERT(Sentence-BERT:SBERT)[7]는 기존 BERT 모델을 변형하여 문장 간의 의미적 유사성을 효율적으로 계산하기 위해 개발된 언어 모델이다. SBERT의 주요 기능은 문장이나 문단을 고정된 크기의 벡터로 변환하는 것이다. 다른 말로 '임베딩을 생성한다.'라고 표현할 수도 있다. 이 벡터들을 활용하여 문서 분류, 클러스터링, 질의응답시스템 등 자연어 처리 과제에서 의미적 유사성을 빠르고 정확하게 계산할 수 있다.

RAG 구조에서 SBERT의 사용은 지식 데이터베이스를 만들고 검색하는 데 있어 매우 중요하다. 문서를 정해진 크기의 문단으로 나누고, 이 문단들을 고정된 크기의 벡터로 변환하여 저장함으로써 지식 데이터베이스를 생성한다. 여기서 SBERT는 문장이나 문단을 의미가 풍부한 벡터로 바꾸는 역할을 한다. 문장이나 문단 간의 유사도를 정확하게 비교하기 위해서는, 이렇게 생성된 벡터들을 코사인 유사도와 같은 방법을 사용하여 서로 비교한다. 질의응답시스템에 적용한다면 이용자의 질문을 벡터로 변환한 후, 지식 데이터베이스 내에 저장된 문단의 벡터들과 비교한다. 이를 통해 의미상 가장 유사한 문단을 찾아낼 수 있다. 이 문단은 이용자의 질문과 함께 생성형 언어 모델에 전달되는데, 질문에 대한 관련성 높고 정확한 답변을 생성하는 데 중요한 역할을 하게 된다.

1.3. Open-source Generative Language Model

최근 몇 년간, Llama2, Mistral, Vicuna 등 다양한 생성형 언어 모델이 오픈소스로 공개되었다. 관련 학계나 개인 등은 이러한 모델을 기반으로 특정 작업이나 도메인에 높은 적합성을 확보하기 위해 파인튜닝 작업을 진행하였고 이 중 일부는 공개되어 사용할 수 있다. 특히, 한국지능정보사회진흥원과 국내 한 기업이 공동 주최하는 Open Ko-LLM 리더보드[8]에는 '24년 3월 기준, 1,100여 개의 한국어 생성형 모델이 등록되어 있다. 추론능력, 상식능력, 언어이해력 등의 평가지표를 평가하여 실시간으로 순위 경쟁이 진행되고 있다. 이에 따라, 시스템 구축을 고려하는 조직에서는 구축 시점에, 라이선스 정책 등을 고려하여 리더보드 상위권의 모델을 선택하는 방법도 구축 기간을 단축하고 성능을 높일 수 있는 대안이 될 수 있겠다.

1.4. Prompt engineering

프롬프트-엔지니어링(Prompt engineering)은 생성형 언어 모델에게 어떻게 질문 또는 지시를 해야 원하는 결과를 얻을 수 있는지를 연구하는 분야이다. ChatGPT 개발사인 OpenAI에서는 좋은 프롬프트를 작성하기 위에 다음과 같이 6가지 가이드를 제시하고 있다. ① 질문을 세부적으로 작성 ②모델에게 특정 역할 부여 ③ 가독성을 높이기 위한 구분자 사용("", <> 등) ④ 작업 처리 절차 명시 ⑤예시 제공 ⑥ 출력의 길이 명시이다[9]. 질문과 함께 유용한 정보가 제공되는 RAG 구조를 사용할 경우, 지시, 질문, 참고정보 등 구성요소를 효과적으로 활용하여 프롬프트를 설계하여야 한다.

III. The Proposed Scheme

1. System Architecture

제안하는 질의응답시스템 설계의 주안점은 한국어 문장 임베딩 모델을 사용하여 기업이 보유한 문서로 지식 데이터베이스를 구축하고, 최적의 검색기법을 통해 질문과 연

관된 정보를 찾아 생성형 언어 모델에게 전달하여 정확한 답변을 생성을 돕는 것이다. 또한, 답변의 출처라고 할 수 있는 생성형 언어 모델이 참고한 정보를 사용자에게도 제공하여 답변의 신뢰성을 높이도록 설계한다.

