Study on Knowledge Map Construction of Fu Xing Jue Based on OCR-NLP Collaborative Resolution Technology

Bo Hou [1], Hao Luo [1, 3, 4, *], Yang Li [2], HanYu Shi [1]

[1] School of Medical Information Engineering, Gansu University of Chinese Medicine, China, 730101

[2] School of Public Health, Gansu University of Chinese Medicine, China, 730101

[3] School of Electronic and Information Engineering, Lanzhou Jiaotong University, Gansu, China, 730000

[4] Key Laboratory of Media convergence Technology and Communication, Gansu Province, Gansu, China, 730000

* Corresponding Author Email: luoh804@gszy.edu.cn

Abstract. Against the backdrop of the digitization and intelligence of traditional medicine, this study employs the collaborative technology of optical character recognition and natural language processing to address the inheritance challenge of the Dunhuang medical literature "fu Xing Jue Zang-fu Organ Medication Method Yao", and constructs a "prescription - drug" knowledge graph. It mainly expounds the processes of data preprocessing, knowledge extraction and fusion, as well as graph construction. Through the application of the two techniques, the role of the two techniques in the transformation of Dunhuang medicine was revealed. The research results show that the integrated architecture of the two technologies not only has significant value in medical research, but also provides a new paradigm for the modernization, inheritance and innovation of Dunhuang medicinet.

Keywords: OCR technology, NLP technology, knowledge atlas, Fu xing formula Zangfu medicine essentials, Dunhuang medicine.

1. Introduction

Dunhuang medicine contains ancient medical theories, prescriptions and other wisdom [1], but its literature faces the problems of fragmentation and inheritance due to its long history. Domestically, in the early days, the focus was mostly on literature collation. Such early research was mostly qualitative analysis, and there were deficiencies in the mining of quantitative rules for the compatibility of prescriptions[2].Currently, OCR and NLP technologies have made initial attempts in the processing of traditional Chinese medicine texts. However, when dealing with content like the Dunhuang medical literature, which contains a large number of ancient and heterogeneous versions, An effective collaborative analysis method has not yet been formed, and the precise calibration of multi-version semantics is even a research shortcoming[3].However, foreign research is scarce and mostly focuses on the analysis of cross-cultural medical techniques. In this study, OCR technology and NLP technology are integrated to extract the relevant information of "prescriptions-drugs", and a knowledge graph is constructed to mine the rules of prescriptions and drugs. Based on this, the main purpose of this study is to solve the following two core problems:

(1)Against the backdrop of the complex original content and numerous versions of the "Auxiliary Practice Formula", how can we scientifically extract and integrate knowledge from it to prevent content errors? This study aims to conduct specific entity extraction and relation extraction through BILSTM related models to ensure the accuracy of data layer construction[4].

(2)In the field of Dunhuang medical transformation, considering the vast and diverse attributes of Dunhuang medicine, The purpose of this paper is to enable OCR-NLP collaborative analysis architecture to digitize Dunhuang medicine, and provide replicable experience and inheritance

paradigm for other related Dunhuang medical classics atlas construction through the relevant models used in this paper, so as to assist the modern inheritance and innovation of Dunhuang medicine.

2. Methods

2.1. Data acquisition and preprocessing

In order to construct the data layer of the system, visit the National Library and Gansu University of Traditional Chinese Medicine Library respectively and representative versions and annotated literature were selected as core data. By fully digitically collecting paper literature and using image enhancement algorithms to improve the recognition of text[5], and referring to multi-version reasoning to restore the details of blurred text. After cleaning and confirmation, there was no missing or duplicate data, and the integrity and reliability met the standards, laying the foundation for subsequent research to reveal the role of the OCR-NLP technology fusion architecture in the medical transformation of Dunhuang.

2.2. Method introduction

(1)OCR technology: It is a technique that uses optical equipment to scan characters in paper documents or images, and then converts them into electronic text through algorithm analysis and processing. It transforms the text of the ancient book "Auxiliary Fu Xingjue" into processable text, providing data for subsequent analysis.

