mathematics

MDPI

Article

Dense-to-Question and Sparse-to-Answer: Hybrid Retriever System for Industrial Frequently Asked Questions

Jaehyung Seo [1], Taemin Lee [2], Hyeonseok Moon [1], Chanjun Park [1], Sugyeong Eo [1], Imatitikua D. Aiyanyo [2], Kinam Park [2], Aram So [2], Sungmin Ahn [3] and Jeongbae Park [2,*]

[1] Department of Computer Science and Engineering, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea; seojae777@korea.ac.kr (J.S.); glee889@korea.ac.kr (H.M.); bcj1210@korea.ac.kr (C.P.); djtnrud@korea.ac.kr (S.E.)
[2] Human Inspired Artificial Intelligence Research (HIAI), Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea; taeminlee@korea.ac.kr (T.L.); titi@korea.ac.kr (I.D.A.); spknn@korea.ac.kr (K.P.); aram@korea.ac.kr (A.S.)
[3] O2O Inc., 47, Maeheon-ro 8-gil, Seocho-gu, Seoul 06770, Korea; acando72@gmail.com
* Correspondence: insmile@korea.ac.kr

Abstract: The term "Frequently asked questions" (FAQ) refers to a query that is asked repeatedly and produces a manually constructed response. It is one of the most important factors influencing customer repurchase and brand loyalty; thus, most industry domains invest heavily in it. This has led to deep-learning-based retrieval models being studied. However, training a model and creating a database specializing in each industry domain comes at a high cost, especially when using a chatbot-based conversation system, as a large amount of resources must be continuously input for the FAQ system's maintenance. It is also difficult for small- and medium-sized companies and national institutions to build individualized training data and databases and obtain satisfactory results. As a result, based on the deep learning information retrieval module, we propose a method of returning responses to customer inquiries using only data that can be easily obtained from companies. We hybridize dense embedding and sparse embedding in this work to make it more robust in professional terms, and we propose new functions to adjust the weight ratio and scale the results returned by the two modules.

Keywords: deep learning; artificial intelligence; natural language processing; frequently asked questions; dense and sparse embedding; industrial system; information retrieval

MSC: 68T50

Citation: Seo, J.; Lee, T.; Moon, H.; Park, C.; Eo, S.; Aiyanyo, I.D.; Park, K.; So, A.; Ahn, S.; Park, J. Dense-to-Question and Sparse-to-Answer: Hybrid Retriever System for Industrial Frequently Asked Questions. *Mathematics* **2022**, *10*, 1335. https://doi.org/10.3390/math10081335

Academic Editor: Victor Mitrana

Received: 21 March 2022
Accepted: 15 April 2022
Published: 18 April 2022

1. Introduction

The traditional frequently asked questions (FAQ) system is used in the industrial field and provides accurate answers to pre-determined user queries and keywords [1–4]. These techniques rely significantly on the database structure that has been set in advance and demonstrate poor response performance without a clear understanding of the concept of free speech.

After the advent of the transformer [5], the FAQ system faced a significant paradigm shift based on pre-trained language models (PLMs). In solving various natural language processing downstream tasks, PLMs outperform existing machine learning models based on context and meaningful information. They also lead to significant progress in semantic research, which returns results based on the context and relationship inference of input queries [6–8]. This advancement significantly alters the FAQ paradigm, allowing the FAQ system to handle spontaneous speech in a conversational format [2,9,10].

Recently, a few domestic conglomerates developed a chatbot-based FAQ model that required massive resources by utilizing the ability to model human-like sentence expressions

and the ability to understand context. As a result, many domestic companies have begun investing significant resources to replace customer service, including FAQ, with PLM-based dialogue systems such as chatbots.

However, most small and medium enterprises and national institutions find it difficult to build training data at high costs or to maintain chatbot-based FAQ systems on a continuous basis without investing significant resources. Furthermore, due to its low generalization, it is difficult to involve the learned model outside of the same industrial domain despite their significant investment.

Moreover, to ensure the higher performance of deep-learning-based retriever models and chatbot-based FAQ systems, additional training datasets and architecture modifications are required to acquire knowledge of specific domains [10]. This is because most languages used in industrial domains require proficient knowledge or appear at extremely low probabilities in the general corpus. In addition, a domain gap occurs in the commonsense knowledge of the language and the knowledge of the language required in a specific industrial domain. Furthermore, for a language model to understand a given context within an industrial domain and return the required results, the ability to retrieve or comprehend prior knowledge is more important than inferring commonsense knowledge.

