Intelligent Computing Technology and Automation
Z. Hou (Ed.)

doi:10.3233/ATDE231193

Research on Named Entity Recognition and Knowledge Graph Construction Based on BERT-BiLSTM-CRF

Zhanlin LI [a,b], Hui YAN[c,1], Yunxin LONG[d], Yaqi SHI [a,b], Huangbin GUO[b], Ping YU[b]

[a]*Jilin Institute of Chemical Technology, Jilin, 132022,China*

[b] *Jilin Province S&T Innovation Center for Physical Simulation and Security of Water Resources and Electric Power Engineering, Changchun Institute of Technology, Changchun, 130012, China*

[c] *Suqian University, Jiangsu, 223800, China*

[d]*College of Traditional Chinese Medicine, Changchun University of Chinese Medicine, Changchun, 130117, China*

Abstract. With the continuous improvement and development of question answering systems, combining them with knowledge graphs in relevant professional fields can relatively improve their practicality. Most existing knowledge graph models are general knowledge graphs with wide coverage, but their quality is poor, leading to issues such as loose data and low coverage, resulting in low coverage of professional knowledge matching in the Q&A process and often unsatisfactory answers. To address this drawback, this article proposes a named entity recognition method based on the BERT-BiLSTM-CRF model. This method can fully combine BERT's context aware embedding, BiLSTM's sequence modeling ability, and CRF's label dependency modeling to improve the accuracy of identifying and annotating entities in text. Comparative experiments with different models have shown that the BERT-BiLSTM-CRF model outperforms the BERT-BiLSTM and BERT-CRF models in terms of accuracy. Therefore, the combination of knowledge graph and BERT-BiLSTM-CRF model applied to question answering systems can greatly improve the accuracy of knowledge extraction, make question answering scenarios more anthropomorphic, and reduce the probability of non answering situations.

Keywords. Knowledge graph; Q&A system; Named entity recognition; BERT BiLSTM CRF model; Annotate Entities

1.Introduction

With the rapid development of related technologies, knowledge graph has become an important research hotspot in recent years. It is essentially a vast knowledge base that links knowledge in the form of a graph structure, using visualization technology to describe knowledge resources and their carriers, and can intuitively mine, analyze, construct, and display knowledge and its connections [1-2]. The vigorous development of knowledge graph is attributed to the concept of knowledge graph officially proposed

[1] Corresponding Author: Hui Yan; School of Information Engineering, Suqian University, Suqian, 223800, China; 17151@squ.edu.cn

by Google in 2012, and its popularity is attributed to its excellent human-machine win-win ability [3].

Q&A system is an advanced form of information retrieval system that can answer users' questions in accurate and concise natural language. A knowledge graph based question answering system can understand natural language questions input by users, and then find accurate answers through entities, relationships, etc. in the knowledge graph and return them to users, thereby improving the efficiency of information search. In recent years, knowledge graph based question answering systems have become a hot research and application topic in academia and industry[4].

However, most of the knowledge graph models constructed by existing technologies are general knowledge graphs with wide coverage but poor quality, which can lead to loose data and low coverage, resulting in low coverage of professional knowledge matching during the Q&A process. Many times we don't get satisfactory answers. Therefore, this article will take the medical consultation problem graph as an example and implement precise named entity recognition in user questioning based on the BERT-BiLSTM-CRF model, achieving intelligent question answering in professional fields. Figure 1 shows the system architecture of a common knowledge graph based question answering robot.

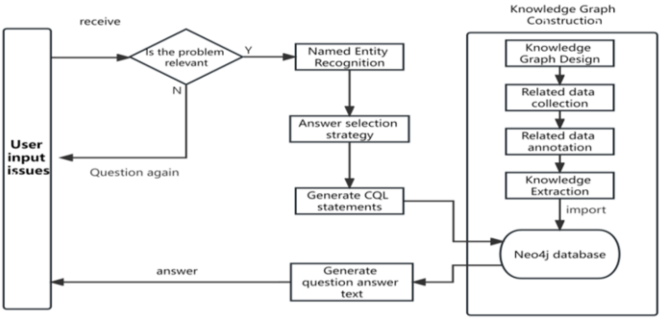

Figure 1. System architecture diagram of common knowledge graph based question answering robots

