

Development of Question-Answering System Based on Regulatory Documents for Korea Atomic Energy Research Institute

Hongbi Kim ^a, Yonggyun Yu ^{ab*}

^a Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, 34057, Republic of Korea

^b University of Science & Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Republic of Korea

*Corresponding author: ygyu@kaeri.re.kr



I. Introduction

- 배경 : BERT, GPT-3와 같은 대규모 언어 생성 모델이 등장하면서 챗봇이나 질의응답(Question-Answer, QA) 시스템 등에 혁신적인 변화를 가져옴
- 문제 상황 : 전문 도메인에 대한 이해 부족
- 목적 : 연구원들이 **규정에 대한 질문에** 빠르고 정확한 답변을 얻을 수 있도록 함
- 한국원자력연구원의 규정 문서를 기반으로 답변을 제공하는 QA 시스템을 제시

II-1. Preprocessing Regulatory documents

- 질문과 관련 있는 단락을 빠르고 정확하게 추출하기 위한 사전 처리 과정

1 데이터 전처리

- 전처리는 질문에 관련된 단락을 효과적으로 추출할 수 있는 첫걸음

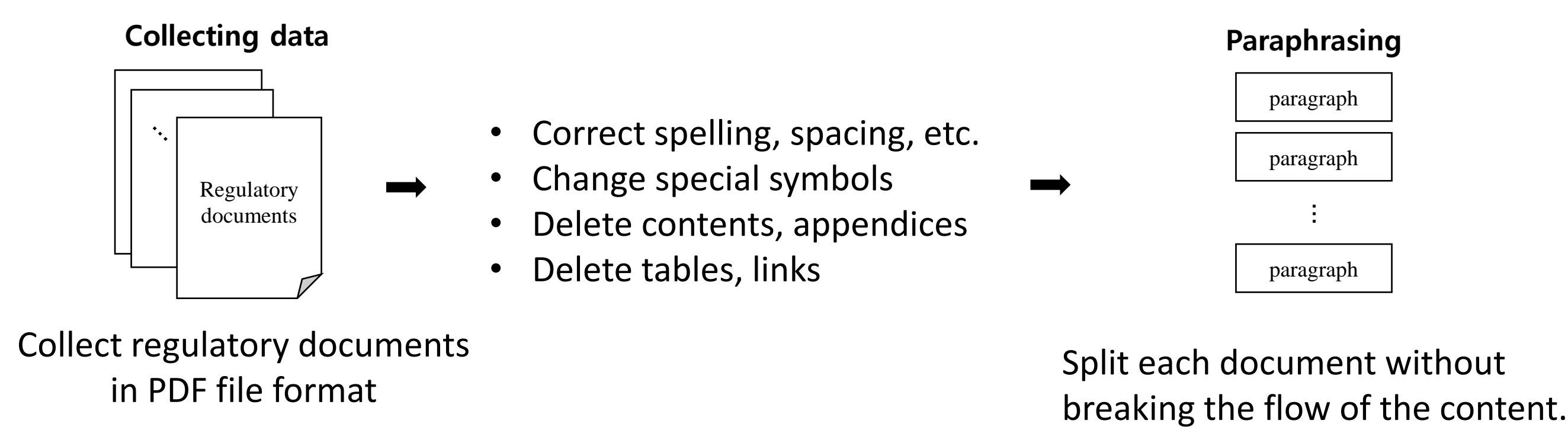


Fig. 1. Structure of the data preprocessing

2 임베딩 계산

- OpenAI에서 제공하는 text-embedding-ada-002 모델 사용
- 텍스트 → 벡터로 변환
- 질문과 관련 있는 단락을 빠르게 추출하기 위한 사전 수행 과정

