TCMKG: A Deep Learning Based Traditional Chinese Medicine Knowledge Graph Platform

1st Ziqiang Zheng
Knowledge and Data Engineering Laboratory of Chinese Medicine
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
zhengzq@std.uestc.edu.cn

2nd Yongguo Liu
Knowledge and Data Engineering Laboratory of Chinese Medicine
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
liuyg@uestc.edu.cn

3rd Yun Zhang
Knowledge and Data Engineering Laboratory of Chinese Medicine
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
yunzhangwww@163.com

4th Chuanbiao Wen
College of Medical Information Engineering
Chengdu University of Traditional Chinese Medicine
Chengdu, China
wcb@cdutcm.edu.cn

***Abstract*—As an effective and novel knowledge management technology, knowledge graph can provide a new way for the inheritance and development of traditional Chinese medicine (TCM). However, the construction of the knowledge graph of TCM is still mainly based on structured data at present. With the accumulation of literatures and electronic medical records, a large amount of knowledge is stored in unstructured texts which urgently needs to be extracted for learning. In this study, we extract TCM core concepts and build ontology layer by analyzing the process of TCM diagnosis and treatment. Then we use deep learning to extract entities and their relations for building TCM knowledge graph from unstructured data. Finally, we build an end-to-end platform TCMKG based on knowledge graph, which can provide functions such as knowledge retrieval, visualization and data management for helping the learning and sharing of TCM knowledge.**

Keywords—knowledge graph, platform, traditional Chinese medicine, deep learning, ontology

I. INTRODUCTION

In the past few decades, with the release of more and more semantic data, the number of semantic data sources has proliferated and linked open data is growing in size. Different researches have been proposed many scalable knowledge acquisition techniques for building large-scale knowledge graphs (KGs). Knowledge graph is based on triples (two entities and their relation). The general knowledge graph is generally constructed from the bottom up. The domain knowledge graph is usually constructed from the top down by defining the ontology to organize the knowledge structure. Large knowledge graphs are public and available, such as Dbpedia [1], YAGO [2], Freebase [3], BabelNet [4], NELL [5], Baidu "Zhixin" and Sogou "Zhishilifang". Currently, there are many researches try to build KGs in various fields [6-8]. For example, in medical domain, many researchers have published multiple linked data and some linked data platforms have integrated most of these data [9,10]. The most well-known biomedical dataset platforms include Linked Life Data [11] and Bio2RDF [12]. Chinese medicine carries the experience and theoretical knowledge of doctors for treating diseases over thousands of years of history [9]. Knowledges graphs are gradually accepted by Chinese medicine doctors because graphs have the characteristics of knowledge semantics, data association and easiness to understand. However, current knowledge graphs are mainly constructed based on structured data. However, there is a great quantity of tacit knowledge are in the literature and books.

With deep learning achieved remarkable results in the field of video, image, voice, and so on, researchers have begun to apply deep learning to text extraction, including entity recognition, attribute extraction, relationship extraction, event extraction, etc [13-15]. Deep learning provides a convenient way to construct knowledge graph triples by extracting knowledge from unstructured data [16]. Due to the heterogeneous nature of internet big data, the knowledge extracted from unstructured data is highly repetitive, which causes difficulties in knowledge management [17]. Based on knowledge graph, these information resources can be semantically labeled and linked to establish a knowledge-centric resource semantic integration service to organize data more conveniently.

In this paper, we analyze the basic theories and diagnosis and treatment process of TCM to define fuzzy relations among disease, symptom, syndrome, therapy, prescription and herb. An ontology layer is created to represent the knowledge-based diagnosis and treatment process. Then we use a named entity recognition (NER) model to extract TCM entities (knowledge) from unstructured data. We align the extracted knowledge with knowledge graph and create triples based on the ontology layer. The fusion of triples and the knowledge graph of TCM can provide auxiliary reference for the diagnosis and treatment of Chinese medicine doctors. Finally, we aim to build an end-to-end platform from unstructured data to knowledge graph to provide some functions to assist doctors.

The rest of the paper is organized as follows. Section 2 outlines related work. Section 3 gives a detailed introduction of the overall architecture of platform and how to build the platform. In addition, we also give a demonstration of a web-based TCMKG platform and realizes the functions of crawler, text parsing, data management, knowledge retrieval and visualization. Section 4 gives a summary of our research.

