T-Know: A Knowledge Graph-based Question Answering and Information Retrieval System for Traditional Chinese Medicine

Ziqing Liu[1], Enwei Peng[2], Shixing Yan[2], Guozheng Li[3]✉, Tianyong Hao[4]✉
[1]Second School of Clinic Medicine, Guangzhou University of Chinese Medicine, Guangzhou, China emma.liu_el@hotmail.com
[2]AI Center, Green Valley, Shanghai, China pengenwei@jindengtai.cn, yanshixing@green-valley.com
[3]Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Science, Beijing, China gzli@ndctcm.cn
[4]School of Computer, South China Normal University, Guangzhou, China haoty@126.com

Abstract

T-Know is a knowledge service system based on the constructed knowledge graph of Traditional Chinese Medicine (TCM). Using authorized and anonymized clinical records, medicine clinical guidelines, teaching materials, classic medical books, academic publications, etc., as data resources, the system extracts triples from free texts to build a TCM knowledge graph by our developed natural language processing methods. On the basis of the knowledge graph, a deep learning algorithm is implemented for single-round question understanding and multiple-round dialogue. In addition, the TCM knowledge graph also is used to support human-computer interactive knowledge retrieval by normalizing search keywords to medical terminology.

1 Introduction

Traditional Chinese Medicine (TCM) is one of precious intangible cultural heritages of the Chinese nation. After thousands of years of development, it has been evolved as a distinct and unique theoretical medical system. Compared with disease treatment only, TCM pays more attention to living conditions and advocates timely adjustment of diet and rest to deal with physical discomforts. This philosophy has an advantage in dealing with sub-health status and chronic disease management (Wang, et al., 2007). With the increasing attention of health from public, the demands of reliable and convenient TCM knowledge services are increasing.

In recent years, there are great progresses being made in domain-specific knowledge graph construction. On TCM, a number of literature databases have been established. These digital resources containing rich TCM knowledge can be utilized to capture knowledge elements to serve public and benefit health management (Gao, et al., 2012). There are several existing TCM knowledge service platforms, such as TCMKS (Yu, et al., 2014). However, most of the systems target for medical professionals rather than public. A typical situation is that, their search functions allow formal TCM terms only, causing common users without TCM background difficulties to obtain required information without rich TCM background. Therefore, how to utilizing natural language methods to analyze informal or even vague queries for more convenient public services is an essential issue (Xu, et al., 2016).

To that end, we propose the T-Know, a user-friendly knowledge service system based on a TCM knowledge graph. The overall system architecture is shown as Figure 1.The system has two major modules: a question answering module and a knowledge retrieval module. The TCM knowledge graph integrates diversified data to enrich knowledge search and usage. The question answering module utilizes deep learning models to understand questions by analyzing question intents. The question answering module provides an interface for common users in both single-round question answering and multi-round dialogue ways. The knowledge retrieval module integrates TCM terminology and synonym dictionaries to extend search keywords semantically. The module can also navigate common users to use the TCM knowledge retrieval in an interactive way.

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 15–19
Santa Fe, New Mexico, USA, August 20-26, 2018.

