Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

Intelligent question answering system for traditional Chinese medicine based on BSG deep learning model: taking prescription and Chinese materia medica as examples

LI Ran[a], REN Gao[a], YAN Junfeng[a], ZOU Beiji[a, b], LIU Qingping[a*]

a. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

b. School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China

ARTICLE INFO

Article history
Received 13 December 2023
Accepted 01 March 2024
Available online 25 March 2024

Keywords
Traditional Chinese medicine (TCM)
Deep learning
Knowledge graph
Intelligent question answering system
BERT+Slot-Gated (BSG) model

ABSTRACT

Objective To construct a traditional Chinese medicine (TCM) knowledge base using knowledge graph based on deep learning methods, and to explore the application of joint models in intelligent question answering systems for TCM.

Methods Textbooks *Prescriptions of Chinese Materia Medica* and *Chinese Materia Medica* were applied to construct a comprehensive knowledge graph serving as the foundation for the intelligent question answering system. In the study, a BERT+Slot-Gated (BSG) deep learning model was applied for the identification of TCM entities and question intentions presented by users in their questions. Answers retrieved from the knowledge graph based on the identified entities and intentions were then returned to the user. The Flask framework and BSG model were utilized to develop the intelligent question answering system of TCM.

Results A TCM knowledge map encompassing 3 149 entities and 6 891 relational triples based on the prescriptions and Chinese materia medica was drawn. In the question answering test assisted by a question corpus, the F1 value for recognizing entities when answering 20 types of TCM questions was 0.996 9, and the accuracy rate for identifying intentions was 99.75%. This indicates that the system is both feasible and practical. Users can interact with the system through the WeChat Official Account platform.

Conclusion The BSG model proposed in this paper achieved good results in experiments by increasing the vector dimension, indicating the effectiveness of the joint model method and providing new research ideas for the implementation of intelligent question answering systems in TCM.

1 Introduction

As medical informatization continues to progress in China, there is a growing research emphasis on technological innovation and the widespread dissemination of traditional Chinese medicine (TCM) knowledge. However, how to visualize and intelligently retrieve complex information in the field of TCM remains a hard problem yet to be solved [1]. In recent years, massive medical information on the Internet has been utilized by researchers for

*Corresponding author: LIU Qingping, E-mail: liuliu@hnucm.edu.cn.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2024.04.006

Citation: LI R, REN G, YAN JF, et al. Intelligent question answering system for traditional Chinese medicine based on BSG deep learning model: taking prescription and Chinese materia medica as examples. Digital Chinese Medicine, 2024, 7(1): 47-55.

study analysis [2]. However, the vast quantity and lack of relevance of available medical information pose challenges to ensure the scientific validity and accuracy of the medical knowledge application [3].

The knowledge graph serves as a sophisticated knowledge repository that could effectively store the relationship between data and knowledge [4, 5]. With the rapid development of natural language processing in artificial intelligence, the question answering system has gradually entered the stage of intelligence research. In 2012, Google introduced a question answering system on the basis of the knowledge graph, offering a pivotal roadmap for the advancement of intelligent question answering systerms in the field. Research on the application of the intelligent answering system in TCM has been carried out as well. Some scholars constructed a TCM question answering system by leveraging a TCM knowledge graph, TAN et al. [6] used knowledge graphs to address challenges related to the lack of Chinese corpus and complexities in labeling in the medical field, and the integration of rule matching technology has improved the accuracy of answers in question answering systems. LI [7] found that most of the question answering systems were based on templates with poor flexibility. In order to improve such defects, knowledge graphs have been introduced as the foundation of constructing the systems. Some scholars utilized the Aho-Corasick (AC) algorithm to optimize the automatic question answering system grounded in a knowledge graph, LI et al. [8] demonstrating AC algorithm superiority over conventional machine learning approaches in extracting entities, and based on semantic similarity calculation to obtain disease symptom entities, implemented a disease question answering system. HONG et al. [9] developed an intelligent question answering system for marine TCM based on AC algorithm to address the public's lack of understanding on marine TCM. Some scholars implemented question answering systems based on the bidirectional encoder representation from transformers (BERT) deep learning model, ZHANG et al. [10] designed a question answering system focusing on children’s diseases, which integrated term frequency-inverse document frequency (TF-IDF), BERT and deep structured semantic model (DSSM), effectively improved answer accuracy. QIAO et al. [11] proposed a knowledge graph and keyword attention mechanism (KK)-BERT method to address the scarcity of high-quality question answering data in the contemporary TCM field, enhanced the interaction between question and answer sentences. XU et al. [12] established the RoBERTa-bidirectional long short-term memory (BiLSTM)-conditional random field (CRF) (RBC) model specifically for the identification of entities in the online diabetes health community, the result was better than the performance of the BiLSTM model. SUN et al. [13] introduced an intelligent question answering approach based on BERT in combination with BiLSTM and bidirectional gated recurrent unit (BiGRU), and at the same time employed deep learning and a knowledge graph, to better solve questions in diagnosis.