II. RELATED WORK

The construction of the ontology in medicine field requires in-depth analysis of the structure and concept of medical terminology to express obscure and cross-lingual medical knowledge effectively. Gu et al. [18] created an ontology-oriented diagnosis system to address the knowledge-based diagnosis based on a well-defined syndrome ontology. Zhang et al. [19] proposed the representation and construction method of core knowledge graph based on the ontology of TCM. Yu et al. [10] proposed the technology of construction

978-1-7281-8156-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICBK50248.2020.00084

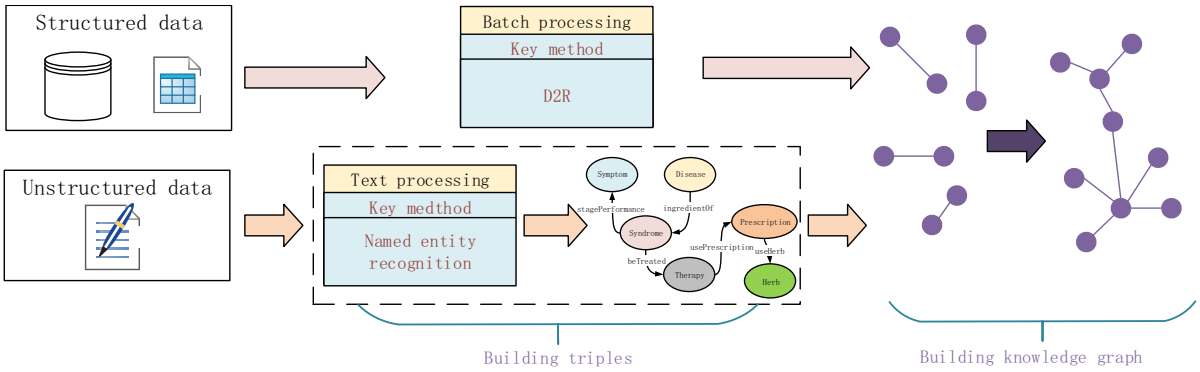

Fig. 1. The overall process.

a large-scale TCM knowledge graph with traditional Chinese medicine language system. These methods demonstrate the benefits of building specific ontology by enabling the standardized terminology description of scientific domain and smooth data interchange on the web.

The construction of TCM knowledge graphs needs to deal with the multi-source of the medical knowledge among the webs, relational databases, clinical information, books and literatures. Jia et al. [20] discussed the data sources, research contents and application prospects of TCM knowledge graph architecture. Ruan et al. [21] proposed the semi-automated construction process of TCM knowledge graph by a multi-strategy learning approach and realized the intelligent application for assisted prescribing. Miao et al. [9] analyzed large amounts of data from different sources and combined the existing TCM knowledge bases to construct a TCM prescription knowledge graph. In this paper, we build TCMKG platform by combing deep learning techniques to achieve end-to-end application of unstructured data to knowledge graph.

III. Construction of TCMKG

This section introduces the construction of TCMKG. The overall process is shown in Fig. 1. We first design the ontology layer, then select high-quality data sources and manually label data to train a NER model to discover TCM entities. Then we align the extracted data with KG and finally built the triples by the ontology layer to complete the construction of the graph.

A. Ontology Construction

Ontology was originally a philosophical concept. Domain ontology defines the scope of a specific research area, which is a common knowledge used to describe specific professional field and define core concepts, entities, events, hierarchical relationships between concepts and sub-domain relationships. The domain ontology can capture core concepts in specific field and the interrelationships among these concepts. Building domain ontology can standardize the relationships among core concepts. We select Protégé as an ontology modeling tool and write a program to implement automatic data filling. As shown in Fig. 2, we define six basic classes of disease, syndrome, symptom, therapy, prescription, and herb and create five semantic relations under the guidance of the professional physician. The semantic relationships are shown in table I.

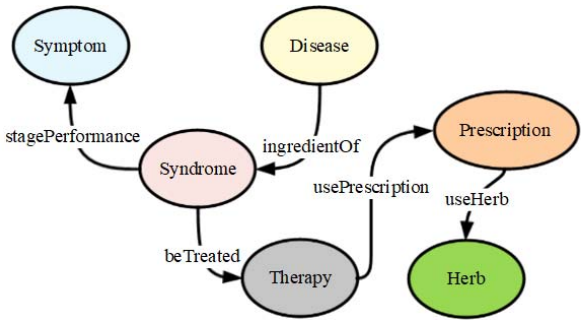

Fig. 2. Ontology layer of TCM diagnosis and treatment.