그리고 시스템은 주요 구성요소인 문장 임베딩 모델, 지식 데이터베이스, Retrieval, 생성형 언어 모델이 인터넷 등 외부망과 통신이 필요하지 않도록 설계하여 보안성을 높인다. 시스템의 전반적인 기능은 그림3과 같이 번호 순서에 따라 동작하도록 설계한다.

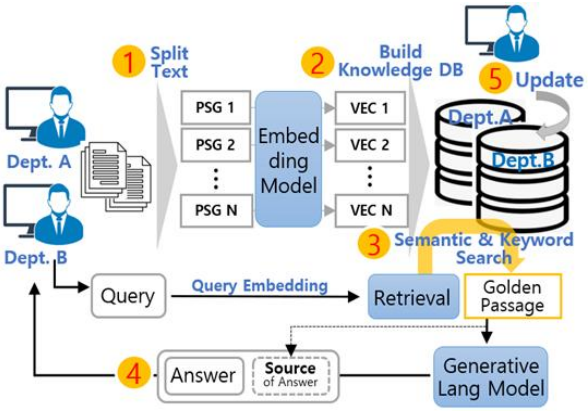

Fig. 3. System Architecture

① 지식 데이터베이스로 만들고자 하는 문서를 일정한 크기의 문단으로 쪼갠다. ② 임베딩 모델을 이용하여 각 문단을 벡터로 변환하고 부서별 지식 데이터베이스에 저장한다. ③ 이용자의 질문을 벡터로 변환하고 Retrieval를 이용하여 지식 데이터베이스에서 유사도가 높은 문단을 찾는다. ④ 이용자의 질문과 검색된 문단을 생성형 언어 모델에 보내 답변을 받는다. ⑤ 업무 정보가 변경되면 이용자가 직접 부서별 지식 데이터베이스를 업데이트한다.

2. Application Design

2.1 Sentence Embedding

본 시스템에서는 문장의 의미적 유사도를 비교하는 작업에 효과적인 Sentence-BERT 기반의 한국어 문장 임베딩 모델을 사용하도록 설계하였다.

'ko-sroberta-multitask[10]' 는 사전에 학습된 BERT 모델을 바탕으로, '한국어 자연언어 추론(KorNLI)과 문장 유사도(KorSTS) 데이터셋[11]'을 추가 학습한 한국어 문장 임베딩 모델이다. 해당 모델을 이용하여 문장 또는 문단을 입력값으로 임베딩하면 768차원의 고정된 크기의 벡터값을 생성할 수 있다.

본 시스템에서 임베딩 모델은 이용자가 문서를 문단으로 쪼개 벡터로 변환·저장하여 지식 데이터베이스를 생성하는 작업과 이용자의 질문을 벡터로 변환하여 지식 데이터베이스에서 유사 문단을 검색하는 작업에 사용된다.

2.2 Data Structure

지식 데이터베이스는 처리하는 데이터의 형태에 따라 효과적인 데이터의 생성(insert), 검색(select), 변경(update), 삭제(delete)를 위해 두 종류의 데이터베이스를 중첩하여 사용하도록 설계하였다. 텍스트 데이터 처리에는 Mongo DB[12], 벡터 데이터 처리는 Chroma DB[13]를 사용한다. Chroma DB는 오픈 소스 임베딩 데이터베이스로 벡터의 저장, 색인, 검색 등 벡터 데이터의 처리에 특화된 DB라고 할 수 있다. 문서를 문단으로 쪼개어 문단의 원문인 텍스트는 Mongo DB에, 문단의 벡터값은 Chroma DB 각각 저장된다.

두 DB 간의 연계(relation)는 각 문단의 공통된 고유번호를 통해 이루어진다. Chroma DB에서 유사 문단(벡터)을 검색하고 해당 문단의 고유번호로 Mongo DB에서 문단의 원문(텍스트)을 찾아내는 구조이다.

