Original Article

Using Semantic Text Similarity calculation for question matching in a rheumatoid arthritis question-answering system

Meiting Li[1]^, Xifeng Shen[1]^, Yuanyuan Sun[1]^, Weining Zhang[1]^, Jiale Nan[1]^, Jia'an Zhu[2], Dongping Gao[1]^

[1]Institute of Medical Information, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China; [2]Department of Ultrasound, Peking University People's Hospital, Beijing, China

Contributions: (I) Conception and design: D Gao; (II) Administrative support: D Gao, J Zhu; (III) Provision of study materials or patients: J Zhu; (IV) Collection and assembly of data: M Li, X Shen, Y Sun, W Zhang, J Nan; (V) Data analysis and interpretation: M Li, X Shen, Y Sun, W Zhang, J Nan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Dongping Gao. Institute of Medical Information, Chinese Academy of Medical Sciences, Peking Union Medical College, No. 3 Yabao Road, Chaoyang District, Beijing, China. Email: gaodp_2022@126.com.

Background: When users inquire about knowledge in a certain field using the internet, the intelligent question-answering system based on frequently asked questions (FAQs) provides numerous concise and accurate answers that have been manually verified. However, there are few specific question-answering systems for chronic diseases, such as rheumatoid arthritis, and the related technology to construct a question-answering system about chronic diseases is not sufficiently mature.

Methods: Our research embedded the classification information of the question into the sentence vector based on the bidirectional encoder representations from transformers (BERT) language model. First of all, we calculated the similarity using edit distance to recall the candidate set of similar questions. Then, we took advantage of the BERT pretraining model to map the sentence information to the corresponding embedding representation. Finally, each dimensional feature of the sentence was obtained by passing a sentence vector through the multihead attention layer and the fully connected feedforward layer. The features that were stitched and fused were used for the semantic similarity calculation.

Results: Our improved model achieved a Top-1 precision of 0.551, Top-3 precision of 0.767, and Top-5 precision of 0.813 on 176 testing question sentences. In the analysis of the actual application effect of the model, we found that our model performed well in understanding the actual intention of users.

Conclusions: Our deep learning model takes into account the background and classifications of questions and combines the efficiency of deep learning technology and the comprehensibility of semantics. The model enables the deep meaning of the user's question to be better understood by the intelligent question answering system, and answers that are more relevant to the original query are provided.

Keywords: Intelligent question-answering system; Semantic Text Similarity (STS); rheumatoid arthritis; classification; deep learning

Submitted Jul 18, 2022. Accepted for publication Jan 26, 2023. Published online Mar 15, 2023.
doi: 10.21037/qims-22-749
View this article at: https://dx.doi.org/10.21037/qims-22-749

^ ORCID: Meiting Li, 0000-0003-1555-7728; Xifeng Shen, 0000-0003-3446-6741; Yuanyuan Sun, 0000-0001-7894-865X; Weining Zhang, 0000-0003-2326-9400; Jiale Nan, 0000-0002-0577-4510; Dongping Gao, 0000-0002-8699-8195.

Introduction

An increasing number of people are using the internet to look up health information. For example, patients with rheumatoid arthritis disease may be interested in the duration of joint symptoms and effective long-acting medicines. Traditional search engines often return too many relevant web pages or data items, which makes it difficult for the required information to be located quickly and accurately (1). An intelligent question-answering system based on frequently asked questions (FAQs) can enable users to ask questions in their natural language. It returns short and accurate results, enhancing users' information acquisition experience.

According to epidemiological reports, as of 2019, the number of patients with rheumatoid arthritis in China reached more than 6 million, with about one-tenth of these being severe cases (2). From the perspective of the user experience, an automatic question-answering system has long been an area of intense research interest in the field of information retrieval. However, the current question-answering system for specific fields has yet to be perfected, especially in the medical specialty field.

Semantic Text Similarity (STS) refers to the degree of a match between two texts. STS has a wide range of applications in many fields of natural language processing tasks, such as text summarization generation, question-answering systems, plagiarism detection, duplicate checking, and similar content recommendations. The present study aimed to evaluate STS calculation for questions of the FAQ-based question-answering system to make improvements that apply to the field of rheumatoid arthritis disease. In the Chinese language, many synonyms, acronyms, negative words, and complex and diverse sentence structures increase the difficulty of calculation. Due to the short length of the question, the absence of context, and the diversity and complexity of users' questions, a Chinese language question-answering system has certain requirements for its accuracy and reasoning ability. Presently, the related research in the field of Chinese question answering is not mature.

Traditional Chinese Semantic Textual Similarity algorithms use a number of methods. The keyword-based method extracts and analyzes the word frequency and part-of-speech information in a sentence (3). The method based on semantic dependence analyzes the deep syntactic structure by analyzing the dependence relationship between the core words of the sentence and other components (4). The method based on semantic analysis calculates the similarity of the words in the sentence using a constructed semantic knowledge base to calculate the similarity value of the whole sentence (5). However, the difficulty of this method lies in the improvement and expansion of some words. The method based on edit distance counts the number of editing operations required to transform two sentences into the same sentence, including addition, deletion, and modification (6).