TABLE I. Semantic Relation

Semantic Relationship	Express
stagePerformance	Syndrome-Symptom
ingredientOf	Disease-Syndrome
beTreated	Syndrome-Therapy
usePrescription	Therapy-Prescription
useHerb	Prescription-Herb

B. Data Processing

As there are no public TCM publication dataset [13], we build a TCM dataset. We obtain 500 medical records from Chinese medicine books as a corpus and annotate these medical records under the guidance of medical professionals by the annotation software provided in [22]. We define 13 types of entities and assertions, including the six entities shown in Fig. 1. An example of text annotation is shown in Fig. 3.

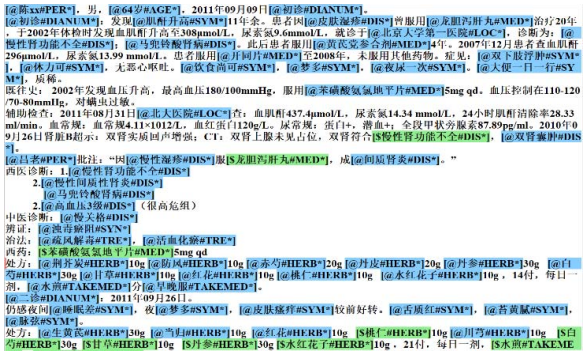

Fig. 3. An example of text annotation.

After text annotation, we convert the data into sequence annotation data and divide it into training set, test set and validation set according to the ratio of 7:2:1. We use the character-based BiLSTM-CRF model to identify named entities. F1 score of the NER model is above 97%. Then we built triples by writing a script in Python to implement automatic data filling based on the ontology later. For the structured data such as knowledge base, we use D2R tools to achieve batch processing of data cleaning and conversion.

C. Data Fusion

Data fusion is the most critical step in process of building KGs and is related to the accuracy of domain knowledge. Our target is to align the entities of the same category with the same meaning. There are multiple meanings of named entities in Chinese medicine. For example, "Obesity" is a disease entity as well as a symptom entity. In order to avoid entity ambiguity, we use Chinese word embedding model ssp2vec [23] to map entities in TCM to vectors. Then a similarity threshold is set to align entities based on the similarity between two entity vectors. When the similarity is greater than the threshold, we merge the two entities, otherwise they are considered as two different entities. An example of knowledge graph for TCM diagnosis and treatment constructed after data fusion is shown in Fig. 4.

D. Platform Architecture

The objective of this platform is to implement end-to-end applications from unstructured data to knowledge graphs. Therefore, the performance, agility and adaptability of the platform are crucial. To achieve these abilities, the platform has multi-layer architecture, such as data collection and storage, parsing and knowledge graph applications, which is shown in Fig. 5.