2.3 Retrieval Design

본 시스템은 유사한 문단을 효과적으로 검색하기 위해 벡터 간 연산을 통한 의미 검색과 전통적인 키워드 검색기법을 결합하여 사용하도록 설계하였다. 관련 연구에 따르면, 이 두 방법을 결합하는 것이 각각 단독으로 사용하는 것보다 검색 성능이 더 우수였다[14].

구체적인 방법은 이용자가 지정한 부서에 해당하는 Chroma DB의 검색을 통해 의미상 유사도가 높은 4개의 후보 문단을 추출하고, BM25[15] 알고리즘을 이용한 키워드 검색을 통해 또 다른 4개의 후보 문단을 추출한다. 이렇게 추출된 문단 중 중복되는 항목을 제거한 후, 최종적으로 최소 4개에서 최대 8개의 후보 문단을 선정한다.

이 후보 문단들은 생성형 언어 모델로 전송되고 실제로 필요한 정보를 판단하여 답변을 생성하게 된다. 이러한 방식은 생성형 언어 모델이 이용자의 질문에 대해 더 관련성 있고 정확한 답변을 만들 확률을 높일 수 있다.

2.4 Generative Language Model

본 시스템은 성능이 향상된 생성형 언어 모델이 지속적으로 공개되는 여건을 고려하여 그림4과 같이 간단한 설정 변경을 통해 원하는 모델과 통신할 수 있도록 설계하였다. 생성형 언어 모델이 구동되는 서버 주소를 운영자가 설정할 수 있도록 하여 해당 주소로 API를 호출함으로써 이용자가 질의응답을 주고받는 방식을 사용한다. 실제 생성형 언어 모델은 논문 작성 시점에 “Open Ko-LLM 리더보드” 상위권의 모델을 사용하였다.

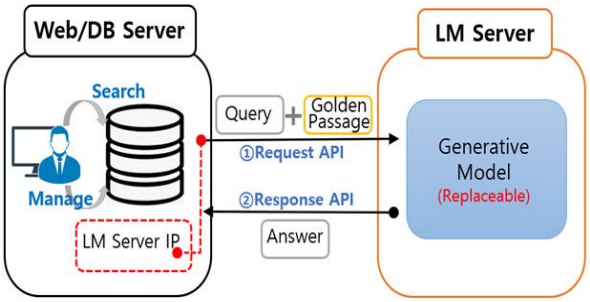

Fig. 4. Communication with Generative Language Models

2.5 Prompt Design

프롬프트는 이용자가 생성형 언어 모델이 특정 작업을 수행하도록 지시하는 자연어 메시지이다. 생성형 언어 모델로부터 정확한 답변을 얻기 위해서 RAG 구조의 특성을 고려하여 프롬프트를 설계하였다.

프롬프트는 일반적으로 지시, 문맥, 입력 데이터, 출력 지시자로 구성된다[16]. 지시항목에는 생성형 언어 모델에게 개인 비서 역할(role)을 부여하고 다음 두 가지 유형으로 모두 답변하도록 지시한다. ① 제공하는 유용한 문단을 활용하여 답변하고 ② 제공하는 정보를 무시하고 생성형 언어 모델이 학습한 지식으로만 스스로 답변하도록 한다. 프롬프트의 문맥항목은 Retrieval을 통해 검색한 4~8개의 후보 문단을 제공하고, 입력 데이터 항목은 이용자의 질문을 그대로 사용한다. 출력 지시자 항목은 지시항목의 내용에 따라 양식을 준수하여 한국어로 답변하도록 한다.

2.6 Development Environment

표1은 제안 방법에 대한 개발환경을 보여준다. 웹서비스와 문장 임베딩 모델을 구동하는 서버와 생성형 모델을 구동하는 서버로 개발환경을 구성하였다.

Table 1. Development Environment

Specification	Server A	Server B
Service	Web(UI, DB), Embedding model	Generative language model
CPU	8th Gen Intel(R) Core(TM) i7-8565U	AMD EPYC 7B12
RAM	DDR4 8GB	DDR4 51GB
GPU	-	NVIDIA A100-SXM4-40GB
OS	Windows 10	Ubuntu 22.04.3

IV. Result

질문과 함께 관련된 정보를 생성형 언어 모델에게 적절히 제공하면, 정보를 제공하지 않았을 때보다 정확한 답변을 하는 것이 RAG 모델의 구조적 특징이다. 이에 따라, 본 논문에서는 두 가지 대표 질문유형을 시스템에 질의하였다. 그 결과, 시스템이 RAG 구조와 프롬프트의 설계 목적에 부합하게 동작함을 확인하였다.

