ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

A Review of Knowledge Graph-based Question and Answer System Research and Its Application in Chronic Disease Diagnosis

Zhaoyang Cao, Lin Ni*, Lirong Dai

University of Science and Technology of China, Hefei 230027, Anhui, China
Email: nilin@ustc.edu.cn
**Corresponding Author*

Abstract: *Question and answer systems have a long history of development, and with the maturity of knowledge graph technology in recent years, knowledge graph-based question and answer systems are gradually applied to many fields. In this paper, we first discuss the concept of knowledge graph and question and answer system, and then analyze the key technologies used in it. Before dealing with linguistic problems, questions need to be structured and represented by semantic parsing and space vector-based modeling are common approaches. The question and answer system can be divided into three parts: question classification, entity recognition, and relationship extraction, for each of which a large number of techniques have been studied. Finally, a question and answer system based on the knowledge graph of chronic diseases is designed to provide a proven solution for this field, in view of the problem that there are many patients with chronic diseases but lack of sufficient knowledge of the diseases.*

Keywords: *Knowledge graph, Question and answer system, Chronic disease*

In recent years, with the continuous development of information technology and artificial intelligence, data has achieved explosive growth in various fields [1]. The term big data was originally used to describe huge data sets. Compared with traditional data sets, it covers a large amount of valuable unstructured data and semi-structured data [2], while traditional data storage and analysis methods are no longer available. To handle this kind of data efficiently, people began to explore a new way of data storage — knowledge graph.

The concept of Knowledge Graph was first formally proposed by Google in 2012 [3]. It is a special semantic network, which can be regarded as a graph-based data structure to describe the relationship between entities, relationships, and attributes [4]. Its original purpose is to improve the search efficiency of search engines [5]. Since 2013, with the continuous development of intelligent information services, knowledge graph has been widely popularized and applied in academia and industry, and it plays an important role in the fields of question and answer system, smart cities, recommendation systems, and big data risk control [6].

Question and Answer System (QAS) is a computer system that uses natural language to interact with users. Compared with traditional web search, on the one hand, users can use daily spoken language to query, and on the other hand, the system finally returns the user with a certain result, which meets the needs of netizens for convenient and accurate information acquisition, and reflects the intelligence of computers [7].

Chronic disease is the abbreviation of chronic non-communicable disease. With the changes in residents' lifestyle, diet and work style, it has become an important factor threatening human health [8]. According to the global burden of disease (GBD) data in 2019, there are 56.5 million deaths worldwide. The highest risk factor attributed to global death is hypertension, which accounts for 19.2% (10.8 million) of the total deaths, followed by smoking (second-hand smoke), accounting for 15.4% (8.71 million) [9]. With the continuous increase of work pressure, chronic diseases have a trend of younger generation [10].

The number of patients with chronic diseases is increasing, but most patients have certain deficiencies and deviations in their understanding of chronic diseases. At present, some existing chronic disease websites cannot provide answers according to the patient's intention, so that the patient's diagnosis time is wasted. Therefore, developing a question and answer system with high quality and high convenience is currently the most effective solution.

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

This article focuses on the technology and development of question and answer system based on knowledge graph in recent years. First, it introduces the concept of knowledge graph, and then summarizes the key technologies involved in the question and answer system to provide information reference for more researchers. At the end of the article, a self-help question and answer system for chronic disease diagnosis will be implemented based on related technologies. This question and answer system can answer questions such as the overview of chronic disease-related diseases, prevention and treatment methods, and related medical guidelines, helping users to more fully understand chronic disease-related knowledge, alleviating the pressure on medical institutions, and providing a suitable solution for related fields.

1. Research status

1.1. Knowledge graph

Knowledge graph is a semantic knowledge network that describes concepts and their relationships in the physical world. The network is generally represented in the form of triples, that is, "head entity-relationship-tail entity" [4]. Of course, some scholars have recently tried to incorporate time information into triples to form a Temporal Knowledge Graph containing quadruples [11]. This type of knowledge graph not only includes causality, timing, etc. between events, but also describes the law and evolution mode between events, so it has more powerful application value. In the corporate world, it is more to use knowledge graph to provide personalized intelligent services for enterprises, such as IBM's Watson Health [12] in the medical field, and Alibaba's "Qianmo Data Management Platform" in the e-commerce field.

According to the different areas of knowledge coverage, knowledge graph can be roughly divided into open domain knowledge graph and vertical domain (specific domain) knowledge graph [13]. The open domain knowledge graph usually contains general domain knowledge, which is similar to encyclopedia knowledge, contains a lot of common sense content, and emphasizes the breadth of knowledge. For example, the link database YAGO [14] developed by Max Planck Institute in Germany, Freebase [15] based on Wikipedia, and Satori knowledge graph on Bing, etc. The vertical domain knowledge graph is oriented to a specific domain, and the knowledge content is more professional, and it emphasizes the depth and relevance of knowledge. This type of knowledge graph has specific use objects and scenarios, such as medical knowledge graph and financial domain knowledge graph, legal domain knowledge graph and social domain-oriented microblog knowledge graph, etc.