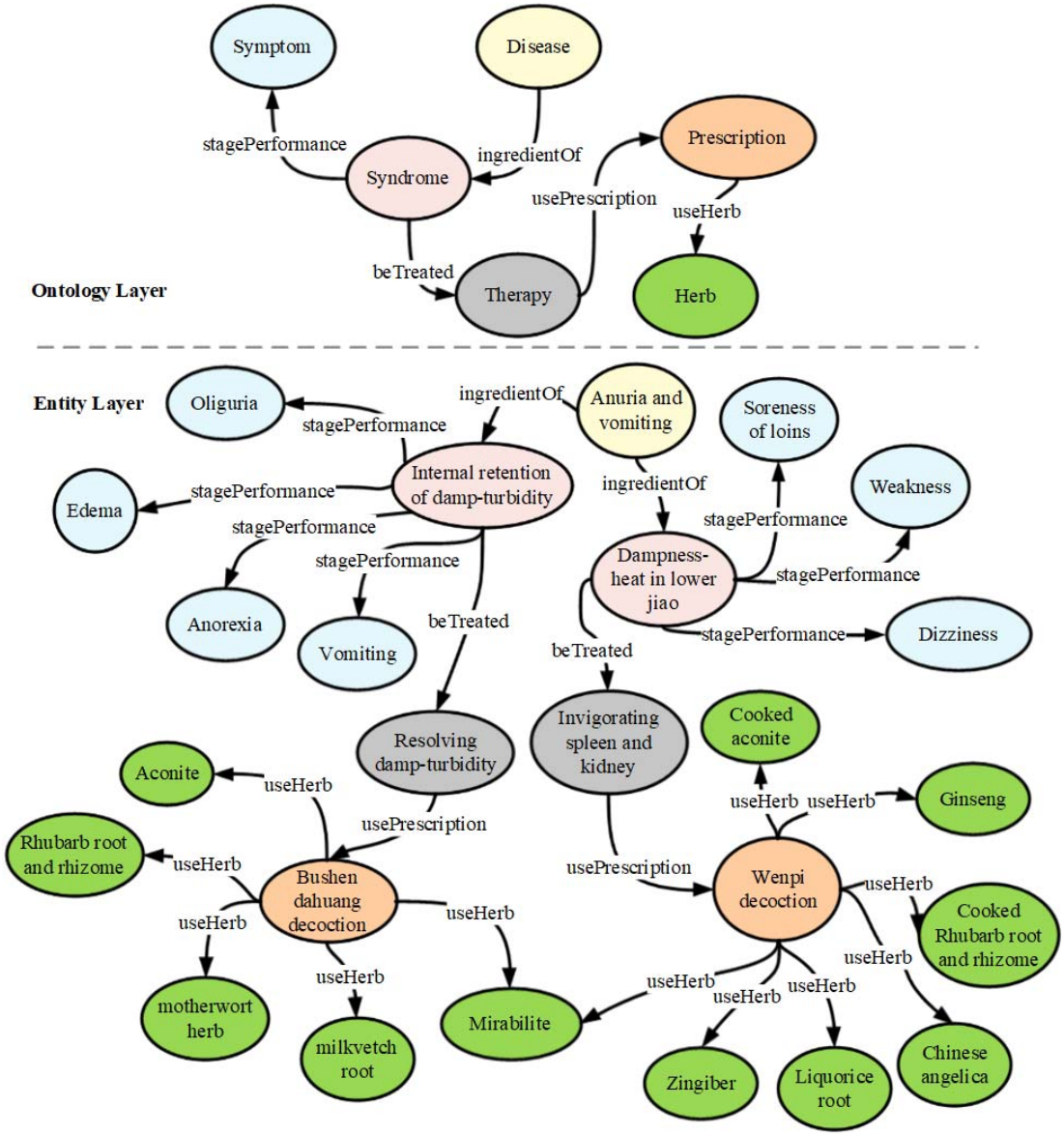

Fig. 4. An illustrative example of knowledge graph for TCM diagnosis and treatment.

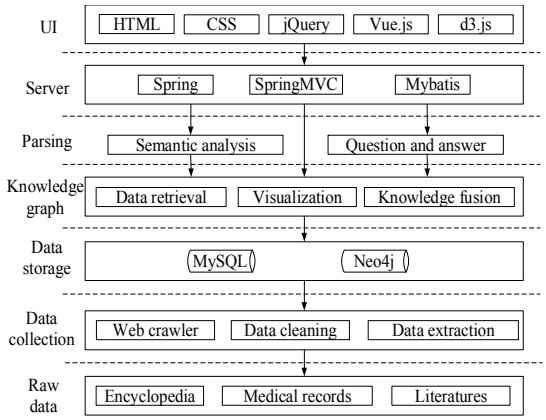

Fig. 5. The overview of platform architecture.

In Fig. 5, the lowest layer represents raw data, including encyclopedia data, electronic medical records, literatures and books. In data collection layer, we use web crawler to crawl encyclopedia data and process them as structured data. Then, we use NER model to extract entities from electronic medical records, literatures and books. In order to facilitate the implementation of the platform, we simply perform data fusion by string matching. The integrated knowledge is stored in neo4j and MySQL. Based on the two databases, we have implemented several applications, such as data retrieval, visualization and knowledge fusion.

E. Platform Implementation

As shown in Fig. 5, the platform is developed based on the SSM framework and all functions are implemented in the environment of java 1.8 and python 3.6.