To address these problems, we propose a specialized FAQ system for industrial domains that combines the advantages of traditional approaches and current study. The proposed FAQ system returns an optimal response to a given query based on a PLM and information retrieval. Our system has three advantages, which are as follows. (i) It outperforms spontaneous speech in terms of semantic similarity by combining dense and sparse embedding for a given query. (ii) To reduce model production costs, raw textual strings, such as previous conversion records and response manuals, are used as databases for small- and medium-sized enterprises and national institutions, and training strategies are organized with published benchmark datasets. (iii) Two additional score functions are proposed to optimize the retrieved results and disentangle the gap of embedding spaces while maintaining high performance.

This paper is organized as follows. Section 2 explains related FAQ retrieval systems. Section 3 shows the proposed methods, the dense embedding retriever, the sparse embossing retriever, and the score function. Section 4 describes specific experimental settings and comparative experimental results. Section 5 discusses the limitations of the paper and its further considerations. Finally, Section 6 concludes the proposed method and plans for future research.

2. Related Work

FAQFinder [1] and Auto-FAQ [11] are early FAQ studies based on representative information retrieval. FAQFinder drives appropriate responses from the frequency of words using a heuristic statistical-based natural language processing technique. Auto-FAQ proposes a natural language processing system based on statistics and rules that makes use of features from the shallow language understanding and question answering (QA) domains.

Rule- and statistics-based FAQ retrieval research introduced reducing the lexical gap [3], template-based word matching systems [4,12], and automated template-based FAQ systems [13] using regular expressions. Subsequently, semantic-similarity-based FAQ systems and methods of evaluating models in dialogue are suggested in [2,14].

In the advent of deep learning, convolutional neural networks (CNN) [7,15], long short-term memory (LSTM) [16], and PLMs are used to measure similarities between input queries and target databases [6,8]. Sakata et al. [6] calculate the similarity between the user's query and the question of the target FAQ set using BM25 and compute the similarity between the user's query and the answer with BERT. BERT trains with FAQ datasets labeled with binary classification. In addition, the retrieval ranking is rearranged based on the BM25 score, and the values of the two modules are heuristically combined into one scalar. Mass et al. [8] also match the user's query to the FAQ dataset's question

and answer. However, Mass et al. [8] use BM25 to re-rank the retrieved candidates and search for questions and answers with high scores in two BERT models. In addition, the GPT-2 model generates a paraphrase for the answer, allowing BERT to learn similarities for weak unlabeled FAQ datasets. Furthermore, as with the dialogue system and the massive corpus-based QA system, the FAQ system, as one of the subtasks, uses the same approach [9,10,17].

In this paper, we concentrate on integrating two distinct methods for application to a real-world industrial problem. We use information retrieval systems with PLMs in a manner similar to that of [6,8], but we train a model using a published benchmark dataset. However, our proposed system differs in that it trains a model using a non-FAQ published benchmark dataset. Furthermore, in contrast to previous studies, the BERT model was used to compute similarities between queries and questions, enhancing the robustness of the language model for free speech. We concentrate on the role of a contextual language model rather than using the BERT model as a classifier. Our approach uses spontaneous speech as an input query, similar to a conversation-based FAQ system, and significantly reduces the cost of data construction. Furthermore, we propose two score functions to disentangle the gap problem between dense and sparse embedding, which has not been considered in previous research.

3. Proposed Method

We combine the dense embedding retriever and the sparse embedding retriever to construct an FAQ system that retrieves the optimal response from a database built from previous conversation records. In addition, we propose two scoring functions that adjust the scalar values returned by two retriever modules with different embedding spaces.

3.1. FAQ Database

The FAQ system aims to resolve customer complaints and difficulties by providing appropriate answers to frequently asked questions. This serves the same purpose as responding to customers through a counselor, but there is a difference in lowering the cost of human resources and promptly resolving customer complaints before responding to a counselor.

We use the conversation history and summaries with customers as our database, inspired by the fact that the FAQ system serves the same purpose as dealing with customers through counselors. Because most domestic companies that employ counselors are required by law to record conversations between counselors and customers, re-using pre-existing data significantly reduces the cost of developing one's database. In contrast, the customer consultation record data contain unnecessary greeting phrases, incorrect information delivery, specific personal information, and inaccurate answers that do not fit the context.

As a result, we manually refine the counselors' conversation contents based on the FAQ manual that has been built in advance by five or more human annotators working in the same industry domain. In addition, we create a rule-based manual to eliminate unethical issues and specificity. Finally, we collect the information in the form of a single turn, which includes a consumer query and a counselor response.

3.2. Question Matching: Dense Embedding Retriever

Dense-embedding-based information retrieval compares the semantic similarity between a query and a search target database in a latent space expressed as a continuous real value. Since this method exploits high-density expression in a small dimension, it can provide a fast retrieval speed and yield a result that is not excessively biased to a specific term. In other words, the meaning of the given input is interpreted according to the dynamically comprehended contextualized representation and the linguistic knowledge learned by the model during the pre-training process.