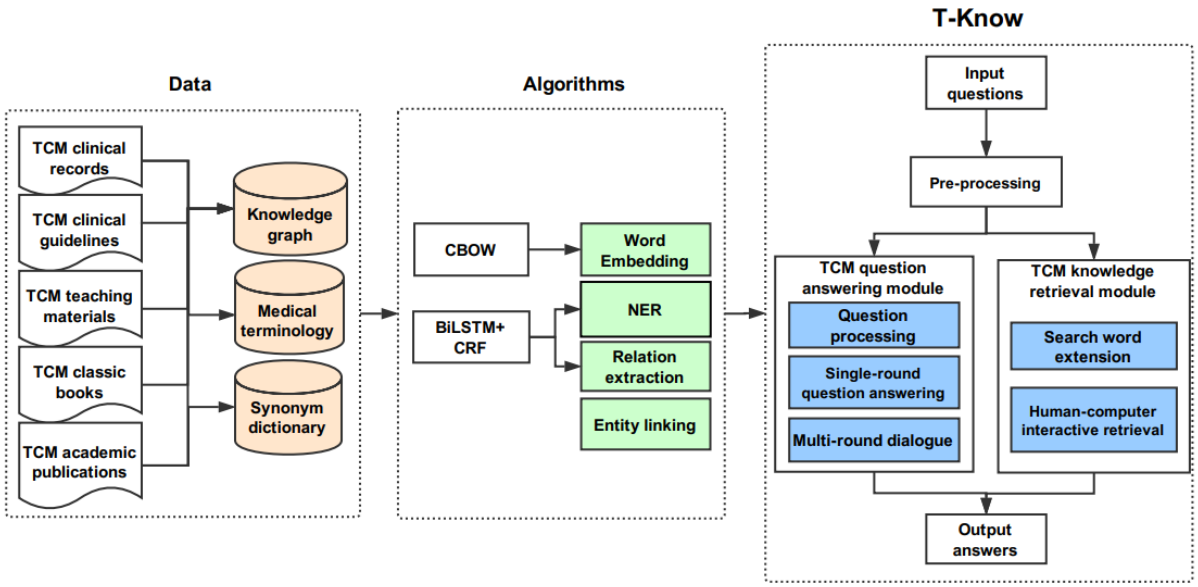

Figure 1: The overall system architecture of the T-Know system

2 The TCM Knowledge Graph

In order to construct the TCM knowledge graph, authorized and anonymized clinical records, clinical guidelines, teaching materials, classic medical books, and academic publications are collected as data resources. The unstructured texts were preprocessed including Chinese word segmentation, stop word removal and word semantic labelling. After that, medical named entity recognition and relation extraction were implemented using a Bi-LSTM+CRF algorithm to obtain <*Entity*, *Relation*, *Entity*> triples. Finally, the knowledge graph was verified and automatically constructed from the extracted triples.

The constructed TCM knowledge graph mainly contains five types of nodes: Diseases, Symptoms, Syndromes, Prescriptions and Chinese Herbals. There are diversified logical relations among the nodes. We also integrate a reasoning function besides the TCM knowledge graph to support entity or relation deduction through of the reasoning of logical relations among entity nodes. The constructed TCM knowledge graph contains more than 10,000 nodes and 220,000 relations currently.

3 The TCM Question Answering Module

1) Question Processing

Medical Named Entity Reorganization: We use a Bi-LSTM+CRF model to achieve named entity recognition, as reported in (Huang, et al., 2015), on different types of TCM texts. In the Bi-LSTM+CRF model, the first layer is a look-up layer, in which each word in a question sentence can be presented as a vector by using a pre-trained or randomly initialized word embedding matrix. The second layer is a bi-directional LSTM, which automatically extracts the characteristics of the sentence. Thus, the word embedding sequence of each word in the sentence can be the input of the bi-directional LSTM. It then splices implicit state sequence output by a forward LSTM in terms of locations to obtain a complete implicit state sequence. The third layer is a CRF layer, which labels sentence-level sequences.

Relation Extraction: Multi-channel convolutional neural networks (CNNs) are utilized to determine the relations between a pair of entities in a given free question (Xu, et al., 2016). Specifically, two CNNs channels are used. One is used to capture syntax information and the other is to capture context information. The convolutional layer of each channel accepts an input of variable length, while returns a vector of fixed length using the Maximum Sampling method. These fixed-length vectors are combined together to form the input of final softmax classifier, whose output vector dimension equals to the total number of relation categories and the value of each dimension equals to the degree of confidence mapped to the corresponding predicates in the knowledge graph.

2) Single-round Question Answering

Entity Linking: Entity linking plays a vital role in the TCM concept association and normalization (Liu, et al., 2016). After medical named entity identification, an entity linking tool named as S-MART is used to obtain associated entities in the TCM knowledge graph.