Many previous studies have researched how to obtain the similarity of two texts. Albitar *et al.* (7) used the term frequency-inverse document frequency (TF-IDF) method to measure the similarity between texts and used the results as the prediction standard for text classification. Ormerod *et al.* (8) applied linear decoding and Representational Similarity Analysis (RSA) to measure the semantic similarity signal in intermediate token representations of 5 popular transformer models, with predictions producing a correlation of 0.87 with ground-truth similarity scores compared with the state-of-the-art correlation of 0.9. These included the bidirectional encoder representations from transformers (BERT)-Large (9), 3 variants of BERT that were fine-tuned on text from the clinical domain, and XLNet-Large (10). The 3 BERT variants were BioBERT (11), ClinicalBERT (1-13), and discharge summary BERT (DS BERT) (12-13). Chinese scholars have developed many methods based on semantic considerations. Li *et al.* (14) presented an approach to compute the similarity and relevance between words based on two kinds of language resources: CiLin (15) and HowNet (16). Yu *et al.* (17) proposed a method to calculate the similarity of question sentences based on the keyword vector space method and semantic conception vector space method. Cui *et al.* (18) further considered the distance and order of words, proposed a new method, and applied it to the network question-answering system. Xiong *et al.* (19) put forward a calculation method for the community question-answering system based on a latent Dirichlet allocation (LDA) model, which separately calculated from the statistical information, semantic information, and topic information of the question sentence, comprehensively obtaining the overall similarity.

Many experts have carried out research on STS based using deep learning. Mahmoud (20) used a convolutional neural network to detect the semantic relevance between suspicious text and source text documents using the open-source Arabic corpus (21). The results showed that the accuracy and recall rates were better compared with those achieved with the TF-IDF method based on statistics

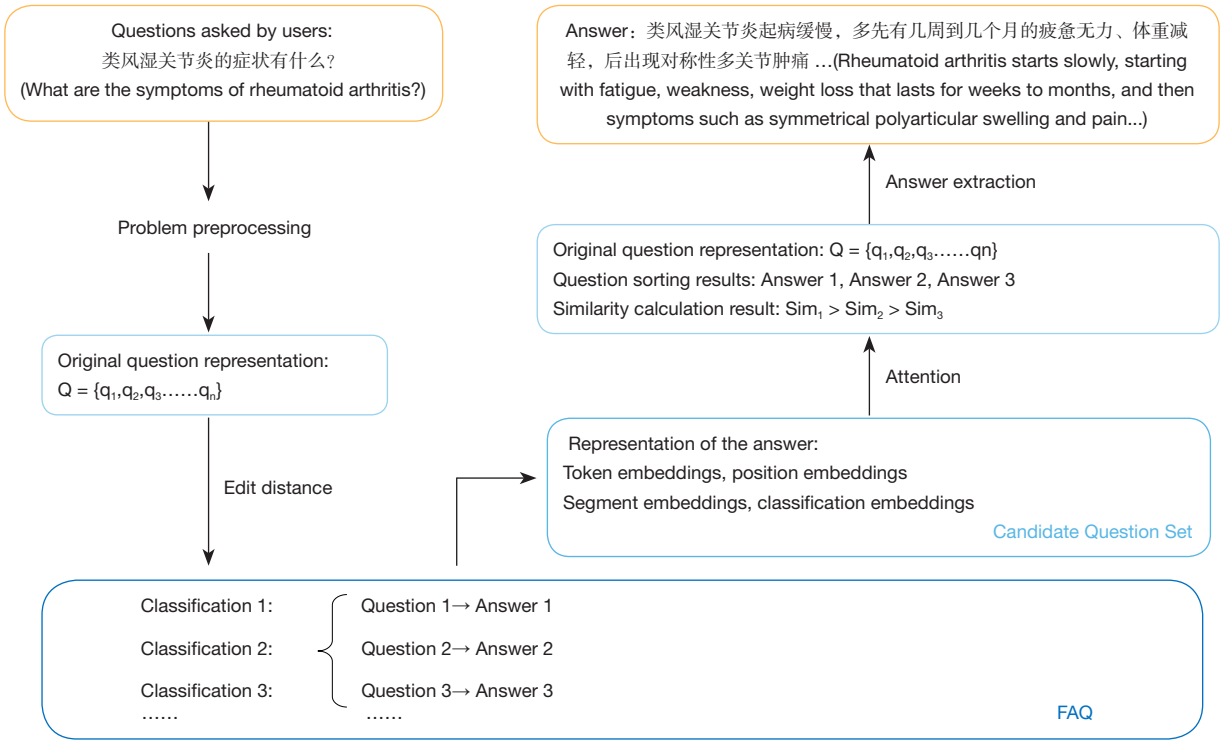

Figure 1 Flowchart depicting the workflow of the answer extraction of this question-answering system. FAQ, frequently asked question.

and a method based on word correlations. Wu *et al.* (22) integrated the BERT and boosted tree models and proved the effectiveness of the ensemble through a correlation analysis. With the ensemble of the two types of models, the F1 score reached 0.90825, which was 3.88% higher than that of the optimal single model. Li *et al.* (23) proposed a combined method based on knowledge-based conceptualization and a transformer encoder that could more effectively capture the semantics of short texts, which was found to be useful in experiments and could be applied to a variety of applications, such as short-text information retrieval and short-text classification.

Most of the current similarity calculation techniques based on deep learning seek to improve performance by changing the model structure, which can often improve the accuracy of the similarity calculation results with a small amount of data. However, achieving a high level of accuracy is not the only challenge. In practical applications, only by thoroughly understanding the user's question and their questioning intention can we better answer the user's query and provide the answer that the user actually needs. Rather than using literal matching, our method can circumvent the drawbacks of literal understanding and matching while understanding deep user needs. The question answering system built in our research not only gives users answers to specific questions, but also enables intelligent understanding.

Our results indicated a strong correlation between the effect of STS calculation of Chinese sentences and the analysis of text structure and the concept of words. This finding suggests that the characteristics of the calculation method and its applicability in a specific field should be fully considered when calculating STS.

Methods

The application of our method in the working process of the question-answering system is shown in *Figure 1*. First, we used the edit distance method to calculate similarity to recall a set of 20 question sentences in the FAQ. Then, we calculated the similarity between the candidate question set and the question and sorted them from largest to smallest. Finally, we selected the answer corresponding to the largest similarity question as the result.