In summary, traditional machine learning methods predominately rely on rule-based template matching, often implemented through AC multi-pattern matching algorithms and sentence vector similarity techniques. While effective at analyzing questions and retrieving answers through rule-based formulations, these methods are limited to handling knowledge explicitly encoded in predefined rules. Moreover, when new rules emerge, these methods become ineffective in addressing subsequent challenges. Even though deep learning-based semantic parsing methods are flexible in coping with challenges proposed by changes in input questions, it is affected by the quality of data annotations in tasks such as recognizing intentions and named entities [14]. Therefore, this paper constructs an intelligent question answering system from the perspective of collecting high-quality TCM data and improving deep learning algorithms, and utilizing structured knowledge graphs as the primary data source. It integrates entity node information from the knowledge graph with question templates, facilitating the automatic construction of training corpora for addressing TCM questions. By implementing intelligent question answering systems for TCM knowledge using the BERT+Slot-Gated (BSG) joint model, users can access accurate and personalized intelligent question answering services. The implementation of these systems holds profound significance and serves as valuable reference for promoting TCM knowledge, supporting decision-making in TCM research and teaching, thereby fostering the inheritance, innovation, and development of TCM.

2 Data and methods

2.1 Data preprocessing

2.1.1 Data collection The present study was undertaken to uphold the scientific validity and accuracy of the medical data by using the *Prescriptions of Chinese Materia Medica* [15] and *Chinese Materia Medica* [16] (the textbooks for higher education in the national traditional Chinese medicine industry listed by the 13th Five-Year Plan) as primary data sources. After obtaining relevant texts from the aforementioned textbooks, thorough checks and proofreading were conducted to ensure the completeness and accuracy of the content. Then, under the guidance of experts, this study refers to standards as the *Chinese Pharmacopoeia* 2020 edition, each prescription and

Chinese materia medica data were manually collected and screened, with the results being stored in an Excel file. Finally, a total of 448 valid pieces of Chinese materia medica data and 233 valid prescription data pieces were obtained.

2.1.2 Data preprocessing Data preprocessing includes several steps, such as data cleaning, deduplication, and standardization, aimed at ensuring data quality and consistency. In this paper, Python-Pandas tools were utilized to preprocess the text in Excel files, after which the preprocessed data were divided into semi-structured JSON files. Taking the JSON format for Mahuang (Ephedrae) as an example, it includes a total of eight key-value pairs: name, taste, meridian, efficacy, attending, usage, consumption, and attention. The specific data are as follows: {"name": "Mahuang (Ephedrae)"; "taste": ["pungent (辛)" "warming (温)" "slightly bitter (微苦)"], "meridian": ["lung" "bladder"]; "efficacy": ["promote sweating to release the exterior (发汗解表)" "lung facilitating for relieving asthma (宣肺平喘)" "inducing diuresis for removing edema (利水消肿)"]; "attending": "for wind-cold superficies without sweating syndrome (用于风寒表实无汗证), for cough and asthma excess syndrome (用于咳喘实证), for edema and superficies syndrome (用于水肿兼表证者)"; "usage": "relieving exterior for raw use, relieving asthma should be honey-fried or raw (解表生用，平喘宜蜜炙用或生用)"; "consumption": "2 – 10 g"; "attention": "this product exhibits potent sweating properties; hence, its usage is contraindicated for patients with conditions such as exterior deficiency spontaneous sweating, Yin deficiency night sweating and kidney deficiency cough and asthma (本品发汗力较强，故表虚自汗、阴虚盗汗及肾虚咳喘者忌用)"}. By preprocessing the data, the subsequent knowledge graph construction process was faster and more effective.

2.2 Knowledge graph construction methods

2.2.1 Knowledge acquisition In this paper, the "domain ontology seven-step method", developed by Stanford University School of Medicine, was employed as the ontology construction method for the knowledge graph [17]. This method proved valuable in organizing the scattered content of TCM into structured knowledge, fostering the creation of interconnected content conducive to TCM applications [18]. The "domain ontology seven-step method" includes the following seven steps: (i) determine the professional field and category; (ii) evaluate the possibility of reusing existing ontology; (iii) list the important terms in the ontology; (iv) define class and class hierarchy; (v) define the attributes of the class; (vi) define the facets of attributes; (vii) create an instance. The ontology layer of the TCM knowledge graph was designed (Figure 1).

