Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

Advances in medical knowledge graph construction techniques and applications

Zhikang Zhang*

School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 065000, China
* **Corresponding author**: Email: Zhikang.ZHANG@university-365.com.

Abstract: Knowledge graph is a way to present knowledge relations using graphics, which includes two factors, nodes and edges, and can visualize and visualize all concepts and connections between multiple entities in real life. With the advantages of clear expression, flexible structure adjustment and high efficiency, knowledge graphs play an important value for explainable artificial intelligence. The theoretical content and application of knowledge graphs have been rapidly developed by high-tech means such as big data and machine learning, and the improvement of technology has helped knowledge graphs gain more attention in various fields and produce many research results. In this paper, the author focuses on the research of medical knowledge graphs and analyzes the research of scholars in detail.

Keywords: Medicine; Knowledge graph; Medical Information; Named Entity Recognition; Relationship Extraction.

1. Background of Medical Knowledge Map

For medical knowledge graph construction, the first task is to use knowledge fusion technology to change semi-structured or unstructured data into structured data, and then collect them into a database, and finally form a semantic search engine, medical question and answer system and decision support system with support. So far, more and more researchers have been using knowledge graphs for research and analysis, and scientific knowledge graph theory has been widely recognized and developed.

At present, the overall usage rate of medical data is still at a low level, and China has invested more efforts in the development of medical information technology, and institutions have huge scientific information, but it is difficult to share data among different institutions, and many data are only shallowly applied, lacking a deep level to advance the use. In the process of medical knowledge mapping, many high-quality training corpora need to be used, but at present, there are only a few publicly available Chinese medical corpora, which brings a great obstacle to the construction of medical mapping. Due to the special nature of medical knowledge, medical knowledge graphs still have some distance to meet medical needs. In this paper, we explore the overall situation of medical knowledge graphs through literature retrieval and visualization analysis of medical-related knowledge graphs.

2. Medical Knowledge Graph Construction

2.1. Named Entity Identification

The definition of named entity recognition was first mentioned in the MUC-6 conference in 1996, and its main function is to identify the various names and locations involved in the text. Its use in the medical field is mainly to identify the names of various diseases and drugs. The initial use of named entity recognition is often based on certain rules and lexicons, so that the names recognized are more accurate, but there are also defects such as difficulty in recall and rule construction.

Deep learning can automate the effective extraction of salient features and greatly reduce the waste of human resources. The BiLSTM-CRF model is the most prominent deep learning tool applied to named entity recognition today. By inputting the word vectors obtained through advance training into the model, the word features are extracted by the forward and backward LSTM layers, and then the label sequences are obtained in the CRF layer. The main problem of this model is the inability to apply global contextual information. Qingxia Zeng [1] improved this model by adding an attention mechanism, which was tested with CCKS2018 and CoNLL data, and the results proved that the addition of the attention mechanism could improve the accuracy. Meisun Chen [2] and other scholars, on the other hand, established a KNN-BERT-BiLSTM-CRF model, selected the content of communication with liver cancer patients for experiments, and launched named entity recognition with the help of transfer learning, which resulted in higher F1 values and only a few annotated words.

2.2. Relationship Extraction

The concept of relationship extraction was first mentioned in the MUC-7 conference in 1998, where three different relationships, namely Location of, Employee of, and Product of, were listed in detail, while the relationships involved in medicine include other complications caused by diseases, tests to confirm diseases, and so on. The original relationship extraction methods were mainly based on co-occurrence and rules. The co-occurrence-based methods have the advantage of simplicity and easy recall, but lack sufficient accuracy; the rule-based methods are the opposite.

Deep learning methods can also achieve relationship extraction between entities in the medical domain. Zhang [3] and other scholars performed relationship extraction of patients' cases by a mechanism combining bidirectional GRU and attention. Ding Long [4], on the other hand, used a BiGRU-CNN model constructed based on the attention mechanism to implement relationship extraction in cases, and found that the F1 value obtained by using this model was the

highest among many models. Qingqing Li [5] used a primary and secondary task model based on Attention to extract medical relationships, and this model can fully combine the relevance of different tasks to achieve an increase in extraction efficiency.