The edit distance is a commonly used distance measurement method that has been widely used in the

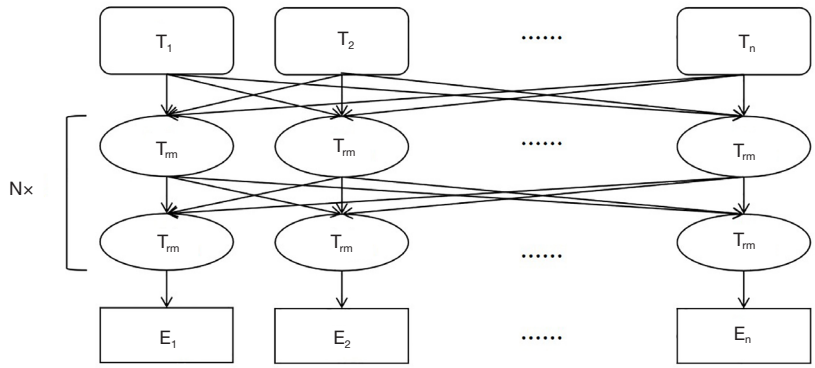

Figure 2 The structure of bidirectional encoder representation from the transformer. E, the vector transformed from each question; T, the implicit output state; Trm, transformer encoding module.

field of text similarity calculation. We used this method to determine the approximate scope of candidate questions in order to improve search efficiency. This method refers to the minimum number of operations required to convert one string to another. The operations include the following: add, which adds a character; delete, which deletes a character; and modify, which modifies a character. The edit distance directly and literally reflects the degree of difference between two texts; that is, the more similar the two texts are, the smaller the edit distance. The main idea of the edit distance technique is to construct a matrix d[m+1, n+1], where m and n are the lengths of the texts d_1 and d_2, respectively, and then calculate ED(d_1,d_2) = d[m+1, n+1], where ED(d_1,d_2) is the edit distance between d_1 and d_2.

The formula of the edit distance algorithm is as follows:

$$\text{Similarity}\left(d_1,d_2\right)=1-\frac{ED\left(d_1,d_2\right)}{\max\left(m,n\right)} \quad [1]$$

$$ED\left(d_{1_i},d_{2_j}\right)=\begin{cases}\max\begin{cases}\left(d_1,d_2\right) & d_1=0\,\|\,d_2=0\\ d\left[i-1,j-1\right] & d_1=d_2\end{cases}\\ \min\begin{cases}d\left[i-1,j\right]+1 & \text{delete } d_{1_i}\\ d\left[i,j-1\right]+1 & \text{add } d_{2_j}\\ d\left[i-1,j-1\right]+1 & \text{modify } d_{2_j}\end{cases}\end{cases} \quad [2]$$

First, the matrix is initialized with values 0…n in the first row and 0...m in the first column, and then each character from 1 to m in d_1 is compared with each character from 1 to n in d_2. The comparison formula is shown in Eq. [2]. When d_1 or d_2 is 0, the edit distance is the length of another nonzero text; when d_{1_i} and d_{2_j} are equal, the edit distance is the edit distance d[i-1,j-1] of the previous character of the current character; when d_{1_i} and d_{2_j} are not equal, the edit distance takes the minimum value of three cases.

The edit distance approach only calculates the similarity from the literal text and does not combine the semantic features for further judgment. Question recall based on edit distance can only obtain a set of candidate questions, and the similarity value needs to be calculated in combination with a transformer for deep learning.

BERT

Sentence vectorization is an important step in the process of sentence feature extraction. With the development of word vector technology, sentence information can be mapped into vectors, as is done with static word vectors word2vec (24) and glove (25). For BERT (9), the transformer is the core module, and it can solve the long-distance dependency problem caused by the traditional recurrent neural network (RNN) model and can be calculated in parallel to improve computational efficiency. The structure of BERT is shown in *Figure 2*.

The training process of BERT is divided into the pretraining and fine-tuning stages. In the pretraining stage, we used the BERT pretraining model provided by the Harbin Institute of Technology (26). Questions related to rheumatoid arthritis were used in the fine-tuning stage. Questions in a question-answering system are typical short texts. Due to the lack of context information, short text without rich background information is more likely to produce ambiguity and be difficult to understand. Therefore, we added a layer of sentence classification information into the BERT embedding layer, which was

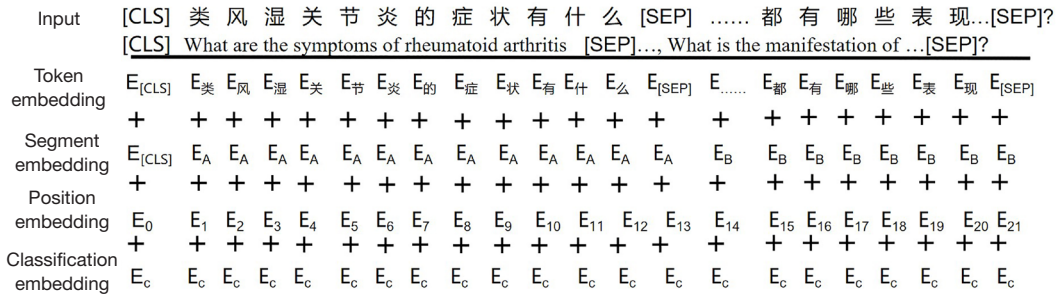

Figure 3 Four input vector representations of BERT. CLS, a classifier token; SEP, a sentence separator; BERT, bidirectional encoder representation from transformers.

developed in advance according to the basic information of questions under the supervision of professional doctors. Finally, the BERT embedding layer consisted of token embedding, segment embedding, position embedding, and classification embedding (*Figure 3*). Token embedding recorded the information of each word; sentence embedding recorded the overall information of the sentence; position embedding used the sine and cosine functions to encode the position of the word, and the position information of the word was encoded into a feature matrix; and classification embedding integrated the semantic information of sentences into the generation of vectors. The model converts each word x_i in the input sentence sequence $X=(x_1,x_2,...,x_t,...,x_n)$ into the embedding vector e_i and output, which obtains the embedding representation $E = (e_1,e_2,...,e_t,...,e_n)$ of the sentence X.