1.2. Question and answer system

The question and answer system has a long history of development. In 1950, Alan Turing, the father of computer science, proposed the concept of human-computer interaction using natural language [16]. In the 1960s, with the help of artificial intelligence technology, researchers tried to build an intelligent system that could answer people’s questions, so the first batch of the question and answer system appeared one after another. Typical representatives are BASEBALL [17], LUNAR [18] and ELIZA [19] etc. From the 1970s to the 1980s, the rise of computer linguistics made people focus on how to reduce the cost and difficulty of building a question and answer system, and the question and answer system became more complicated. The representative system is Berkeley Unix Consultant (UC) [20].

In the 1990s, especially after the Text Retrieval Conference (TREC) introduced QA track into the conference theme, it greatly promoted the development of the question and answer system, and the question and answer system entered the open field and based on free text in the new period [6]. Nowadays, more and more enterprises and universities have participated in the technical research of the question and answer system, and have also put forward Question Answering over Knowledge Bases (KBQA), Community-based Question Answering (CQA), Web-based question and answer (WQA) [21], etc. With the birth of large-scale knowledge graph, as an emerging data storage method, the question and answer system has a new research direction.

2. Key technology

2.1. Structured representation of the problem

The structured representation of the problem can be roughly divided into two ways, one is through

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

semantic analysis, and the other is based on space vector modeling.

The method of semantic analysis mainly analyzes the natural language question components and converts them into corresponding logical expressions, and then uses the semantic information of the knowledge graph to map the logical expressions into structured query sentences that the computer can understand. Berant [22] et al. proposed a dictionary-grammar-based semantic analysis method, which relies on λ-DCS (Lambda dependency-based compositional semantics) grammar for syntactic analysis. Bast [23] et al. proposed a template-based semantic analysis, which generates different query sentences by making different templates, and then uses a sorting algorithm to sort the query candidate answers for analysis. With the popularity of neural networks, traditional semantic analysis methods are gradually being replaced. Yin [24] et al. incorporated an attention mechanism into the Convolutional Neural Network (CNN), and obtained the correlation between the question and the knowledge of the knowledge base by calculating the similarity of the embedding vector to obtain the answer to the question. Golub [25] et al. proposed the use of a character-level neural network coding and decoding framework to learn the representation of sentences and reduce the interference in obtaining the correct question entity mentions. Lukovnikov [26] et al. designed an end-to-end attention coding and decoding network model based on two levels of words and characters that can be trained at the same time, which can deal with text problems outside the lexicon and capture the semantics of the text level.

Based on the space vector modeling method, the candidate answers in the question and knowledge graph are mapped to the low-dimensional vector space in a distributed manner to generate a space vector. The correlation score between the question vector and the corresponding correct answer is calculated through the model, and finally the knowledge with the highest association score is used as the basis for determining the answer. This method was first proposed by Bordes [27] et al. Subsequently, Bordes [28] and others combined the memory network to store the knowledge in the knowledge base in the memory module, and then convert the problem into a distributed expression through the input module, and then select the memory with the highest relevance to the problem through the output module, and finally output the object in the triplet through the answer module. Guu [29] et al. regarded the vector space model of the knowledge graph as an edge (relation) traversal operation, and used multi-step reasoning to make the learned vector representation more helpful to the reasoning and prediction of entities and relationships.

2.2. Key technology of the question and answer system

The question and answer system generally consists of three parts: question classification, entity recognition, and relationship extraction. This section will give an overview of the three modules.

2.2.1. Problem classification

Question classification is a huge system. At present, most of the question classification uses a classification system with the answer type as the main body. The most authoritative internationally is the UIUC problem classification system [30], which classifies English, which contains 6 major categories and 50 sub-categories, as shown in Table 1. On the basis of this classification system, domestic researchers have formulated a set of Chinese question classification system [31], which has 7 major categories and 60 sub-categories, as shown in Table 2.