These ways of extracting relations have homogeneity, all of them are entity extraction first and then their relations, and this way may lead to errors in propagation, and the correlation of multiple tasks cannot be used effectively. Yangzi Mu [6] used the BiLSTM model to extract the entity relationships from patient cases to reach the task more smoothly. Luo Ling scholars [7] researched a new annotation method for extracting repetitive relations of medical texts, based on the Att-BiLSTM-CRF model to achieve the extraction of relations, and the results are more scientific and accurate than the traditional methods.

2.3. Entity Alignment

For example, "Parkinson's disease" can also be called "Parkinson's syndrome" and "PD", etc. The construction of medical knowledge graph often has multiple names for medical terms. The entity alignment task allows for the summarization of knowledge and the acquisition of higher quality knowledge content. The main task of entity alignment is to identify the same entities in many databases that refer to the real world, and then link them well to realize the effective combination of multiple sources of knowledge. Entity alignment is mainly accomplished with the help of pairwise entity alignment with similar properties and collective entity alignment that measures entity relationships. Pairwise entity alignment can use probabilistic models, machine learning methods, etc.; collective entity alignment mainly uses vector space models, similarity propagation methods, etc.

The approach based on knowledge representation learning, which can reasonably use the semantic relations contained in the graph to achieve high efficiency of entity alignment, has received more attention from scholars today. Qiannan Sun [9] used TransE algorithm to achieve entity and relationship integration, so as to complete the task of aligning the database entities in several respiratory departments of hospitals. Teng Fei [10] proposed to incorporate word root sets and rules in entity alignment based on learning and comprehensive consideration of the characteristics of medical knowledge to achieve high accuracy of the results. With the support of graph convolutional networks, the relevant relational and structural information is implemented to model the entity alignment task with the help of TransE to model the attribute information and then combine them to complete the entity alignment task to obtain better results.

3. Medical Knowledge Graph Application

3.1. Medical Information Search Engine

Past search engines in the medical field often had to implement searches, storage, etc. for a large number of relevant web pages, but just could not have accurate localization of user input words. The search engine based on medical knowledge graph can provide users with textual links of hyperlinks between web pages on the one hand, and cover multiple semantic relationships existing between multiple entities on the other.

The content of the knowledge graph to the previous search engine for innovation focuses on the expansion of the query, through the knowledge graph to retrieve the entities involved in the query content and the connectivity between the entities, and also to achieve the provision of more attribute content, so as to meet the user's query needs. Struck Adam and other scholars [11] proposed that medical knowledge graphs have become a common method for representing biomedical knowledge and is to achieve effective combination of information retrieval and UMLS to achieve query expansion, which is then widely applied to the medical field. Some scholars achieved query expansion by incorporating medical ontology MeSH into the search engine, involving entities of synonymy, proximity and other concepts as well as linkage, so that the efficiency of information retrieval can be effectively improved. Based on the medical oncology, LSA is used to automate the exploration of the semantic linkage of entities, such as the relationship between drugs leading to complications and the co-action relationship between multiple drugs, to complete the query expansion of entities and relationships. Our scientists constructed the medical language system in 2002 and built a medical knowledge map with more than 120,000 concepts, 600,000 terms, etc. Its main purpose is to add knowledge cards and knowledge maps to the content of the search system to realize the concretization of all concepts, and users independently select concepts to implement relevant searches. (See Figure 1 for details). There are already a number of relatively mature knowledge bases for the medical field abroad, such as DrugBank as a comprehensive, freely accessible bioinformatics and chemical information repository, which has contained 10,513 drug entries and 4,774 non-redundant protein sequence information as of November 2017, and is widely used by the pharmaceutical industry and medical education industry for simulated drug target discovery, drug docking screening, drug interaction prediction, and many other research areas.