For constructing a dense embedding retriever, we adopt the KoBERT (https://github.com/SKTBrain/KoBERT, accessed on 20 March 2022) model that is pre-trained with the Korean corpus. The embedding size of the dense vector is set to 768, which is the same as the embedding size of the model. In addition, we use a sentence transformer structure that connects two KoBERT autoencoders as a pair of bi-encoders [18]. The value returned by each encoder represents structural and contextual information for a given query and the target customer's question.

We use the open KorSTS dataset for training the information retrieval model [19]. KorSTS is a natural language understanding downstream task consisting of a D training dataset for calculating semantic similarity between two sentences. Unlike previous studies, we use open datasets to reduce the burden of building datasets to train information retrieval models and ensure versatility within industrial domains.

Specifically, KoBERT takes two input sequences composed of the query sh and the target question sp. We add a pooling layer P to the output of the KoBERT for converting variable-length vectors sh and sp into fixed-length embedding vectors with size n. σ denotes activation unit and W, b are trainable parameters of a pooling layer P. Subsequently, as in Equation (1), the values of all output vectors of the pooling function $P(\cdot)$ are averaged to obtain a fixed-size query embedding vector m_q and a search target embedding vector m_c. y in Equation (2) corresponds to the existing correct label representing the relationship between two sentences as a real number. We additionally define y' as a min-max normalized vector of y for comparison with results $\hat{y}$ in Equation (3). As in Equation (4), the difference between two values is optimized using mean-squared-error $\mathcal{MSE}$ loss, which is a loss function that meets the regression objective.

$$m_{sh}, m_{sp} = P(\sigma(W(sh) + b)), P(\sigma(W(sp) + b)) \tag{1}$$

$$y' = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \tag{2}$$

$$\hat{y} = \cos(m_{sh}, m_{sp}) = \frac{\sum_{i=1}^{n}(m_{sh_i} m_{sp_i})}{\sqrt{\sum_{i=1}^{n}(m_{sh_i})^2}\sqrt{\sum_{i=1}^{n}(m_{sp_i})^2}} \tag{3}$$

$$\mathcal{L}_{\text{dense}} = \mathcal{MSE}(y', \hat{y}) = \frac{1}{D}\sum_{j=1}^{D}(y'_j - \hat{y}_j)^2 \tag{4}$$

The trained dense embedding retriever calculates the semantic similarity as a comparison group of the contents corresponding to the customer questions in the FAQ database built in Section 3.1. Considering that the KorSTS dataset used for fine-tuning consists of a pair of short dialogues, we set the history of the customer question as a target sentence (input-query-to-customer question). The semantic similarity value converges to 1 point if the input query and the customer question in the database have a similar context and 0 point if the relevance is low.

The retrieved questions are arranged in descending order based on the similarity score, including the result of the counselor's response. The PLM-based information retrieval method has the advantage of retrieving a list that can respond to spontaneous speech as well as strictly pre-designed queries. Furthermore, the bi-encoder provides superior efficiency by calculating performance in linear time even when the retrieval target is asymmetric or the database is too large.

3.3. Answer Matching: Sparse Embedding Retriever

Even a retrieved question with a high degree of similarity does not always yield an optimal response. For example, if a user query intends to block adult content, similar customer questions, such as removing adult content or blocking entertainment content, can have high scores despite returning incorrect responses.

To address these issues, we do not rely solely on the semantic similarity calculated by averaging the sentence-embedding values using the sparse embedding retriever, but we also include the inclusion of a specific keyword in the score. The ranking function BM25 is used as the sparse embedding retriever [20]. We use mecab-ko (https://github.com/hephaex/mecab-ko, accessed on 20 March 2022) to segment the given input query and the counselor response corresponding to the database into morpheme units [21]. We then extract retrieval target tokens having practical meanings, such as nouns and verbs, among these segmented tokens. The remaining morphemes that play a grammatical role are removed.

We suppose the content morphemes are $\{q_1, \ldots, q_n\}$ in a user query Q and a counselor response R, where n is the number of tokens. The average response length of all indexed responses is $avgRL$ and length of the given response is $|R|$. k and b are free hyper-parameters which are initialized by the default values 1.2 and 0.75, respectively. In particular, k is adopted for refining the effect that a single query token can have on the document score, and b is utilized for penalizing the long documents. We rank highly relevant responses by computing the term frequency $f(q_i, R)$ and the inverse document frequency $IDF(q_i)$. The ranking function BM25 is denoted as follows:

$$s_{\text{sparse}}(R, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, R) \cdot (k+1)}{f(q_i, R) + k \cdot (1 - b + b \cdot \frac{|R|}{avgRL})} \tag{5}$$

Similarly, the counselor response is organized in descending order, beginning with the customer's question and ending with the corresponding response. Because the counselor response is longer than the customer's question, we use BM25, which has a regularization method based on the length of the search document. Furthermore, we compensate for the dense embedding method's shortcomings, in which the token embedding value varies with context by retrieving keywords that must be included in the counselor response with a fixed token embedding value.