Table 1: UIUC problem classification system

Coarse	Fine
DESC	definition, description, manner, reason
ABBR	abbreviation, expansion
NUM	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight
ENTY	animal, body, color, creation, currency, disease, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
LOC	city, country, mountain, other, state
HUM	description, group, individual

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

Table 2: Chinese question classification system

Large category	Small category
Location	Planet, city, continent, country, province, river, lake, mountain, ocean, island, location list, address, location, etc.
Description	Abbreviation, meaning, method, reason, definition, description other
Time	Year, month, day, time, time range, time enumeration, time other
Characters	Specific characters, organization, character description, character enumeration, character other
Numbers	Number, price, percentage, distance, weight, temperature, age, area, frequency, speed, range, order, number list, number other
Entity	Animals, plants, food, colors, currency, languages, materials, machinery, vehicles, religion, entertainment, entity enumeration, entity other
Unknown	Unknown

Regarding the problem classification algorithm, machine learning methods were initially used, such as naive Bayes classification algorithm, Support Vector Machine (SVM), k-nearest neighbor model, etc. Li [32] et al. proposed combining part of speech, bag of words and syntactic dependency tree on the basis of SVM, and analyzed the structure of the problem by calculating the value of its kernel function. For the classification of Chinese problems, Yu [33] et al. used high-frequency words as features, and trained classifiers on the labeled and unlabeled data sets through a collaborative training algorithm, and achieved 88.9% and 78.2% in the classification of large and small problems respectively. Zhang [34] and others made full use of the syntax structure of the problem, using the bag-of-words model and the kernel function based on the problem grammar tree to train SVM, and also got a higher classification accuracy.

The problem classification method based on deep learning has higher fault tolerance and accuracy than traditional machine learning. Kim [35] and others first proposed a CNN network classification model — TextCNN. This method uses word2vec to turn question sentences into word vectors, and achieves good results on multiple English sentence classification data sets by adjusting the parameters of the CNN model. Based on the Recurrent Neural Network (RNN) model, Liu [36] et al. proposed using three different information sharing mechanisms to train text with specific tasks, and finally tested it on four benchmark text classification tasks, and achieved good results. For problem classification tasks, Yao [37] and others, combined with more flexible Graph Convolutional Networks (Graph Convolutional Networks, GCN), proposed a text graph convolutional network — textGCN for corpus, which is stronger robustness for less training data.

2.2.2. Entity recognition

Named Entity Recognition (NER) is the core task of question and answer system construction. Its earliest research is based on the rule method, that is, relying on domain experts to manually formulate a rule template, and then use the rule template to match and search the text to identify entity information [38].

Methods based on statistical machine learning mainly include Maxmium Entropy (ME) [39], Hidden Markov Mode (HMM) [40], SVM [41] and Conditional Random Fields (CRF) [42] and other related models.

Later, Collobert [43] and others proposed a neural network-based named entity recognition method, which uses the entire sentence as the input of the current predicted word, and adds the relative position information of each word in the sentence, and then uses the convolutional neural network to extract the sentence characteristic information. Huang [44] et al. proposed the BiLSTM-CRF model, which uses a bidirectional long short-term memory network to capture the past and future features of the current time ***t***. It is applied to sequence tagging tasks such as named entity recognition and part-of-speech tagging, but for Chinese entity recognition, the effect is average, and there is still a lot of room for improvement in accuracy. The proposal of BERT (Bidirectional Encoder Representations from transformers) [45] has made a breakthrough in the research of named entity recognition. It uses a bidirectional transformer encoder to generate context-dependent word vectors, and successfully obtained eye-catching results in 11 text analysis tasks that year. Later on, more and more researchers have improved BERT accordingly for different application scenarios, such as Souza [46] et al. used the BERT-CRF-based model architecture for Portuguese entity classification tasks; Zhang [47] et al. applied BERT-BiLSTM-CRF to Chinese electronic medical record entity recognition; Yang [48] et al. built the BERT-BiGRU-CRF model for the shortcoming that traditional word vector representation method cannot represent the ambiguity of words, which enhanced the semantic representation of the word, and achieved a good effect on the MSRA corpus.

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

2.2.3. Relation extraction

Relation extraction in question and answer systems is generally used to capture the relationship between subject entities and answer entities. There are three main methods of relation extraction: based on supervised learning, based on semi-supervised learning and based on unsupervised learning. After deep learning has matured, it has also been applied to relation extraction and has become a new research focus [49].

Relation extraction based on supervised learning mainly uses manual intervention, that is, training models on manually labeled data, and then applying them in specific fields. Kambhatla [50] et al. adopted a feature vector-based method, and used the ME classifier to model the entity relationship extraction problem for the first time, and combined the entity type, entity word, syntax analysis tree, dependency relationship and other features to construct the feature vector. Good results can be achieved by using few vocabulary features. Zelenko [51] and others used a kernel-based method to convert nonlinear problems into linear problems through kernel functions. Using shallow analytic tree kernels combined with SVM classifiers can achieve good results on hundreds of news corpus data sets.

Semi-supervised learning is based on a small number of artificially added entity pair seeds and a large number of unlabeled samples, using pattern learning methods for iterative training to obtain a classification model. Bootstrapping method is one of the more commonly used algorithms [52]. Brin [53] and others established the DIPRE (Dual Iterative Pattern Relation Expansion) system based on this algorithm. The system uses a small number of book titles and author names as seed entity relationship pairs at the beginning of the iteration. Through continuous iteration, the template can automatically discover new entities relationship pairs and join them. Combined with SVM, Zhang [54] et al. proposed a BootProject algorithm based on random feature projection, which can reduce the dependence on labeled training data.