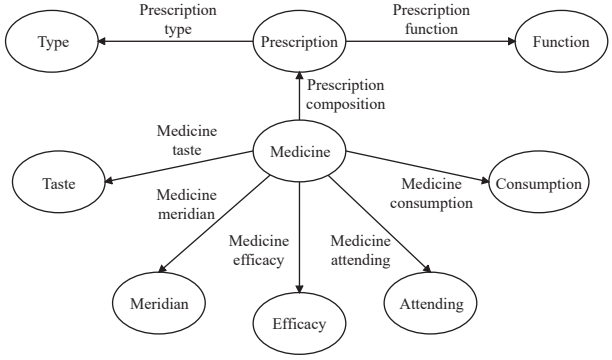

Figure 1 Design of ontology layer

Prescriptions consist of Chinese materia medica components, which mainly exert therapeutic, regulatory, and preventive effects. Chinese materia medica is the basic unit and internal structure of prescriptions. These are two indispensable elements in the treatment system of TCM, which together constitute important means in TCM for disease treatment. Under the guidance of basic theories in TCM, this study first took *Prescriptions of Chinese Materia Medica* [15] as the starting point, and extracted the knowledge of medicines, type and function contained in each prescription. Subsequently, with this as the foundation, relevant knowledge about the taste, efficacy, consumption, and other aspects of each medicine were extracted from the *Chinese Materia Medica* [16]. Entities are the most basic elements in the knowledge graph, entity types of knowledge graphs are shown in Table 1, constructed 3 149 entities and generated entity dictionaries for convenient subsequent data analysis.

2.2.2 Knowledge representation Indeed, triples serve as the fundamental unit of knowledge representation in a graph. Typically, they consist of entity 1, a relationship, and entity 2, or may include attributes along with their corresponding attribute values. Different TCM entities showcase different relationships. In this paper, according to the ontology layer design of the knowledge graph, the TCM knowledge was processed into triple information. As shown in Table 2, 6 891 entity relationships were constructed.

Attributes primarily refer to the characteristics of TCM entities; attribute value mainly refers to the content of the corresponding attributes of TCM entities [7]. As shown in Table 3, information such as source and application are additional attributes and helpful for subsequent applications of the knowledge graph.

2.2.3 Knowledge storage Typically, the knowledge is stored using two main methods: RDF and graph databases, with the latter being categorized under Not only SQL (NoSQL) database. In this study, Neo4j, a leading graph database, was selected as the preferred knowledge storage tool due to its exceptional performance, reliability, and scalability [19].

Table 1 Entity types in the knowledge graph

Entity type	Entity quantity	Example
Type	68	Formula for relieving superficies syndrome with pungent and warm natured drugs (辛温解表剂)
Prescription	233	Daqinglong decoction (大青龙汤)
Function	350	Promote sweating to release the exterior, clear internal heat (清里热)
Medicine	690	Mahuang (Ephedrae)
Taste	18	Pungent, warming, slightly bitter
Meridian	12	Lung, bladder
Efficacy	543	Promoting sweating to release the exterior, lung facilitating for relieving asthma, inducing diuresis for removing edema
Attending	1 027	For wind-cold superficies without sweating syndrome, for cough and asthma excess syndrome, for edema and superficies syndrome
Consumption	208	3 – 10 g

Table 2 Entity relationship types in the knowledge graph

Entity relationship type	Relationship quantity	Example
Prescription_type	233	<Daqinglong decoction, type, formula for relieving superficies syndrome with pungent and warm natured drugs>
Prescription_function	428	<Daqinglong decoction, function, promote sweating to release the exterior> <Daqinglong decoction, function, clear internal heat>
Prescription_composition	1 612	<Daqinglong decoction, composition, Mahuang (Ephedrae)> <Daqinglong decoction, composition, Guizhi (Cmnamomi Mmulus)>
Medicine_taste	1 018	<Mahuang (Ephedrae), taste, pungent> <Mahuang (Ephedrae), taste, slightly bitter> <Mahuang (Ephedrae), taste, warming>
Medicine_meridian	959	<Guizhi (Cmnamomi Mmulus), meridian, lung> <Guizhi (Cmnamomi Mmulus), meridian, bladder>
Medicine_efficacy	1 106	<Fangfeng (Saposhnikoviae Radix), efficacy, expelling wind to relieve superficies (祛风解表)> <Fangfeng (Saposhnikoviae Radix), efficacy, eliminating dampness (胜湿)> <Fangfeng (Saposhnikoviae Radix), efficacy, analgesic (止痛)> <Fangfeng (Saposhnikoviae Radix), efficacy, spasmolysis (解痉)>
Medicine_attending	1 148	<Xinyi (Magnoliae Flos), attending, wind chill headache nasal congestion (风寒头痛鼻塞)> <Xinyi (Magnoliae Flos), attending, thick rhinorrhea headache (鼻渊头痛)>
Medicine_consumption	387	<Zhuye (Folium Phyllostachydis Henonis), consumption, 6 – 15 g>