II-2. Extract paragraphs related to a question

- 질문과 가장 관련 있는 문서 단락을 추출하기 위해 2개의 검색 알고리즘을 활용

Case1) 코사인 유사도(Cosine similarity) 알고리즘

각 단락의 텍스트 벡터와 질문 텍스트 벡터 사이의 유사도 계산

Case2) BM25 모델 : 질문과 단락 사이의 연관성을 평가

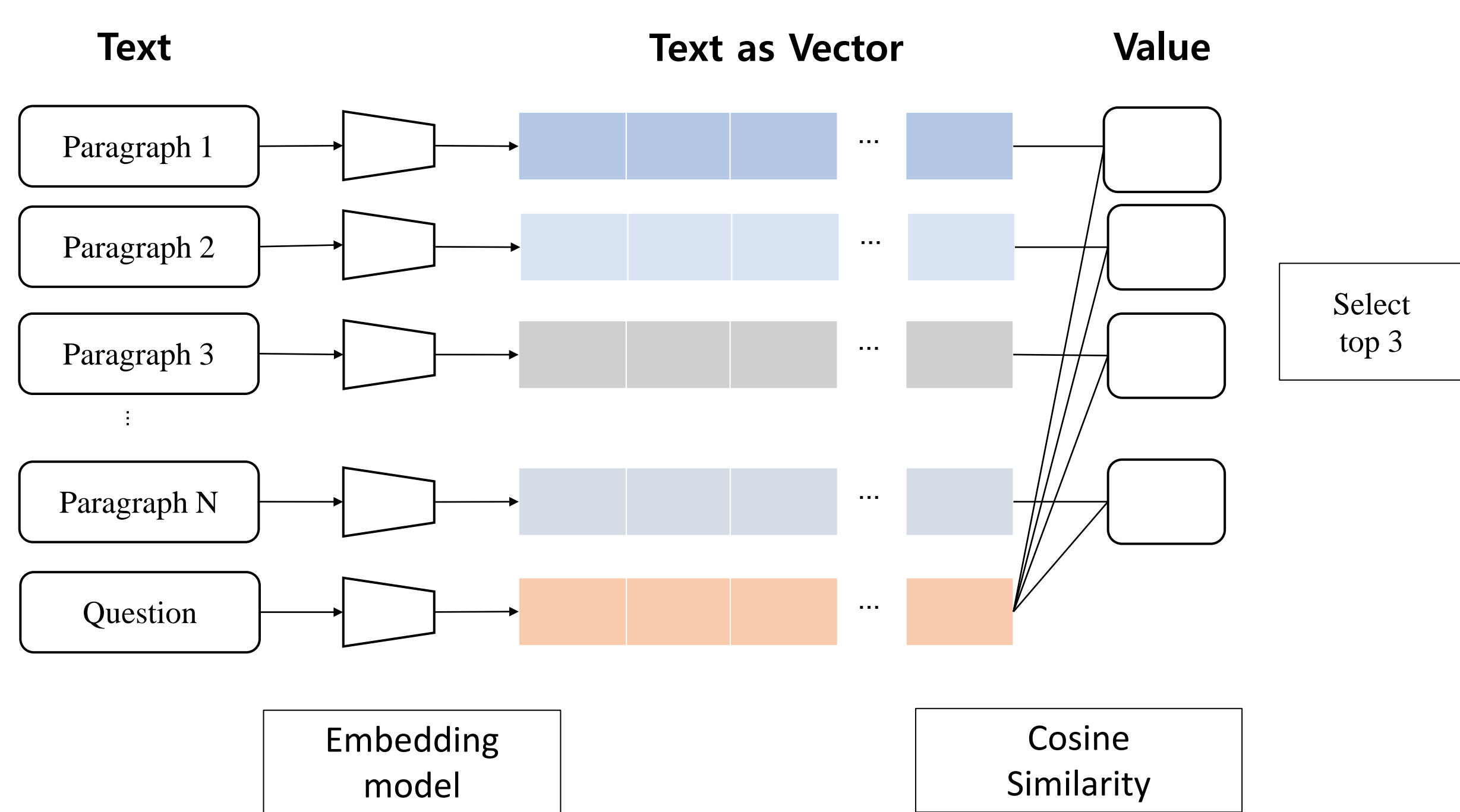


Fig. 2. Structure of the process of selecting the top three paragraphs using embedding model.

III. Generating answers based on documents

- OpenAI의 대화형 언어 모델인 GPT-3.5-turbo 모델 사용 : 자연어 이해 및 생성 가능
 - 프롬프트(prompt)를 통해 모델에 역할을 부여하고, **질문과 문서 정보를 이해 및 활용**할 수 있도록 함
- 본 논문을 게재한 당시의 역할과 가이드라인은 다음을 따름:

“You are a large language model whose expertise is **understanding company regulatory documents.** ”
 “You are given a query and a series of text embedding from a paper in order of their cosine similarity to the query. ”
 “You must take the given embedding and return a very accurate answers the query. ”
 “Given the question: **user’s question and the following embedding as three extracted paragraphs.** ”
 “Return a detailed answer in Korean based on the documents: ”