Clustering is mainly used based on unsupervised learning. Hasegawa [55] et al. set the threshold of repeated occurrences to identify latent semantic relationships and cluster them. Hassan [56] et al. proposed an unsupervised information extraction method based on graph mutual enhancement, which achieved excellent performance in automatic content extraction (ACE) relation detection and characterization (RDC) tasks.

Regarding the use of deep learning-related relation extraction, Socher [57] and others took the lead in using the RNN model to learn word semantics and propositional logic, so as to obtain a variety of syntax types and different lengths of phrase vectors and sentence vectors. Lin [58] and others improved the max pooling layer of the traditional CNN model and proposed the PCNN (Piece-Wise-CNN) model, and used the attention mechanism to reduce the generation of false labels. The final F-score was 5 percentage points higher compared to the machine learning-based approach. Xu [59] et al. proposed a new neural network model SDP-LSTM for relation classification. It can combine the key information on the shortest dependency path for iterative learning, thereby ignoring irrelevant information and providing a new solution to the problem of relationship extraction.

3. Construction of question and answer system

This chapter designs and implements a Chinese question and answer system based on the chronic disease knowledge graph to provide users with self-service queries of chronic disease-related diseases. The system enhances the timeliness and convenience of question and answer, and can answer user questions in real time without waiting for doctors to go online. This system can not only strengthen users' awareness of related diseases, but also relieve the pressure on medical institutions to a certain extent. The system construction is mainly divided into two parts: (1) data collection and knowledge graph construction; (2) the realization of the question and answer system based on the knowledge graph. The following will be discussed separately.

3.1. Data collection

The knowledge of the knowledge map mainly comes from the semi-structured Chinese Chronic Disease Encyclopedia (http://jbk.39.net/zq/manxingbing/) data. Use crawler tools to focus on structured data and disease as the center to crawl the introduction, symptoms, etiology, prevention, and diet information in the corresponding disease pages. The single data storage format is: {"id": "data", "name": "data", "symptom": "data", "category": "data", "food":{"good": " data", "bad": "data"}, "treat": "data", "drug": "data"}.

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

3.2. Construction of knowledge graph

There are generally two ways to construct knowledge graph, top-down and bottom-up. The top-down is constructing data model firstly. The data model is constructed from the top level concept, and then gradually refined downwards to form a good hierarchical structure, and finally entities are added to the concept. It is suitable for the knowledge graph of the vertical industry field. The bottom-up is just the opposite. First, the entities are summarized to form the underlying concepts, and then gradually abstract upwards, and finally the upper-level concepts are formed. It is suitable for the knowledge map of the general domain. What this article needs to build is a knowledge map for chronic diseases, which belongs to the vertical field, so it adopts a top-down approach, that is, build the model layer first, and then build the data layer. The specific construction process is shown in Figure 1.

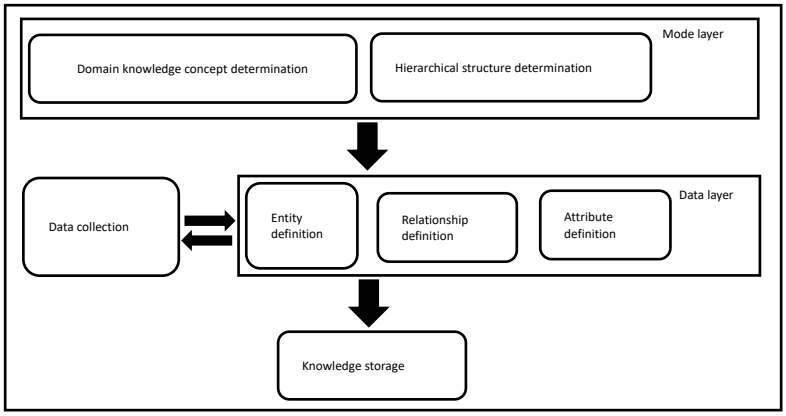

Figure 1: Top-down construction process of knowledge graph

The pattern layer mainly stores abstracted knowledge, which is the conceptual induction of the ontology of the data in the data layer. Specifically, it is to collect relevant ontology knowledge in the field of chronic diseases and determine the ontology framework.

The data layer mainly analyzes and manipulates data, and extracts entities, relationships and attributes in the data by programming. The chronic disease-based knowledge graph constructed in this paper mainly involves 5 entity types, 6 relationship types, and 6 attribute types, as shown in Table 3, Table 4, and Table 5, respectively.