Table 3 Entity attributes in the knowledge graph

Entity attribute	Example
Source	*Shang Han Lun* (《伤寒论》)
Apply	Water decoction and taking medicine warm (水煎温服)
Indication	Summer heat dampness, heat-related thirst, urinary retention, enterorrhea (暑湿证，身热烦渴，小便不利，或泄泻)
Characteristic	Warming concurrent antipyresis, exterior interior resolving, emphasis on pungent warm diaphoresis (寒温并用，表里同治，重在辛温发汗)
Utilization	This prescription serves as the fundamental treatment for addressing summer heat dampness. The primary syndrome differentiation focuses on symptoms of heat-related thirst and urinary retention (本方为治疗暑湿证之基础方。以身热烦渴，小便不利为辨证要点)
Usage	The prescription shouldn't be cooked in water for long. Long leaves are conducive to dispelling cold and releasing the exterior, while long stem is beneficial for regulating Qi and harmonizing the top, stabilizing the fetus, and detoxification (不宜久煎。叶长于发表散寒，梗长于理气宽中、安胎、解毒)
Attention	This product is pungent and warm in nature, so patients with endogenous heat due to Yin deficiency and heat are not allowed to use (本品辛温，故阴虚内热及热盛者忌用)

2.3 Question answering methods

2.3.1 Question identification The question identification of the intelligent question answering system involves classifying and comprehending user queries to determine their type, intention, and pertinent details. This capability enables the system to deliver precise answers or solutions tailored to users' needs [20]. The question

identification process of the intelligent question answering system usually includes named entity recognition and intention recognition. Named entity recognition aims to identify and extract named entities from users' queries. Through employing named entity recognition, the system can capture the key information contained in the question. For example, through named entity vecognition, the system can identify essential elements in the question, including the prescription name, the medicine name, the medicine taste, and the medicine efficacy in the question. Intention recognition evaluates the intention of users' questions by analyzing the semantics and context of the questions, assists the TCM intelligent question answering system in extracting key information from the question, understanding users' intentions accurately, and provides a foundation for subsequent question answering products and research.

2.3.2 Pattern matching algorithm Pattern matching algorithms represent text-matching methods that rely on predefined rules or patterns to identify corresponding segments in the input text. They have been employed to locate portions of text that match specific predefined patterns or rules. In the task of question named entity recognition, the AC multi-pattern matching algorithm [13] can aid in entity extraction. It relies on the entity dictionaries generated during the construction of the TCM knowledge graph, such as those for prescriptions and medicines. In the task of question intention recognition the AC multi-pattern matching algorithm mainly classifies questions by extracting the interrogative words as the algorithm matching words.

The primary advantage of the AC multi-pattern matching algorithm is its high efficiency and exceptional performance in matching multiple pattern strings. However, maintaining the accuracy of this approach necessitates regular updates to the entity dictionaries and interrogative words. Additionally, this method may struggle to identify new acquired knowledge not yet included in the existing dictionaries.

2.3.3 BSG model The deep learning approach employed by the previous medical question answering system treats named entity recognition and intention recognition as separate tasks, each with its own distinct methodologies. However, this approach fails to consider the inherent interdependence between these tasks. GOO et al. [21] introduced the Slot-Gated joint model, which integrated entity recognition and intention recognition tasks for simultaneous training. By leveraging the inherent correlation between these key tasks, this method enhanced the effectiveness of question answering. However, the model employs a word vector coding approach, which cannot effectively record the TCM information. BERT has demonstrated its efficacy in accurately classifying TCM records [22]. By pre-training on extensive corpora, BERT has consistently realized superior performance compared to other models. Therefore, this paper utilized the BERT model as the vector input for sentence encoding in the Slot-Gated joint model, and introduced a new model of BSG for the construction of the intelligent question answering system. It converted each word in the users' questions into a 768-dimensional vector, where the distance between two words in the vector space reflected the semantic similarity between the corresponding words in the original text. By using the BSG model to concurrently train two tasks, the complexity of task processing was streamlined without compromising the accuracy of the question answering. This approach could introduce novel research directions for intelligent question answering in the medical field.

As shown in Figure 2, each word in the discourse is assigned a distinct slot label, while the entire discourse is associated with a specific intention. Slot filling is regarded as a sequence labeling task that maps the input word sequence $X = (x_1, \cdots, x_T)$ to the corresponding slot label sequence $Y^S = (Y_1^S, \cdots, Y_T^S)$. Intention detection is regarded as a classification problem that maps the input word sequence $X = (x_1, \cdots, x_T)$ to the corresponding intention label.

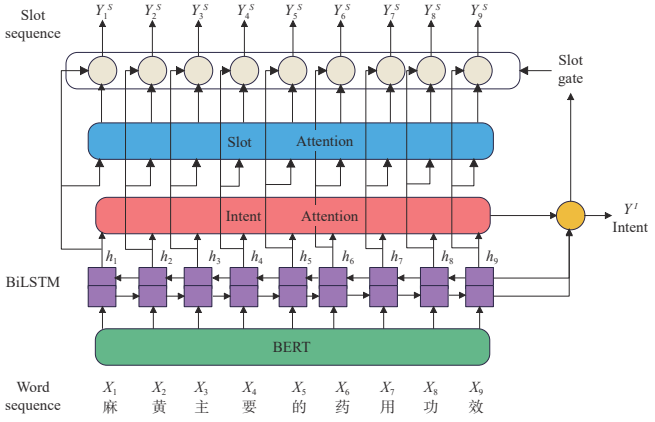

Figure 2 The BSG model structure

In Equation (1), c_i^S is used as the weighted sum of the hidden states of BiLSTM. $\alpha_{i,j}^S$ is the weight of learning attention.