(2) Named Entity Recognition technology: It is a fundamental task in natural language processing, aiming to automatically identify and label entities and their categories with specific meanings from text. Accurate identification of the physical entities of traditional Chinese medicines and prescriptions is the foundation of knowledge graph construction [6].

(3) Dependency syntactic analysis: It is a task in natural language processing to analyze the grammatical structure of sentences, revealing the syntactic levels and semantic associations of sentences by determining the dependency relationships between words. Explore the "composition" relationship between drugs and prescriptions in the composition sentences of prescriptions.

(4) Entity extraction based on BiLSTM-CRF [7]:

$$\left[\mathrm{score}(\mathrm{x},\mathrm{y}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}_{\mathrm{i},\mathrm{y}_\mathrm{i}} + \sum_{\mathrm{i}=1}^{\mathrm{n}+1} \mathrm{A}_{\mathrm{y}_{\mathrm{i}-1},\mathrm{y}_\mathrm{i}}\right] \quad (1)$$

In the formula, P_{i,y_i}:indicates that the its word in the text is labeled as (y_i) category, and its emission probability measures the direct association between the word and the category[8] (A_{y_{i-1},y_i}): represents the transition probability from the category $(\mathrm{i}-1)$ of the (y_{i-1}) word to the category (y_i) of the (i) word, reflecting the logical order between categories. The overall $score\ (x,y)$ reflects how well the label sequence y fits the input text.

$$\left[\mathrm{P}(\mathrm{y}|\mathrm{x}) = \frac{\exp(\mathrm{score}(\mathrm{x},\mathrm{y}))}{\sum_{\mathrm{y}'} \exp(\mathrm{score}(\mathrm{x},\mathrm{y}'))}\right] \quad (2)$$

This formula converts the $score\ (x,y)$ into a probability for the label sequence, normalizing the scores for all possible label sequences (y'). In knowledge extraction, the conditional probability of a label sequence y under a given text x is calculated. The higher the probability, the more reasonable the label.

$$\left[\log P(y^{\mathrm{x}}|x) = score(x,y^{\mathrm{x}}) - \log\left(\sum_{y'} \exp\left(score(x,y')\right)\right)\right] \quad (3)$$

This is the training objective to optimize the model parameters by maximizing the logarithmic probability of correctly labeling the sequencey'. Adjust $\left(P_{iy_j}\right)$ and $\left(A_{y_{i-1},y_j}\right)$ to make the model more inclined to output labels that conform to the knowledge system of "Essentials of Fu xing Jue Zangfu Medicine"[8].

$$[y^* = \text{argmaxscore}(x, y')] \tag{4}$$

For the new auxiliary line formula text x, calculate the scores of all possible label sequences $(y)'$, and select the highest score (y^*) as the prediction label[9]. By comparing the scores of different labels, the model outputs the label sequence that best accords with the auxiliary knowledge, and completes the information extraction.

(5) Relation Extraction Based on BilSTM:

$$\text{it} = \sigma(\text{Wxixt} + \text{Whiht} - 1 + \text{Wcict} - 1 + \text{bi}) \tag{5}$$

Forget gate filters out irrelevant information and retains memories of "constituent drugs."

$$\text{f t} = \sigma(\text{W x f x t} + \text{W h f h t} - 1 + \text{W c f c t} - 1 + \text{b f}) \tag{6}$$

When processing "ginger three liang", the input gate focuses on "ginger" and "three liang", thus ignoring the interference of quantifier "each".

$$\text{ct} = \text{i t g t} + \text{f t c t} - 1 \tag{7}$$

This is the cell state renewal formula, which gradually accumulates the information of the constituent drugs of "Xiao Xie Gan Tang" into the cell state for subsequent relationship judgment.

$$\text{ot} = \sigma(\text{Wxoxt} + \text{Whoht} - 1 + \text{Wcoct} + \text{bo}) \tag{8}$$

This is the output formula. When encountering the character "Treat", the output gate activates the related features of the "main therapeutic relationship", providing a basis for subsequent classification.