1. Domain-Specific Q&A

이용자가 그림5의 표식Ⓐ와 같이 질문에 참조할 부서 단위 지식 데이터베이스를 설정하고 질문을 하면 답변의 근거와 함께 답변을 받을 수 있다. 사용된 질문은 특정 업계의 고유 업무영역에 관한 구체적인 질문이다.

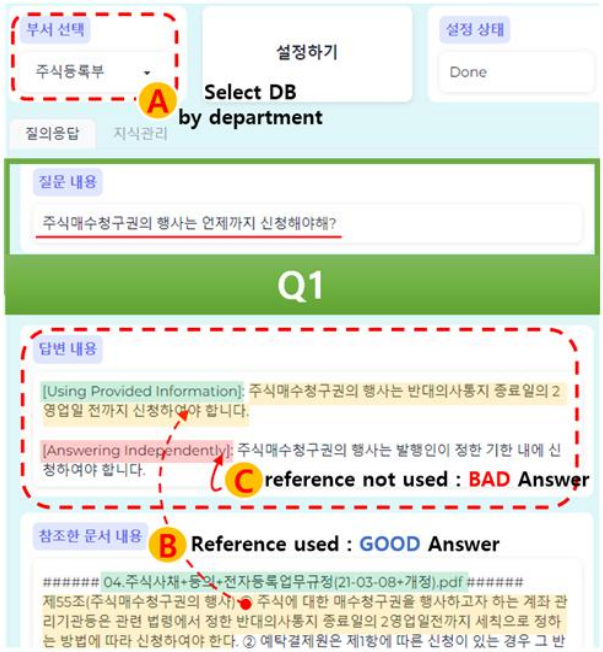

Fig. 5. Domain-Specific Question and Answer ❶

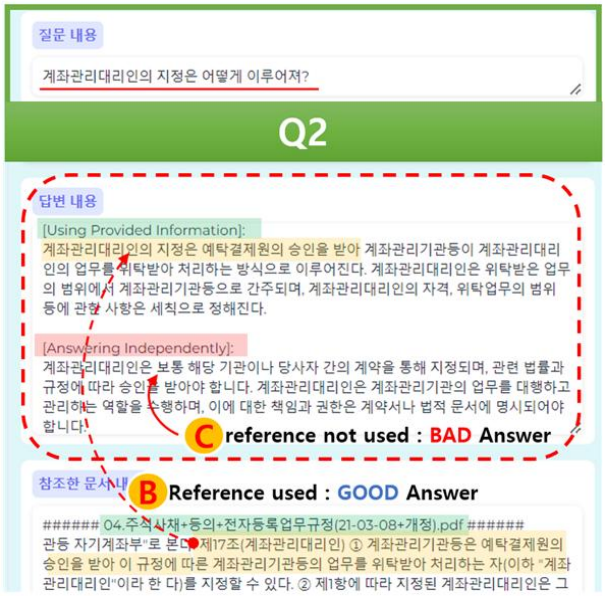

Fig. 6. Domain-Specific Question and Answer ❷

그림5, 6과 같이 미리 설계된 프롬프트에 따라서 두 가지 유형으로 대답을 한다. 첫 번째는 지식 데이터베이스에서 검색된 유용한 정보를 활용하여 답변하는데, 그림5, 6의 표식Ⓑ와 같이 제공된 정보를 참조하여 정확한 답변을 제시하였다. 두 번째는 제공된 정보를 무시하고 생성형 언어 모델의 학습 지식을 바탕으로 스스로 답변하는데, 그림5, 6의 표식Ⓒ와 같이 부정확한 답변을 하였다.

RAG 구조의 시스템이 제대로 동작함을 확인하였고 인터넷 등에 공개된 정보로만 학습한 생성형 언어 모델 단독으로는 특정 업계의 고유 업무영역에 대한 정확한 답변이 어려움을 알 수 있다.