$$c_i^S = \sum_{j=1}^{T} \alpha_{i,j}^S h_j \tag{1}$$

$$\alpha_{i,j}^S = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{T} \exp\left(e_{i,k}\right)} \tag{2}$$

$$e_{i,k} = \sigma\left(W_{he}^S h_k\right) \tag{3}$$

In Equation (2), $e_{i,k}$ calculates the relationship between e_k and input vector e_i. In Equation (3), σ is the

activation function, and W_{he}^{S} is the weight matrix of the feed forward neural network. Subsequently, the hidden state and the slot context vector are used to fill the slot.

$$Y_i^S = softmax\left(W_{hy}^S\left(h_i + c_i^S\right)\right) \tag{4}$$

In Equation (4), Y_i^S is the slot label of the ith word in the input layer, and W_{hy}^S is the weight matrix. Among them, h_i and c_i^S are used to do $softmax$ to obtain the predicted value of the corresponding label of the ith word.

$$Y^I = softmax\left(W_{hy}^I\left(h_T + c^I\right)\right) \tag{5}$$

In Equation (5), the intention context vector c^I is calculated in the same way as c^S, but the intention detection part only takes the last hidden state of BiLSTM.

$$p\left(Y^S, Y^I|X\right) = p\left(Y^I|X\right)\prod_{t=1}^{T} p\left(Y_t^S|X\right) = p\left(Y^I|X_1,\cdots,X_T\right)\prod_{t=1}^{T} p\left(Y_t^S|X_1,\cdots,X_T\right) \tag{6}$$

In Equation (6), $p\left(Y^S, Y^I|X\right)$ is the conditional probability of understanding slot filling and intention recognition when the input word sequence is given, from which the final result of model prediction can be obtained.

2.3.4 Experimental dataset This paper proposes the development of an automatic medical question generator. The entity information from the knowledge graph was used as a slot to fill in the question. By combining question intention with slot filling, the dataset of TCM knowledge was generated [23]. The corpus of users' questions is shown in Figure 3. Constructing category feature words based on TCM knowledge, a pre-set question classification system comprising 20 types of questions was designed for the intelligent question answering system (Table 4). A total of 8 938 medical question datasets were subsequently generated by combining entity dictionaries with corresponding category feature words.

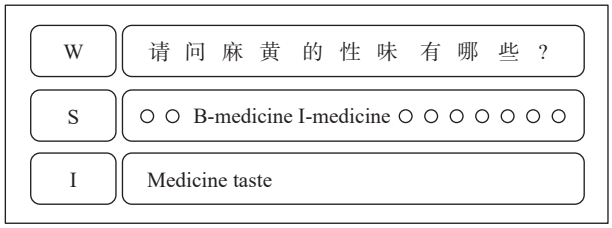

Figure 3 An example of the generation of the user's question corpus

W, the words in the question. S, slot identification. I, the intent information of the sentence. O, a non entity. B-, the beginning of the entity. I-, the inside of the entity.

3 Results

3.1 Knowledge graph

The Neo4j graph database includes two basic data types: nodes and relationships. Each node represents an entity, which can have zero or more relationships and attributes. The relationship represents a connection between two nodes. This paper used Python language to execute the Cypher CREATE statement via the py2neo module. Subsequently, it operated on the Neo4j graph database to create entity nodes, entity attributes, and entity relationships. The sorted TCM triplet data were transmitted to the graph database Neo4j to achieve the storage and visualization of the knowledge graph. Part of the TCM knowledge graph is shown in Figure 4.

3.2 Algorithm analysis

3.2.1 Experimental environment The hardware environment for model training is NVIDIA RTX 3060 and Intel Core i7-10750H. The software environment is Python 3.6, Torch 1.10.0, Tensorflow 1.14.0, and Keras 2.2.5. The BERT version is Bert-Chinese-base. The transformer layer of the model is 12 layers, with a hidden state of 768 dimensions. In the training phase, the initial learning rate is 5×10^{-5}, optimized using the Adam optimizer with a dropout rate of 0.2, a batch size of 32, and a training batch size of 20. This study conducted relevant experiments according to the above parameter settings.