3. Data layer construction

3.1. Knowledge extraction

Knowledge extraction mainly covers two major dimensions:entity extraction and relation extraction[3], The entity extraction stage focuses on the identification of entities named after traditional Chinese medicine and those named after prescriptions[10]. Among them, the extraction of traditional Chinese medicine entities adopts the BiLSTM-CRF architecture model to achieve the precise extraction of drug names[11]. The extraction of prescription entities is based on the semantic analysis of the original text context of "Auxiliary Fu Xingjue", comprehensively considering the text context information to improve the accuracy and completeness of the extraction.Relation extraction, on the other hand, explores the logical connections between" prescriptions and drugs" and analyzes the complex relationships among entities. As shown in Figure 1, by using dependency syntactic analysis techniques, with the association type as the horizontal axis and the proportion of relations as the vertical axis, it indicates that the proportion of "constituent relations" is the highest, providing a core basis for relation alignment.

Figure 1. The main relationship between prescription drugs and their proportion

3.2. Knowledge fusion

In terms of entity disambiguation [12], to address the issue of multiple expressions of the same entity or similar expressions of different entities in the "Auxiliary Practice Formula", through comparison with professional dictionaries of traditional Chinese medicine and semantic analysis of the text context, the entity representations are unified. When aligning relationships, in view of the situation where the relationships among entities are complex, the expressions are different but the semantics are similar, based on the knowledge ontology in the field of traditional Chinese medicine, ensure that the relationship structure of the knowledge graph is clear and standardized.

In summary, through knowledge extraction and fusion, simply taking a category in Table 1 as an example, the structured correlation information of "category-prescription-drug-treatment" is presented and displayed from four dimensions, providing triplet data for the subsequent process.

Table 1. Knowledge table for extracting attributes, prescriptions, medicines, and treatments

Category of Prescriptions	Name of Decoction	Ingredients of Decoction	Processing Method
Prescriptions for Differentiating Liver Diseases	Xiao Xiegan Decoction	Immature Bitter Orange (Zhishi), Peony (Shaoyao), Fresh Ginger (Shengjiang)	Wash the herbs, soak for 30-60 minutes, add appropriate water to decoct. Boil with strong fire first, then turn to gentle fire and decoct for 20-30 minutes. Take about 300 ml of the decoction, warm and take in 2-3 times, one dose per day.
Prescriptions for Differentiating Liver Diseases	Da Xiegan Decoction	Immature Bitter Orange (Zhishi), Peony (Shaoyao), Fresh Ginger (Shengjiang), Scutellaria (Huangqin), Rhubarb (Dahuang), Licorice (Gancao, roasted)	First, soak Zhishi, Shaoyao, Shengjiang, Huangqin and Gancao for 30-60 minutes, then add Dahuang and decoct together. Boil with strong fire first, then turn to gentle fire and decoct for 15-20 minutes (Dahuang should be added later to avoid weakening the purgative effect due to prolonged decocting). Take about 300-400 ml of the decoction, warm and take in 2-3 times, one dose per day.
Prescriptions for Differentiating Liver Diseases	Xiao Bugan Decoction	Cinnamon Twig (Guizhi), Dried Ginger (Ganjiang), Schisandra (Wuweizi), Jujube (Dazao, broken)	Wash the herbs, break the Dazao, soak for 30-60 minutes, add appropriate water to decoct. Boil with strong fire first, then turn to gentle fire and decoct for 30-40 minutes. Take about 300-350 ml of the decoction, warm and take in 2-3 times, one dose per day.
Prescriptions for Differentiating Liver Diseases	Da Bugan Decoction	Cinnamon Heart (Guixin), Dried Ginger (Ganjiang), Schisandra (Wuweizi), Inula Flower (Xuanfuhua, one formula uses Moutan Bark - Mudanpi), Hematite (Daizheshi, burned red, quenched in vinegar three times, crushed), Bamboo Leaf (Zhuye), Jujube (Dazao, broken)	First decoct Daizheshi for 20-30 minutes, then add the rest of the herbs which have been washed and soaked for 30-60 minutes, and decoct together. Boil with strong fire first, then turn to gentle fire and decoct for 30-40 minutes. Take about 350-400 ml of the decoction, warm and take in 2-3 times, one dose per day.