Specifically, the combination of embedding layers is as follows.

Token embedding

In this layer, each character is converted into a 768-dimensional vector. Before the conversion, two special tokens (CLS) (classification) and (SEP) (separation) are added to the beginning and end of the sentence text, which are regarded as signs of dividing the sentence.

Segment embedding

In the similarity calculation task, the input is a concatenated sentence pair. In this layer, there are only two vector representations. For the position of the first sentence, the value is assigned to 0; the position of the second sentence is assigned to 1.

Position embedding

The position of the characters is critical for the meaning sentences written in Chinese.

The transformer has a different structure from RNN and cannot use the order information of words. Therefore, the position information of the words is encoded in the manner shown in Eq. [3] and Eq. [4]:

$$PE_{(POS,2i)} = \sin\left(\frac{pos}{\alpha^{\frac{2i}{d_{model}}}}\right) \quad [3]$$

$$PE_{(POS,2i+1)} = \cos\left(\frac{pos}{\alpha^{\frac{2i}{d_{model}}}}\right) \quad [4]$$

pos represents the position of the character in the sentence, d_{model} is the dimension of the vector, and i is the position of the vector. The position relation of subsequent characters relative to the current character can be obtained through this position encoding, which is beneficial to expressing the semantic relation.

Classification embedding

We input the sentence pair into the original BERT model to obtain the CLS vector of the sentence. The classification information of sentences obtained with one-hot encoding is passed through an embedding layer and added to (CLS) to obtain the final representation vector.

After the above process, the four different vector representations are averaged to synthesize a vector representation with a batch size of 768, which is the input

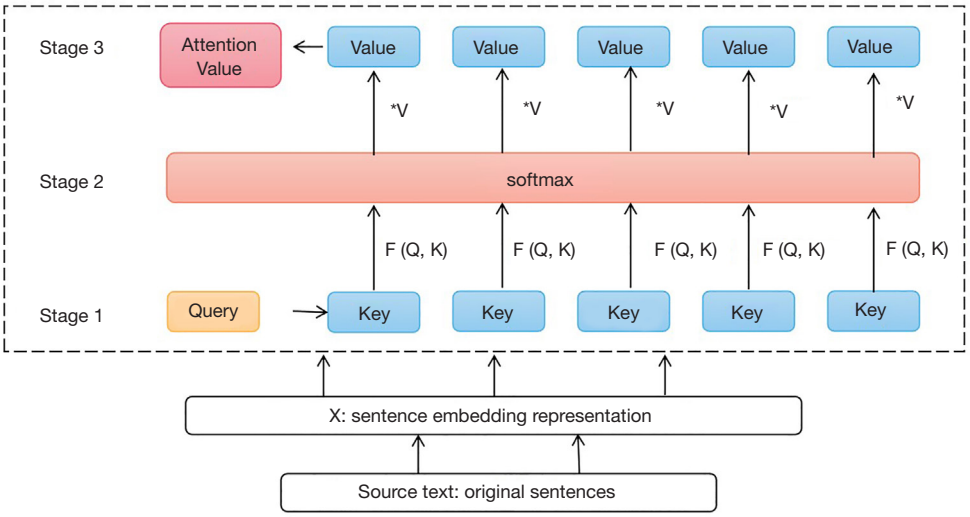

Figure 4 The calculation process is divided into the 3 stages. *, the matrix used in the calculation.

of the transformer encoder part.

The BERT pretraining model is a 2-stage model. The first stage is a language pretraining model, which employs a masked language model and next sentence prediction to capture the textual representation at the word and sentence level.

Specifically, 15% of the token locations are randomly selected for masking, and 80% of these 15% locations use (MASK) tokens, 10% use any other random token, and the last 10% keep the original token unchanged. The training process of the next sentence prediction task is that for example sentences A and B, there is a 50% probability that B is similar to A, and the other 50% probability is irrelevant. In this study, in the second stage of BERT, we used the prediction result of the next sentence prediction task as the calculation result of the similarity of the two questions and continuously adjusted a large number of parameters in the BERT network through the backpropagation mechanism to minimize the loss function. The loss function used in the training phase of this article is shown in Eq. [5], where F represents the true label, s_j represents the probability, and T represents the total.

$$L = -\sum_{j=1}^{T_n} y_i \log_{s_j} \quad [5]$$

Self-attention

The encoder of the transformer is composed of 6 identical stacked basic layers, each of which contain two sublayers: the multihead attention layer that uses the three matrices of Q, K, and V for the calculation of the attention mechanism and the fully connected feedforward layer that introduces nonlinearity and transforms the output space of the attention layer to increase the model's characterization ability. After the addition of the multihead attention layer and fully connected feedforward layer, residual connection and normalization layers are added, respectively, to standardize the optimized space and accelerate the convergence process, preventing problems such as the disappearance of gradients.

The function of the multihead attention layer is to enable the encoder to fully integrate the context information when extracting each position feature, thereby capturing the internal structure information of the input text. First, this part decomposes the information contained in the input source text into a series of key-value pairs (key, value). Next, the degree of correlation between each query and each key is calculated to obtain the value corresponding to each key in each source text. Finally, the value is weighted and summed to obtain the final attention value. Essentially, attention is the weighted summation of the value of the elements in the source, and the query and key are used to calculate the weight coefficient of the corresponding value. The calculation process is shown in *Figure 4*. The first stage involves calculating the weight coefficient based on query and key, and the second stage involves normalizing the original score of the first stage using softmax, as shown in Eq. [6] (27).

