

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

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

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

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

*Corresponding author: ygyu@kaeri.re.kr

1. Introduction

In recent years, with the development of natural language processing (NLP) technology, large-scale language generation models such as BERT [1] and GPT-3 have emerged [2]. These models have been able to comprehend and generate human-like responses based on the context provided, thereby revolutionizing the way we interact with machines. As a result of these advancements, automatic response systems such as chatbots and Question-Answer (QA) systems are now being applied in various fields. In particular, there is a growing interest in developing QA systems tailored for specialized domains, as these technologies have the potential to significantly enhance work efficiency within organizations. However, the QA system applying the current language generation model has a limitation in that it lacks expertise in specialized domains. Obtaining accurate answers can be challenging, and in some models can generate incorrect responses instead of admitting a lack of knowledge. As a way to solve this problem, new specialized QA services such as Researchgpt [3] and chatpdf have been introduced. These services generate contextually relevant answers when provided with a single document containing domain-specific information.

In this paper, we present a QA system that generates answers based on hundreds of regulatory documents from the Korea Atomic Energy Research Institute (KAERI), in order to provide accurate and fast answers in specialized fields involving information such as regulations of a company. Currently, when research staffs directly conduct searches, the process can be considerably time-consuming, as it entails locating pertinent documents, reviewing their contents, and extracting relevant information. The QA system generates answers based on hundreds of regulatory documents, enabling research staffs to obtain accurate and quick answers to specialized questions. This contribution not only saves time and effort but also reduces the potential for human error. Furthermore, our QA system accepts multiple documents as inputs, so we can overcome the limitations of existing document-based QA services that accept only a single document.

In this study, we utilize an embedding model and language generation API (Application Programming Interface) provided by *OpenAI* for the QA systems. The embedding model is used to identify relevant paragraphs from a document and the user's question,

and then a language generation model generates accurate answers based on this information. The model is designed to understand regulatory documents and generate accurate answers.

As a follow-up study if such a service is created by documenting frequently asked questions by research staffs, it is expected to contribute to sharing knowledge and improving work efficiency within the researcher. Furthermore, when these QA technologies are applied globally, companies will be able to improve their work processes and strengthen their competitiveness more actively.

2. Preprocessing Regulatory documents

In this chapter, we described the process of preprocessing the text and selecting the paragraphs that are most relevant to the question. We extracted text from 225 regulatory documents in PDF file format and performed preprocessing. During the preprocessing, we used regular expressions to remove special characters and unimportant information such as tables of contents, appendices, and deleted regulations. We also used the *PyKoSpacing* package to automatically correct spacing errors in Korean text data. Next, we divided the text into paragraphs of about 500 characters. This preprocessing is the first step in effectively extracting the important information contained within the regulatory documents to generate accurate answers.

Text embedding is performed for each preprocessed document paragraph. We used the *text-embedding-ada-002* model [4] provided by *OpenAI*. The model significantly outperforms previous embedding models in text search, code search, text classification, and sentence similarity.

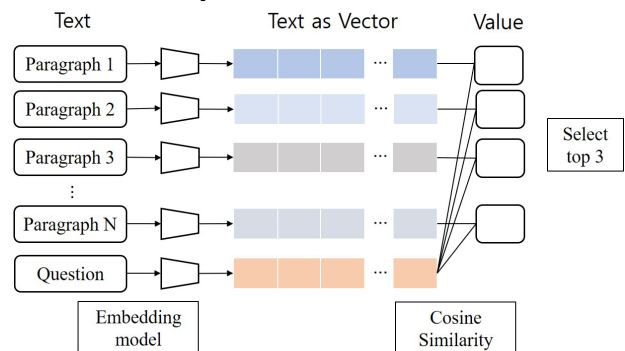


Fig. 1. Structure of the process of selecting the top three paragraphs using embedding model.

Compared to previous embedding models, the context length increased to 8912 and the embedding size decreased to 1536 dimensions.

Fig. 1. shows the process of selecting the three most relevant paragraphs to the user's question among all the paragraphs in the regulatory documents. First, the *text-embedding-ada-002* embedding model is utilized to convert all paragraphs in the documents and the text of the input question into vectors. Next, computes the cosine similarity between the vectors to extract the three paragraphs in the documents that are most relevant to the question, and generates an accurate answer based on them.

3. Generating answers based on documents

This chapter describes the process of generating answers using the *gpt-3.5-turbo* model [5], a chat-based large-scale language model provided by *OpenAI*. GPT-3.5 models can understand and generate natural language, and the *gpt-3.5-turbo* model is pre-trained on a large dataset, making it the most capable and cost-effective GPT-3.5 model optimized for chat.