3.3. Atlas visualization

In summary, through the construction of the data layer and with the help of the Neo4j graph database, the triplet data has been transformed into a visual and interactive knowledge graph, as shown in Figure 2. Nodes of different colors represent the prescriptions and drugs in the "Auxiliary Practice Formula", and the lines connecting the nodes indicate the relationship associations. This diagram can visually present the relationships among various elements within the knowledge system of "Fu Xing Jue", and assist in analyzing the rules of drug compatibility, disease associations, etc.

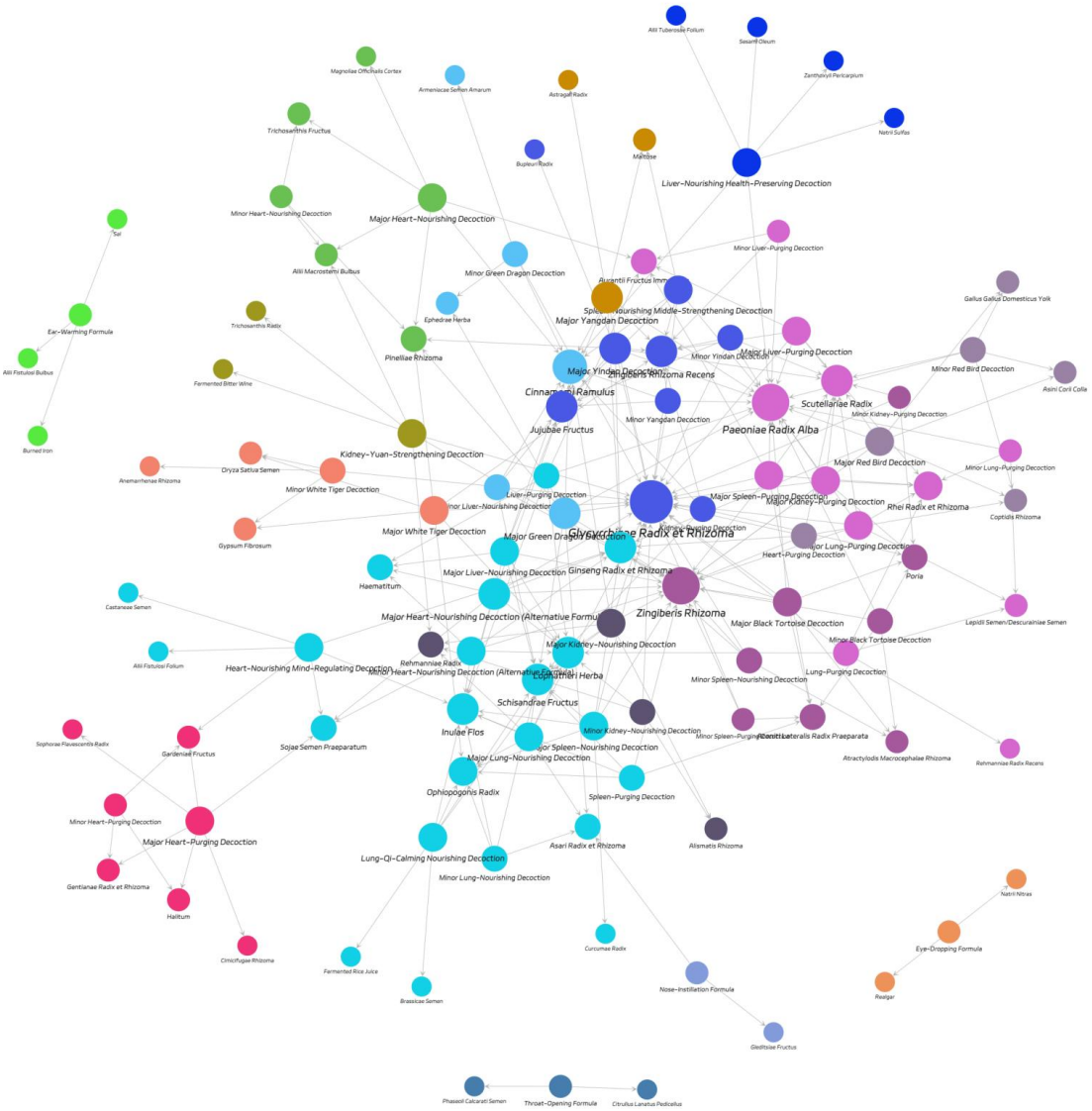

Figure 2. "Fuxing Formula" prescription drug compatibility knowledge atlas

4. Conclusion

The rise of digitization in traditional medicine is closely related to emerging digital technologies. This study analyzed the "Prescription-Drug" knowledge graph of "Fu xingjue" constructed through two techniques, QCR-NLP, and deeply analyzed the compatibility rules among drugs in the prescription. Meanwhile, the research process indicates that the integrated architecture based on the two technologies can be diffused throughout the transformation process of Dunhuang medicine, providing replicable experiences and paradigms for its inheritance and innovation, and promoting the dissemination and inheritance of traditional Chinese medicine culture. At the same time, this paper expounds the process of constructing the atlas of ancient books, hoping that it can be applied to more Dunhuang medical books and help to construct digital knowledge of traditional Chinese medicine.

Acknowledgements

This work is supported by the Natural Science Foundation for Young Scientists of Gansu Province(Grant No. 22JR5RA595), the Youth Project of Social Science Planning of Gansu Province(Grant No. 2021QN027), and the Scientific Research and Innovation Fund Project of Gansu University of Chinese Medicine(Grant No. 2021KCYB-10).

References

[1] Zhang L. Research on "General Micro-Prescriptions for Pediatric Health" of the Southern Song Dynasty [D]. Zhengzhou University, 2018.

