A DIK-based Question-Answering Architecture with Multi-sources Data for Medical Self-Service (KG)

Mengxing Huang[1,2],Menglong Li[1,2], YuZhang[1,2], Wenglong Feng[1,2]
1 State Key Laboratory of Marine Resource Utiliza tion in South China Sea, Hainan University, Haikou, China
2 College of Information Science &Technology, Hainan University, Haikou, China
Corresponding author: Yu Zhang (Email: yuzhang_nwpu@163.com)

Abstract—**Medical data is amplified in terms of speed and capacity in a very fast way, which creates obstacles for users to quickly access valid information. We present a DIK-based Question-Answering Architecture for Medical Self-Service. In addition, we propose a model based on the attention mechanism to extract high-quality medical entity concepts from the Chinese Electronic Medical Records (EMR). Then we modeled the medical data based on the DIK architecture (Data graph, Information graph, and Knowledge graph), construct a Question-Answering model (DIK-QA) for medical self-service that meets the needs of users to quickly and accurately find the medical information they need in massive medical data. Finally, we have realized this approach and applied it to real-world systems. The experimental results on our medical dataset show that the DIK-QA can effectively handle 4W (who/what/why/how) questions, which can help users find the information they need accurately.**

Keywords—Data graph, Information graph, and Knowledge graph (DIK); Question-Answering (QA); Electronic Medical Records (EMRs); attention mechanism

I. INTRODUCTION

Recently, knowledge graph have increasingly attracted attention in various fields. Medical Knowledge Graph (MKG) is also very popular in the medical field due to the unique expression of knowledge graph. Researchers explore with various approaches to construct the MKG. Reference [1] used deep learning method to extract 4 entities and 9 entity relationships from Chinese electronic medical records (EMR), then adopted the triple form to import data into the Neo4j and visualized the data into a knowledge graph. Reference [2] built a medical field table by using Bootstrapping and Conditional Random Fields (CRF) and solved the actual problem based on the obtained knowledge graph.

At present, research on knowledge graph is endless. Cowie et al. divided the knowledge graph into Data Graph (DG), Information Graph (IG) and Knowledge Graph (KG) according to the concept of Data, Information, Knowledge, and Wisdom (DIKW) [3].

However, there is a gap to fill for combining the MKG with the QA model. There are basically two major challenges. The first challenge is no effective framework for the union between QA and MKG for medical services. The second challenge the reliability of medical knowledge base. By addressing two major challenges, we solve and successfully construct a DIK-based QA for medical self-services. The complete structure of the DIK-QA is shown in Fig. 1.The contributions of our work are: 1) According to [4], we constructed the medical self-service DIK-QA based on the DIK. It can model massive amounts of data and qucikly and accuratelly find the inforamtion users need and provide services to users in a friendlier manner, 2) We guarantee the data reliability of the medical knowledge base from two perspectives: clinical knowledge and health information. For clinical knowledge, we use the model BiLSTM-Attended-CRF model to obtain a higher quality medical entity concept. For health information we adopt the distributed crawlers to collect richer information, such as dietary advice and rehabilitation exercises.

II. FRAMEWORK

We divide the work into three parts: a) designing a DIK-QA framework (including three layers of knowledge graph), b) data acquisition, c) DIK-QA implementation. As shown in Fig. 1.

A. DIK-QA Architecture

According to the DIKW [3] idea, the medical knowledge graph can be divided into three categories: 1) Data graph, 2) Information graph, 3) Knowledge graph.

Data Graph Data graph can record the frequency of the data, including spatial frequency and structure frequency. We refer to the definition of data frequency in [4] to define the medical data frequency as a two-tuple.

$$DFreq =< f_{spatial}, f_{structure} > \quad (1)$$

Where the spatial frequency, structure frequency of data are represented by $f_{spatial}$, and $f_{structure}$, respectively. The frequency of structure and spatial are the medical department of the disease and the treatment, respectively. As shown in Fig. 2.

