Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4600404, 6 pages
https://doi.org/10.1155/2022/4600404

Research Article

Design and Implementation of a Medical Question and Answer System Based on Deep Learning

Yun Hu,[1] Guokai Han,[2] Xintang Liu,[2] Hui Li,[2] Libao Xing,[2] Yong Gu,[2] Zuojian Zhou,[1] and Haining Li[3]

[1]*School of Information Technology, Nanjing University of Chinese Medicine, Nanjing, Jiangsu, China*
[2]*School of Computer Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu, China*
[3]*Department of Neurology, General Hospital of Ningxia Medical University, Yinchuan, Ningxia, China*

Correspondence should be addressed to Haining Li; 2002000051@jou.edu.cn

Received 25 July 2022; Accepted 23 August 2022; Published 21 September 2022

Academic Editor: Lianhui Li

Medical services play a pivotal role in people's lives and in the national economy. Although the number of healthcare facilities is currently growing every year, there are still major problems in terms of access and pressure on the flow of people. Therefore, there is an urgent need for complementary medical services to alleviate the flow of patients and their psychological burden and to enable them to receive timely medical advice. This article designs and implements a medical Q&A system based on deep learning. We took a retrieval-based approach, using crawler technology that has been manually reviewed to build the Q&A database, and the Seq2Seq algorithm and the TF-IDF model to build the answer generation model. The medical question and answer system developed enable effective Q&A and relevant medical advice to be given. The algorithm proposed in this paper can quickly provide users with accurate answers compared to conventional search methods in real datasets.

1. Introduction

Artificial intelligence technologies, represented by machine learning, are currently being used in large numbers in various industrial sectors. Deep learning is a key technology in machine learning that has made great breakthroughs in computer vision, natural language processing, and speech processing, some of which surpass those of human professionals. However, the medical industry is more traditional and outdated than social media and e-commerce, and the development of modern medical technology can greatly improve people's quality of life.

With the rapid development of medical information technology, technologies such as deep learning and natural language processing have also developed rapidly, and intelligent diagnostic technology based on artificial intelligence will bring about a huge change to the medical industry. Answering users' questions quickly, accurately, and concisely in natural language is an important problem to be solved in medical Q&A systems. Traditional database search methods are unable to meet the demands of search efficiency and accuracy, and Q&A using natural language can quickly provide users with accurate answers compared to conventional search methods. This algorithm will improve the efficiency of medical services, promote the development of medical question and answer systems, and provide users with more accurate and faster answers to everyday medical questions.

The rest of the paper is organized as follows: Section 2 introduces the relevant materials and methods. The general design, key technologies of the medical Q&A system, and application results and data analysis are explained in Section 3. Finally, Section 4 concludes the work of this paper.

2. Materials and Methods

2.1. Related Studies. In the late 1980s, the discovery of new neural network propagation algorithms gave impetus to the development of machine learning and sparked a machine learning frenzy based on statistical models. This frenzy

continues to this day. In the 1990s, machine learning models such as the vector machine, boosting, and maximum entropy methods were developed and achieved good results in both theory and practice. Research on shallow artificial neural networks has been in limbo during this time due to the difficulties of theoretical analysis and the fact that training methods also take time to hone. The rapid growth of the Internet since 2000 to date has greatly necessitated intelligent parsing and prediction of massive amounts of information, but shallow learning models have achieved good results online. Some of the most relevant applications are as follows: CTR prediction, content-oriented recommendations, web search sorting, spam filtering, etc. Deep learning is the most active field of artificial intelligence and has achieved fruitful results in the fields of speech recognition, computer vision, and natural language processing in recent years [1]. One of them is artificially intelligent customer service.