[2] Lin SY, Qu YQ, Liu C, et al. Review of the Development of Artificial Intelligence in Traditional Chinese Medicine and Discussion on the Trend of Technology Integration. [J]. Chinese Journal of Traditional Chinese Medicine, 2020, 35(11): 5384–5389.

[3] Xie T, Yang JA, Liu H. Fusion of Multiple Features BERT Model for Chinese Entity Relation Extraction. [J]. Computer System Applications, 2021, 30(5): 253–261.

[4] Zhou LJ, Li ZA, Li X, et al. Knowledge Graph Construction for Pressure Pipeline Safety Inspection Using BERT-BiLSTM-CRF. [J]. China Petroleum and Chemical Standards and Quality, 2024, 44(8): 182–186.

[5] Zhou X. Enhancement of License Plate Character Image Clarity Based on Analogical Thought. [D]. Xidian University, 2012.

[6] Li C, Zhen KH, Tang DX, et al. Research on the Construction of Knowledge Graph Based on Prescription Dataset. [J]. World of Traditional Chinese Medicine, 2024, 19(9): 1329–1333.

[7] Wei L, YaJun D, XianYong L, et al. UDBBC: Named Entity Recognition in Social Network Combined BERT-BiLSTM-CRF With Active Learning. [J]. Engineering Applications of Artificial Intelligence, 2022, 116: 115123.

[8] Feng WK. Design and Implementation of a Triple Semi-Automatic Labeling System. [D]. Beijing University of Posts and Telecommunications, 2023.

[9] Lin WY. Research and Design of a Question-Answering System Based on the Knowledge Atlas of Special Local Products. [D]. Xiamen University, 2020.

[10] Su SS, Yang Y, Cheng MT, et al. Research on Electronic Medical Record Information Extraction Based on Rule Base. [J]. Chinese Digital Medicine, 2014, 9(7): 12–13+51.

[11] Jiang LH, Zhao RX, Dong CY, et al. Construction and Verification of a Visual Knowledge Graph of Aquatic Diseases Based on Deep Learning. [J]. Transactions of the Chinese Society of Agricultural Engineering, 2023, 39(15): 259–267.

[12] Ander B, Aitor S, Eneko A. Towards Zero-Shot Cross-Lingual Named Entity Disambiguation. [J]. Expert Systems With Applications, 2021, 184: 115584.