3.4. Score Function

Dense embedding retriever and sparse embedding retriever return output vectors from different embedding spaces. Accordingly, to calculate the score considering the two values, it is necessary to map these vectors to an integrated space. Thus, we adopt the arctangent function for re-ranging the output of each architecture. In particular, for adjusting the mapped value to $-1 < y < 1$, we multiply $\frac{2}{\pi}$ and the function output. Specifically, we utilize the scaling function g as shown in Equation (6).

$$g(x) = \frac{2 \times \tan^{-1}(x)}{\pi} = \frac{i \times \log(1 - i \times x)}{\pi} - \frac{i \times \log(1 + i \times x)}{\pi} \tag{6}$$

The scaling function sets a significant increment for the scores returned by the retriever, making the distinction between each score clear within the range $-1 < y < 1$. Furthermore, when the response contains the same keyword but has a completely different meaning, we can impose a significant penalty by allowing the dense retriever's output to be a negative value. This scaling method is critical in addressing the shortcomings of a method that returns a similar response and is traditionally appropriate for approaches such as simple n-gram overlapping.

Furthermore, as it approaches 1.0, the increase is significantly reduced, preventing one retriever from receiving an excessively high score and ranking regardless of the score of the other retriever.

However, summing the scaled scores in equal proportions does not guarantee optimal results. In addition, each retriever must calculate a score based on other retrieval targets and adjust the reflected ratio in the final score. We apply the $q_{\text{blending}}(\cdot)$ function that calculates the weight for each score to adjust the reflection degree of the scores returned by the dense

embedding retriever and the sparse embedding retriever. The weight adjustment function formula is as follows.

$$q_{\text{blending}}(s_{\text{dense}}, s_{\text{sparse}}) = (\lambda \times g(s_{\text{dense}})) + ((1-\lambda) \times g(s_{\text{sparse}})) \quad (7)$$

The λ value determines the weight ratio. λ is set to 0.75 by applying grid search using Optuna [22]. Grid search sets the optimal ratio by initializing the ratio to 0.5 and increasing or decreasing it by 0.05. We empirically learn that a higher performance can be obtained by imposing a high weight on the result s_{dense} of the dense embedding retriever and minimizing the weights of the results derived by the sparse embedding retriever s_{sparse}.

4. Experimental Results

This section presents an open training dataset and database setting. We demonstrate the effectiveness of each retriever and score function through comparative experiments.

4.1. Data Details

We leverage released KorSTS to train the dense embedding retriever and present the number of train, validation, and test datasets in Table 1. The FAQ database shown in Table 2 is provided by O2O (http://www.o2o.kr/, accessed on 20 March 2022), a company to develop customized conversational AI solutions, and consists of transcription and processing results of customer–counselor inquiries collected through various consulting organizations. We filter out inadequate pairs containing more than 15% of special characters, overly short answers, unnecessary tags that hinder the natural language context, and no mapping of response from the counselor. Subsequently, 9687 refined pairs of customer questions and counselor responses out of 70,000 raw descriptions remain.

Table 1. Statistics for KorSTS. We use the open Korean sentence textual similarity dataset KorSTS to train a highly generalized model in industry domain.

KorSTS	Train	Validation	Test
# Examples	5749	1500	1379
Average # words	7.5	8.7	7.6

Table 2. Statistics for **FAQ database**. We construct the **FAQ database** to retrieve dialogue history of customers and counselors based on sentence textual similarity.

FAQ Database	Customer Question	Counselor Answer
# Examples	9687	9687
Average # words	7.89	30.95
Average # sentences	1.21	3.58

4.2. Training Details

We conduct all training and testing based on pytorch-lightning (https://github.com/PyTorchLightning/pytorch-lightning, accessed on 20 March 2022). We use KoBERT, a language model trained on the BERT [23] model in Korean, as a language model in a dense embedding retriever. We experiment with consultation records and training datasets organized in Korean, so a PLM specialized in Korean encoding is appropriate. The model is a transformer architecture with 12 encoder layers, with a hidden size of 768 and a vocab size of 8002.

We train the KoBERT model by dividing the encoding strategy into a cross-encoder and a bi-encoder for the KorSTS dataset [24]. The cross-encoder is an approach to fine-tuning a single encoder according to the training objective of the downstream task. This method jointly encodes two input sequences and computes a single concatenated vector using a special token such as $[CLS]$. On the other hand, a bi-encoder is a method of fine-tuning the two encoders. This approach converts the embedding vector values returned by the

two encoders into a fixed-size vector by averaging them and then calculates the cosine similarity. We also adopt the BM25 algorithm for the sparse embedding retriever, as no additional training step is required.

