Medical Knowledge Question Answering System Based on Knowledge Graph

Junwei Li(✉)

Computer Science and Technology, Gannan University of Science and Technology, Ganzhou 341000, Jiangxi, China
ljw687520@126.com

Abstract. Knowledge graph has broad application prospects in the field of medical question answering. The research first builds Neo4J knowledge graph based on Chinese OpenKG.cn structured medical common sense data, and then builds a question answering system based on the constructed knowledge graph. The steps include constructing a complete and usable graph from data collection, data representation, and implementing a general graph construction tool, and then using a pipelined question answering system model to build a medical commonsense question answering system, and finally look forward to the application prospects and industry development trends of question answering systems in different medical fields.

Keywords: Knowledge graph · Question answering system · Medical

1 Introduction

1.1 Background and Significance of the Study

NO issue in recent years has drawn as munch attention as the debate surrounding Knowledge Graph. In the context of today's technological advancement, knowledge graphs have become a digital resource to help various industries play a more effective role in their respective fields. Among the well-known general knowledge graphs are Google's "Knowledge Graph" [1], Sogou's "Knowledge Cube", YAGO [2] and DBpedia [3]. In the field of Chinese knowledge graph construction, the more famous one is Open.CN [4] built by OpenKG, a Chinese open knowledge graph consortium. In the medical field, medicine is one of the most widely used vertical fields for knowledge graphs, such as the Knowledge Graph of Chinese Medicine built by Shanghai Shuguang Hospital [5]. However, many medical knowledge is kept as unstructured data, especially many experts in the medical field mostly give their opinions on diagnosis and treatment based on the industry's own experience. Therefore, it is particularly important to construct knowledge graphs of structured data for the medical field.

However, there are many shortcomings in the current research on the construction of knowledge graphs in the medical field mentioned above. The research areas at this stage are too broad, and the variety of knowledge graphs for building common sense

G. Ali et al. (Eds.): ISEMSS 2022, ASSEHR 687, pp. 2268–2273, 2022.
https://doi.org/10.2991/978-2-494069-31-2_267

is not rich enough, and the impact caused by this will be a lack of effective results on the application level. On the other hand, there is a difference in the target audience for mapping application, as most of the current research is in the area of niche mapping for doctors, and the general knowledge mapping for patients and the general population is not comprehensive enough. In fact, this is where the current research needs to be expanded, as many data about medical segments are not comprehensive and transparent, and the development of knowledge graphs in the medical field is relatively slow at this stage, considering the specificity of the industry and the expertise of the practitioners in the field of knowledge graphs. Therefore, this thesis will focus on two main areas of research: 1) the acquisition of common-sense medical domain knowledge and as comprehensive a set of typical entities as possible, and the transformation of this unstructured knowledge into structured knowledge and the construction of a knowledge graph; 2) the implementation of a question and answer system for its application.

1.2 A Review of National and International Research

At the same time, building application scenarios based on knowledge graphs is also worth studying at present and in the future. As a typical application scenario of knowledge graphs [6], question and answer systems can facilitate the screening and retrieval of knowledge. This method extracts natural language and parses it into logical expressions [8]; and vector space modelling, which uses machine learning, deep learning and other methods to process entities and relationships in the knowledge graph to generate question and answer models, and is also used in different areas of medicine [9]. A wealth of structured knowledge is very useful for artificial intelligence applications. But it is still a challenge to integrate this knowledge into the computational framework of real-world applications. Some of the main downstream applications involved here are those in Natural Language Understanding (NLU), Recommendation Systems, and Question Answering.

2 Methods

2.1 Construction of Medical Knowledge Graphs

In the process of building the knowledge graph, there are several main steps: 1. Requirement analysis, to forecast the functions to be achieved by the graph and the expected results to be achieved by the question and answer system at a later stage. 2. Data collection, to obtain structured data from major medical or knowledge graph data websites and to collect questions at a later stage. 3. Build the code for building the graph, Structured data is data that can be represented and stored using a relational database and is expressed in a two-dimensional form. The general characteristics are that the data is in rows, a row of data represents information about an entity, and the attributes of each row of data are the same. Thanks to the fact that most of the data is structured data, using graph building tools to achieve Neo4j Knowledge Graph.