Table 3: Types of knowledge graph entities

Entity	Disease	Food	Symptom	Drug	Department
Examples	Endometritis	Bitter gourd	Abdominal bloating	Jinji capsule	Gynecology

Table 4: Types of knowledge graph relationships

Relationship	comment_drug	do_eat	no_eat
Examples	<Coronary heart disease, commonly used, Xuesaitong effervescent tablets>	<Coronary heart disease, suitable for food, soy milk>	<Coronary heart disease, not to eat, pepper>
Relationship	has_symptom	belong_to	acompany_with
Examples	<Coronary heart disease, symptoms, chest distress>	<Coronary heart disease, belongs to, cardiovascular medicine>	<Coronary heart disease, accompany with, myocardial infarction>

Table 5: Types of Knowledge Graph Attributes

Attribute	name	desc	site
Examples	Hypertension	Hypertension is based on the arteries of the systemic circulation...	Blood and vessels
Attribute	cured_prob	cure_lasttime	cure_way
Examples	0.0001%	1-3 months	Medication

Common methods of knowledge storage are based on Resource Describle Framwork (RDF) storage and graph database storage. Since the graph database stores data in an unstructured form, that is, it uses

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

the nodes and edges in the graph structure to organize the data, which has strong flexibility, so this article uses the currently popular open source graph database Neo4j for data storage. At the same time, the Cypher sentence provided by Neo4j can be used to retrieve the data in the knowledge base. Cypher is a descriptive graph query language with simple syntax and powerful functions. The visual display of some nodes based on the chronic disease knowledge map constructed by Neo4j in this paper is shown in Figure 2.

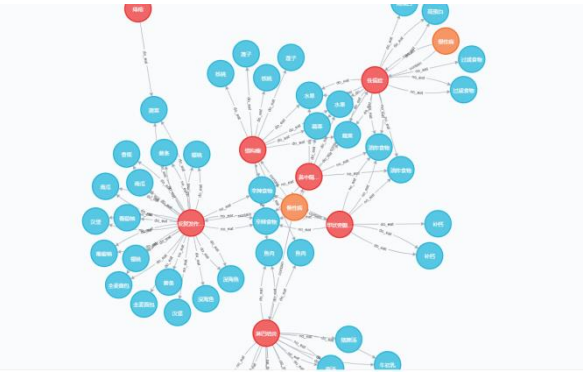

Figure 2: Visual display of knowledge graph

3.3. Implementation of the question and answer system

This article divides the design of the question and answer system into the following modules: question classification module, question analysis module, answer query module and answer module. The design process is shown in Figure 3.

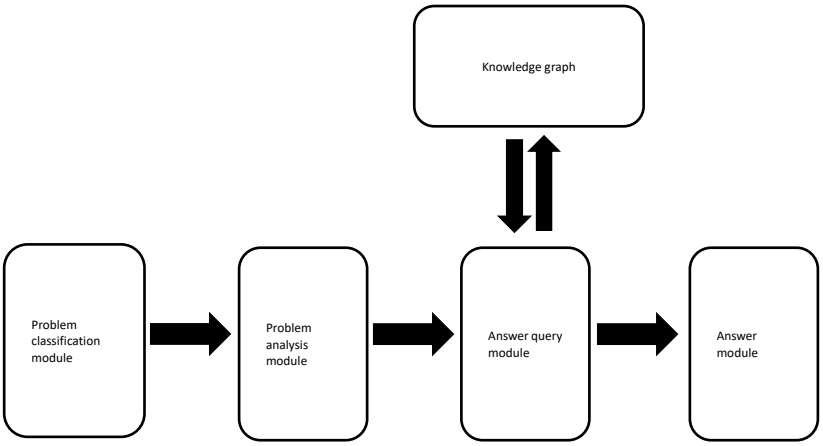

Figure 3: Flowchart of question and answer system design

3.3.1. Problem classification module design

The question classification module first filters the question sentences. After the user enters the question sentence, the Aho-Corasick (AC) matching algorithm is used to match the entity name that appears in the question sentence, and the corresponding entity type label is added to each entity name. Then manually create question feature words, such as prevention question feature words ['prevention', 'resistance', 'prevent', 'evade', 'avoid', 'Lest', 'Run away', ...] etc. Finally, according to the question feature words matched in the question sentence and the entity type detected in the question sentence, the question sentence type is deduced. For example, if you enter the question "What to eat for hypertension", the entity "hypertension" will be detected first, and its type is "disease", and then combined with the question feature word "eat" in the sentence, the question type can be deduced as "disease_do_food".