Considering the characteristics of Korean, we segment sentences into morpheme units using mecab-ko for the sparse embedding retriever. The input query and the retrieval target database are guided to return results based on segmented frequencies. For dense embedding retriever training, we exploit hydra's Optuna to perform grid research on batch size $\{32, 64, 96, 128\}$, sequence maximum length $\{76, 96, 128\}$, learning rate $\{10^{-5}, 10^{-4}, 10^{-3}\}$, drop rate $\{0.1, 0.2, 0.3, 0.4\}$, and warm-up ratio $\{0, 0.1, 0.2, 0.3\}$.

4.3. Implementation Details

The user query sequence is fed into the dense and sparse embedding retrievers. Firstly, the dense embedding retriever stores the histories of the previous customer queries in the database as fixed-size embedding vectors. Subsequently, the cosine similarity with the user query sequence converted into the embedding vector through the dense embedding retriever is calculated, as shown in Equation (3). Secondly, the counselor responses in the database are expressed discretely through morpheme segmentation, and a score based on the keyword frequency is calculated through a BM25, as indicated in Equation (5). The two scores obtained through the two different retrievers are scaled and expressed as a single scalar, as described in Equations (6) and (7). The detailed flowchart is described in the Appendix A Figure A1.

4.4. Evaluation Details

As the evaluation metrics of the retriever model, we use the Pearson and Spearman correlations. The Pearson correlation is an indicator of the linear relationship strength between two variables by calculating the normalized cosine similarity. Based on the two variables moving to each other, we induce variables to converge to 1.0 in strong positive correlations and 0 in negative correlations.

The Spearman correlation is a measure for assessing monotonicity by ranking values between two variables. We return the correlation coefficient in the same range as the Pearson correlation because two variables increase or decrease with a constant magnitude.

In addition, to verify the retrieval performance in the FAQ database, we exploit a hit at k and the mean reciprocal rank (MRR). In accordance with the scaled score of two retrieval module values, the hit at k returns "True" if there is a correct answer among k candidates ranked in descending order. The MRR is an information retrieval metric considering priority and is obtained by taking the correct answer position as a reciprocal in descending order. This method considers the relative ranking and can be applied as a key performance criterion when only the top rank results are shown to the user, such as in the FAQ system.

4.5. Main Results

We compare the inference speed and performance of the dense embedding retriever, which is part of the proposed retrieval system, on the KorSTS test dataset. The performance of the information retrieval in the FAQ database is also measured in order to compare the performance of the industrial FAQ retriever for arbitrarily given queries. To evaluate the performance, we parse the FAQ database's customer questions and reconstruct 200 new question–response pairs. To demonstrate the significance of the scaling effect of the two score functions, we apply them to the retriever module with the best performance.

In general, the cross-encoder method performs better in most natural language understanding downstream tasks than the bi-encoder method but shows relatively slow inference speed. As shown in Table 3, we attempt to maintain a performance close to the cross-encoder's performance while gaining the benefit of the speed of the bi-encoder. To maintain the model's performance, we proceed with normalization between 0 and 1 point, considering the cosine similarity calculation of the scores labeled between 0 and 5, as shown

in Equation (2). In addition, we operate the optimal hyper-parameter search framework Optuna (https://github.com/optuna/optuna, accessed on 20 March 2022) to make the bi-encoder model have a performance similar to that of the cross-encoder model, even if it is more than 100 times faster in inference. Moreover, unlike other encoding methods, our bi-encoder approach has a significant advantage from commercialization employing only CPU resources.

Table 3. Experimental results of dense embedding retrievers on the KorSTS test dataset.

Model	Encoding Method	Runtime	(Pearson + Spearman)/2
KoBERT	Cross-encoder	3533 ms	77.76
KoBERT	**Bi-encoder**	**24 ms**	**77.77**

As shown in Table 4, we perform a comparative analysis using various settings for the dense and sparse embedding retrievers. Overall, the context-based information retrieval with the dense embedding retriever outperforms the simple keyword mapping retrieval method with the sparse embedding retriever. Figure 1 clearly depicts the performance gap between the single information retriever model and the hybrid information retriever model. As demonstrated, increasing the number of retrieved answers (from one to five) consistently improves retrieval accuracy. The single information retrieval model shows a rate of 0.65 or less in Hit @ 1. However, combining dense and sparse embedding retrievers shows a rate of at least 0.8 or more when deducing the optimal answer. The information retrieval system using solely one model shows a relatively low score in Hit @ 1. However, in Hit @ 5, the performance gap between solely one model and combining the two information retrieval models gradually decreases. The proposed model has a robust performance in retrieval of the correct answer, natural language understanding narrowing the gap between Hit @ 1 and Hit @ 5 to 0.07. These findings show that the single model extracts comparable responses as candidate groups for input queries, but it still struggles to determine which is the most reasonable response to the input query.