Now, companies are launching their own intelligent Q&A systems. Examples include Google's GoogleAssistant and Apple's Siri. They can answer some basic natural language questions and can also follow simple user instructions. In China, many manufacturers have developed their own smart quiz software, such as Huawei's HiAssistant and Xiaomi's "Xiaoai classmate." Another commonly used Q&A system is a type of intelligent voice audio, such as the Tmall Genie developed by Alibaba, which solves basic natural language problems and can perform some basic commands. This voice interaction-based Q&A can solve some of the problems of everyday life, but they rely more on their own experiences on the Internet and face open-ended questions rather than medical ones.

Most of the current medical Q&A uses knowledge graphs to store medically relevant knowledge in a nonrelational database in an entity-relational way [2] and to present medical advice in a search and reasoning manner. Izcovich et al. [3] developed a graphical GRADE-based medical Q&A system based on GRADE. Oyelade et al. [4] collected relevant information based on the patient's symptom profile in order to conduct an initial specialist consultation. In addition, there have been many achievements in this field in China in recent years. For example, Xin [5] built a community health Q&A system using natural language processing techniques and various machine learning methods, while Chao [6] used big data analysis and deep learning techniques to build a disease guidance system that can be very helpful for patients for consultation and guidance. Elytai et al. [7] used a joint learning model to perform knowledge extraction and a stack-propagation framework to recognise medical input interrogatives and quickly feed the user with accurate medical answers. Hu [8] implemented CMQA, a Chinese medical Q&A system that understands user semantics well and generates SPARQL queries.

However, natural language processing techniques in Chinese are complex, and existing theories and industrialized results are not yet well suited to address the intolerance that exists in medical problems. Therefore, further research in the field of medical Q&A is still to be conducted. The Seq2Seq model is often used in machine translation, chat robots, text summarization, automatic generation of picture descriptions, and creation of ancient poems. In addition, the Seq2Seq model can also be applied to speech recognition, search intent completion, and recommendation. In the search recommendation scenario, when the user inputs the first half of the keyword, through the idea of interesting writing, the user inputs the first half of the vocabulary as the model input, predicts the possible search content in the second half, and improves the search efficiency. On this basis, we propose a medical Q&A system based on deep learning and build a sequence-order (Seq2Seq) architecture.

2.2. General Design of the Medical Q&A System. The medical Q&A system developed in the thesis uses a hierarchical architecture consisting of four layers: data layer, model layer, functional layer, and interaction layer. The advantages of using a layered architecture are that it reduces the correlation between layers and facilitates the standardisation of work; specific layers can be replaced, and analysis can be carried out from one level without too much knowledge of other levels, thus enhancing the repeatability and modifiability of the system. The architecture of the medical Q&A system based on deep learning [9–11] is shown in Figure 1.

The data layer managed and processed the data for the training corpus, laying a solid foundation for the design of an appropriate training set for the model layer. The training corpus was recorded this time in the form of questions with the title being the symptom of a condition and the answer being the name of a condition, combining the symptoms of these conditions sequentially and in reverse order to form a series of question responses.

At the model layer, secondary processing of the completed training set is completed with segmentation of the text, similarity operations, and feature extraction of the text. The Seq2Seq model is learnt so that the value of loss gradually decreases to achieve better accuracy.

At the functional layer, using natural language processing technology, the medical Q&A system is achieved by extracting text features from the input text and analysing the results to output predictive text, which is then subjected to secondary operations on the generated text via TF-IDF to improve accuracy.

The interaction layer provides access to the underlying layers for the purpose of meeting user requirements. This part consists mainly of the front-end interface and the human-computer interaction, which displays and receives the user's interactive actions through the terminal. The interaction layer is the top layer of the whole system and is the level that is directly accessible to the user.

