

Development and Implementation of Intelligent Q&A System in Medical Domain

Yan Yan^{1, a}, Yuexin Li^{1, b}, Jiancheng Wen^{1, c} and Baohang Sun^{1, d}

¹Hong Kong Polytechnic University, Hong Kong, China.

^a19101145g@connect.polyu.hk, ^b19094903g@connect.polyu.hk,

^c19098145g@connect.polyu.hk, ^d19092969g@connect.polyu.hk

Abstract

The spread of COVID-19 in 2020 has become a global hot issue. Because of the high infectivity and the limited medical resources, people need to obtain the consultation service of medical service quickly and accurately. This paper focused on using natural language processing technology to build a medical knowledge graph and improve the key algorithms in the existing medical Q&A system.

Keywords

Q&A, NLP, Knowledge Graph, Data Mining, Medical System.

1. Introduction

With the improvement of the quality of people's living standard, people's health awareness has been constantly improved. On the other hand, with the development of Internet technology, the level of informatization and intelligence of medical services has been continuously improved. The interaction and timeliness of medical services are relatively strong. When users need medical services, they should not only find the answers accurately, but also quickly. Therefore, users often choose to use the general search engine to search the content they want to know. To some extent, general search engine can help people to get information quickly, but it also has some limitations, which are mainly manifested in: search engine search keywords only stay on the surface of language, do not really understand the user's intention, and cannot interact according to the search content; search engine search results data are mixed, it is difficult for users to distinguish useful information, once not used When it does, it can cause panic (such as looting drugs under an epidemic).

In this case, people urgently need a technology product to meet people's demands for high-quality medical services. The intelligent question answering system for medical field can solve these problems well. The intelligent question answering system is based on semantic analysis of the user's question content, intention recognition, search and match the answer according to certain rules and algorithms, and finally return the answer to the user. The intelligent Q & a system for medical field can not only provide accurate medical services to users, but also interact with users in time, which largely meets the needs of users for accurate and rapid access to medical information. Chinese style

However, the development of medical Q&A system at present is facing many difficulties due to its high professional knowledge requirements. The main obstacle is how to build a perfect and reliable corpus and provide training database. This paper will extract entities, relationships and attributes from a large number of structured and unstructured medical data and establish a scientific knowledge graph as the basis of question answering function. This paper will optimize the medical Q&A system by choosing the most accurate problem classification algorithm.

2. System Design

In this paper, the process of building a medical knowledge graph based Q & A system is divided into knowledge graph building stage and knowledge graph based Q & A stage. The system is divided into front-end Q & A, data query and back-end server-side management. This chapter will introduce the system structure and process in detail from the overall architecture, detailed design of each functional module, medical knowledge graph entity, relationship, attribute definition and knowledge scale.