2. General Knowledge Q&A

사용된 질문은 대한민국의 수도는 어디인지에 관한 일반적인 상식 영역의 질문이다. 이번에도 미리 설계된 프롬프트에 따라서 두 가지 유형으로 대답을 한다.

첫 번째는 지식 데이터베이스에서 검색된 유용한 정보를 활용하여 답변하는데, 그림7의 표식Ⓐ와 같이 제공된 정보로 판단할 수 없다는 답변을 하였다. 제공된 정보는 증권업계에 관한 정보로 대한민국의 수도가 어디인지에 관한 정보는 없다. 환각(hallucination) 현상 예방 측면에서 모르면 모른다고 답하는 것이 중요한데, 설계된 프롬프트를 준수한 적절한 답변이라고 할 수 있다.

두 번째는 제공된 정보를 무시하고 생성형 언어 모델의 학습 지식을 바탕으로 스스로 답변하는데, 표식Ⓑ와 같이 적절한 답변을 하였다. 일반영역에 관한 질문에는 생성형 언어 모델 단독으로 답변 가능함을 알 수 있다.

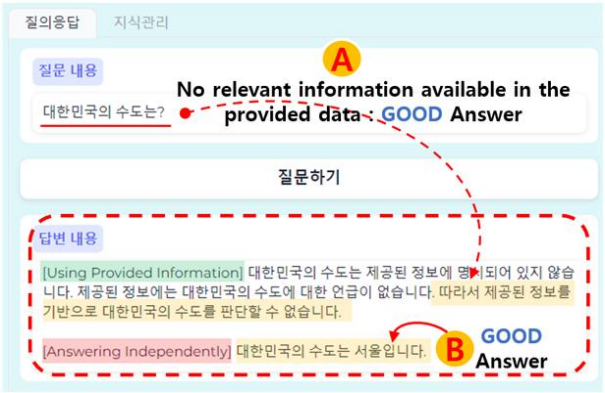

Fig. 7. General Knowledge Question and Answer

3. Knowledge Database Management

이용자가 부서별로 지식 데이터베이스를 등록할 수 있는 인터페이스를 제공한다. 그림8과 같이 이용자는 부서를 선택하고 한 개 이상의 문서를 업로드 가능하고 '등록'버튼을 누르면 문서는 문단 단위로 쪼개어져 지식 데이터베이스에 등록된다. 등록처리 후 '등록상태'에 처리결과를 반환한다.

Fig. 8. Knowledge Database Registration

또한, 지식 데이터베이스를 최신의 상태로 유지하기 위해 이용자가 부서별로 지식 데이터베이스에 등록된 지식을 변경할 수 있는 인터페이스를 제공한다. 그림9와 같이 이용자는 삭제하고자 하는 문서를 선택하고 '삭제'버튼을 누르면 삭제처리가 된다. 이후 상단의 문서등록 메뉴를 통해 수정된 문서를 다시 등록하면 변경처리는 완료된다

Fig. 9. Knowledge Database Update

V. Conclusion

본 연구에서는 국내 기업과 기관의 생성형 AI 도입에 있어 주요 우려 사항인 부정확한 정보제공, 정보유출 등 보안 문제를 조명해보고 이러한 문제점을 개선하기 위해 RAG 구조 기반의 질의응답시스템을 설계하고 구현하였다.

제안 방법은 기업의 고유 업무영역에 대한 생성형 언어 모델의 답변 정확성과 신뢰성을 높이기 위해, 한국어 문장 임베딩 모델을 이용하여 기업이 보유한 문서로 지식 데이터베이스를 구축하고, 의미 검색과 키워드 검색을 조합한 검색기법을 통해 질문과 연관된 정보를 찾도록 하였다. 이러한 정보를 생성형 언어 모델에게 전달하여 정확한 답변

생성에 도움을 주었다.

또한, 문장 임베딩 모델, 생성형 언어 모델 등 시스템의 모든 구성요소가 인터넷 등 외부망과 통신이 불필요하도록 구성하여 기밀 정보의 유출 가능성을 낮추었다.