Table 1 Experiment environment configuration

Experiment environment	Configuration
Operating system	Ubuntu16.04
CPU	Intel(R) Xeon(R) CPU E5-2603 v4 at 1.70 GHz
Random access memory	32 G
Development language	Python
Development framework	PyTorch
GPU	NVIDIA GeForce GTX 1080

CPU, central processing unit; GPU, graphic processing unit; NVIDIA GeForce GTX 1080, video card product of NVIDIA.

The calculation idea for the entire text is as follows: $Attention\left(query, source\right)=\sum_{i=1}^{lx} similarity\left(query, key_i\right)*value_i$.

$$a_i = Softmax\left(Sim_i\right)=\frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_j}} \quad [6]$$

where a_i is the attention weight vector, and L_x is the length. The third stage involves performing a weighted summation of the value according to the weight coefficients normalized in the second stage. The calculation process of the three stages is shown in *Figure 4* (25).

The calculation formula is shown in Eq. [7]. Three vectors, Q, K, and V, are calculated through model learning: the Q vector represents the current character, the K vector represents each character in the context, and the vector V represents the relationship of the target word and each context. In the calculation, the weight expressed by the similarity between the Q vector and each K vector is weighted and then fused with the V vector. The scaling factor $\sqrt{dim_k}$ is used to achieve softmax normalization, making the model more stable during training.

$$Attention\left(Q,K,V\right)=Softmax\left(\frac{QK^T}{\sqrt{\dim_k}}\right)V \quad [7]$$

The process of calculating the scaled dot-product of the query and key is also the process of focusing. The focus of important information is reflected in the weight coefficient, which represents the importance of the information and corresponds to value. Specifically, multihead attention combines and linearly transforms the enhanced semantic vectors of each word in the sentence in different semantic spaces to obtain a final enhanced semantic vector with the same length as the original word vector, which is the result of the multihead attention layer. The calculation formula is shown in Eq. [8] and Eq. [9] (28). Multihead attention correlates different attention results to obtain contextual semantic information from multiple dimensions, which reflects the relevance and importance of different characters to a certain extent and obtain richer semantic feature expression.

$$Multihead\left(Q,K,V\right)=Concat\left(head_1,\cdots\cdots,head_h\right)W^O \quad [8]$$

$$head_i = Attention\left(QW_i^Q, QW_i^K, QW_i^V\right) \quad [9]$$

W_i^Q, W_i^K, and $W_i^V \in R^{d_{model}\times d_k}$ represent the weight matrix Q, K, and V of the i-th head (i=1,2,……,h, h=12), h represents the number of attention, and d_{model} is the dimension of the element vector, $d_k = d_v = d_{model}/h$. Each attention captures the information of a subspace in the question, and the multihead attention value is obtained by splicing h attention heads through matrix $W^O \in R^{d_{model}\times d_{model}}$.

The second sublayer of the encoder's basic structure is the fully connected feedforward layer, which was composed of a rectified linear activation function (ReLU) activation function layer and the linear function layer. The overall formula is expressed as follows: $FFN(x) = \max(0, xW_1+b_1)W_2+b_2$. Here, W_1 and W_2 are two trainable weight matrices, b_1 and b_2 are the 2 offset vectors, and x is the input of the activation function.

Attention imitates the activity mechanism of the human brain. In the question sentence "What are the symptoms of rheumatoid arthritis in the active period?", "in the active period" and "symptoms" are highly correlated. To answer this question about "symptoms," the limiting condition of "in the active period" must be taken into consideration. The attention mechanism is used to distinguish the effects of different parts of the input on the output.

Experiment

We conducted our experiments on two models: a baseline model and an improved baseline model. The specific experimental environment settings are shown in *Table 1*. The experimental data were divided into the training, validation, and test sets according in a ratio of 8:1:1.

Data set

There is no public question-pair corpus in the field of rheumatoid arthritis disease. Therefore, we first built a question-answering corpus for rheumatoid arthritis. We manually classified the resulting interrogative sentences according to the meaning of the question sentences and the characteristics of the disease. We first searched for relevant questions and corresponding answers with "rheumatoid arthritis" on several well-known Chinese question and answering websites and then marked the similarity of the questions. We labeled questions with similar meanings or the same category as similar and the rest as dissimilar. We treated questions with the same answer requirements as similar. All of our work was conducted under the supervision of professional sonographers, following the actual clinical situation and the actual needs of patients, and the questions and classification were professional and authoritative. Since our research is still in progress, the code and data have not yet been released, we will consider sharing these after we complete our research. The final question classification is shown in *Table 2*.

For each question in the question-answering pairs, 5 dissimilar questions and 1 similar question with their corresponding category information were randomly selected. This was used for the training set data input to the model. Each sentence corresponded to 6 pieces of data, including 5 wrong answers and 1 correct answer. The validation set was constructed in the same way as the training set. For each piece of test data, the top 100 pieces of data were selected in descending order according to the edit distance of all questions that were paired with the question and answer pair. Then, their similarity to these 100 questions was calculated in the prediction stage.

Results

We evaluated the effect of the question and answer system from the following aspects: accuracy, comprehensiveness, the speed of the system to return answers, and the fluency of answer sentences, which should be concise and conform to the characteristics of natural language.

We used two ways to measure the performance of our question-answering system.