Table 4 Partial exhibition of manual labeling

Question category	Category feature word
Prescription_type	What is the type, what is the type of drug
Prescription_composition	What is the Chinese materia medica included, what are the traditional Chinese medicines, what are the compositions
Composition_prescription	What are the prescriptions included, what are the compositions of the prescriptions
Medicine_taste	What is the taste
Taste_medicine	What is the taste of Chinese materia medica
Medicine_efficacy	What are the main efficacy, medicinal efficacy, main medicinal efficacy, what are the conditioning efficacies
Efficacy_medicine	What are the efficacies of Chinese materia medica
Medicine_usage	The specific usage, the common usage, the general usage, the usage method, the correct usage, the main usage, the application method
Medicine_attention	What is the use attention, which people can not use, what are the disadvantages, use precautions

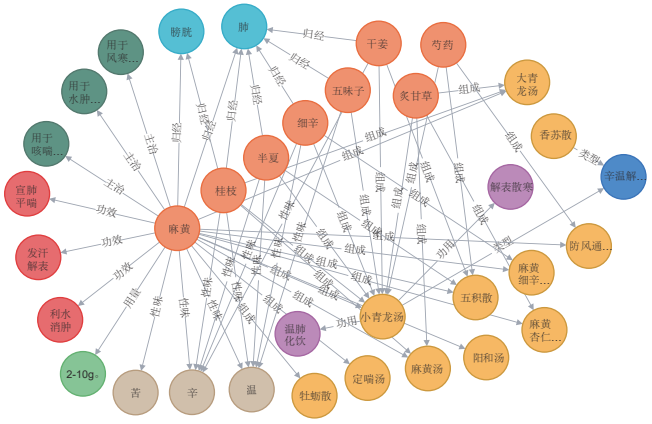

Figure 4 Partial exhibition of the TCM knowledge graph

Orange, medicine name. Green, medicine attending. Red, medicine efficacy. Celeste, medicine meridian. Grey, medicine taste. Yellow, prescription name. Blue, prescription type. Purple, prescription function.

3.2.2 Experimental results The medical question dataset was partitioned as follows: 70% of the data was earmarked as the training set, 20% was allocated to the validation set, and the remaining 10% served as the test set for model training. The BSG model was compared with the AC multi-pattern matching algorithm [13], BiLSTM model [15], Slot-Gated model [21], and BERT model [22]. The precision (P), recall (R) and F1 score in the multi-classification task were used as evaluation indicators to compare and analyze the results of the named entity recognition task [Equation (7)]. Exact Match (EM) was used to evaluate the percentage of correct answers matched in the prediction, which served as the comparison result of the algorithm intention recognition task.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

As shown in Table 5, the results obtained from the deep learning algorithm have greatly improved compared to the prediction results from pattern matching. The pattern matching approach struggled to identify TCM entities containing typos in questions, and was restricted to recognizing only those entities present in the limited TCM entity dictionary. The comparison between the BiLSTM model and the Slot-Gated model showed that the joint model effectively improved the results of both tasks. The BERT model was better than the Slot-Gated model, indicating that the BERT model demonstrated a stronger ability to understand the semantic information of TCM questions. The BSG model introduced in this paper enhanced the vector dimension of the input layer. The BERT model was utilized to encode TCM questions, which identified more entity information and question intentions. The F1 score of the improved model increased by 0.87%, and the EM by 0.13%. Compared with the static word vector method used in the original model, BERT functions as a dynamic word vector. It computed the contextual representation of text comprehensively during the pre-training process, resulting in densely represented semantic information and thereby improving the model's effectiveness.

Table 5 Evaluation on the model's effects in different named entity recognition and intention recognition tasks

Algorithm	Precision	Recall	F1 score	EM
AC multi-pattern matching	0.8962	0.9046	0.9004	89.63%
BERT	0.9913	0.9962	0.9938	99.87%
BiLSTM	0.9742	0.9863	0.9802	98.91%
Slot-Gated	0.9876	0.9888	0.9882	99.62%
BSG	0.9963	0.9975	0.9969	99.75%

3.3 Question answering system

The design of the intelligent question answering system effectively harnesses the rich structured semantic information in the knowledge graph, enhancing the efficiency of interaction between humans and machines. Users ask questions according to their own intentions. The BSG algorithm was employed to identify the entities and intentions of users' questions. The entity was extracted from the users' questions, and the type of the question was determined accordingly. The system conducted model analysis to identify both the entity name and intention classification in the question. Subsequently, it matched the required question template according to the obtained question intention classification and filled the entity name in the question template to generate a Cypher query statement [24]. The query engine performed operations in the Neo4j graph database which stored knowledge, and the results corresponding to the users' questions were retrieved. The overall framework of the intelligent question answering system, with TCM knowledge graph as the foundation constructed (Figure 5).

In the backend of the WeChat Official Account, the system accessed the Flask server. It utilized the chat window of the WeChat Official Account as the frontend to interact with users. The user inputs the question through the chat window, and clicks "Send" to transmit the question to the pre-set Flask interface. After receiving the questions, the backend invokes the trained BSG model for identification. It extracted the entity and intention information from the user's question, fills in the entity according to the corresponding intention template, generates the Cypher statement to retrieve answers from the knowledge graph, and then returns the results to the WeChat Official Account platform for display. For example, what is the taste of Mahuang (Ephedrae) in natural language? The intelligent question answering system in the WeChat Official Account operates as follows. After parsing the question by the BSG algorithm, it would identify the "Mahuang (Ephedrae)" as a drug name, and the