Joint Disambiguation of Entities and Relations: Under normal circumstances, both named entity recognition and entity relation extraction are independently predicted. The errors generated in the processes are usually difficult to avoid. We use a joint optimization model to select a globally optimal 'entity-relation' configuration from the candidate results of entity linking and relation extraction. The process of optimizing the globally configuration can be treated as a sorting problem in essence. To find 'reasonable' entity- relation configuration, TCM knowledge is applied to sort entity-relations, that is, the 'reasonable' entity-relation configuration should be more common in the TCM knowledge graph so as to efficiently locate answers to users' questions using the TCM knowledge graph.

3) Multi-round Dialogue

Multi-round dialogue refers to the management of multiple-round interactions while keeping context linkage. It contains entity linking for recognizing entities and joint disambiguation, through which, a topic and the certain slot are resolved and then stored. After that, the scope of the topic is identified. Given a new question, the system judges whether it follows the previous topic. If it follows, the question is regarded as in the same dialogue. Otherwise, it is treated as a new multi-round dialogue.

Forward facing centers strategy was adopted for anaphora resolution. When context information is insufficient to identify specific entity, a ranked list of discourse entities will be displayed to users along with inquiry. User's input straight after this list will be preferentially considered as the decision of entity assignment. If none of listed candidates is mentioned, then the entity of input question will be treated as constrain of last question to joint next round disambiguates. The screen snapshots of the whole TCM question answering module is presented as Figure 2.

Figure 2: The screen snapshots of the TCM question answering module (left: the multi-round dialogue; right: the single-round question answering)

4 The TCM Knowledge Retrieval Module

1) Search Word Extension

Search word extension provides users with a guided retrieval, that is, the entity node in the knowledge graph can be located according to search words and corresponding entity attributes can be returned as a retrieval guide. For example, if a user asking about a specific disease, the module returns with disease interpretation, the disease property and other relevant information. In addition to retrieving entities, the user can also retrieve relations. For example, if the user asking a syndrome of a specific disease, all the associated syndromes of the disease are regarded as relevant information and are returned. Moreover,

to assistant common users in the usage of the module, a list of TCM terminology and synonym dictionaries are integrated. The module will display matched synonyms of search words when a user simply type an informal medical term as a search word. The search word extension function is demonstrated as Figure 3.

Figure 3: An example of search word extension of TCM syndromes of headache

2) Human-computer Interactive Retrieval

The Human-computer interactive retrieval module consists of entity-relation visualization, relevant illustration and knowledge association. The entity-relation visualization assists common users to view all associated entities and relations for a specific entity. When a user clicks an entity node, the module automatically presents related entities whose distances below a specific threshold. The relations between the nodes also are displayed. When a node or link is clicked, related knowledge information will be extracted and be displayed as the detailed explanation of the node to users as relevant illustration. The knowledge association provides users with relevant knowledge to search content according to the classical TCM logic of 'Theory-Approach-Prescription-Medicine'. From the expertise, contents are associated to different categories. For instance, associated contents are similar disease or frequently co-occurrence disease when search content is a disease, and the associated contents are co-occurrence Chinese herbal or same efficacy Chinese herbal when search content is a Chinese herbal. Through this TCM association, the interactive retrieval module can serve common and professional users who has partial knowledge about TCM better. The interactive retrieval module is shown as Figure 4