Table 2 Classification information of questions

Number	Primary classification	Secondary classification	Tertiary classification
1	Disease profile	Awareness of disease	Concept
			Genetics
			Classification
			Hazards and effects
			Cure
			Self-cure
			Influence factor
			Complication
		Epidemiology	Basic statistical indicators
			Prone population
2	Pathogeny and predisposing factors	Pathogeny	Disease causes
			Causes of accompanying other symptoms or illnesses
			Causes of symptoms
		Predisposing or aggravating factors	Predisposing or aggravating factors

Table 2 (*continued*)

Table 2 (*continued*)

Number	Primary classification	Secondary classification	Tertiary classification
3	Diagnosis and evaluation of disease	Symptoms	Symptoms of the disease itself
			Symptoms associated with other symptoms or diseases
		Diagnostic criteria	Diagnostic indicators
			Differential diagnosis
			Active diagnosis
		Means of inspection and inspection	Laboratory inspections
			Physical examinations
			Imageological examinations
			Other examinations
4	Treatment principles and protocols	Principle of treatment	Principle of treatment
		Therapeutic schedule	Drug therapy
			Operative treatment
			Other therapeutic substances
			Folk prescription
		Treatment of symptoms	Treatment of symptoms
		Treatment of accompanying other symptoms or diseases	Treatment of accompanying other symptoms or diseases
		Traditional Chinese medicine therapy	Traditional Chinese medicine therapy
		Naturopathy	Naturopathy
5	Matters needing attention	Disease control considerations	Disease control strategy
		Other matters needing attention	Dietary contraindications and dietary recommendations
			Exercise and movement patterns
			Matters needing attention merged with other diseases and symptoms
			Nursing methods
6	Prognosis	Prognostic effect	Prognostic effect
		Prognostic factors	Prognostic factors
7	Precaution	Disease prevention	Disease prevention
		Symptoms of prevention	Symptoms of prevention
8	Others	Economics	Expenditure
			Health insurance
		Medical establishment	Hospital
			Department

Table 3 Performance of the baseline and improved models

Model	MRR	Top-5 precision	Top-3 precision	Top-1 precision
Bert[a]_Baseline_model	0.562	0.676	0.614	0.500
Bert_Improved_model	0.660	0.813	0.767	0.551
RoBerta[b]-base-Baseline_model	0.306	0.489	0.307	0.188
RoBerta-base-Improved_model	0.517	0.653	0.580	0.420
RoBerta-large-Baseline_model	0.252	0.386	0.284	0.153
RoBerta-large-Improved_model	0.342	0.534	0.426	0.210
ALbert[c]-large-Baseline_model	0.719	0.852	0.801	0.614
ALbert-large-Improved_model	0.712	0.865	0.831	0.588
ALbert-xlarge-Baseline_model	0.571	0.761	0.665	0.432
ALbert-xlarge-Improved_model	0.622	0.807	0.756	0.483
MacBert[d]-base-Baseline-model	0.722	0.858	0.830	0.608
MacBert-base-Improved-model	0.739	0.863	0.807	0.636
MacBert-large-Baseline-model	0.618	0.863	0.795	0.438
MacBert-large-Improved-model	0.705	0.841	0.818	0.591

[a]Bert, bidirectional encoder representations from transformers; [b]RoBerta, robustly optimized bidirectional encoder representations from transformers approach; [c]ALbert, A LITE bidirectional encoder representations from transformers; [d]MacBert, MLM as correction bidirectional encoder representations from transformers; MRR, mean reciprocal rank.

First, precision indicated whether the result returned was what we expected. We checked the precision of the top 1, 3, and 5 answers. Mean reciprocal rank (MRR) indicated the average of reciprocal ranks of the desired items, which was computed as the sum reciprocal of all the correct answers, and rank (q) was the position of the first correct answer in the returned list. The results are shown in *Tables 3*,*4*. In addition to the basic BERT model, we also improved several classic BERT variant models, including robustly optimized bidirectional encoder representations from transformers approach (RoBerta); a LITE bidirectional encoder representations from transformers (ALbert), and MLM as correction bidirectional encoder representations from transformers (MacBert). It should be noted here that the ratio of positive and negative samples was fixed at 1:5, the learning rate of experiments was performed for 1e-7, and the batch size was 32.

$$MRR = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{rank(q)} \quad [10]$$

Second, we chose several representative sentences as examples to compare the calculation effects of the baseline and improved model. The sorted sentences in the top 3 and the corresponding similarity values returned by the baseline model and the improved model are shown in *Table 5*. The baseline model tended to match literally, such as asking “类风湿关节炎可以采用免疫治疗吗？” (“Can rheumatoid arthritis be treated with immunotherapy?”), for which the baseline model matches, “类风湿关节炎可以打疫苗吗？” (“Can rheumatoid arthritis be vaccinated against?”) and “类风湿关节炎可以做沙疗吗？” (“Can I do sand therapy?”). We considered that the users’ needs were the treatments of rheumatoid arthritis and knowledge about whether immunotherapy was one of them. Therefore, we matched the question “类风湿关节炎要怎么治疗最好”？ (“How best to treat rheumatoid arthritis?”). The answers returned to users were common treatments for rheumatoid arthritis, not other vaccine-related issues.