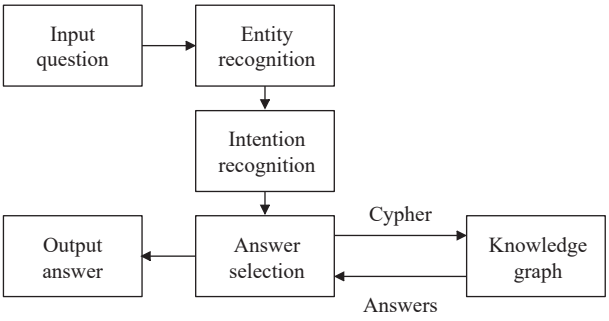

Figure 5 The overall framework of the system

intention recognition result is “Medicine_taste”. According to the result of intention recognition, the matched question template is returned, and the “Mahuang (Ephedrae)” is filled in the question template. Executing the corresponding Cypher statement in Table 6, the information returned in the Neo4j graph database is “the taste of Mahuang (Ephedrae) include: pungent, slightly bitter, warming”. The system realizes the rapid retrieval and accurate answer of various types of questions related to TCM, thereby delivering authoritative and high-quality TCM knowledge according to users’ needs.

Table 6 Examples of correspondence between question template and query statement

Question example	Question template	Cypher query statement
What are the medicines contained in Daqinglong decoction?	Prescription_ composition	MATCH(m:Prescription)-[r:Prescription_composition]->(n:Medicine) where m.name = "大青龙汤" return n.name
What is the taste of Mahuang (Ephedrae)?	Medicine_taste	MATCH(m:Medicine)-[r:Medicine_taste]->(n:Taste) where m.name = "麻黄" return n.name
What are the Chinese materia medica with the efficacy of inducing diuresis for removing edema?	Efficacy_medicine	MATCH(m:Medicine)-[r:Medicine_efficacy]->(n:Efficacy) where n.name = "利水消肿" return m.name
Can you introduce the use of Danggui (Angelicae Sinensis Radix)?	Medicine_attention	MATCH (m:Medicine) where m.name = "当归" return m.name, m.attention

4 Discussion

The TCM knowledge graph intuitively presents various knowledge elements and their interrelationships, providing data support for the discovery of potential connections and laws within knowledge [17]. This paper focuses on leveraging the knowledge of prescriptions and Chinese materia medica to carry out research on intelligent question answering. The study has systematically and efficiently constructed a TCM knowledge graph and applied deep learning technology to develop an intelligent question answering system. This system facilitates more efficient and comprehensive learning and understanding of TCM, thereby enhancing individuals’ health awareness and preventive measures. Moreover, it contributes to promoting the inheritance and innovation of TCM culture. In essence, the system offers new solutions for promoting the development of TCM culture.

This paper introduced an innovative approach by utilizing the BSG joint model to simultaneously address entity recognition and intention recognition challenges in the field of TCM. This methodology enhanced the efficiency of the intelligent question answering system in identifying user queries effectively. This research idea and approach can more efficiently implement intelligent question answering systems in the medical field. The proposed model has demonstrated remarkable effectiveness. Of course, the intelligent question answering system constructed in this study still has improvement [22, 24]. For instance, enhancing the retrieval speed of the knowledge graph can accelerate the system's response time. Furthermore, automating and efficiently expanding the question template information to encompass a broader range of potential user inquiries, can enhance the recognition accuracy of text-based questions. Additionally, employing multiple rounds of dialogue and more sophisticated knowledge reasoning methods can further elevate the system’s level of intelligent .

5 Conclusion

Based on pertinent TCM knowledge, this paper has extracted knowledge triples to establish a high-quality TCM knowledge graph. Building upon the knowledge graph and addressing the shortage of Chinese question answering training corpora, this study has proposed an automatic medical question generator. By leveraging TCM entities and category feature words, this generator has encompassed a dataset of TCM-related questions. This system has employed the BERT pre-training model as the vector input for the Slot-Gated algorithm’s sentence encoding. As a result, it has improved the accuracy of the answer selection in the intelligent question answering system, assist medical students, hospital interns, as well as grassroots physicians in learning and understanding TCM knowledge comprehensively.

Fundings

National Key R&D Program of China (2018AAA0102100), Hunan Provincial Department of Education Outstanding Youth Project (22B0385), and 2022 Disciplinary Construction “Revealing the List and Appointing Leaders” Project (22JBZ051).

Competing interests

The authors declare no conflict of interest.