The AC algorithm is a multi-pattern matching algorithm, that is, multiple strings can be matched from the text at one time, and the target string can be located. The algorithm is based on a *Trie* pattern tree. It has the following characteristics: (1) the value of each edge on the tree represents a character; (2) the values of any two edges starting from the same node are different; (3) for any pattern string ***P***, it can be found a node ***v*** makes ***L(v)=P***, where ***L(v)*** represents the sequence of values of all edges on the path from the root node to node ***v***; (4) For any leaf node ***v'***, a pattern string ***P*** can be found to make ***L(v')=P***. For the

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

pattern string ***P***={he, she, hers, his}, the *Trie* tree representation is shown in Figure 4, where the circle represents the node, the black solid circle is each matched word node, the solid line is the hierarchical structure of the tree, and the dashed line is the *fail* pointer. The *fail* pointer is similar to the pointer in *next* array. When a character in the pattern string is found to be inconsistent with the character of a branch node in the current tree, it jumps to another branch node position pointed to by the *fail* pointer, and then performs string matching. The pseudo code is shown in Figure 5.

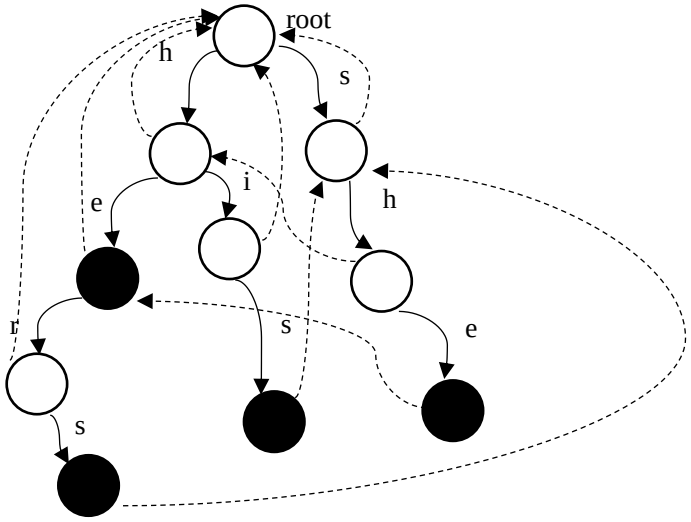

Figure 4: Trie tree

Input: string $x=a_1a_2 \ldots a_n$**, where** a_i **is the input character; goto function; failure function; output function.**
Output: the position of the keyword in the string x
begin
$state \leftarrow 0$
for $i \leftarrow 1$ **until** n **do**
begin
while $g(state, a_i) = fail$ **do** $state \leftarrow f(state)$
$state \leftarrow g(state, a_i)$
if $output(state) \neq empty$ **then**
begin
print i
print $output(state)$
end
end
end

Figure 5: String matching process

3.3.2. Problem analysis module design

The question analysis module is to convert the corresponding Cypher sentence for the different question type results generated in the previous section, which is convenient for answer query in Neo4j. For example, if the query "what to eat for hypertension", the question type is "disease_do_food", then the query sentence is converted to ["MATCH (m:Disease)-[r:do_eat]->(n:Food) where m.name = ' hypertension' return m.name, r.name, n.name"].

3.3.3. Answer query module

The answer query module is about to call the Cypher statement generated at the end of the previous step in Neo4j, and the returned result is output and returned through the manually set answer template.

3.3.4. Answer module

This module is the last part of the question and answer system, which integrates all the previous modules to complete the system construction.

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

4. Experimental analysis

After completing the system design and implementation, the function of the system was tested and investigated. The test was mainly to verify the correct rate of the question and answer system. A total of 200 chronic disease-related questions with similar semantics to the template question were prepared for this test.

In the end, 79% of the questions could return the correct answer. Analyze questions that did not return the correct answer for the following reasons. (1) Some keywords in the set questions did not match the template. (2) Unable to identify semantic problems beyond the scope of knowledge of the knowledge graph.

In order to compare the advantages and disadvantages of the question and answer system and manual diagnosis, this study cooperated with local community hospitals to convene a total of 30 volunteers with a history of chronic diseases to experience the difference between the two through offline promotion. Related comparisons are shown in Table 6. Repeat consultation rate refers to the proportion of reconsultation due to poor initial diagnosis due to special reasons. The evaluation score is to allow patients to score 1-100 points for each experience item. The higher the score, the better the diagnosis experience.

Table 6: Question and answer system and manual consultation survey form

	Average time per patient	Cost required	Repeat consultation rate	Evaluation score
Question and answer system	1.3min	Material cost	13.33%	83
Manual sitting	5.6min	Labor cost	6.67%	93

It can be seen from the statistical table that the question and answer system has a very high efficiency compared with manual consultation, which greatly reduces labor costs, but it is not as good as manual consultation in terms of repeated consultation rate and the evaluation score. Analyzing the reason may be that the patient's problem has not been completely resolved. Question and answer systems still have room for improvement in terms of semantic understanding and output answers.

The main deficiencies of this system are as follows. (1) The user's question is relatively single, and satisfactory answers cannot be returned to some questions with complex semantics. (2) Since each entity needs to be extracted separately and all saved, the constructed template is too large, which is not conducive to later maintenance and transplantation.