2.3. Key Technologies for Medical Q&A Systems

2.3.1. The Seq2Seq Model. The intelligent quiz developed in this thesis uses the Seq2Seq model based on the encoder-decoder architecture [12, 13]. The Seq2Seq model is essentially an encoder-decoder construct: it transforms a series of long variables of data into a fixed vector; the decoder converts this fixed-length vector into a larger sequence of data. The difficulty of obtaining the true meaning in the case

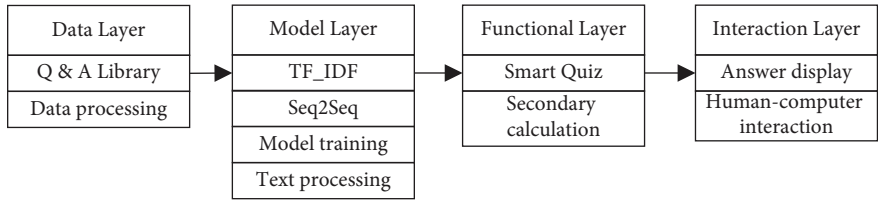

Figure 1: Medical Q&A system architecture.

of long input sequences can be well overcome by introducing the attention mechanism. The workflow of the Seq2Seq model is shown in Figure 2.

When doing the experiment, the first step is to obtain names, aliases, sites, infectiousness, population, symptoms, complications, departments to which they belong, and clinical management, treatments, common drugs, etc., of different diseases in the emergency department from one of the health sites, using crawlers. The data were analysed and filtered to obtain 3600 sets of questions and answers, which were then put into lists and then saved in data files for easy correction and training.

The model was built using the TensorFlow 2.0 framework. This model is a variant of the RNN that improves the neural network's ability to extract long text information [14], achieving better results than using the LSTM alone.

2.3.2. TF-IDF Model. TF-IDF is a statistical method for assessing the importance of words to a document set or corpus. TF-IDF has two values, one for TF (term frequency) and the other for IDF (inverse document frequency). It is calculated as follows:

$$TF - \mathrm{IDF} = tf * \mathrm{idf},$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{1}$$

$$i\,df_i = \log \frac{|D|}{\left|\left\{j : t_i \in d_j\right\}\right|}.$$

The basic concept of IDF is that if the smaller number of files includes t, which means that n is smaller and the IDF is larger, it means that t has a better classification function. If the number of documents including t is m and the number of all documents of other classes containing t is k, it is clear that the number of all documents containing t, $n = m + k$, is large when m is large and n is also large, then the smaller value will be obtained by the IDF, indicating that t does not have good classification performance. In practice, however, when a word is used multiple times in a category of documents, it indicates that it can denote the character of the text, and such words should have a higher weight and be used to distinguish other documents.

2.3.3. Bahdanau Attention. TensorFlow provides two attention mechanisms, a Bahdanau attention mechanism (additive accumulation) and a Luong attention mechanism

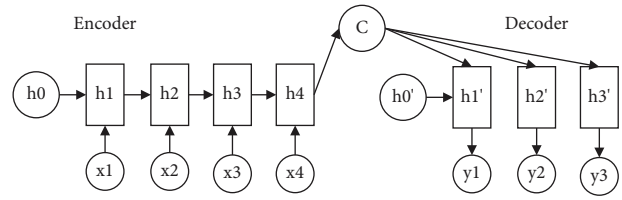

Figure 2: Workflows.

(multiplicative multiplication), and the former is used in this system [15, 16].

Bahdanau, an additive attention mechanism, uses a linear combination to output the hidden layer of the decoder and the full position of the encoder, thus improving the decoding pattern of the queue [17, 18]. It is essentially a two-layer fully connected network with an activation function of tanh and an output layer of dimension 1. The advantages are that the encoder generates a hidden state vector for each input vector; the alignment score is calculated at each encoder output xi using the hidden state st-1 of the previous moment; this alignment score can be converted into a probability distribution vector by softmax; according to the probability distributed alignment score, the context vector ct can be derived by weighing the encoder outputs at each position; the context vector ct and the embedding corresponding to the previous moment's encoder output ŷt-1 are spliced as the current moment's encoder input, and the new output and hidden state are generated by the RNN network, with the real target sequence y=(y1,ym) in the training process, and more yt-1 is used instead of ŷt-1 as the decoder input at moment t. At time t, the hidden state of the decoder is denoted as st = f(st-1, ct, yt-1), and the attention fraction of the hidden state st-1 for all outputs X of the encoder at each t is

$$\alpha_t s_{t-1}, X = \mathrm{softmaxtanh}\ s_{t-1} W_{\mathrm{decoder}} + X W_{\mathrm{encoder}} W_{\mathrm{alignment}}, c_t = \sum_i \alpha_{ti} x_i. \tag{2}$$