KAERI Question-Answering System

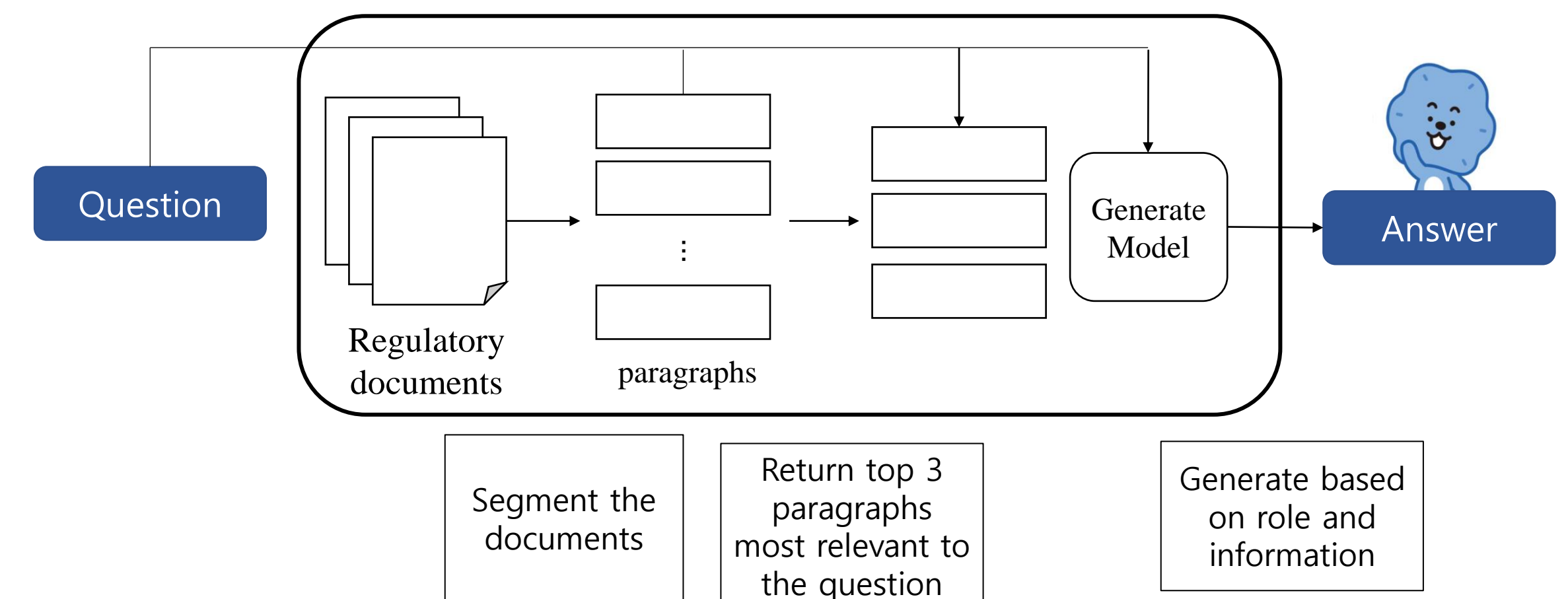


Fig. 3. Diagram of Question-Answering system for Korea Atomic Energy Research Institute.

IV. Results

1 답변 검토 : 아래의 표 1은 검토한 질문과 답변의 그 일부임

Table I: Example question-answer

	Question	Answer
#1	법인카드 사용 한도 알려줘.	법인카드 사용 한도에 대한 정보는 이 문서에서 찾을 수 없습니다.* [중간 생략] 법인카드 사용 한도와 관련된 정보를 찾고자 하는 경우 해당 기관의 재무 관리 규정을 참조하는 것이 좋습니다. *관련 정보가 문서에 없는 경우
#2	법인카드 사용 한도 알려줘.	법인카드 사용 한도는 500만 원으로 설정되어 있으며, 증액이 필요한 경우 사용 담당자가 회계 부서의 검토를 거쳐 요청할 수 있습니다. [이후 생략]
#3	휴직 기간 알려줘	휴직기간은 제46조에 따라 다양하게 규정되어 있습니다. 휴직기간은 근무기간에 따라 다르며, 근속 1년 미만은 3개월 이내, 3년 미만은 6개월 이내, 5년 미만은 12개월 이내, 10년 이상은 24개월 이내로 정하고 있습니다. [이후 생략]

2 평가 예정 : 본 QA 시스템의 정확성과 사용성 향상 목적

- 한국어 이해 평가(KLUE) 벤치마크 사용
- 시스템을 사용하는 연구자들의 직접적인 피드백 확보

V. Conclusion

- 질문과 관련 있는 문서를 기반으로 답변을 생성하는 한국원자력연구원의 QA 시스템 소개
 - ✓ 시멘틱 검색을 통해 질문과 관련된 문서의 단락 추출
 - ✓ 대화형 질의응답 인공지능을 통해 정확하고 구체적인 답변 제공
- 항상 정확한 답변을 제공하는 것이 아니기 때문에 계속해서 개선 필요

Extra

- 한국원자력연구원 이외의 사람들도 자유롭게 활용할 수 있는 **대화형 챗봇 개발**
- 파동이와 원자력에 관해 대화 가능

한국원자력연구원의
규정 문서 QA 시스템을
이용해보세요

*본 시스템은 학회 기간에만
공개됩니다.

LINE APP을 설치하고
파동이봇을 친구로 추가해서
원자력에 대한 대화를 나눠보세요.

파동이를 위해
피드백을 주세요.

ACKNOWLEDGMENTS

This research was supported by a grant from Korea Atomic Energy Research Institute (KAERI) R&D Program (No. KAERI-524450-23)