2.Construction of a knowledge graph based on medical related issues

2.1 Knowledge Graph Data Processing

1) **Data Sources**

The quality of question answering systems is closely related to the quality of knowledge graphs, and the most important factor affecting the quality of knowledge graphs is the quality of data sources. Through the comparison of multiple data sources, this paper selects the open source data of Professor Liu Huanyong from the Software Research Institute of the Chinese Academy of Sciences on github, and selects semi-structured json medical data as the data set of this knowledge map.

2) **Knowledge Extraction**

The knowledge graph is essentially a vast knowledge base, where knowledge is connected in the form of a graph, consisting of nodes and edges, representing entities, and the edges connecting entities are relationships. Entities are the basic units of the knowledge graph, carrying important information in the text. Different relationships connect independent entities to form a knowledge graph. The data layer of the knowledge graph is mainly composed of a series of facts, presented in the form of entity relationship entity and entity attribute attribute value triplets [5].

3) **Data Storage**

Generally speaking, there is no unified standard for knowledge storage. There are currently three main ways in the industry to store knowledge. The first type is RDF storage in the form of triplets; Secondly, traditional relational database storage; The third one is graph database storage [6]. The most commonly used method currently is graphic database storage. This article uses the current mainstream Neo4j graph database as the storage database for knowledge data. As a graph database, Neo4j has the advantages of powerful graph structure, fast query performance, flexibility, and scalability. Import the triplets obtained from knowledge extraction into the Neo4j graph database to obtain the knowledge graph corresponding to the semi-structured data [7].

2.2 Construction of hierarchical relationships in knowledge graphs

The construction methods of knowledge graphs vary depending on the type of data. It is mainly divided into two schemes: bottom-up and top-down. Most unstructured data is built through a bottom-up architecture. The bottom-up construction method is based on the process of extracting data from triples, and is generally used for structured information data. This requires defining entities and relationships in the data first, and then integrating the data into a knowledge base. This construction method is usually suitable for knowledge structure in the industrial field, where the content and organization of the data are relatively easy to find. This article uses a top-down approach to construct a knowledge graph [8].

The entity types of the knowledge graph mainly include diagnostic examination items, medical subjects, diseases, drugs, food, sales of drugs, disease symptoms, etc.

The entity type table is shown in Table 1.

Table 1. Entity Type Table

entity type	The meaning of Chinese	Entity quantity	give an example
Check	Diagnostic examination items	3,353	Bronchchiography; arthroscopy
Department	Medical subjects	54	Plastic surgery department; burn department
Disease	disease	8,807	Thromboangiitis obliterans; aneurysm of the descending thoracic aorta
Drug	drug	3,828	Jingwanhong hemorrhoid cream; brinzamide eye drops
Food	food	4,870	Tomato and beef ball soup; mutton with bamboo shoots
Producer	On sale of drugs	17,201	Pharmaceutical penicillin V potassium tablet; Qingyang dexamethasone acetate tablet
Symptom	Disease symptoms	5,998	Hypertrophic breast tissue; hemorrhage in the deep brain parenchyma

The entity relationship table is shown in Table 2.

Table 2. The entity relationship table

Type of entity relationship	The meaning of Chinese	Relationship quantity	give an example
belongs_to	belong to	8,844	<Gynecology, belong to, obstetrics and gynecology>
common_drug	Commonly used drugs for diseases	14,649	<Yang strong, commonly used, phentolamine mesylate dispersion tablet>
do_eat	Food is appropriate for diseases	22,238	<Thoracic vertebra fracture, appropriate to eat, black fish>
drugs_of	Drugs are on sale	17,315	<Penicillin V potassium tablets, on sale, all pharmaceutical pharmaceutical penicillin V potassium tablets>
need_check	Examination required for disease	39,422	<Unilateral emphysema, required examination, bronchography>
no_eat	Avoid food for disease	22,247	<Lip disease, avoid eating, almond>
recommand_drug	Drug recommended for disease	59,467	<Mixed hemorrhoids, recommended medication, Jingwanhong hemorrhoid cream>
recommand_eat	Disease Recommended Recipes	40,221	<Hydrocele, recommended recipes, tomato and beef ball soup>
has_symptom	Disease symptoms	5,998	<Early breast cancer, disease symptoms, breast tissue hypertrophy>
acompany_with	Disease concurrent disease	12,029	<Inclosed closure of communicating venous valve of lower limbs, concurrent disease, thromboangiitis obliterans>