As shown in Figure 3, the blue one is an encoder and the red one is a decoder. Based on traditional encoding and decoding algorithms, the attention mechanism requires more context vectors to generate the corresponding context vectors. Each context vector is a weighted sum of each word_x of Input_Sentence, where the weight vector is the attention vector, indicating the importance of each word_x of Input_Sentence at this point in time when word_y produces Output_Sentence. Eventually, the current text vector is combined with the current y, and it is taken as the final result.

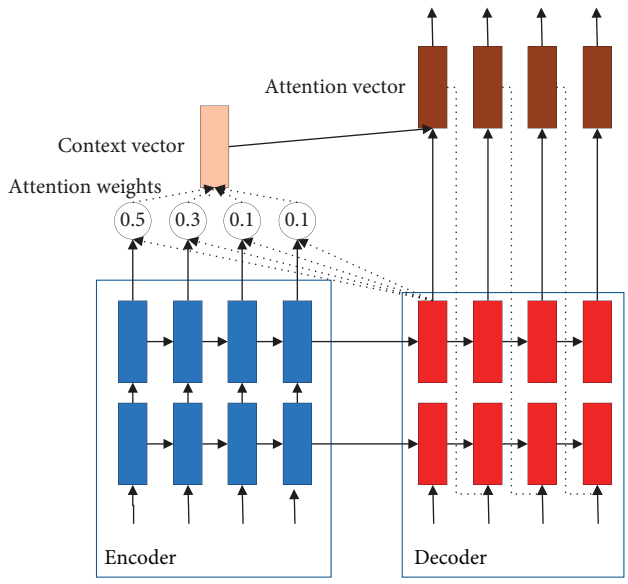

Figure 3: Attention mechanism.

The rules for calculating attention mechanisms are as follows:

$$\begin{aligned} \alpha_{ts} &= \frac{\exp\left(score\left(h_t, \overline{h}_s\right)\right)}{\sum_{s'=1}^{S} \exp\left(score\left(h_t, \overline{h}_s\right)\right)}, \\ c_t &= \sum_s \alpha_{ts} \overline{h}_s, \\ a_t &= f\left(c_t, h_t\right) = \tanh\left(W_c\left[c_t; h_t\right]\right), \\ \text{Score}\left(h_t, \overline{h}_s\right) &= \begin{cases} h_t^T W \overline{h}_s \\ v_a^T\left(W_1 h_t + W_2 \overline{h}_s\right) \end{cases}. \end{aligned} \tag{3}$$

2.4. Application Results and Data Analysis

2.4.1. Experimental Evaluation Indicators. The trained model can be used to predict new text, and it is not possible to determine whether the results are satisfactory. No model is as good as it should be, and we are always looking for better. When the model has finished training, in order to determine how good or bad it is, we can make predictions based on the available information and compare the predictions with reality as a way to judge how good the model is [19]. Therefore, we need some metrics to measure how similar the actual predicted results are to the expected results.

Common evaluation metrics for text classification tasks include accuracy, precision, recall, and F1-score to name a few.

(1) Accuracy. Accuracy is the most basic evaluation metric, which is the percentage of correctly classified test samples of the total test samples. The advantages are that it is simple to calculate, easy to understand, and can be used for both dichotomous and polyphenolic classes. However, when the data are unbalanced, it is not a good measure of how good the model is. The formula is as follows:

$$\text{Accuracy} = \frac{\text{answer}_{\text{true}}}{\text{answer}_{\text{all}}}. \tag{4}$$

(2) Precision. In the classification model, there exists an outcome with an output, which is a prediction. Assume that A is predicted by Class_1. There are only two cases of A: A is Class_1 (prediction is correct) and A is not Class_1 (prediction is wrong). If all data are predicted, then Class_1 data will appear to be incorrectly predicted in relation to Class_1 and the other non-Class_1 will be considered to be Class_1. The confusing evidence is shown in Table 1.