2.1. Overall System Architecture

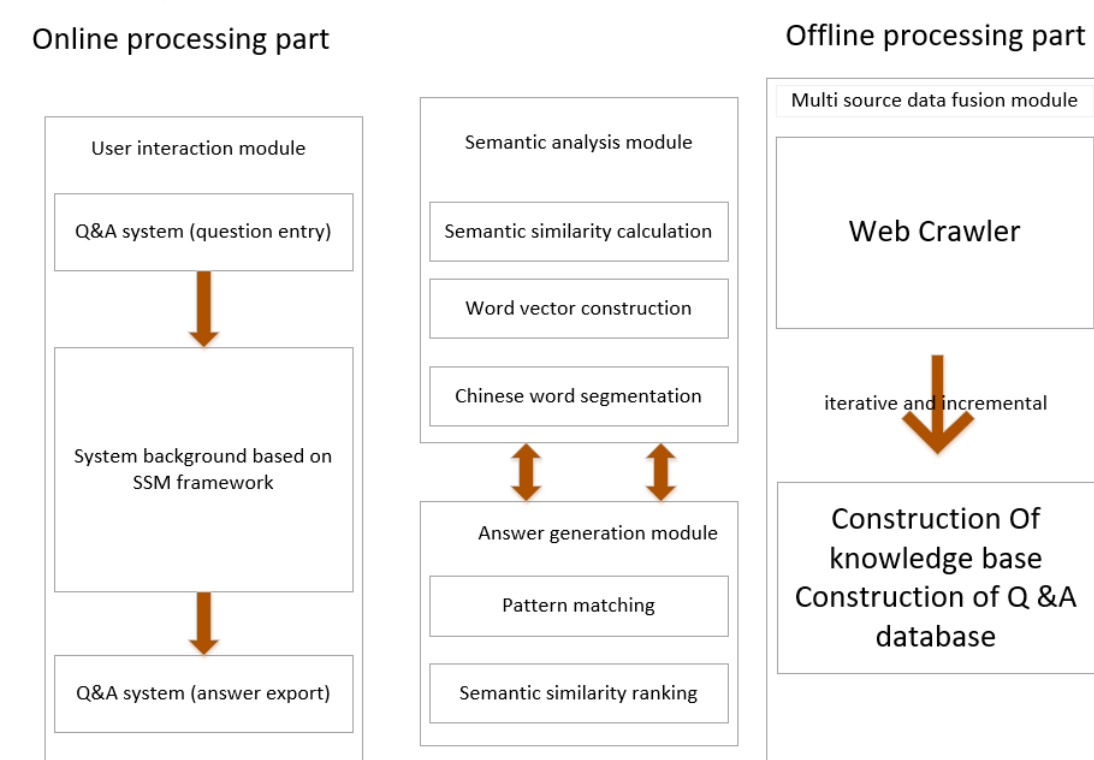


Figure 1. Expected system structure

The overall architecture of the system is shown in Figure 1. The overall construction process of the system is divided into four stages: Data crawling, data processing, medical knowledge graph storage and question answering system construction. The detailed design of the relevant process will be described in the next section. In the design of the system, the data interaction between the front end and the back end is completed based on the SSM framework. The front end includes question and answer page, data query display page and knowledge graph management page. The back end server is written in Python language and calls py2neo software package to realize the database query interface of neo4j graph and the relevant interface of model call. The data storage uses neo4j non relational graph Database to store entities, relationships and attributes in knowledge graph.