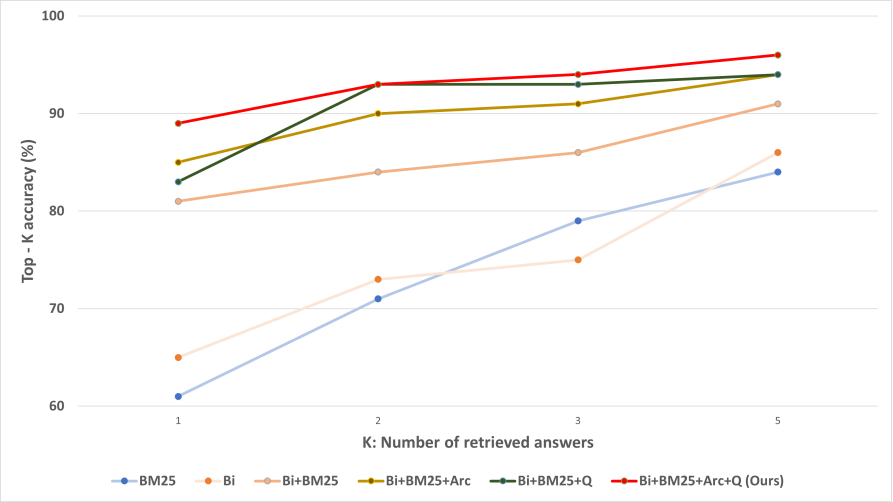

Figure 1. Top-K accuracy with different models used in sparse embedding retriever, dense embedding retriever, and score function. The results are measured on the paraphrased new question–response pairs. Our proposed hybrid retriever FAQ system with score function outperforms other models. Bi, Arc, and Q indicate bi-encoder, arctangent, and Q-blending, respectively.

In addition, we run the experiments proposed in Section 3.4's score functions. The highest performance is obtained by scaling with arctangent and setting a weight λ in Q-blending for the influence rate on the results of each model. Arctangent scaling and Q-blending have been shown to improve performance. Closer examination of the experimental results reveals that arctangent scaling has an advantage in selecting the most optimal answer and

Q-blending has an advantage in determining high-rank candidates. Finally, it can be seen that the retriever model performs best when hybridizing dense and sparse embedding and two scaling and weighting functions are combined.

Table 4. Performance of dense and sparse embedding retrievers for FAQ database.

Dense Embedding Retriever	Sparse Embedding Retriever	Score Function	Hit @ 1	Hit @ 2	Hit @ 5	MRR
-	BM25	-	0.61	0.71	0.84	0.72
Bi-encoder	-	-	0.65	0.73	0.86	0.75
Bi-encoder	BM25	-	0.81	0.84	0.91	0.85
Bi-encoder	BM25	Arctangent	0.85	0.90	0.94	0.91
Bi-encoder	BM25	Q-blending	0.83	0.93	0.94	0.92
Bi-encoder	**BM25**	**Arctangent + Q-blending**	**0.89**	**0.93**	**0.96**	**0.93**

4.6. Qualitative Results

Table 5 depicts the results of a qualitative analysis of the FAQ retrieval system. The experiment is carried out with the paraphrased question mapped to the counselor's response as an input query, resulting in sentence compositions that differ but have the same contextual meaning. Based on the context, the FAQ system proposed in this paper retrieves the most optimal response. All three cases have adj. scores close to 1.0 and return the correct answers as intended by the original questions. The first is an input query that corresponds to available content regardless of industry domain and returns an appropriate response based on the stored database's domain. This is the result of learning from the KorSTS data and has a significant advantage for queries that are similar to natural conversations. The second case is an input query that is limited to a single industrial domain and produces a response with a similar contextual meaning even when a different sentence structure is used. Finally, the third case is an input query that has the most transformation in the sentence and uses a completely different vocabulary. Despite the high degree of transformation in the sentence, the proposed model compensates for the shortcomings of each retriever module and returns a stable answer with a high score.

Table 5. Actual example sentences of the FAQ database and paraphrased query. Paraphrased Sent. refers to the reconstructed input query sentence, and Adj. score means the sum of the scores of the two retriever models with arctangent and Q-blending function.