The denominator of precision is all test samples classified as Class_1, and the numerator is the number of test samples that are predicted to be Class_1 test samples that are actually positive classes.

$$\text{Precision} = \frac{TP}{(TP + FP)}. \tag{5}$$

(3) Recall. Recall is also derived from the above table, and its denominator is all test samples that are positively true to Class_1; its numerator is predicted to be a test sample of Class_1, as is precision.

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{6}$$

(4) F1-Score. In practical experiments, we all want both precision and recall to be high, but in reality, you cannot have both, so there is something that combines them both, the $F1$ value. The larger the $F1$ value, the better the model.

$$F1\text{Score} = \frac{2PR}{P + R}. \tag{7}$$

The similarity is the result of a similarity calculation between the predicted answer by the model and the original test data.

2.4.2. Algorithm Description. When designing the Q&A system, we initially used the TF-IDF algorithm, which calculates the similarity between user input and existing statements in the database and then returns the answer corresponding to the value with the highest similarity, but the algorithm takes longer to calculate as the data grow. Later, after reviewing the information, the Seq2Seq model, which is currently the most used, was chosen for implementation. Seq2Seq is generative, predicting possible output values from a trained model and getting one word or text per prediction.

Since the most likely value is predicted each time, the value is not necessarily what the user wants or the correct value. If the predicted result is a value within a certain range, it is possible to control the content of the output and also

Table 1: Confusing evidence.

Actual/Forecast	Predicted as Class_1	Predicted for other categories
Actual class is Class_1	TP: the number of test data items that are actually predicted correctly even for Class_1	FN: the number of test entries that are actually Class_1 but are predicted to be other classes
Actual class is other class	FP: the number of data items that are not actually Class_1 but are predicted to be Class_1	TN: the number of data entries that are not actually Class_1 and are not predicted to be Class_1

improve the accuracy of the output. Thus, the two methods can be combined, and result A predicted by Seq2Seq can be then calculated by TF-IDF to output result B that is most similar to A from the existing database, as shown in Figure 4.

2.4.3. Results. The model first classifies the dataset questions with a batch_size of 128 and an epoch of 7. The questions are classified into two categories, symptom and name, and the results of recall, precision, and F1 of the classification results are shown in Table 2.

Secondly, all test datasets were manually tested, and a total of 410 test datasets were counted; it was found that a total of 361 datasets could be predicted by the model, a rate of 88.0488%.

Because the Seq2Seq model is called based on the textual properties of the sentence, if the information in the training set is not precise enough, or if the user's text does not correspond to the input to the model, then this can lead to inaccurate results. On this basis, a simple algorithm of TF-IDF is introduced to perform text similarity analysis. We take the output of Seq2Seq and recall TF-IDF to perform a secondary operation so that the output belongs to the data already used in the database. For example, the Seq2Seq model predicts "exercise-induced asthma" for "coughing and dry cough after strenuous exercise" (a symptom of cough variant asthma), although this condition is not named in the database, and then the TF-IDF algorithm outputs "cough variant asthma" to improve the accuracy of the output. The comparison results are shown in Table 3.

As the information used in this system is taken from the emergency department, it is very useful for emergency management of emergencies. The medical information provided to the user by the system comes from the medical knowledge base, and this database is rigorously hand-selected, its answers are output based on the data available in the database, so the accuracy of its answers can be guaranteed. As calling the model consumes a long time, the files for model calculation are placed on Tencent Cloud servers, which improve the rate of model calculation and reduces the time for Q&A. Simulations of the calculations show that the method gives a relatively good answer.