2.3 Import of knowledge graph data

In this case, the above data is successfully imported into neo4j database by connecting Python to neo4j database. The screenshot of neo4j database is shown in Figure 2 and Figure 3.

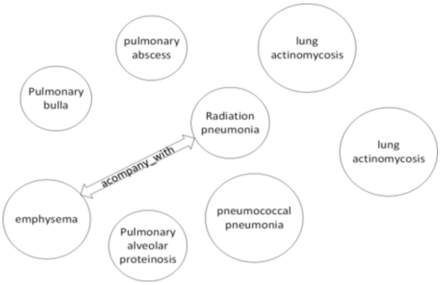

Figure 2. Disease screenshot of entity part

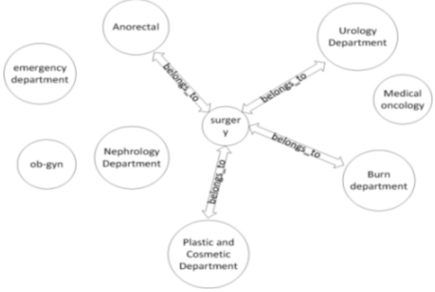

Figure 3. screenshot of Department entity part

3.Implementation and verification of the named entity recognition model

3.1 Entity extraction based on BERT-BiLSTM-CRF

LSTM is an RNN variant commonly used for sequence modeling, which is used to solve the problems of gradient vanishing and gradient explosion in traditional RNNs. It controls the flow of information by introducing gating units such as forgetting gates, input gates, and output gates, effectively capturing and transmitting long-term dependencies in the sequence. BiLSTM has two directional LSTM units, one for forward processing of sequence data and the other for reverse processing. Forward LSTM processes data step by step from the beginning to the end of the sequence, while reverse LSTM processes data step by step from the end of the sequence to the beginning. In this way, BiLSTM can capture both past and future contextual information, rather than just current moment information [9]. The main advantage of BiLSTM is its ability to effectively capture contextual information for each time step in the sequence. This is very important for natural language processing tasks, as the meaning of words usually depends on the words around them [10]. The structure diagram of BiLSTM is shown in Figure 4.

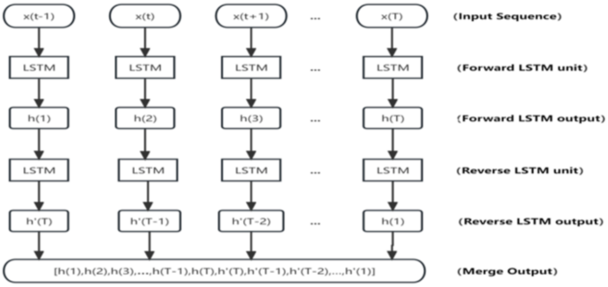

Figure 4. BiLSTM structure diagram

The LSTM unit consists of three key parts: input gate, forgetting gate and output gate. The input gate determines which parts of the current input need to be updated or retained. It calculates a value between 0 and 1 based on the current input and previously hidden states, which represents the importance of each input element.

$$\mathrm{i}_t = \delta(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{1}$$

The forgetting gate determines which information needs to be discarded from the cellular state. Similar to the input gate, it calculates a value between 0 and 1 using the current input and previously hidden states, representing the importance of each cell state element.

$$\mathrm{f}_t = \delta(W_\mathrm{f} \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

Cell state is the core component of LSTM used to store and transmit information. It is updated with control of the input and forgetting doors to determine which information needs to be retained.