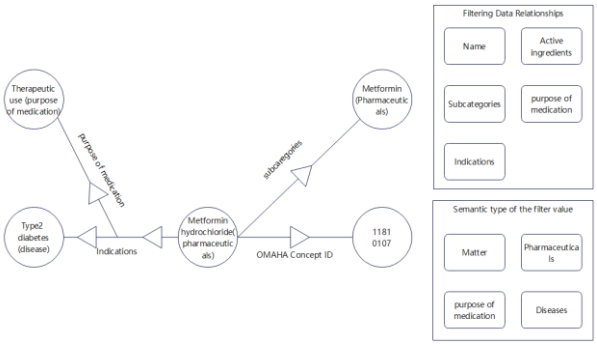

Figure 1. View Visual Data

3.2. Medical Q&A System

Medical question and answer systems are also a specific aspect of the application of knowledge graphs. The main means of constructing Q&A systems based on knowledge graphs today are: first, information-based extraction, where information about the question is integrated with resources in the database to select appropriate answers; second, semantic analysis-based approach, where linguistic questions are decomposed into logical expressions, and then the final answers are explored based on such expressions into the knowledge base; third, vector space modeling-based approach, where vector space is used to develop discourse on linguistic questions and entity relationships, and answers are constructed based on question models supported by machine learning and deep learning. The third approach is based on vector space modeling, which uses vector space to develop a

discourse on the relationship between linguistic questions and entities, and constructs a question model with the support of machine learning and deep learning to answer.

In the past, scholars have mainly explored medical question and answer systems in terms of information retrieval and summarization techniques, but after the emergence of the knowledge graph concept, scholars' research has shifted to knowledge graph-based question and answer systems. Fecho Karamarie et al [12] introduced the Biomedical Map of Evidence (BMEG), a graphical database and query engine for cancer biology discovery and analysis. BMEG is unique from other biological data maps in that sample-level molecular and clinical information is correlated with a reference knowledge base. It combines gene expression and mutation data with drug response experiments, pathway information databases, and literature-derived associations. Another study indicated that after synthesizing the two knowledge bases, UMLS and WordNet the ten major dimensions of medical problems are summarized, and the logical order of the problems is realized with the help of natural language processing technology, and then the results are searched in the huge knowledge base. After researching the medical ontology-based Q&A system, medical oncology, medical-related knowledge, and NLP technology are integrated to automate the Q&A system. Another scholar cooperates with Shanghai Shuguang Hospital to build a knowledge map of Chinese medicine covering many aspects such as diseases, symptoms, and herbal medicines, and then realizes problem solving and prescription prescribing based on this map, which can solve more problems using word separation, templates anyway and matching, etc. It can also transform the relevant information material into an appropriate way to provide a basis for prescription prescribing based on patient-specific information uploaded by doctors.

3.3. Medical Decision Support System

With the aid of knowledge mapping, the medical decision system can make intelligent treatment plans from the recorded patient symptoms and laboratory results, and also analyze the plans made by the doctors to make up for the shortcomings and avoid the problems of treatment errors. (See Figure 2 for details).