Given the specific nature of the system's model training set, which requires user input of symptoms to most accurately invoke the model, a guided input module was added to the system. The initial values in the module are the twenty most common symptoms for the user to select. For each symptom selected, the symptom is added to the input box, and the value in the guided input module change to all the remaining symptoms that have the symptomatic disease. The input box monitors the user's input in real time, and if the user enters symptoms themselves, the value of the guided input module will change accordingly. It is recommended that the user selects 3–5 symptoms before choosing to send, as this will make the answers more accurate. After sending, the symptoms on the right revert to their initial values.

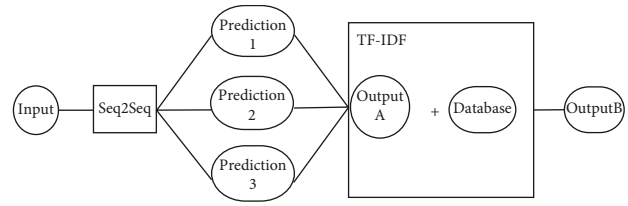

Figure 4: Algorithm description diagram.

Table 2: Classification results.

Type of question	Recall	Precision	*F*1
Symptom	**0.894737**	**0.772727**	**0.829268**
Name	**0.842105**	**0.727273**	**0.780488**
All questions	**0.804878**	**0.804878**	**0.804878**

Table 3: Models comparison results.

Models	*F*1
TF-IDF	**0.731707**
Seq2Seq	**0.804878**
Fusion model	**0.880488**

3. Conclusions

This article designs and implements a medical Q&A system based on deep learning. The algorithm proposed in this paper can quickly provide users with accurate answers compared to conventional search methods in real datasets. The system has been validated through several experiments and has achieved excellent results [20]. The added guided input enables the user to select the information accurately, which in turn helps the user to quickly locate disease information and know how to administer medicine. However, this system has some limitations in certain aspects. Due to the specificity and high quality requirements of the medical question and answer content and the system's ability to learn autonomously is not yet optimal, in practice, the user is currently limited to selecting or entering symptoms in order to use the system most effectively. In terms of modelling, the accuracy of the model has not been maximised, and modifications to certain values could be considered in the future. In terms of data, the quality of the training set would need to

be improved by consulting a professional and modifying it manually due to the small amount of first aid data; in addition, a library of common question and answer statements could be added to diversify user input [21]. Intelligent diagnostic techniques based on artificial intelligence will bring great changes to the healthcare industry [21]. The intelligent diagnostic technology based on artificial intelligence will bring about a huge change to the medical industry [22, 23].

Data Availability

No data were used to support the findings of this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 72174079, No. 1210050123, No. 72101045), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 21KJB520033), Qinglan Project Teaching Group of Jiangsu Province, Lianyungang "521 project," and the Research Project of Graduate Education and Teaching Reform of Jiangsu Ocean University (YJG202201).

References

[1] Q. Yao, Q. Wang, Q. Shi, M. Zhang, and Wu. Deng, "Deep learning in modern medical applications," *Computer Systems Applications*, vol. 31, no. 04, pp. 33–46, 2022.

[2] H. Liu, Li. Yang, H. Duan, L. Yao, and Z. Qin, "A review of knowledge graph construction techniques," *Computational Research and Development*, vol. 53, no. 3, pp. 582–600, 2016.

[3] A. Izcovich, J. M. Criniti, J. I. Ruiz, and H. N. Catalano, "Impact of a GRADE-based medical question answering system on physician behaviour: a randomised controlled trial," *Evidence-Based Medicine*, vol. 20, no. 3, pp. 81–87, 2015.

[4] O. N. Oyelade, A. A. Obiniyi, S. B. Junaidu, and S. Adewuyi, "Patient symptoms elicitation process for breast cancer medical expert systems: a semantic web and natural language parsing approach," *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 72–81, 2018.