2.1.1 Access to Medical Knowledge

The entities and relationships in the medical field are widely distributed, most of them exist as unstructured data, and along with the emergence of new entities and relationships, the ability to obtain clear data in a relatively professional manner is a major issue to be considered at present. Therefore, it is more accurate and reliable to obtain the required entity-relationship data from professional databases. In this study, we mainly obtain data from several sources: SNOMED-CT, a local medical knowledge base, IBM Watson Health, LinkBase, a medical concept knowledge base, and OpenKG.cn, an open Chinese knowledge graph. Thanks to openKG, the Chinese open database, the study was able to obtain standard structured data from it more easily.

2.1.2 Medical Knowledge Representation

Knowledge representation is a process of structuring, symbolising and formalising natural language, and the common form of knowledge representation used in the field of knowledge mapping is the triad. The so-called triad is an ontology representation, in which Entity 1, Relation, Entity 2, as well as Concept, Attribute and Attribute Value are used as the basic representation of the triad. Entities are the most basic elements in the knowledge graph, while different inter-entity relationships exist between different entities. A concept is mainly a collection of entities that may share some common features, which are called attributes. There are also more diverse description languages for ontologies, and the more representative ones are RDF and RDF-S, OWL and DAML. Therefore, in building medical knowledge graphs, the use of ontology representations for medical knowledge can greatly improve the availability of data.

2.1.3 Knowledge Graph Construction

At this stage, there are two common storage solutions for graph structures: RDF storage and Graph Database. Among them, graph databases are more generic in their display and representation than RDF databases, and achieve graph data presentation of nodes, edges and attributes in graph structures. Therefore, this study uses Neo4j (http://neo4j.com/), a popular open source graph database in the field, for storing knowledge graphs, and also supports importing through various data.

In Neo4j, it provides many statements to support querying data as well as importing data, for example Cypher is one of the descriptive graph query languages to import data and query graph data. For larger data, Neo4j provides a tool called neo4j-import, which allows for relatively fast import of large numbers of nodes as entities and edges as relationships into the database. Thanks to the publicly available structured data provided by openKG, this study collated and imported the Neo4j graph database via the neo4j-import tool. Table1 illustrates some of the relational triples in the medical knowledge graph.

2.2 Based on a General Medical Knowledge Q&A System

This study runs a Chinese question and answer system based on the OpenKG open database question and answer system model, which provides a more convenient way for

Table1. Example of a partial entity representation

Type of entity	Type of entity relationship	Type of property
Check	common_drug	name
Department	belong_to	casuse
Disease	recommend_drug	prevent
Durg	recommend_eat	desc
Food	has_symptom	cured_prob
Producer	accompany_with	cure_way
Symptom	need_check	easy_get

users to find structured data on medical general knowledge. The question and answer system mainly consists of the following modules: 1. Classification of the question by Classifier 2. Parsing of the question 3. Conversion of the query statement of the database. The specific process is as follows.

Input questions: Enter the concept of entities of general characteristics in the dialog box, such as diseases, drugs, food and other representations related to Chinese medical general knowledge questions: "What is a cold?" "What should I eat when I have a cold?".

2.2.1 Pre-processing of Issues

In the medical field many named entities are identified using individual words as units, but the characters of some statements can have the effect of misinterpreting the problem parsing, thus reducing the accuracy of entity identification. Therefore pre-processing of the question for disambiguation is required.

2.2.2 Analysis of Problem Entities

When classifying questions for parsing, we need to collect the types of entities involved in the question and distinguish which ones and translate the identified text into query statements. This is done by using the driver module in Neo4j in Python to query the corresponding entities or attributes. Once the results have been returned, the Q&A system generates a natural language return output window that conforms to the language syntax.

3 Result

After the Q&A system was built, we manually designed 45 corpus questions with similar grammar to the template and evaluated the output to assess the performance of this knowledge graph Q&A system. In addition, we collected 51 scientific questions related to general medical knowledge from the medical Q&A website (http://www.drugs.com) for experimentation.

As a knowledge graph question and answer system built from structured data, when a question is entered, if the entity of the question is accurately identified by the system

and in the question template, the result is considered correctly identified after the system identifies the output.