Original Sent.	Paraphrased Sent.	Hit @ 1 Result	Adj. Score
I want to know about today's weather.	Tell me about the weather.	Click the [MIC] button on the remote control and say it as follows: "Tell us about today's weather". We are informing you of the weather information within 10 days including today.	0.875
Please tell me how to apply to number change service.	I would like to change my number, so let me know how to apply.	If the area you signed up for and the area you are currently using are the same, you can apply for a number change service. Once this service is completed, the previous agency will be automatically terminated, and if it is terminated within one year of subscription, a number transfer fee may be charged.	0.812
What gifts or benefits can I get when I sign up?	Do I get a gift if I sign up for a new one?	If you sign up, you can get various benefits depending on the plan and service. For more information, please contact the website or customer service center.	0.792

5. Discussion

The proposed FAQ retriever is an effective system that combines the benefits of existing information retrieval systems and can be implemented with limited resources. We discover that if we design the system with the attributes of the textual input and the passages to be retrieved in mind, we can maximize performance in information retrieval. Previous studies concentrated on calculating the similarity between answer candidates and input queries using language-model-based dense embedding retrievers. However, we believe that contextualized representation is more effective in understanding the intent of user queries, and we present the dense-to-question and sparse-to-answer systems. We also attempt to train with an open dataset that can be used universally rather than the FAQ system's narrow-domain dataset.

Since a published training dataset is used, it is a domain-agnostic method if only existing FAQ history remains. However, there are limitations to conducting experiments solely in the IT and mobile domains in this paper. Although KorSTS training data demonstrate some domain-agnostic characteristics, further testing is required. Furthermore, we present a computational approach for bridging the gap between the dense embedding retriever and the sparse embedding retriever. This results in the robustness of our proposed system even at top-1 accuracy. On the other hand, λ in Q-blending is a hyper-parameter that needs to be searched anew according to the domain to be changed. Therefore, in future studies, λ should be automatically set as a trainable parameter.

6. Conclusions and Future Works

This paper addresses the shortcomings of traditional rule- and statistics-based FAQ system approaches while lowering data construction costs for the most recent deep-learning-based FAQ system. To demonstrate robust performance on specialized domain terms as well as free-conversation-based utterances commonly used in the industrial domain, we propose a hybrid model that combines the strengths of dense and sparse embedding retrievers. In addition, the scaling and weight ratios are adjusted to optimize the score value returned by the hybrid retrieval system via the arctangent and Q-blending score functions. We demonstrate the validity of the proposed model through comparative experiments and qualitative analysis. We intend to diversify the types of databases that can be used within the industry domain in the future as well as conduct improvement research. In addition, we will apply methods such as cumulative probability sampling to the candidate group returned by the model as a retrieval result in order to find more appropriate responses from the user's perspective.

Author Contributions: Methodology, J.S.; project administration, J.P. and T.L.; software, T.L. and J.S.; conceptualization, J.S.; formal analysis, H.M.; validation, A.S. and I.D.A.; writing—review and editing, J.S., H.M., C.P. and S.E.; supervision, J.P. and K.P.; funding acquisition, S.A.; database annotation, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information& Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-00988, Development of AI contact center using MRC-based automatic question-answering system). This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Many thanks to KU NMT Group for taking the time to proofread this article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Technical details of proposed FAQ retriever system. GPU resource is used only for training the KoBERT model, not required in the inference process.

Operating System (OS)	Ubuntu 18.04 (Linux x64)
Programming language	Python 3.7
Rest API	flask 1.1.4, Swagger UI
Deep learning framework	Pytorch-lightning 1.4.2, hydra 1.1.0, optuna 2.4.0
Dense embedding retriever	KoBERT (Bi-encoder)
Sparse embedding retriever	BM25
CPU	18-core Intel Xeon Gold 6230 CPU
GPU (optional, only train)	Nvidia A6000
CUDA version (optional)	11.1