2.2. Function Module Design and Implementation

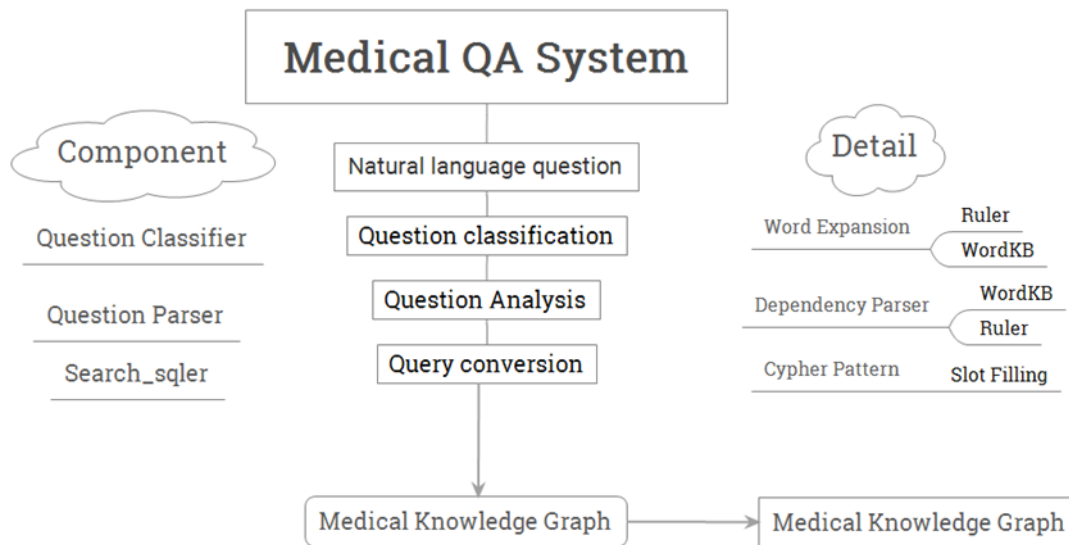


Figure 2. Expected functional structure

The general process of building the knowledge graph is to analyze the domain that the knowledge graph needs to be built for, for the open domain knowledge graph, data can be obtained by acquiring open knowledge; for the specific domain knowledge, field related data need to be found to obtain domain data, this process is called the data collection stage. After the collection of the original data of knowledge required by the knowledge graph, it is necessary to clean and format the collected original data, define the entity, relationship and attribute of the processed data, and store the processed data in the neo4j map database. This process is called the stage of data processing and knowledge graph building in this paper. After the construction of the knowledge graph, users can put forward questions in the front-end interface of the Q & A system. According to the questions raised by users, the system analyzes the questions, finds the knowledge base, and provides reasonable and accurate answers. This process is called the Q & A stage in this paper. Therefore, we divide the system into four modules: data collection and data processing, knowledge graph construction, question and answer query.

2.3. Entity Relationship Design of Medical Knowledge Graph

The question answering system based on knowledge graph needs knowledge graph as the data base of the system. Knowledge graph stores knowledge such as entity, relation and attribute. The medical knowledge graph constructed in this paper mainly involves 7 types of knowledge graph entity, 10 types of knowledge graph relationship and 8 types of knowledge graph attribute. The specific definition of each data in the medical knowledge graph constructed in this paper are shown in Table 1.

3. Algorithm and Key Technology

Through the design of Chapter 2, we find that in the process of data processing, we match the corpus information in the data through bilstm-crf, and generate the word vector with word2vec. In the design of Q & A module, we introduce the method of AC tree plus question template matching to better consider the character information in the question. Next, we will specifically introduce and analyze the principle of the algorithm.

Table 1. Knowledge Graph Definition Type

Question Type	Entity type	Entity relationship type
disease_symptom	Check	belongs_to
symptom_disease	Department	common_drug
disease_cause	Disease	do_eat
disease_acompany	Drug	drugs_of
disease_not_food	Food	need_check
disease_do_food	Producer	no_eat
food_not_disease	Symptom	recommand_drug
food_do_disease	/	recommand_eat
disease_drug	/	has_symptom
drug_disease	/	acompany_with
disease_check	/	/
check_disease	/	/
disease_prevent	/	/
disease_lasttime	/	/
disease_cureway	/	/
disease_cureprob	/	/
disease_easyget	/	/
disease_desc	/	/

3.1. Data Processing Stage

For entity detection, the general work is to first use ngram question words to search for entities whose names and questions have a common substring, and then use neural networks and other syntactic indicators (such as Edit Distance, longest continuous common subsequence (LCCS)) Together to capture similar matches between questions and entity names. We use a heuristic expansion method based on bi-directional LSTM-CRF network (BiLSTM-CRF) to obtain entity candidates, without the need to build additional neural networks, which can significantly reduce complexity and improve efficiency, while eliminating some Noise generated by unrelated entities of the same substring. For a given question, it can be regarded as a binary classification problem, which can predict whether each word in a sentence belongs to The mentioned entity. Then we use heuristic algorithm to modify the entity in the extended question without additional training of neural network. We will put the extracted structured data into the graph database Neo4j and display it visually

Then use a text vectorization tool based on Word2Vec represented by neural network, and use the tool to generate Chinese word vectors. Word2Vec is a toolkit for obtaining word vectors that Google launched open source in 2013. It is optimized on the basis of traditional neural

networks and other language models, including CBOW and Skip-Gram models. We chose CBOW (Continuous Bag-of-Words Model) to predict the occurrence probability of the current word at a certain position according to the words in the context of the input text. Its essence is to know the context information and estimate the language model of the current word. The schematic diagram is shown in Figure 3.