The final experimental results show that 81% of the manually designed questions can be output accurately, and most of the questions can still be matched semantically with the same pattern after the question vector is matched. In contrast, only 17% of the questions collected on the medical Q&A website were answered relatively correctly. In terms of manual questions there is mainly more of a bias towards explanations of entities and approaches to common knowledge diseases, for example, "What is asthma?" type of general knowledge question. In cases where the semantics of entities are relatively well recognised, the accuracy improvement is greater. Conversely, for some general knowledge questions in specialist areas, similar questions collected by medical Q&A websites are less accurate in this Q&A system. For example, "Under what conditions is short-acting insulin stored at room temperature." is relatively incompetent when it comes to questions of this type that are detailed to specific areas of expertise.

The current Knowledge Graph Q&A system can be in a basic state of operation, in that it can identify relevant entities, relationships and attributes in the process of implementing a knowledge graph based on general medical knowledge, and return relevant outputs after being transformed into query statements. Therefore, there is much room for further research and extension. 1. to expand the richness of entities, attributes and relationships in the specialised field. 2. to introduce the form of voice question and answer in the Q&A system to improve the convenience of users. 3. to use relevant algorithms, machine learning and deep learning to analyse the questions in more depth.

4 Conclusion

This paper focuses on the construction of a question and answer system based on a knowledge graph of general medical knowledge, in which structured data is obtained from relevant database sites and a knowledge graph is constructed using Neo4j. The question and answer system was built to achieve the functional requirements of the standard process, and was able to achieve basic functionality in terms of entity recognition and analysis. In terms of experimental results, relatively simple standardised corpus questions can be identified accurately. However, the system is not as capable in sub-discipline and complex grammatical structures. The next research will focus on the use of machine learning and deep learning in question parsing to improve the accuracy of the question and answer system. There is also room for improvement in the acquisition and construction of graphs of entities, attributes and relationships of professional data. Finally, some creative question and answer interaction modes, i.e. natural language speech dialogue interaction, will be included. In the future, knowledge mapping industry participants such as knowledge mapping vendors, big data vendors, NLP vendors, Internet majors and information vendors will continue to deepen the development of industry knowledge mapping business from the perspective of strengthening technical strength and deepening industry cognition, combining their original business advantages. The business scenarios of knowledge mapping will also continue to be iterated, the boundaries of industry application scenarios will be broadened, and vertical application scenarios will be deepened and penetrated. The knowledge mapping ecology will also continue to

be built by the integration of regulatory guidance, supply side, demand side, investment side, universities and research institutes to gather the construction synergy and promote the growth and expansion of the industrial ecology.

References

1. Singhal A. Introducing the knowledge graph: things, not strings [EB/OL]. Official google blog, 2012. https://googleblog.blogspot.co.za/2012/05/introducing-knowledge-graph-things-not.html.
2. Amarilli A, Galárraga L, Preda N, et al. Recent Topics of Research around the YAGO Knowledge Base [M]// Web Technologies and Applications. Springer International Publishing, 2014: 1–12.
3. Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data [M]// The Semantic Web. Springer Berlin Heidelberg, 2007
4. Qi Guilin, Gao Huan, and Wu Tianxing, Advances in knowledge graph research [J]. Intelligence Engineering, 2017, 3(1): 4-25.
5. Construction and application of knowledge graphs in Chinese medicine [J]. Journal of Medical Informatics, 2016, 37 (4): 8-13.
6. Yuan Kaiqi, Deng Yang, Chen Daoyuan, Zhang Bing, Lei Kai. Advances in medical knowledge graph construction techniques and research [J]. Computer Application Research, 2018, 35(7): 1929-1936
7. Yao X, Durme B V. Information Extraction over Structured Data: Question Answering with Freebase [C]// Proc of Meeting of the Association for Computational Linguistics. 2014: 956-966.
8. Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs [C]. Proc of EMNLP, 2013
9. Cao Mingyu, Li Qingqing, Yang Zhihao, Wang Lei, Zhang Yin, Lin Hongfei, Wang Jian. Knowledge graph-based knowledge quiz system for primary liver cancer [J]. Chinese Journal of Informatics, 2019, 33(6): 88-93