DOI reference number: 10.18293/SEKE2019-112

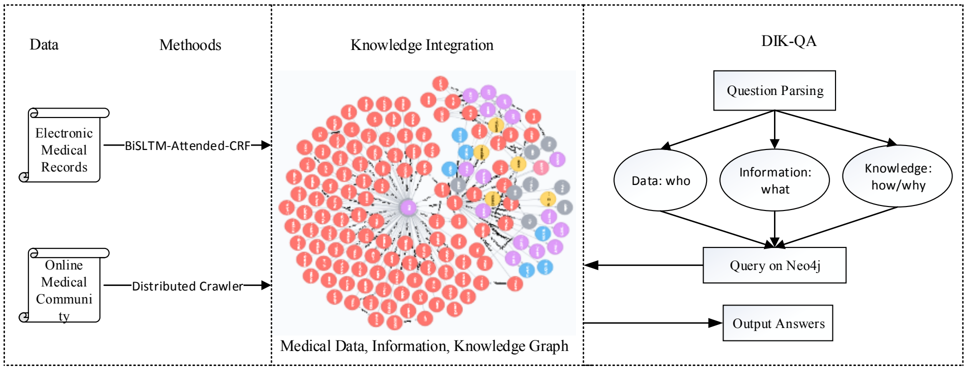


Figure 1. Overview of the DIK-QA Architecture for Medical Self-Service

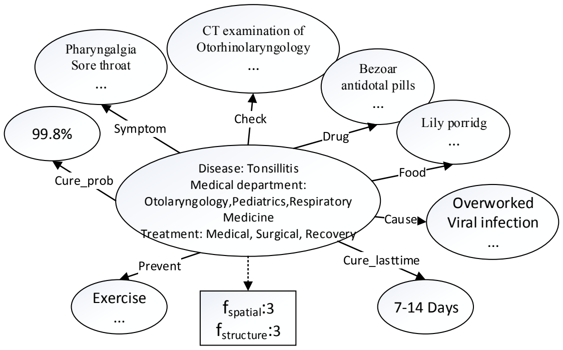


Figure 2. Statistics on $f_{spatial}$ and $f_{structure}$ of medical data.

Information Graph Information is extracted from the data for analysis and interpretation. We store the collected data as the medical knowledge base in dictionary format, as shown in Fig. 3. We define the medical knowledge base as the directed graph $G(V,E)$. The nodes and edges are denoted by V, E, respectively.[4]

$$Com\_\deg ree = \sqrt{\deg^{+} \times \deg^{-}} \tag{2}$$

$$\mathrm{Im}\,por = \alpha DFreq \times \beta Com\_\deg ree \tag{3}$$

The in-degree and out-degree of the node are represented by $\deg^{+}$, $\deg^{-}$, respectively. In order to prevent the loss of the information, we adopt (3) to further measure the importance of the nodes. The weights of the *DFreq* and the $Com\_\deg ree$ is denoted by α , β .

Knowledge Graph Knowledge is the overall understanding and awareness gained from the accumulated information.

$$Cr(E_1,R,E_2) = \frac{\sum_{\pi \in Q} P(E_1 \to E_2)\theta(\pi)}{|Q|} \tag{4}$$

We use (4) to calculate the correctness of the relationship between E_1 and E_2. Before we construct a medical knowledge graph, we need to design the knowledge graph: entity type nodes and entity relationship nodes, as shown in TABLE 1, 2.

Where all path between E_1 and E_2 are represented by Q, the weights of single path and path are represented by $\theta(\pi)$, π .

We comprehensively evaluate the importance of the nodes on the knowledge graph according to (5). N is the number of the relationship types. The weight of $\mathrm{Re}\,l_i$ is represented by λ_i .

$$Final\_\mathrm{Im}\,por = \mathrm{Im}\,por \times \gamma \frac{\sum_{i=1}^{n} \lambda_i \times \mathrm{Re}\,l_i}{n} \tag{5}$$

After completing the design of the DIK-based medical knowledge graph, we use the py2neo module in Python to import the dictionary type data into the Neo4j. As shown in Fig.1.

TABLE 1. The definition of the entity nodes

Entity nodes	**Meaning**
Department	Disease department
Disease	Disease name
Drug	Drugs for treating the disease
Cause	Cause of the disease

TABLE 2. The definition of the relationship nodes

Relation node	**Meaning**
do_eat	Recommended food for the disease
recommend_drug	Recommended drug for the disease
has_symptom	Disease corresponding symptoms
cure_way	Treatment for disease

B. Data Acquasition

In [1], we have extracted five entity types and nine entity relationship types by the BiLSTM-CRF model. However, we propose a novel model based on the Chinese EMR to improve the accuracy of the entity recognition, combining BiLSTM-CRF model with attention mechanism, as shown in Fig.4. Please refer to [1] for the BiLSTM-CRF model.