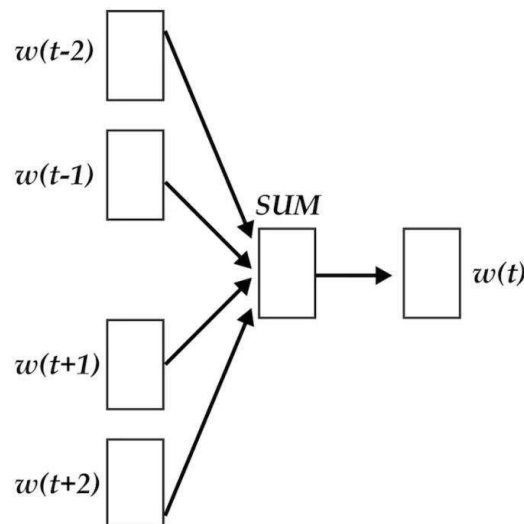


Figure 3. Schematic diagram of CBOW model

3.2. Q & A Stage

After the system accepts the user questions of the question and answer system, it needs to classify the questions and distinguish the specific question types of the questions.

First of all, we have to classify the questions. Here we build AC tree and keyword dictionary based on the structured data extracted in advance.

Then, for the user's question input, we will perform NER and record the corresponding relationship dictionary data = {'headache': ['disease', 'symptom']}

Then fill in the default cypher statement based on all question types and dictionaries, cyps = {}
[{"question_type": 'disease_desc', 'sql': ["MATCH (m: Disease) where m.name = 'headache' return m.name, m.desc "," MATCH (m: Disease) where m.name = 'influenza' return m.name, m.desc "]]}]

Finally, accept the question analysis result cyps, for each question type, link the database and get the answer of each query into the list final_answers and return the list.

After the model is built, a manually constructed labeled user question data set was used to test the effect of the model, and the F1 value reached 0.85.

4. Conclusion

The construction of question answering system based on medical knowledge map in this paper includes the construction of medical knowledge map and question answering module. The system completes the collection, processing and storage of medical data, defines the entity, relationship and attribute according to the existing data, constructs the medical knowledge map, and completes the construction of question answering system. In the stage of data collection and processing, we use Python language to write a crawler program to analyze and capture the data of a specific medical website, and store the captured original data in Mongo DB non-relational database; later, we use BiLSTM-CRF algorithm to extract a total of about 44000 entities in seven categories and about 30 Entities in ten categories Medical knowledge

map of 10000 relation data scale. In the construction stage of Q & a system, word2vec is used to train the word vector of Q & a data. The problem classification algorithm of AC tree + keyword matching is used to pre classify the user's problems, and the answer to the problem is found by combining with the graph knowledge base neo4j. The F1 value is 0.85 in the manual data set.

Acknowledgements

Thanks to my supervisor Dr. Li Jing for giving me a lot of teaching and help in my usual study. Under the influence of the teacher's serious and rigorous work attitude, I have continuously deepened and expanded my understanding of scientific research. The study of various theoretical knowledge points out the direction, and also lays the foundation for the realization of the system, and will also become a valuable asset that will benefit a lifetime. At the same time, I need to thank the teacher for my careful guidance and revision of my paper. Every time I discuss the paper with the teacher, I will always benefit a lot. May the teacher be in good health and work smoothly.

References

- [1] Dai Z, Li L, Xu W. CFO: Conditional Focused Neural Question Answering with Large-scale Knowledge Bases, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Berlin, Germany, August 7-12, 2016). Vol. 1, p. 800.
- [2] F. Miao, X. Wang, P. Zhang and L. Jin: Question-Answering System Based on the Knowledge Graph of Traditional Chinese Medicine, 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (Hangzhou, China, August 24-25, 2019) p. 264-267.
- [3] Mohammed S, Shi P, Lin J. Strong: Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (New Orleans, June 1-6, 2018). Vol 2, P.291.
- [4] Jens L, Robert I, Max J, et al: A large-scale, multilingual knowledge base extracted from Wikipedia, Semantic Web, Vol. 2, (2015), p.167-195.
- [5] Shvaiko P, Euzenat J: Ontology matching: state of the art and future challenges. IEEE Trans on Knowledge & Data Engineering, Vol. 1, (2013), p.158-176.
- [6] Information on <https://arxiv.org/abs/1901.08746>