The *gpt-3.5-turbo* model generates accurate answers based on some paragraphs extracted from the document and questions from the user. The maximum number of tokens in the model is 4,096, which means the maximum input text length that the model can process. If a longer sentence is input, an error occurs. Therefore, in this study, we do not input all the text in the document to the model as shown in Fig. 2., but only provide the information of the three extracted document paragraphs to the model. When using the model, we provide a prompt that specifies the information and guidance that the model needs to perform a given task by assigning roles and informing it of the document paragraphs that are relevant to the question. The role and guidelines granted are as follows: "*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:*"

If the prompt for this model is crafted appropriately, the model should be capable of providing sufficient information, as it comprehends the user's intentions and possesses the knowledge to respond accordingly. This answer generation process is also based on the assumption that both the input question and the extracted paragraphs contain important information within the document. Therefore, if the preprocessing and extraction process is not handled correctly, the accuracy of the generated answers may be affected.

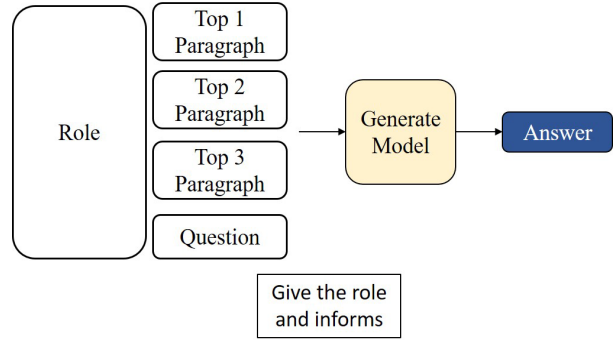


Fig. 2. Structure of the process of generating answers based on the role and information provided

4. Results

To validate the answers generated by the model, we created a list of questions and reviewed the answers to each question. Questions and answers in actual verification are conducted in Korean, and Table 1 shows the translation of them into English. First, we checked how the model answers questions that are not in the regulatory documents. We asked the model about the usage of corporate cards without informing it about the limit of using the corporate card, and as shown in Table 1 #1, the model answered that the information was not found in the document. In the second case, we asked the model about the use of the corporate card without informing it of the previous content, and the model generated an answer based on the content of the document, as shown in Table 1 #2. Third, we verified that the model generated the correct and specific answer to the question. As shown in Table 1 #3, the model was able to utilize the clauses and numerical expressions in the document to generate accurate and specific answers based on the information in the document. Finally, we checked whether the model generates answers based on previous conversations. As shown in Table 1 #4, when we didn't specify what the question was about, the model remembered that the previous conversation was about "vacation periods" and inferred that the current conversation was about that as well. However, the model is not always able to provide the correct answer, both the discovery of the problem and the solution are important for model performance.

Our future work will continue to focus on identifying and fixing the situations where problems arise, and we should take a closer look at the following processes. (1) preprocessing the text to make it easier for the model to understand. (2) extracting the three paragraphs within documents most relevant to the question. (3) assigning roles to the model to match the intent. Furthermore, the performance of the proposed QA system will be evaluated through two stages. The first stage involves evaluating the QA performance using the Korean Language Understanding Evaluation (KLUE) benchmark, which assesses the system's ability to handle Korean language tasks of language models. The second stage will involve evaluating the system's