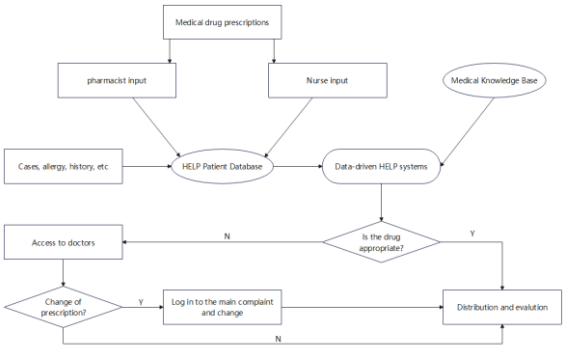

Figure 2. Knowledge graph assistance

ODDIN, an ontology-driven medical diagnosis system based on logical reasoning and probabilistic optimization, has been independently developed. The system as a whole consists of two knowledge bases: logical rules and medical oncology, which includes several rule contents and medical resources, allowing ontology to present a variety of states, RDF, RDFS, OWL is all possible, and diagnosis is expanded based on Bayesian theory. A study shows that independent research and development has built the iOSC3 system, which can implement automated monitoring and diagnosis for acute heart disease patients, and it can develop scientific and reasonable treatment plans for patients' specific conditions, and the knowledge base as a whole is composed of OWL ontology and SWRL rules together. The use of knowledge graphs in medical decision making can be considered as a major part of current research, but in practice there are two shortcomings, namely, there is no sound knowledge graph for general medicine and the accuracy of medical decision making. For example, IBM's Watson Health provides decision help for two diseases, oncology and cancer, with a large knowledge base and a high level of cognitive computing, enabling primary care physicians to develop targeted treatment plans more quickly. Medical decisions have a direct impact on the physical and mental health of the people who use them, and the use of AI technology in the medical decision-making process can achieve more accurate and reliable results. The current application of knowledge graphs in medical decision making is only of secondary value.

4. Conclusion

Knowledge graphs have stronger value in semantic processing and can be considered as an expansion of semantic web and knowledge base. Due to the continuous improvement of artificial intelligence technology level, it promotes faster construction of medical knowledge graph as well as more accurate inference of knowledge and reduces construction cost, which brings development opportunities as well as obstacles for medical industry. The future research of medical knowledge mapping should solve the problem of inability to share data among institutions and use crowdsourcing technology to obtain more Chinese medical annotated corpus; professional institutions also need to actively compile dictionaries to provide more support for medical knowledge mapping construction. By using artificial intelligence technology to build an automated, closed-loop system capable of autonomous learning, the construction cost of medical knowledge map can be reduced.

References

[1] Zeng QH, Xiong WP, Du JQ, et al.(2021). Combining self-attentive BiLSTM CRF for named entity recognition in electronic medical records, 38(3):159-162,242.

[2] Chen MS, Xia CX.(2019) .A study of named entity recognition for online questions from liver cancer patients:a migration learning-based approach, 3(12):61-69.

[3] Zhang ZC, Zhou D, Zhang RF, et al.(2020). Fusion of bidirectional GRU and attention mechanism for medical entity relationship recognition, 46(6):296-302.

[4] Ding L.(2020) .Research on information extraction technology for electronic medical records, South China University.

[5] Li QQ, Yang ZH, Luo L, et al.(2019) .Multi-task learning-based relationship extraction for biomedical entities[J]. Chinese Journal of Informatics,2019,33(8):84-92.

[6] Mu YZ.(2018). Research on entity recognition and entity relationship extraction of Chinese electronic medical records based on semi-supervised learning, Hainan University.

[7] Luo L.(2.019) .Research on some key technologies for biomedical text mining, Dalian University of Technology.

[8] Sun QN.(2019) .Knowledge extraction and alignment for respiratory diseases, Harbin Institute of Technology.

[9] Teng F, Zhong W, Xu Q, et al.(2020) .A medical knowledge graph entity alignment method based on representation learning,CN111309930A[P].

[10] Cheng R. (2020).Research and Application of Entity Alignment Method for Chinese Medical Knowledge Graph, Beijing University of Posts and Telecommunications, 17.

[11] Struck A, Walsh B, Buchanan A, Lee J A, Spangler R, Stuart J M, Ellrott K.(2020). Exploring Integrative Analysis Using the BioMedical Evidence Graph. 4:147-149.

[12] Fecho K, Balhoff J, Bizon C, Byrd W E, Hang S, Koslicki D, Rensi S E, Schmitt P L, Wawer M J, Williams M, Ahalt S C. (2021). Application of MCAT questions as a testing tool and evaluation metric for knowledge graph-based reasoning systems. 14(5):1719-1724.