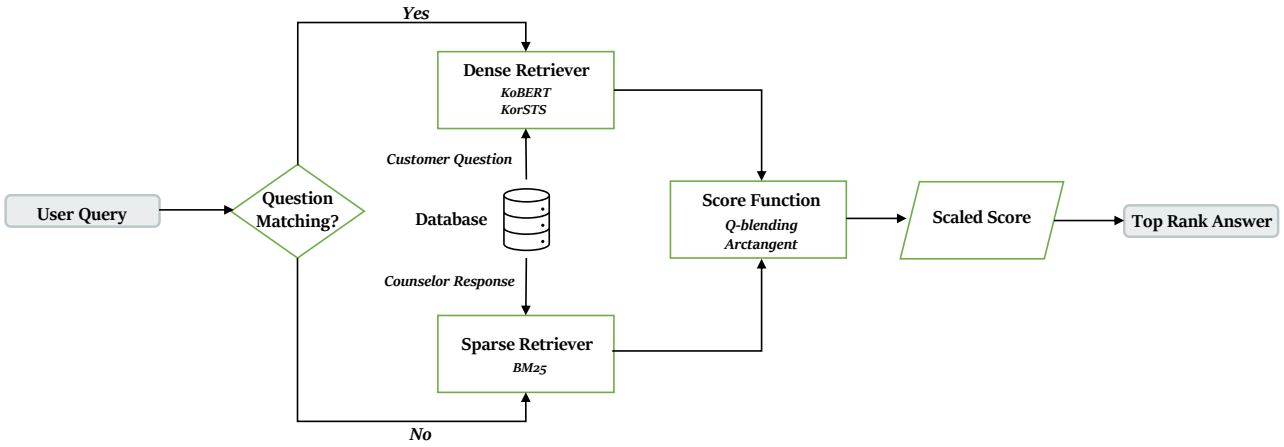

Figure A1. Flowchart of proposed FAQ retriever system. The proposed system uses a different model of whether to retrieve the counselor response or the customer question in the database.

References

1. Hammond, K.; Burke, R.; Martin, C.; Lytinen, S. FAQ finder: A case-based approach to knowledge navigation. In Proceedings of the 11th Conference on Artificial Intelligence for Applications, Los Angeles, CA, USA, 20–23 February 1995; pp. 80–86.
2. Karan, M.; Žmak, L.; Šnajder, J. Frequently asked questions retrieval for Croatian based on semantic textual similarity. In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 8–9 August 2013; pp. 24–33.
3. Kim, H.; Seo, J. High-performance FAQ retrieval using an automatic clustering method of query logs. *Inf. Process. Manag.* **2006**, *42*, 650–661. [CrossRef]
4. Sneiders, E. Automated FAQ answering with question-specific knowledge representation for web self-service. In Proceedings of the 2009 2nd Conference on Human System Interactions, Catania, Italy, 21–23 May 2009; pp. 298–305.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
6. Sakata, W.; Shibata, T.; Tanaka, R.; Kurohashi, S. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 1113–1116.
7. Karan, M.; Šnajder, J. Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. *Expert Syst. Appl.* **2018**, *91*, 418–433. [CrossRef]
8. Mass, Y.; Carmeli, B.; Roitman, H.; Konopnicki, D. Unsupervised FAQ Retrieval with Question Generation and BERT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 807–812. [CrossRef]
9. Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.t.; Choi, Y.; Liang, P.; Zettlemoyer, L. QuAC: Question answering in context. *arXiv* **2018**, arXiv:1808.07036.

10. Oliveira, H.G.; Ferreira, J.; Santos, J.; Fialho, P.; Rodrigues, R.; Coheur, L.; Alves, A. AIA-BDE: A Corpus of FAQs in Portuguese and their Variations. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 5442–5449.
11. Whitehead, S.D. Auto-FAQ: An experiment in cyberspace leveraging. *Comput. Netw. ISDN Syst.* **1995**, *28*, 137–146. [CrossRef]
12. Sneiders, E. Automated question answering using question templates that cover the conceptual model of the database. International Conference on Application of Natural Language to Information Systems, Saarbrücken, Germany, 23–25 June 2002; pp. 235–239.
13. Moreo, A.; Eisman, E.M.; Castro, J.L.; Zurita, J.M. Learning regular expressions to template-based FAQ retrieval systems. *Knowl. Based Syst.* **2013**, *53*, 108–128. [CrossRef]
14. Caputo, A.; de Gemmis, M.; Lops, P.; Lovecchio, F.; Manzari, V.; Spa, A.P.A. Overview of the EVALITA 2016 question answering for frequently asked questions (QA4FAQ) task. In Proceedings of the Final Workshop, Naples, Italy, 7 December 2016; p. 124.
15. Karan, M.; Šnajder, J. FAQIR–a frequently asked questions retrieval test collection. In Proceedings of the International Conference on Text, Speech, and Dialogue, Brno, Czech Republic, 12–16 September 2016; pp. 74–81.
16. Gupta, S.; Carvalho, V.R. FAQ retrieval using attentive matching. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 929–932.
17. Reddy, S.; Chen, D.; Manning, C.D. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 249–266. [CrossRef]
18. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084.
19. Ham, J.; Choe, Y.J.; Park, K.; Choi, I.; Soh, H. Kornli and korsts: New benchmark datasets for korean natural language understanding. *arXiv* **2020**, arXiv:2004.03289.
20. Svore, K.M.; Burges, C.J. A machine learning approach for improved BM25 retrieval. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 1811–1814.
21. KUDO, T. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Available online: http://mecab.sourceforge.net/ (accessed on 20 March 2022).
22. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
24. Humeau, S.; Shuster, K.; Lachaux, M.A.; Weston, J. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv* **2019**, arXiv:1905.01969.