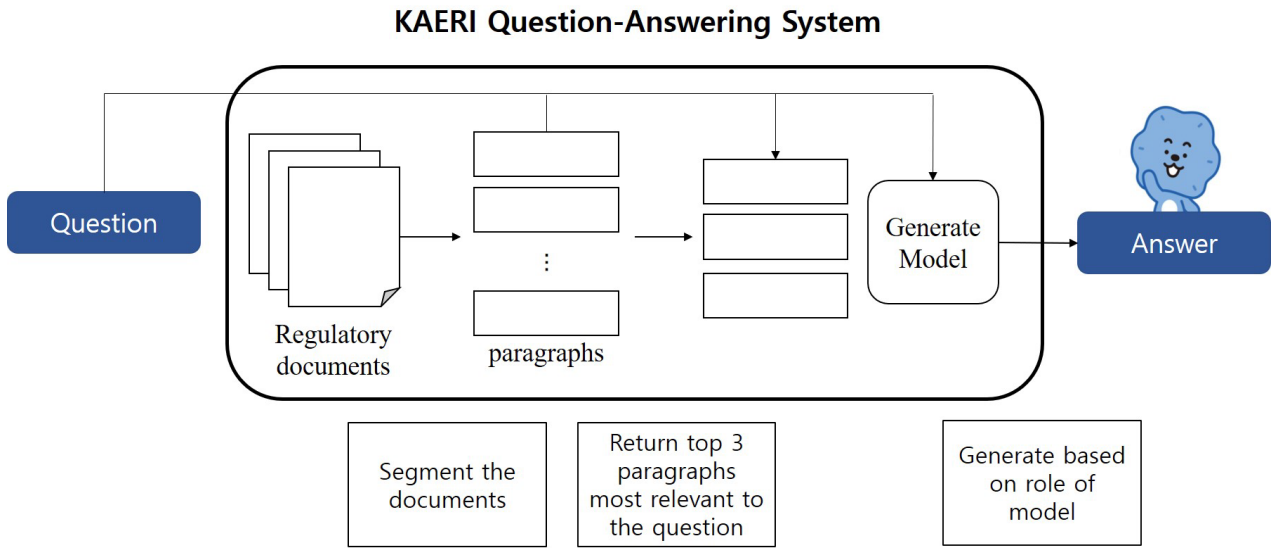


Fig. 3. Diagram of Question-Answering system for Korea Atomic Energy Research Institute.

effectiveness through direct feedback from researchers who use the system. These evaluations will help to enhance the accuracy and usability of the proposed system.

	3 months if I've been here for three months?	absence can be within three months. Therefore, if you've been here for three months, you can apply for a leave of absence within three months. [omitted for brevity]
--	--	--

Table I: Problem Description

	Question	Answer
#1	Tell me the limit of using the corporate card.	Information on corporate card usage limits is not found in this document. [omitted for brevity] If you want to find information related to the corporate card usage limit, it would be better to refer to the institution's financial management regulations.
#2	Tell me the limit of using the corporate card.	The limit of using the corporate card is set at 5 million won, and if an increase is required, the person in charge of use can request it after reviewing by the accounting department. [omitted for brevity]
#3	Tell me about the period of leave of absence.	The period of leave of absence is stipulated in various ways in accordance with Article 46. The period of leave of absence varies depending on the period of service, and it is set to be within 3 months for those with less than 1 year of service, within 6 months for those with less than 3 years of service, within 12 months for those with less than 5 years of service, and within 24 months for those with more than 10 years of service. [omitted for brevity]
#4	Then can I do it within	If the period of service is less than one year, the period of leave of

5. Conclusions

In this paper, we introduce a QA system that employs utilizes the *OpenAI* API model to generate precise responses. The system draws upon KAERI's regulatory documents as its knowledge base. For this purpose, we performed text preprocessing, embedding, similarity calculation, and answer generation, as shown in Fig. 3.. Analyzing the answers generated by the model, we found that the model was able to generate accurate and specific answers by utilizing paragraphs from documents related to the question, and that the model remembered and connected previous conversations. However, it does not always give accurate answers, so there is a need to continue to improve the system performance. Moreover, the proposed QA system has potential applications beyond regulatory documents because it can generate responses to a large number of documents containing intricate and specialized content. Consequently, research staffs are expected to obtain answers to their questions rapidly and accurately, thereby enhancing overall work efficiency.

ACKNOWLEDGMENTS

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

REFERENCES

- [1] D. Jacob Devlin, C. Ming-Wei, L. Kenton, T. Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv, abs/1810.04805, 2019.
- [2] K. Enkelejda, S. Kathrin, K. Stefan, B. Maria, D. Daryna, F. Frank, G. Urs, G. Georg, G. Stephan, H. Eyke, K. Stepha, K. Gitta, M. Tilman, N. Claudia, P. Jürgen, P. Oleksandra, S. Michael, S. Albrecht, S. Tina, S. Matthias, W. Jochen, K. Jochen, K. Gjergji, ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences, Vol.103, 2023.
- [3] Patnaik, Mukul. "Research GPT." GitHub, 2019, <https://github.com/mukulpatnaik/researchgpt>.
- [4] "New and Improved Embedding Models." OpenAI Blog, 14 May 2021, <https://openai.com/blog/new-and-improved-embedding-model/>.
- [5] OpenAI. "GPT-3.5 (Turbo) - API Documentation." OpenAI Platform, 2021, <https://platform.openai.com/docs/models/gpt-3-5>.