Candidate cell stateS:

$$\tilde{C}_t = \tanh(W_\mathrm{c} \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

Cell state update: $C_t = \mathrm{f}_\mathrm{t} \cdot c_{t-1} + i_t \cdot \tilde{C}_t$ (4)

The output gate calculates a value between 0 and 1 based on the current input and the previously hidden state. It determines which contents in the cell state will be exported to the next time step.

$$O_t = \delta(W_\mathrm{o} \cdot [h_{t-1}, x_t] + b_\mathrm{o}) \tag{5}$$

wherein，i_t，f_t，$\tilde{\mathrm{C}}_\mathrm{t}$，$\mathrm{C}_\mathrm{t}$，$\mathrm{Q}_\mathrm{t}$ Represents the values of input gate, forgetting gate, candidate cell state, cell update state, and output gate at time t, respectively. W_t，

W_f , W_c , W_o The weight matrices represent input gates, forgetting gates, candidate cell states, cell update states, and output gates, respectively. b_i , b_f , b_c , b_o Representing the bias of input gate, forgetting gate, candidate cell state, cell update state, and output gate, respectively.

BERT is a pre training model proposed by Google in 2018, with the core idea of utilizing large-scale text data for pre training and then fine-tuning on specific tasks to achieve excellent performance in various natural language processing tasks. BERT implements bidirectional context modeling, which means it can consider both the left and right contexts of a word, rather than only considering the left or right contexts like traditional language models such as recurrent neural networks. This enables BERT to better understand the meaning of words in sentences, thereby improving the performance of various NLP tasks [11].

CRF is a probabilistic graph model, widely used in natural language processing and computer vision, especially in sequence annotation, named entity recognition, word segmentation, part of speech annotation and other tasks. CRF annotate or classifies by modeling the conditional dependence between individual elements in the sequence data. The CRF can be viewed as a discriminative model that models the conditional probability between a given input sequence and an output sequence. Compared to hidden Markov models, CRF can better model the dependencies between labels, because CRF considers the context information of the entire sequence when calculating the conditional probabilities[12].

This paper uses the BERT-BiLSTM-CRF method, and the BERT-BiLSTM-CRF is a model architecture for named entity recognition tasks, which combines multiple deep learning components to improve the performance of entity extraction. The context representation capability of BERT, sequence modeling capability of BiLSTM and label dependency modeling capability of CRF were fully utilized to achieve higher performance in the named entity recognition task. This model architecture is useful in addressing entity recognition problems in natural language text, especially for handling complex entity-nested and context-sensitive situations.

3.2 Analysis of model comparison results

To verify the high performance of the model, we used the OntoNotes 5.0 dataset for model contrast validation. In terms of model selection, BERT model and BERT-BiLSTM model, BERT-CRF model and BERT-BiLSTM-CRF model were selected for two sets of comparative experiments. In the comparison experiment of BERT model and BERT-BiLSTM model, the accuracy of BERT-BiLSTM model is slightly improved compared with BERT model, but the effect is not particularly obvious; in the comparison experiment of BERT-CRF model and BERT-BiLSTM-CRF model, the accuracy of BERT-BiLSTM-CRF model is correspondingly improved compared with BERT-CRF model. The model comparison data are shown in Table 3.

Table 3. Model comparison results

model	accuracy	P1
BERT	73.48	75.13
BERT-BiLSTM	74.24	75.49
BERT-CRF	73.76	75.54
BERT-BiLSTM-CRF	75.81	76.31

Among them, P1 was used as the criterion for the comparative analysis. Therefore, this study adopts the BERT-BiLSTM-CRF model to fully use the context representation ability of BERT, the sequence modeling ability of BiLSTM and the tag-dependent modeling ability of CRF to complete the named entity identification more reliably.

3.3 Implementation of the question and answer system

The intelligent question and answer system cited in this paper mainly allows users to ask questions about common diseases and their symptoms, symptomatic drugs, diagnostic examination items, and food that should be avoided before medical treatment freely, and the system can answer them accordingly. That is, users use natural language input to ask questions about diseases, the system analyzes and processes them, and then returns the accurate answer to the questions to the user. The process architecture of the intelligent question answering system is shown in Figure 5.