1) Crawler: In order to achieve knowledge acquisition, we integrate a crawler script in the platform for real-time access to literatures and provide download function. The crawler function is shown in Fig. 6.

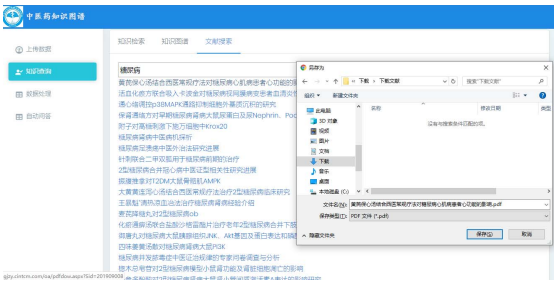

Fig. 6. The interface of crawler.

2) Parsing: The function of parsing is to convert books or literatures into structured data and add knowledge graph data dynamically. We unify all unstructured data into PDF format. Then we use OCR text recognition technology to read the contents of PDF and convert them into an operable string sequence. As shown in Fig. 7, we use the NER model to process string sequence and provide parsing results and download functions.

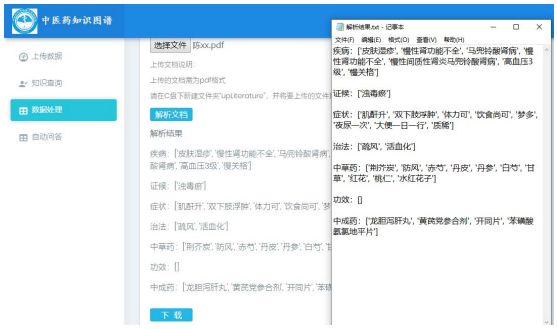

Fig. 7. The interface of parsing results.

3) Knowledge retrieval and Visualization: Data visualization is used to visualize the query results based on knowledge graph, including TCM knowledge query and the diagnosis and treatment path of TCM. As shown in Fig. 8, this platform can show the entire diagnosis and treatment path of Disease-Syndrome-Therapy-Prescription in TCM. Users can click on the nodes in knowledge graph to view the corresponding attributes, which can help users to learn the complete knowledge of TCM diagnosis and treatment.

Fig. 8. An visualization example of diagnosis and treatment path of TCM.

4) Data Management: For graph database, the relationship must be added to existing nodes. Meamwhile, the relationship is modified by deleting the old relationship and adding a new relationship. Therefore, the data management function of knowledge graph can be divided into the addition, deletion, modification of nodes and the addition and deletion of relationships. We use Cypher query to achieve neo4j data management. This platform also provides addition of batch triples as shown in Fig. 9.

Fig. 9. The interface of addition of batch triples.

IV. Conclusions

In this paper, we deploy a deep learning-based TCMKG platform to implement the end-to-end application of unstructured data to knowledge graph. In addition, the platform also can provide batch processing of structured data. We integrate the ontology layer as the logical framework of knowledge graph, NER model as a bridge for data conversion, and crawler script as a real-time data acquisition tool as a multi-layer architecture. This platform not only provides knowledge retrieval, but also provides a knowledge path display of the complete system structure of TCM diagnosis and treatment, which is of great help to user learning. In the future work, we are going to use more corpus to train the NER model to improve the accuracy of entity recognition. In addition, we want to design new algorithms for data fusion to improve the quality of knowledge graph. Finally, based on the constructed TCMKG, we will develop more practical functions, such as question answering, knowledge recommendation, etc.

Acknowledgment

This research was supported in part by the National Key R&D Program of China under grants 2019YFC1710300 and 2017YFC1703905, and the Sichuan Science and Technology Program under grants 2020YFS0372 and 2018SZ0065.

References

[1] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, and P. Mendes, "DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia," Semantic Web, vol. 6, pp. 167-195, January 2015.

[2] F.M. Suchanek, G. Kasneci, and G. Weikum. "Yago: a core of semantic knowledge," in Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, May 8-12, 2007, pp. 697-706.

[3] K.D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in Proceedings of the ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, June 10-12, 2008, pp. 1247-1250.

[4] R. Navigli and S.P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, July 11-16, 2010, pp. 216-225.

[5] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R.H. Jr., and T.M. Mitchell, "Toward an architecture for never-ending language learning," in Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, Georgia, USA, July 11-15, 2010, pp. 1306-1313.