TABLE 3. The labeling rules of the Chinese EMR

Label Prefix	**Label Suffix**	**Meaning**
B	DISEASE,TREATMENT, CHECK, ,BODY, SIGNS	Head of the entity
I	DISEASE,TREATMENT, CHECK, ,BODY, SIGNS	Middle and tail of the entity
O	None	Other words

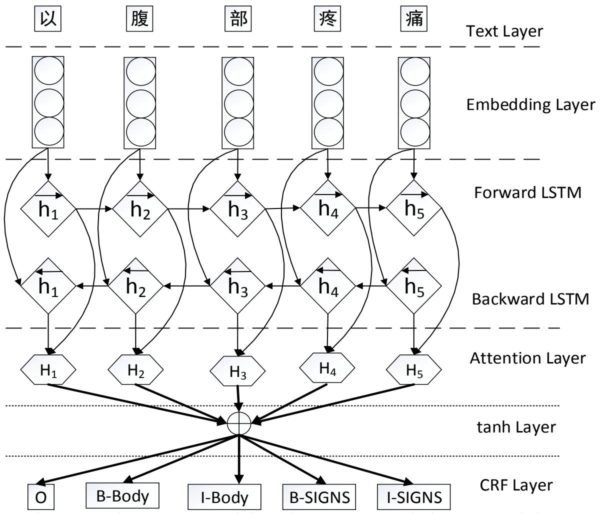


Figure 4. The structure of the BiLSTM-Attended-CRF model

BiLSTM-Attended-CRF Attention mechanism is a mechanism to simulate the attention of the human brain. The core idea is to draw on the attention of the human brain to things at a certain moment, it will focus on a certain key point, and ignore other key points [5]. We introduce the attention mechanism between the BiLSTM layer and the CRF layer to solve the challenge of extracting more precise semantic features for named entity recognition of Chinese EMR. Specifically, the attention mechanism is applied to the hidden layer of BiLSTM and then produces the newly hidden layer vectors. The implementation of the attention mechanism is as shown in (6)-(8).

$$e_{ij} = V \tanh\,(U_1\overrightarrow{h}_i + U_2\overleftarrow{h}_j + b_\alpha) \tag{6}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1}^{T}\exp(e_{ij})} \tag{7}$$

$$\overrightarrow{H_\iota} = \sum_{j=1}^{T}\alpha_{ij}\,\overrightarrow{h_J} \tag{8}$$

Where e_{ij} represents energy value of the jth word to the ith word, that is to say, the relationship between the two words. We use α_{ij} represents the attention weight of the jth word to the ith word, reflecting the influence of each vocabulary in the input text X. V , U_1 , U_2 , b_α is the parameter trained with the model. The new feature representation after the attention mechanism in Forward-LSTM Layer and Backward-LSTM is represented by $\overrightarrow{H_\iota}$, respectively. We combine two new feature representations to form the final feature representation$H_i = [\overrightarrow{H_\iota}, \overleftarrow{H_\iota}]$. After the attention layer, it will pass the activation function tanh layer, then pass the mapping results of tanh layer to the CRF Layer. Finally, the model can output the predicted label sequence for input text, as shown in Fig. 4.

Disease-Oriented We constructed a medical entity dictionary for recognition of the BiLSTM-Attended-CRF model. We take the knowledge fusion based on the medical concept in the dictionary to eliminate some ambiguous entities. The Disease-Oriented Knowledge base has the advantages: the disease can integrate various information such as symptoms, checks, causes, foods and drugs et al. into one line, as shown in Fig. 2.

C. DIK-QA Implementation

The DIK-QA model is consist of three parts: Analysis question, Question division, Query knowledge. As show in Fig.1.

Analysis question The DIK-QA can analysis the question in plain language way by the BiLSTM-Attended-CRF model. In this part, we can get the medical entity in the question, then we submit the identified results to the next part-Question division.

Question division After the analysis question, we can get the category of question. By classifying the user's questions, we can accurately get the intent of user. According to the frequency statistics on the types of questions and vocabulary that appear in the Q&A module on the medical website, we summarize two main types of questions: 1) Entity relationship template. It can solve the question of entity relationship query, including disease and symptoms, disease and drug, food, treatment and so on. For example, "what are the symptoms of hypertension?", "Recently dizzy, what disease might I get? ", etc. 2) Entity attribute template. It can solve the question of entity attribute query, including disease prevention, cause and so on. For example, "why do I get the hypertension?", "Precautionary measures for the hypertension?".