References

[1] REN CY. Research on intelligent question answering system based on COVID-19 knowledge graph. Baotou: Inner Mongolia University of Science and Technology, 2021.
[2] LI XL. Research and implementation of traditional Chinese medicine question answering system based on knowledge graph. Qingdao: Qingdao University, 2021.
[3] WANG ZY, YU Q, WANG N, et al. Survey of intelligent question answering research based on knowledge graph. Computer Engineering and Application, 2020, 56(23): 1–11.
[4] JIANG CY, HAN XY, YANG WR, et al. Review of research and application of medical knowledge graph. Computer Science, 2023, 50(3): 83–93.
[5] HUANG XF, ZHANG JX, XU ZS, et al. A knowledge graph based question answering method for medical domain. PeerJ Computer Science, 2021, 7: e667.
[6] TAN W, LIU CL. Development of medical question-answering system based on knowledge graph and model integration. Chinese Journal of Medical Library and Information, 2021, 30(11): 1–9.
[7] LI F. Research and implementation of question answering system based on knowledge graph. Nanjing: Nanjing University of Posts and Telecommunications, 2022.
[8] LI H, LIU JY, LI SY, et al. Optimizing automatic question answering system based on disease knowledge graph. Data Analysis and Knowledge Discovery, 2021, 5(5): 115–126.
[9] HONG HL, LI WL, YANG T, et al. Design and implementation of intelligent question answering system for marine traditional Chinese medicine based on knowledge map. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2023, 25(6): 1935-1941.
[10] ZHANG X, WANG HR, LI ML, et al. Construction of children's disease question answering model based on knowledge graph. Journal of Zhengzhou University, 2022, 54(2): 74–80.
[11] QIAO K, CHEN KJ, CHEN JQ, et al. Chinese medical question answering matching method based on knowledge graph and keyword attention mechanism. Pattern Recognition and Artificial Intelligence, 2021, 34(8): 733–741.
[12] XU Q, ZHOU Y, LIAO BL, et al. Named entity recognition of diabetes online health community data using multiple machine learning models. Bioengineering, 2023, 10(6): 659.
[13] SUN TT. Research on knowledge graph medical diagnosis method based on deep learning. Baotou: Inner Mongolia University of Science and Technology, 2022.
[14] HUANG JY. Research and applications analysis of knowledge base question answering. Highlights in Science, Engineering and Technology, 2022, 16: 16–22.
[15] LI J, LIAN JW. Prescriptions of Chinese materia medica. Beijing: China Press of Traditional Chinese Medicine, 2016.
[16] ZHOU ZX, TANG DC. Chinese materia medica. Beijing: China Press of Traditional Chinese Medicine, 2016.
[17] YU HL, CAO LY, QU YQ, et al. Knowledge map construction and knowledge discovery of Xiaoke Jingfang. Journal of Zhejiang University of Traditional Chinese Medicine, 2022, 46(2): 113–119, 125.
[18] JIANG ZX, CHI CY, ZHAN YY. Research on medical question answering system based on knowledge graph. IEEE Access, 2021, 9: 21094–21101.
[19] YIN YT, ZHANG L, WANG YG, et al. Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis B. BioMed Research International, 2022, 2022: 7139904.
[20] MA ZL, WANG SY, ZHANG HZ, et al. Intelligent question answering intention recognition joint model based on knowledge graph. Computer Engineering and Application, 2023, 59(6): 171–178.
[21] GOO CW, GAO G, HSU YK, et al. Slot-Gated modeling for joint slot filling and intent prediction. Association for Computational Linguistics, 2018: 753–757.
[22] ZHANG FL, HUANG X, WANG RJ, et al. Chinese NER model based on BERT multi-knowledge graph fusion embedding. Journal of University of Electronic Science and Technology of China, 2023, 52(3): 390–397.
[23] WU D, ZHOU ZJ. Intelligent question answering system for cardiovascular diseases based on knowledge graph. Software Guide, 2022, 21(3): 160-164.
[24] XIONG HB, WANG ST, TANG MR, et al. Knowledge graph question answering with semantic oriented fusion model. Knowledge-Based Systems, 2021, 221: 106954.

基于 BSG 深度学习模型的中医药智能问答系统研究：以方剂和中药为例

李冉[a], 任高[a], 晏峻峰[a], 邹北骥[a, b], 刘青萍[a*]

a. 湖南中医药大学信息科学与工程学院, 湖南 长沙 410208, 中国
b. 中南大学计算机学院, 湖南 长沙 410083, 中国

【摘要】目的 基于深度学习的方法，利用知识图谱构建中医药知识库，探寻联合模型在中医药智能问答系统的应用。**方法** 以《方剂学》和《中药学》规划教材为基础建立知识图谱，作为智能问答系统的知识来源，本研究提出一种 BERT+Slot-Gated（BSG）深度学习模型，获取用户自然问题中包含的中医药实体和问句意图，通过实体和意图在知识图谱检索答案返回给用户，运用 Flask 框架和 BSG 模型研发中医药智能问答系统。**结果** 构建了包含 3 149 个实体和 6 891 个关系三元组的中医药知识图谱，通过问题语料测试,本系统在回答中医药 20 类问题的实体识别 F1 值为 0.996 9，意图识别准确率为 99.75%，表明本系统具有较高的实用性和可行性，并通过微信公众号平台实现了用户与系统交互。**结论** 本文所提出的 BSG 模型通过提高向量维度在实验中取得了较好的结果，表明联合模型方法的有效性，可为实现中医药智能问答系统提供新的研究思路。

【关键词】中医药；深度学习；知识图谱；智能问答系统；BERT+Slot-Gated 模型