RAG 구조의 특징에 기인하는 문제지만, 본 논문에서 제안한 질의응답시스템의 성능은 Retrieval의 성능에 크게 의존적이다. 따라서 향후 연구에서는 Retrieval 성능을 더욱 개선하기 위해 다양한 실험을 설계하고 수행할 계획이다. 다양한 방식의 Retrieval 성능을 측정함으로써, 성능 향상의 객관적인 데이터를 확보하고 생성형 언어 모델의 답변 정확성과 신뢰성을 더욱 높이는 방향으로 연구를 진행하고자 한다.

또한, 본 시스템을 실제 기업에 적용하기 위해서는 대규모 데이터 처리, 데이터베이스 관리 등 효율적인 유지보수 방안 마련을 위한 추가적인 연구가 필요하겠다.

ACKNOWLEDGEMENT

This research was supported by the 2024 scientific promotion program funded by Jeju National University.

REFERENCES

[1] Bloomberg Intelligence, Generative AI to Become a $1.3 Trillion Market by 2032, "https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/"

[2] Fasoo, Current Status of Enterprise and Institutional Generative AI Usage, "https://www.fasoo.com/press-release/20230821"

[3] Korea Institute of Public Administration, Survey on Government Sector Generative AI Chatbot Usage and Prospects: Awareness of Public Officials, "https://www.kipa.re.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000017801&fileSn=12"

[4] EBN, Prevent Confidential Information Leakage... Samsung, LG, SK Take Caution against ChatGPT, "https://m.ebn.co.kr/news/view/1577665"

[5] National Intelligence Service, Security Guidelines for Generative AI Usage Including ChatGPT, "https://www.ncsc.go.kr:4018/main/cop/bbs/selectBoardArticle.do?bbsId=InstructionGuide_main&nttId=54340&pageIndex=1"

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Proceedings of the 34th International Conference on NIPS 2020, pp. 9459-9474, December 2020. DOI: 10.48550/arXiv.2005.11401

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of the EMNLP 2019, January 2019. DOI: 10.18653/v1/D19-1410

[8] NIA and Upstage, Open Ko-LLM Leaderboard, "https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard"

[9] OpenAI, Six strategies for getting better results, "https://platform.openai.com/docs/guides/prompt-engineering"

[10] Jhgan00, ko-sroberta-multitask, "https://github.com/jhgan00/ko-sentence-transformers"

[11] J. Ham, Y. J. Choe, K. Park, I. Choi, and H. Soh, "KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding," Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 422-430, November 2020. DOI: 10.18653/v1/2020.findings-emnlp.39

[12] MongoDB, Inc., MongoDB, "https://www.mongodb.com/ko-kr/what-is-mongodb"

[13] J. Huber and A. Troynikov, Chroma DB, "https://docs.trychroma.com"

[14] S. W. Kim and G. R. Park, "Deep Learning Based Semantic Similarity for Korean Legal Field," KIPS Transactions on Software and Data Engineering, vol. 11, no. 2, pp. 93-100, February 2022. DOI: 10.3745/KTSDE.2022.11.2.93

[15] S. E. Robertson and S. Walker, "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval," Proceedings of the 17th Annual International ACM-SIGIR Conference, pp. 232-241, January 1994. DOI: 10.1007/978-1-4471-2099-5_24

[16] DAIR.AI, Elements of a Prompt, "https://www.promptingguide.ai/introduction/elements“

Authors

Gwang-Wu Yi is currently a student in the Department of Computer Engineering, Jeju National University. He has also been working in the IT department of the Korea Securities Depository (KSD) since 2013. He has an interest in artificial intelligence and computer security.

Soo Kyun Kim received Ph.D. in Computer Science & Engineering Department of Korea University, Seoul, Korea, in 2006. He joined the Telecommunication R&D Center at Samsung Electronics Co., Ltd., in 2006 and 2008. He is now a professor at the Department of Computer Engineering at Jeju National University, Korea. Dr. Kim has published many research papers in international journals and conferences. His research interests include multimedia, pattern recognition, image processing, mobile graphics, geometric modeling, and interactive computer graphics. He is a member of ACM, IEEE, IEEE CS, KACE, KMMS, KKITS, and KIIT.