Query knowledge We can take the query on the Neo4j based above part. We take the Aho-Corasick (AC) algorithm [6] as the core method to realize the Query knowledge. AC automaton is the most classic multi-pattern matching algorithm. It uses a plurality of pattern strings to construct a finite state pattern matching automaton. The DIK-QA performs knowledge reasoning on the corresponding knowledge graph (DIK) according to different question types, so that the answer can be quickly found. We can convert the question to cypher query statement and search the corresponding answer, and then return it. Cypher is a graph query language designed for operating Neo4j that efficiently queries and updates knowledge graphs. It has more powerful than SQL in relational capabilities. For example, Question: what are the symptoms of tonsillitis? Cypher statement: Match (m: tonsillitis) –[r: has_symptom] → (n: Symptom) where m.name = '{0}' return m.name, r.name, n.name. Where r, m, n are the variables of the disease name, relation: has_symptom, and symptom node, respectively. Node definition refer to TABLE 1, 2.

III. EXPERIMENTS AND RESULTS

A. Eexperiments on BiLSTM-Attended-CRF

To evaluate the performance of the BiLSTM-Attended-CRF model, we compared several basic model. The EMR data from the local hospital. Other medical data comes from two medical websites[1]. The evaluation indicator adopts accuracy, recall rate, and F value. The comparison algorithm uses common models such as CRF, BiLSTM, and BiLSTM-CRF. We tested 100, 128, 200, 256 and 300 on word vector dimension. The BiLSTM-Attended-CRF model has a good performance when word vector dimension is 128. Learning rate is 0.001, the LSTM layer is 4. Dropout is 0.4. The experimental results are shown in Table 4.

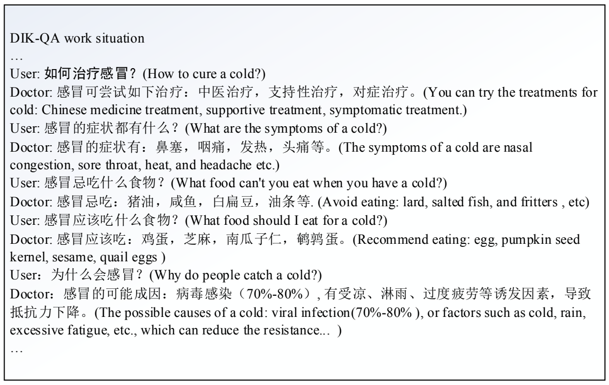
TABLE 4. Experimental results of each model for entity recognition

Model	P (%)	R (%)	F (%)
CRF	60.45	58.72	60.13
BiLSTM	61.42	58.23	57.68
BiLSTM-CRF	83.53	78.38	80.35
BiLSTM-Attended-CRF	89.76	85.51	88.90

The experimental results show that the BiLSTM-Attended-CRF model has better performance than state-of-the-art baselines in entity recognition. F-value and accuracy increased by 6%. In our model, all three evaluation marks have been improved.

B. Eexperiments on the DIK-QA

To evaluate the effectiveness of the DIK-QA, we selected five common diseases as small sample data, including cold, pediatric cold, hypertension, diabetes, cervical spondylosis. Observed from Fig.5, we can know that the DIK-QA can solve effectively some common question.



DIK-QA work situation
…
User: **如何治疗感冒？**(How to cure a cold?)
Doctor: 感冒可尝试如下治疗：中医治疗，支持性治疗，对症治疗。(You can try the treatments for cold: Chinese medicine treatment, supportive treatment, symptomatic treatment.)
User: 感冒的症状都有什么？(What are the symptoms of a cold?)
Doctor: 感冒的症状有：鼻塞，咽痛，发热，头痛等。(The symptoms of a cold are nasal congestion, sore throat, heat, and headache etc.)
User: 感冒忌吃什么食物？(What food can't you eat when you have a cold?)
Doctor: 感冒忌吃：猪油，咸鱼，白扁豆，油条等. (Avoid eating: lard, salted fish, and fritters , etc)
User: 感冒应该吃什么食物？(What food should I eat for a cold?)
Doctor: 感冒应该吃：鸡蛋，芝麻，南瓜子仁，鹌鹑蛋。(Recommend eating: egg, pumpkin seed kernel, sesame, quail eggs)
User：为什么会感冒？(Why do people catch a cold?)
Doctor：感冒的可能成因：病毒感染（70%-80%）, 有受凉、淋雨、过度疲劳等诱发因素，导致抵抗力下降。(The possible causes of a cold: viral infection(70%-80%), or factors such as cold, rain, excessive fatigue, etc., which can reduce the resistance...)
…