Discussion

Due to the limitations of technology and knowledge, our question-answering system based on FAQs had some limitations. First, our study did not consider the relevance of semantic understanding in complex multiround question

Table 4 Performance of the baseline and improved models based on question classification

Classification	Model	MRR	Top-5 precision	Top-3 precision	Top-1 precision
Laboratory inspections	Baseline	0.889	0.889	0.889	0.889
	Improved	1.0	1.0	1.0	1.0
Drug therapy	Baseline	0.440	0.563	0.500	0.375
	Improved	0.573	0.688	0.688	0.5
Dietary contraindications and dietary recommendations	Baseline	0.7	0.7	0.7	0.7
	Improved	0.733	0.9	0.9	0.6
Principle of treatment	Baseline	0.341	0.6	0.367	0.233
	Improved	0.831	0.933	0.9	0.767
Traditional Chinese medicine therapy	Baseline	0.6	0.6	0.6	0.6
	Improved	0.75	1.0	0.8	0.6
Exercise and movement patterns	Baseline	0.333	0.167	0.0	0.0
	Improved	0.5	0.5	0.5	0.5
Symptoms of the disease itself	Baseline	0.659	0.818	0.727	0.545
	Improved	0.727	0.909	0.909	0.545
Disease prevention	Baseline	0.452	0.714	0.714	0.286
	Improved	0.524	0.857	0.857	0.286
Classification	Baseline	0.733	1.0	1.0	0.6
	Improved	0.8	0.8	0.8	0.8
Cure	Baseline	0.567	0.75	0.625	0.5
	Improved	0.588	0.875	0.75	0.375
Nursing methods	Baseline	0.417	0.5	0.5	0.333
	Improved	0.806	1.0	1.0	0.667
Disease control strategy	Baseline	0.875	1.0	1.0	0.75
	Improved	1.0	1.0	1.0	1.0

MRR, mean reciprocal rank.

answering. Second, as user problems become more complex and diverse and people's knowledge of a certain fields becomes more complete, the use and maintenance of the question-answering system will become more difficult. Zhou *et al.* (29) proposed for the first time that the upper and lower questions of the question are also particularly important in the limited field. When the question is incomplete, it is necessary to combine the previous question for a joint semantic understanding, especially in the field of chronic medical diseases. For example, some user questions will add time and context restrictions to describe their symptoms. Questions about how to distinguish these subtle differences and return answers in a targeted manner are the key issues that our intelligent question answering system will aim to solve in its next stage. In the future, an important research direction to consider is the content of sentences and adding more and broader disease information to the similarity calculation method. Mining deeper semantic information of questions can provide more relevant analytical dimensions for the similarity calculation model.

Conclusions

We have researched and improved the text similarity

Table 5 Similarity calculation results of example sentences

Examples	Sorting of the returned results of the baseline model	Similarity scores for the baseline model	Sorting of the returned results of the improved model	Similarity scores for the improved model
膝盖类风湿关节炎能打玻璃酸钠吗？Can rheumatoid arthritis be treated with sodium hyaluronate injection in the knee?	类风湿关节炎能根治吗？Can rheumatoid arthritis be curative?	0.0072	类风湿关节炎要怎么治疗最好？What is the best treatment for rheumatoid arthritis?	0.8960
	类风湿关节炎早期能治好吗？Can inchoate rheumatoid arthritis be cured?	0.0057	类风湿关节炎需要注意什么？What are the precautions for patients with rheumatoid arthritis?	0.8877
	类风湿关节炎会好吗！能怀孕吗？Can rheumatoid arthritis be cured? Can the patient become pregnant?	0.0051	类风湿关节炎有哪些症状？What are the symptoms of rheumatoid arthritis?	0.8869
类风湿关节炎可以采用免疫治疗吗？Can rheumatoid arthritis be treated with immunotherapy?	类风湿关节炎可以打疫苗吗？Can rheumatoid arthritis be vaccinated against?	0.3141	类风湿关节炎要怎么治疗最好 What is the best treatment for rheumatoid arthritis?	0.9053
	类风湿关节炎可以做沙疗吗？Can rheumatoid arthritis be treated with sand therapy?	0.1713	类风湿关节炎的治疗药物有什么危害？What are the dangers of rheumatoid arthritis medications?	0.9016
	类风湿关节炎能根治吗？Can rheumatoid arthritis be curative?	0.1211	严重类风湿关节炎怎么治疗？How to treat patients with severe rheumatoid arthritis?	0.8896
类风湿因子 71，能确诊是类风湿关节炎吗？Rheumatoid factor value is 71; can you be sure it is rheumatoid arthritis?	类风湿关节炎能根治吗？Can rheumatoid arthritis be curative?	0.0249	类风湿因子阳性是类风湿关节炎吗？Can a positive rheumatoid factor diagnose rheumatoid arthritis?	0.9170
	类风湿关节炎会好吗！能怀孕吗？Can rheumatoid arthritis be cured? Can the patient become pregnant?	0.0130	如何诊断类风湿性关节炎？How to diagnose rheumatoid arthritis?	0.9160
	如果有类风湿关节炎，现吃药可以怀孕吗？Can I get pregnant if I have rheumatoid arthritis while taking medicine?	0.00978	什么是类风湿关节炎？What is rheumatoid arthritis disease?	0.9156
类风湿关节炎如何进行康复训练？How do patients with rheumatoid arthritis undergo rehabilitation?	如何诊断类风湿性关节炎？How to diagnose rheumatoid arthritis?	0.4063	类风湿关节炎怎样运动？How do patients with rheumatoid arthritis exercise?	0.8945
	类风湿关节炎应该如何预防 How to prevent rheumatoid arthritis?	0.0779	类风湿关节炎要怎么治疗最好？What is the best treatment for rheumatoid arthritis?	0.1376
	类风湿关节炎如何医治，怎么预防？How to treat and prevent rheumatoid arthritis?	0.0711	类风湿关节炎预后如何？What is the prognosis for rheumatoid arthritis?	0.0339
怎么治好类风湿关节炎？How to cure rheumatoid arthritis?	如何诊断类风湿性关节炎？How to diagnose rheumatoid arthritis?	0.9422	类风湿关节炎要怎么治疗最好？What is the best treatment for rheumatoid arthritis?	0.9066
	类风湿关节炎诊断，怎么治疗？How is rheumatoid arthritis diagnosed and treated?	0.8995	严重类风湿关节炎怎么治疗？How to treat patients with severe rheumatoid arthritis?	0.8860
	怎么治疗类风湿关节炎的肾损害？How to treat kidney damage caused by rheumatoid arthritis?	0.7161	什么是类风湿关节炎？What is rheumatoid arthritis disease?	0.8669