[6] H. Lian, Z.M. Qin, T.K. He, and B. Luo, "Knowledge graph construction based on judicial data with social media," WISA 2017, Liuzhou, Guangxi, China, November 11-12, 2017, pp. 225-227.

[7] Q. Zhang, Y.Q. Wen, C.H. Zhou, H. Long, D. Han, F. Zhang, and C.S. Xiao, "Construction of knowledge graphs for maritime dangerous goods," Sustainability, vol. 11, p. 2849, May 2019.

[8] Q.L. Miao, Y. Meng, and B. Zhang, "Chinese enterprise knowledge graph construction based on linked data," in Proceedings of the 9th IEEE International Conference on Semantic Computing, Anaheim, CA, USA, February 7-9, 2015, pp. 153-154.

[9] F. Miao, H.X. Liu, Y.M. Huang, C.M. Liu, and X.Y. Wu, "Construction of semantic-based traditional Chinese medicine prescription knowledge graph," IAEAC 2018, Chongqing, China, December 18-19, 2018, pp. 1194-1198.

[10] T. Yu, J.H. Li, Q. Yu, Y. Tian, X.F. Shun, L.L. Xu, L. Zhu, and H.J. Gao, "Knowledge graph for TCM health preservation: Design, construction, and applications," Artificial Intelligence in Medicine, vol. 77, pp. 48-52, March 2017.

[11] V. Momtchev, D. Peychev, T. Primov, and G. Georgiev, "Expanding the pathway and interaction knowledge in linked life data," Proc. of International Semantic Web Challenge, January 2009.

[12] F. Belleau, M.A. Nolin, N.Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems," Journal of Biomedical Informatics, vol. 41, pp. 706-716, October 2008.

[13] J.C. Wang and J. Poom, "Relation extraction from traditional Chinese medicine journal publication," IEEE International Conference on Bioinformatics and Biomedicine, Shenzhen, China, December 15-18, 2016, pp. 1394-1398.

[14] Q. Zhang, J.L Fu, X.Y. Liu, and X.J. Huang, "Adaptive co-attention network for named entity recognition in tweets," in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018, pp. 5674-5681.

[15] M.B. Xu, H. Jiang, and S. Watcharawittayakul, "A local detection approach for named entity recognition and mention detection," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, July 30 - August 4, 2017, pp. 1237-1247.

[16] S.C. Zheng, Y.X. Hao, D.Y. Lu, H.Y. Bao, J.M. Xu, H.W. Hao, and B. Xu, "Joint entity and relation extraction based on a hybrid neural network," Neurocomputing, vol. 257, pp. 59-66, September 2017.

[17] B.X. Shi and T. Weninger, "Open-world knowledge graph completion," in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018, pp. 1957-1964.

[18] P.Q. Gu, H.J. Chen, and T. Yu, "Ontology-oriented diagnostic system for traditional Chinese medicine based on relation refinement," Computational and Mathematical Methods in Medicine, vol. 2013, pp. 317803:1-317803:11, February 2013.

[19] D.Z. Zhang, Y.H. Xie, M. Li, and C. Shi, "Construction of knowledge graph of traditional Chinese medicine based on the ontology," Information Engineering, vol. 3, pp. 35-42, January 2017.

[20] L.R. Jia, J. Liu, T. Yu, Y. Dong, L. Zhu, B. Gao, and L.H. Liu, "Construction of traditional chinese medicine knowledge graph," Journal of Medical Informatics, vol. 36, pp. 51-59, August 2015.

[21] T. Ruan, C.L. Sun, H.F. Wang, Z.J. Fang, and Y.C. Yin, "Construction of Traditional Chinese medicine knowledge graph and its application," Journal of Medical Informatics, vol. 37, pp. 8-13, April 2016.

[22] J. Yang, Y. Zhang, L.W. Li, and X.X. Li, "YEDDA: A lightweight collaborative text span annotation tool," in Proceedings of ACL, Melbourne, Australia, July 15-20, 2018, pp. 31-36.

[23] Y. Zhang, Y.G. Liu, J.J. Zhu, Z.Q. Zheng, X.F. Liu, W.G. Wang, Z.J. Chen, and S.Q. Zhai, "Learning Chinese word embeddings from stroke, structure and pinyin of characters," in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, November 3-7, 2019, pp. 1011-1020.