Based on the above shortcomings, the next improvement direction of this research is as follows. (1) Use deep learning methods to train the corpus, and introduce question and answer research on complex user questions. (2) Expand the knowledge source of the knowledge graph, and carry out research in combination with the knowledge fusion of the knowledge graph. (3) Use frameworks such as Flask to encapsulate the system to optimize the data query operation interface.

5. Conclusion

In this paper, the concepts and key technologies of the knowledge graph and question and answer system are demonstrated in detail, and a question and answer system based on the knowledge graph of chronic diseases is established in combination with actual needs. Although a solution is provided for this field, there are still some shortcomings in optimization, which is also the focus of the next step of research.

References

[1] Tsai C, Lai C, Chao H, et al. Big data analytics: a survey[J]. Journal of Big Data. 2015, 2(1).
[2] Chen M, Mao S, Liu Y. Big Data: A Survey[J]. Mobile Networks and Applications. 2014, 19(2): 171-209.
[3] L Qiao, L Yang, D Hong, et al. Knowledge Graph Construction Techniques[J]. Journal of Computer Research and Development. 2016, 53(3): 582-600.
[4] Wang Q, Mao Z, Wang B, et al. Knowledge Graph Embedding: A Survey of Approaches and Applications [J]. IEEE Transactions on Knowledge and Data Engineering. 2017, 29(12): 2724-2743.
[5] Sowa J F. Principles of Semantic Networks: Exploration in the Representation of Knowledge[J].

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

frame problem in artificial intelligence. 1991.
[6] Mao Xianling L X. A survey on question and answering systems. Journal of Frontiers of Computer Science and Technology, 2012, 6(3): 193 − 207[J]. 2012.
[7] Wang Zhiyue, Yu Qing, Wang Nan, et al. Research review of intelligent question and answer based on knowledge graph[J]. Computer Engineering and Applications. 2020, 56(23): 1-11.
[8] Li Wenling. Analysis of the status quo of chronic disease management models[J]. Medical Theory and Practice. 2018, 31(22): 3353-3354.
[9] Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019[J]. Lancet. 2020, 396(10258): 1204-1222.
[10] Hu Shilian, Wang Jing, Cheng Cui, et al. Epidemiological trend analysis of chronic diseases among Chinese residents[J]. Chinese Journal of Clinical Healthcare. 2020, 23(03): 289-294.
[11] Shib Sankar Dasgupta S N R P. HyTE: Hyperplane-based Temporally aware Knowledge Graph Embedding [J]. Association for Computational Linguistics. 2018: 2001-2011.
[12] Chen Y, Elenee Argentinis J D, Weber G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research[J]. Clinical Therapeutics. 2016, 38(4): 688-701.
[13] Yang Bo, Yang Meifang. A Survey of Knowledge Graph Research and Its Application in the Field of Risk Management[J]. Small Microcomputer System.: 1-13.
[14] Fabian M. Suchanek G K G W. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia[C]. 2007.
[15] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. ACM, 2008.
[16] Turing A M. Computing Machinery and Intelligence[J]. Mind. 1950, 59(236).
[17] Green B F, Alice K W, Chomsky, et al. Baseball: an automatic question-answerer[C]. 1961.
[18] Woods W A. LUNAR ROCKS IN NATURAL ENGLISH: EXPLORATIONS IN NATURAL LANGUAGE question and answer.[J]. 1977.
[19] Weizenbaum, Joseph. ELIZA-A Computer Program for the Study of Natural Language Communication between Man and Machines [J]. Communications of the ACM. 1983, 26(1): 23-28.
[20] Wilensky R, Chin D N, Luria M, et al. The Berkeley UNIX consultant project[M]. Springer, 2000, 49-94.
[21] Li Zhoujun, Li Shuihua. Overview of Web-based question and answer System [J]. Computer Science. 2017, 44(06): 1-7.
[22] Berant J, Chou A, Frostig R, et al. Semantic Parsing on {F}reebase from Question-Answer Pairs[Z]. Seattle, Washington, USA: 20131533-1544.
[23] Bast H, Haussmann E. More Accurate question and answer on Freebase[C]. 2015.
[24] Yin W, Yu M, Xiang B, et al. Simple question and answer by Attentive Convolutional Neural Network [J]. 2016.
[25] Golub D, He X. Character-Level question and answer with Attention [J]. 2016.
[26] Lukovnikov D, Fischer A, Lehmann J, et al. Neural Network-based question and answer over knowledge graph on Word and Character Level[C]. 2017.
[27] Bordes A, Weston J, Usunier N. Open question and answer with Weakly Supervised Embedding Models[C]. 2014.
[28] Bordes A, Usunier N, Chopra S, et al. Large-scale Simple question and answer with Memory Networks[J]. Computer Science. 2015.
[29] Gu K, Miller J, Liang P. Traversing knowledge graph in Vector Space[J]. Computer Science. 2015.
[30] Xe J, Silva O, Xed L, et al. From symbolic to sub-symbolic information in question classification [J]. Artificial Intelligence Review. 2011.
[31] Wen Xie, Zhang Yu, Liu Ting, et al. Chinese question classification based on syntactic structure analysis [J]. Journal of Chinese Information Processing. 2006, 020(002): 33-39.
[32] Liu L, Yu Z, Guo J, et al. Chinese Question Classification Based on Question Property Kernel[J]. International Journal of Machine Learning & Cybernetics. 2014, 5(5): 713-720.
[33] Yu Z, Su R, Li R, et al. Question classification based on co-training style semi-supervised learning [J]. Pattern Recognition Letters. 2010, 31(13): 1975-1980.
[34] Zhang D, Lee W S. Question Classification using Support Vector Machines[Z]. 2003.
[35] Kim Y. Convolutional Neural Networks for Sentence Classification [J]. Eprint Arxiv. 2014.
[36] Liu P, Qiu X, Huang X. Recurrent Neural Network for Text Classification with Multi-Task Learning [J]. AAAI Press. 2016.
[37] Yao L, Mao C, Luo Y. Graph Convolutional Networks for Text Classification [J]. 2018.
[38] Liu Liu, Wang Dongbo. Review of Named Entity Recognition Research [J]. Journal of Information. 2018, 037(003): 329-340.
[39] Koeling R. Chunking with Maximum Entropy Models [Z]. 2000.