technology of the FAQ-based intelligent question-answering system, which involves the comprehension and analysis of the content and form of users' questions using natural language processing technology to provide answers described in natural language. STS calculation techniques are key techniques for retrieving and displaying answers to such question answering systems.. Most of the current similarity calculation techniques based on deep learning seek to improve performance by changing the model structure, but the semantic information of the question is also an important part of the calculation result that must be considered. Our research additionally considers the classification information of the question, which can effectively limit the range of questions and answers, making the returned answer as close as possible to the users' query needs and improving the efficiency of the question-answering system. For example, for the question: "手指痛一定是类风湿关节炎吗"? ("Is finger pain necessarily rheumatoid arthritis?"), this study classified it as being related to the diagnosis of rheumatoid arthritis because we believe that patients need information related to the diagnosis of the disease. However, the top 5 classifications of similar sentences given by the baseline model were the following: concept, matters needing attention merging other diseases and symptoms, the principle of treatment, operative treatment, and a cure.

Acknowledgments

Funding: This work was supported by the National Key Research and Development Program of China (No. 2020AAA0104905) and the National Natural Science Foundation of China (No. 82071930).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-749/coif). The authors report that this work was supported by the National Key Research and Development Program of China (No. 2020AAA0104905) and the National Natural Science Foundation of China (No. 82071930). The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. There were no participants in this study, and no ethics approval and informed consent form were required.

References

1. Zheng S, Liu T, Qin B, Li S. Overview of Question-Answering. Journal of Chinese Information Processing 2002;(6):46-52.
2. Feng H. The Study on the Relationship Between Medication Adherence, Empowerment Level and Self-efficacy in Patients with Rheumatoid Arthritis. Dalian Medical University, 2019.
3. Huang C, Yin J, Hou F. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method. Chinese Journal of Computers 2011;34:856-64.
4. Li B, Liu T, Qin B, Li S. Chinese Sentence Similarity Computing Based on Semantic Dependency Relationship Analysis. Application Research of Computers 2003;(12):15-7.
5. Zhang L, Yin C, Chen J. Calculation and Analysis of Chinese Word Similarity Based on Semantic Tree. Journal of Chinese Information Processing 2010,24:23-30.
6. Jiang H, Han A, Wang M, Wang Z, Wu Y. Solution Algorithm of String Similarity Based on Improved Levenshtein Distance. Computer Engineering 2014;40:222-7.
7. Albitar S, Fournier S, Espinasse B. editors. An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification. Cham: Springer International Publishing, 2014.
8. Ormerod M, Martínez Del Rincón J, Devereux B. Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis. JMIR Med Inform 2021;9:e23099.
9. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019;1:4171-86.
10. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: generalized autoregressive pretraining for language understanding. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.
11. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234-40.
12. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. ArXiv. 2019;abs/1904.05342.
13. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv. Jun. URL: https://arxiv.org/abs/1904.03323 [accessed 2021-04-19].
14. Li S. Research of Relevancy between Setences Based on Semantic Computation.Computer Engineering and Applications 2002;(7):75-6, 83.
15. Mei J, Zhu Y, Gao Y. TongYiCi CiLin. Shanghai: Shanghai Lexicographical Publishing House, 1983.
16. Dong Z. Expression of Semantic Relationships and Construction of Knowledge Systems. Language Application 1998;(3):76-82.
17. Yu Z, Gao S, Ji P. Research on Question Sentence Similarity in RDAQAS. Journal of Kunming University of Science and Technology 2004;29:40-4, 71.
18. Cui H, Cai D, Miao X. Research on Web-based Chinese Question Answering System and Answer Extraction. Journal of Chinese Information Processing 2004;(3):24-31.
19. Xiong D, Wang J, Lin H. An LDA-based Approach to Finding Similar Questions for Community Question Answer. Journal of Chinese Information Processing 2012;26:40-5.
20. Mahmoud A, Zrigui A, Zrigui M. A Text Semantic Similarity Approach for Arabic Paraphrase Detection. Conference on Intelligent Text Processing and Computational Linguistics 2017.
21. Available online: http://www.academia.edu/2424592/OSAC_Open_Source_Arabic_Corpora
22. Wu Z, Liang J, Zhang Z, Lei J. Exploration of text matching methods in Chinese disease Q&A systems: A method using ensemble based on BERT and boosted tree models. Journal of Biomedical Informatics 2021;115:103683.
23. Li J, Huang G, Chen J, Wang Y. Short Text Understanding Combining Text Conceptualization and Transformer Embedding. IEEE Access 2019;7:122183-91.
24. Mikolov T, Chen K, Corrado GS, Dean J. Efficient estimation ofword representations in vector space. International Conference on Learning Representations 2013.
25. Pennington J, Socher R, Manning CD, editors. GloVe: Global Vectors for Word Representation. Conference on Empirical Methods in Natural Language Processing 2014.
26. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-Training with Whole Word Masking for Chinese BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings 2020:657-68.
27. Zhao Y. Similarity Calculation of Short Text based on attention mechanism. Harbin Institute of Technology 2020.
28. Liu P. Research on question similarity computation in domain question answering system. Harbin Institute of Technology 2018.
29. Zhou J, Chen Z, Li Z. Methods of Questions Pattern Classification and Similarity Measure for Question Answering System. Computer Engineering and Applications 2014,50:116-20.

Cite this article as: Li M, Shen X, Sun Y, Zhang W, Nan J, Zhu J, Gao D. Using Semantic Text Similarity calculation for question matching in a rheumatoid arthritis question-answering system. Quant Imaging Med Surg 2023;13(4):2183-2196. doi: 10.21037/qims-22-749