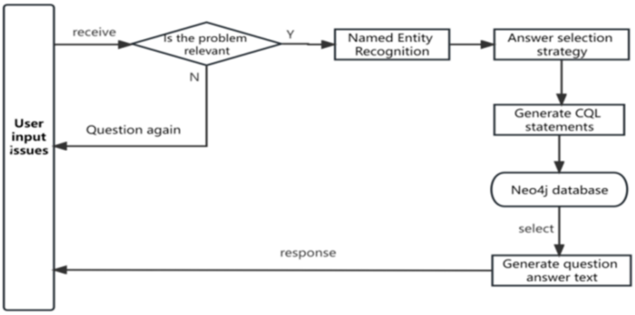

Figure 5. Flow architecture diagram of the intelligent question and answering system

4.Conclusion

Most of the knowledge graph models constructed in existing techniques are general knowledge graphs, with wide coverage but poor quality. They are prone to problems with loose data and low coverage. As a result, the coverage rate of professional knowledge matching in the question and answer process is low, and in many times, satisfactory answers cannot be obtained. This paper takes the medical knowledge graph as an example, uses the open source semi-structured data to build the knowledge graph of the proprietary domain, and uses the deep learning algorithm to identify the entity and finally generates the common sense answer text, and realizes the intelligent question and answer based on the proprietary domain knowledge graph. Compared with the traditional question answering system, this method has concentrated data distribution and high matching coverage of professional knowledge, which improves the accuracy of question answering in professional fields.

Acknowledgments

The authors gratefully acknowledge the funding of this study by Jilin Province Science and Technology (S&T) Department under Grant 20220505004ZP and 20190902011TC, by Ministry of Education China University Industry-University-Research Innovation Fund (2020ITA05035).

References

[1] Wang Xin, Zou Lei, Wang Chaokun, etc. Summary of Knowledge Graph Data Management studies [J].Journal of Software, 2019,30 (07): 2139-2174.

[2] Xu Lulu, Yang Jiale, Kangle Le. Thematic drift and future outlook of artificial intelligence technology in the field of medical information —— Based on the text of this Medical Information Journal [J]. Modern Intelligence, 2022,42 (10): 163-176.

[3] Zhang Yongliang. Research on the intelligent question-answering system of apple diseases and insect pests based on knowledge graph [D]. Northwest A & F University, 2022.

[4] Li Fei. Research and implementation of a question-answering system based on a knowledge graph [D]. Nanjing University of Posts and Telecommunications, 2022.

[5] Tang Xiaobo, Liu Zhiyuan. Study on the joint extraction of text sequence annotation and entity relationship in the financial field [J]. Intelligence Science, 2021,39(05):3-11.DOI:10.13833/j.issn.1007-7634.2021.05.001.

[6]Broscheit S, Ruffinelli D, Kochsiek A, et al.LibKGE-A knowledge graph embedding library for reproducible research[C].Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.2020: 165-174.

[7]Tellería C, Ilarri S, Sánchez C.Text Mining of Medical Documents in Spanish: Semantic Annotation and Detection of Recommendations[C]//WEBIST.2020: 197-208.

[8] Wang Zhiyue, Yu Qing, Wang Nan, etc. Summary of intelligent question-answering research based on the knowledge graph [J]. Computer Engineering and Application, 2020,56 (23): 1-11.

[9]Krisnadhi A.Challenges,Techniques,and Trends of Simple Knowledge Graph Question Answering: A Survey[J].Information, 2021, 12(7):271-303.

[10] Jiang Xiang, Ma Jianxia, Yuan Hui. Named entity identification in the field of ecological governance technology based on the BiLSTM-IDCNN-CRF model [J]. Computer Applications and Software, 2021,38 (03): 134-141

[11] Zhang Qiang. Research and application of key technologies of question answering system based on knowledge graph [D].Shenyang. University of Chinese Academy of Sciences (Shenyang Institute of Computing Technology, Chinese Academy of Sciences).2021.

[12] Wang Hailiang, Li Zhuhuan, Lin Xuming. Intelligent question answering and deep learning [M]. Beijing: Press of Electronic Industry, 2019:3.