Figure 5. The DIK-QA's work situation

IV. RELATED WORK

In 2006, Rowley [7] revisits the data-information-knowledge-wisdom (DIKW) hierarchy and analyses the statement about the nature of data, information, knowledge, and wisdom. Through this process we get a consensus on definitions and transformation about DIKW. It makes the theoretical development of the further development of the DIKW. Duan et al. [8] proposes to clarify the expression of knowledge graph as a whole for lacking a unified definition and standard expression form of knowledge graph. They clarify the architecture of knowledge graph from data, information, knowledge and wisdom aspects respectively. They also propose to specify DIKW-based knowledge graph. Moreover, Song et al. [9] proposed a processing optimization mechanism of typed resources in a wireless-network-based three-tier architecture consisting of DIK mechanism. Simulation results show that the proposed approach improve the ratio of performance over user investment.

V. CONLUSIONS

The main task of this paper is: 1) we propose a highly accurate medical entity recognition model--BiLSTM-Attended-CRF to extract the high-quality medical concepts, 2) we construct the DIK-QA for medical self-service. The DIK-QA is mainly focused on some common diseases and questions such as colds, headaches, fever, ligament strains and some special diseases: pneumonia, etc.

However, there are still some shortcomings in this paper. We can combine DIK-QA with the deep learning method to automatically analyze the problem. Then, we can construct a smarter medical self-service. We can also try to construct a wider application of the DIKW architecture.

ACKNOWLEDGMENT

This work was supported by the Key R&D Project of Hainan province (Grant #: ZDYF2019020), National Natural Science Foundation of China(Grant #: 61662019 and 61862020), Major Science and Technology Project of Hainan province (Grant #: ZDKJ2016015), Higher Education Reform Key Project of Hainan province (Grant #: Hnjg2017ZD-1).

REFERENCES

[1] M.X. Huang, M.L. Li, H.R. Han, "Research on entity recognition and knowledge graph construction based electronic medical records," in press.

[2] Y.B. Zhang, Research on the construction of medical knowledge graph and its application. Ph.D. Thesis, Harbin Institute of Technology. Harbin, China: Haibin Institue of Technology, 2018.

[3] J. Cowie, W. Lehnert, Information extraction. Berlin, Hei-delberg: Springer, 2004.

[4] L.X. Shao, Y.C. Duan, Z.B. Zhou, et al., "Design of recommendation services based on data, information and knowledge graph architecture," Journal of Frontiers of Computer Science and Technology, Biejin, China, vol. 13, 2019, pp.214 - 225.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez et al., "Attention is all you need," *Advances in Neural Information Processing Systems* 2017, pp.5998-6008.

[6] A.V. Aho, M.J. Corasick, "Efficient string matching: an aid to bibliographic search," Communications of the ACM, vol. 18, 1975, pp.333-340.

[7] Rowley, Jennifer (2007). "The wisdom hierarchy: representations of the DIKW hierarchy". Journal of Information and Communication Science. 33 (2): 163–180.

[8] Y.C. Duan, L.X. Shao, G.Z. Hu, "Specifying Knowledge Graph with Data Graph, Information Graph, Knowledge Graph, and Wisdom Graph," IJSI vol.6, 2018, pp.10-25.

[9] Z.Y. Song, Y.C. Duan, S.X. Wan, X.B. Sun, Q. Zou, H.H. Gao, D.H. Zhu, "Processing Optimization of Typed Resources with Synchronized Storage and Computation Adaptation in Fog Computing," Wireless Communications and Mobile Computing , 2018, pp.1-13.

[10] L.X. Shao, Y.C. Duan, L.Z. Cui, Q. Zou, X.B. Sun, "A Pay as You Use Resource Security Provision Approach Based on Data Graph, Information Graph and Knowledge Graph," IDEAL, 2017, pp.444-451.

[11] Y.C. Duan, G.H. Fu, N.J. Zhou, "Everything as a Service (XaaS) on the Cloud: Origins, Current and Future Trends," IEEE 8th International Conference on Cloud Computing (CLOUD). IEEE, 2015.

[1] http://www.xywy.com