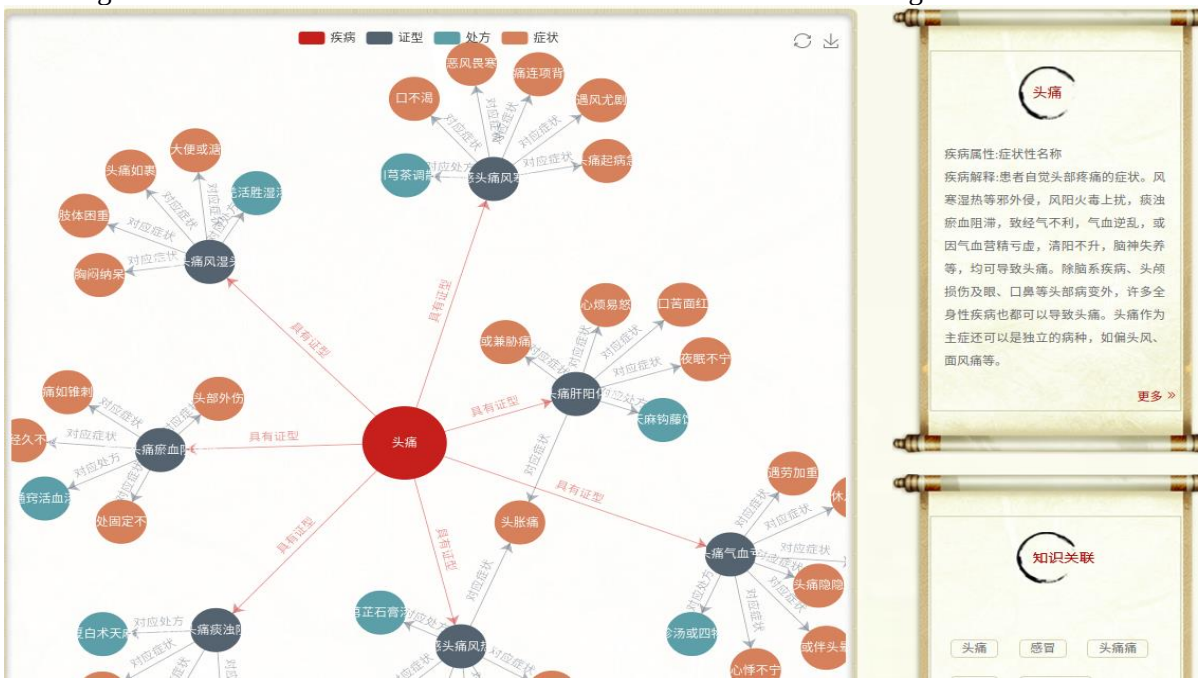

Figure 4: A knowledge graph about headache including TCM symptoms, syndromes, and prescriptions visualized by the human-computer interactive retrieval module

5 Summary

Aiming at improving the accessibility of TCM knowledge for general public. We proposed a knowledge based TCM question answering and information retrieval system named T-Know. Using heterogeneous medical texts as data resources, a TCM knowledge graph was automatically built. Based on the knowledge graph, T-Know delivers TCM question answering and knowledge retrieval services for public users via http://zhishi.jindengtai.cn:9999 .

Reference

Bo Gao, Meng Cui, Shuo Yang, Lirong Jia, Yan Dong, and Ling Zhu. 2012. Knowledge Services of TCM Based on Data. *Library and Information Service*, 56(9):5-9.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang and Dongyan Zhao. 2016. Qustion answering on freebase via relation Extraction and textual evidence. *In Processing of the 54th Annual Meeting of the Association for Computational Linguistics*, (pp. 2326-2336).

Tong Yu, Daming Su, Renfang Yin, Zhulv Zhang, and Ye Tian. 2014. Research on the Construction of Knowledge Services Platforms for Traditional Chinese Medicine. *Medical Innovation of China*, 11(15):120-123.

Yanhui Wang, and Kuanqi He. 2007. Advantages of Traditional Chinese Medicine in Diagnosis and Treatment of Subhealth State. *China Journal of Traditional Chinese Medicine and Pharmacy*, 22(7):473-475.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Zhiyuan Liu, Maosong Sun, Yankai Lin, and Ruobing Xie. 2016. Knowledge Representation Learning: A Review. *Journal of Computer Research and Development*, 53(2):247-261.