[5] Y. Xin, "A natural language processing-based health care question and answer system," *Communication World*, vol. 2018, no. 6, pp. 255-256, 2018.

[6] Li. Chao, *Research and application of Intelligent Disease Guidance and Medical Question and answer Methods*, Dalian University of Technology, Dalian, 2016.

[7] Li-T. Yi, C. Dong, Z. Niu, Si-J. Liu, X. Ni, and S.-K. Luo, "Research and implementation of intelligent medical domain question and answer system," *Information Record Materials*, vol. 22, no. 05, pp. 232–234, 2021.

[8] R. Hu, *Design and Implementation of a Chinese Medical Q&A System Based on Deep Learning*, Huazhong University of Science and Technology, 2020.

[9] X. Wang, *Research on Intelligent Question and answer Model Based on Deep Learning*, Xi'an University of Science and Technology, 2021.

[10] Z. Yao, "Development of a medical question-and-answer system based on deep learning," *China Medical Equipment*, vol. 34, no. 12, pp. 88–91+141, 2019.

[11] T. Xu and T. Qi, "Research on intelligent Q&A for virtual academic communities based on deep learning," *Journal of Intelligence*, vol. 40, no. 04, pp. 163–169, 2021.

[12] S. Ilya, V. Oriol, and V. L. Quoc, "Sequence to sequence learning with neural networks," 2014, https://arxiv.org/abs/1409.3215.

[13] Ke. Sun, T. Qian, X. Chen, and M. Zhong, "Context-aware seq2seq translation model for sequential recommendation," *Information Sciences*, vol. 581, no. 12, pp. 60–72, 2021.

[14] K. Cho, B. van Merrienboer, C. Gulcehre et al., "Learning Phrase Representations using RNN encoder–decoder for statistical machine translation," 2014, https://arxiv.org/abs/1406.1078.

[15] H. Wang, L. Sun, B. Wu, Z. Liu, W. Zhang, and S. Zhang, "Research on RPR fusion model based on intelligent question and answer for reading comprehension," *Computer Application Research*, vol. 39, no. 03, pp. 726–731+738, 2022.

[16] J. Gao, X. Fang, S. Liu, and J. Fu, "Research on knowledge association and intelligent question and answer of cultural heritage resources in collections based on Linked Data," *Intelligence Science*, vol. 39, no. 05, pp. 12–20, 2021.

[17] Li. Hui, H. Li, and S. Zhang, "Intelligent learning system based on personalized recommendation technology[J]," *Neural Computing & Applications*, vol. 31, no. 9, pp. 4455–4462, 2019.

[18] G. R. Reddy, C. Xanthopoulos, and Y. Makris, "On improving Hotspot Detection through Synthetic pattern-based database Enhancement," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 12, pp. 2522–2527, 2021.

[19] H. Li, Z. Zhong, J. Shi, H. Li, and Y. Zhang, "Multi-objective Optimization-based recommendation for massive online learning resources," *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25274–25281, 2021.

[20] L. I. Hui, M. A. Xiao-Ping, S. H. I. Jun, L. I. Cun-Hua, Z. H. O. N. G. Zhao-Man, and C. A. I. Hong, "A recommendation model by means of Trust Transition in complex network Environment," *Acta Automatica Sinica*, vol. 44, no. 2, pp. 363–376, 2018.

[21] L. T. van Eijk, S. Servaas, C. Slagt, and I. Malagon, "Predicting fluid responsiveness," *European Journal of Anaesthesiology*, vol. 38, no. 5, pp. 449–451, 2021.

[22] E. Mutabazi, J. Ni, G. Tang, and W Cao, "A review on medical textual question answering systems based on deep learning Approaches," *Applied Sciences*, vol. 11, no. 12, p. 5456, 2021.

[23] H. Veisi and H. F. Shandi, "A Persian medical question answering system," *The International Journal on Artificial Intelligence Tools*, vol. 29, no. 06, Article ID 2050019, 2020.