ISSN 2616-5775 Vol. 4, Issue 4: 1-11, DOI: 10.25236/AJCIS.2021.040401

[40] Shi Xiaoxing, Wang Taijun, He Zhenya. The learning algorithm of the second-order hidden Markov model and its relationship with the first-order hidden Markov model [J]. Journal of Applied Sciences. 2001(01): 29-32.
[41] Chen Xiao, Liu Hui, Chen Yuquan. Recognition of Chinese organization name based on support vector machine method [J]. Application Research of Computers. 2008(02): 362-364.
[42] Sun Xiao, Sun Chongyuan, Ren Fuji. Biomedical named entity recognition based on deep conditional random fields[J]. Pattern Recognition and Artificial Intelligence. 2016, 29(11): 997-1008.
[43] Collobert R, Weston J, Bottou L E O, et al. Natural Language Processing (Almost) from Scratch [J]. Journal of Machine Learning Research. 2011, 12(76): 2493-2537.
[44] Huang Z, Wei X, Kai Y. Bidirectional LSTM-CRF Models for Sequence Tagging [J]. Computer Science. 2015.
[45] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. 2018.
[46] Souza F, Nogueira R, Lotufo R. Portuguese Named Entity Recognition using BERT-CRF[J]. 2019.
[47] Zhang W, Jiang S, Zhao S, et al. A BERT-BiLSTM-CRF Model for Chinese Electronic Medical Records Named Entity Recognition[C]. 2019.
[48] P Yang, W Dong. 2020, 46(04): 40-45.(In Chinese)
Yang Piao, Dong Wenyong. Chinese named entity recognition method based on BERT embedding [J]. Computer Engineering. 2020, 46(04): 40-45.
[49] Kumar S. A Survey of Deep Learning Methods for Relation Extraction [J]. 2017.
[50] Kambhatla, Nanda. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [J]. 2004: 22.
[51] Zelenko D, Aone C, Richardella A. Kernel Methods for Relation Extraction [J]. Journal of Machine Learning Research. 2003, 3(3): 1083-1106.
[52] Wang Chuandong, Xu Jiao, Zhang Yong. Overview of entity relationship extraction [J]. Computer Engineering and Applications. 2020, 56(12): 31-42.
[53] Brin S. Extracting Patterns and Relations from the World Wide Web [Z]. Berlin, Heidelberg: 1999172-183.
[54] Zhang Z. Weakly-Supervised Relation Classification for Information Extraction [Z]. New York, NY, USA: 2004581-588.
[55] Hasegawa T, Sekine S, Grishman R. Discovering Relations among Named Entities from Large Corpora [J]. Association for Computational Linguistics. 2004.
[56] Hassan H, Hassan A, Emam O. Unsupervised information extraction approach using graph mutual reinforcement[C]. 2006.
[57] Socher R, Huval B, Manning C D, et al. Semantic Compositionality through Recursive Matrix-Vector Spaces[C]. 2012.
[58] Lin Y, Shen S, Liu Z, et al. Neural Relation Extraction with Selective Attention over Instances[C]. 2016.
[59] Xu Y, Mou L, Ge L, et al. Classifying relations via long short term memory networks along shortest dependency paths[J]. Computer science. 2015.