Plant Knowledge Mining and Organization in Pre-Qin Classics from a Digital Humanities Perspective

Authors: Wu Mengcheng, Lin Litao, Qi Yue, Huang Shuiqing, Wang Dongbo, Liu Liu, Liu Liu

Date: 2023-04-13T00:00:00+00:00

Abstract

Purpose/Significance: Conducting knowledge mining of plants in pre-Qin classics and constructing a knowledge graph of plants from pre-Qin classics is of great significance for understanding the social and living conditions of ancient Chinese people. Methods/Process: Detailed annotation and quantitative analysis of plant terms in pre-Qin classics were conducted. A classical Chinese plant named entity recognition model was constructed based on CRF and multiple deep learning models, with comparative analysis of each model's performance to determine the optimal model; a knowledge organization schema for classical Chinese plants oriented towards knowledge graphs was designed. Results/Conclusion: The classical Chinese plant named entity recognition model constructed based on the domain-specific pre-trained language model SikuRoBERTa achieved optimal performance, with a harmonic mean of 85.44%, providing an effective method for entity-based plant knowledge mining; a knowledge graph of plants from pre-Qin classics was constructed, achieving aggregation and visual presentation of plant entities and their associated knowledge in pre-Qin classics.

Full Text

Plant Knowledge Mining and Organization Construction in Pre-Qin Classics from the Perspective of Digital Humanities

Authors: Wu Mengcheng[1], Lin Litao[1], Qi Yue[1], Huang Shuiqing[1], Wang Dongbo[1], Liu Liu[1]
Affiliation: [1]College of Information Management, Nanjing Agricultural University, Nanjing 210095

Abstract: [Purpose/Significance] The knowledge mining of plants in pre-Qin classics and the construction of a pre-Qin plant knowledge graph are of great significance for understanding the society and living conditions of ancient Chinese people. [Method/Process] This paper conducts detailed annotation and quantitative analysis of plant terms in pre-Qin classics. Based on CRF and various deep learning models, we construct a plant named entity recognition model for classical Chinese texts, compare the performance of each model to determine the optimal approach, and design a knowledge graph-oriented organization schema for classical plant knowledge. [Result/Conclusion] The plant named entity recognition model based on the domain-specific pre-trained language model SikuRoBERTa achieves the best performance with a harmonic mean of 85.44%, providing an effective method for entity-based plant knowledge mining. We construct a pre-Qin classics plant knowledge graph, achieving aggregation and visualization of plant entities and their associated knowledge in pre-Qin classics.

Keywords: Digital Humanities; Pre-Qin Classics; Plant Named Entity; Deep Learning; Knowledge Graph

Introduction

Throughout the long history of Chinese civilization, plants have served as important literary imagery and living materials, influencing human life in myriad ways. In the verse "When I left, the willows were swaying softly," willows convey the author's reluctance to part, while "the reeds are lush, the white dew turns to frost" uses reeds to describe an unattainable situation. Even in modern times, plants remain imbued with rich symbolism—cacti represent resilience, roses signify love. Additionally, certain plants possess significant medicinal value and serve as raw materials for traditional Chinese medicine. For instance, honeysuckle clears heat and detoxifies, while jasmine soothes the liver and relieves depression. Thus, plants embody far more than material attributes; they also carry emotional significance and medicinal value.

Current plant research primarily encompasses four perspectives: environmental studies examining plants' impact on atmosphere and soil [**?**]; biological research investigating plant functions and traits [**?**]; traditional Chinese medicine studies exploring medicinal value [**?**]; and onomastic research analyzing plant naming patterns and the origins of corresponding objects [**?**].

The rise of digital humanities has introduced new research paradigms for classical texts, offering novel methods and perspectives for mining and organizing plant knowledge in ancient books. China's massive digital classical resources and recent advances in classical Chinese information processing technology provide robust data and technical support for uncovering hidden knowledge in ancient texts. Among various classical Chinese information processing techniques, named entity recognition forms the foundational step for mining word-

level knowledge units, while knowledge graphs—developed from semantic web technology—serve as effective means for organizing and storing massive knowledge units and providing visualization and retrieval of associated knowledge. Employing named entity recognition and knowledge graph technology to mine, organize, and present plant knowledge embedded in classical texts holds significant importance for promoting excellent traditional Chinese culture and facilitating the creative transformation and innovative development of knowledge contained in ancient books.

This study selects 25 pre-Qin classics as research objects, meticulously annotating plant named entities within them. We conduct comparative experiments based on CRF, Bi-LSTM-CRF, and various deep pre-trained language models to explore effective plant named entity recognition models for classical Chinese, applying the optimal model to identify plant entities in the *Classic of Mountains and Seas* for entity supplementation. We then associate the annotated and recognized plant entities with plant knowledge from external resources such as *Ancient Chinese Plant Names with Illustrations*, constructing a knowledge graph to organize and visualize plant knowledge in pre-Qin classics.

Literature Review

Plant Knowledge Mining Plant knowledge mining based on classical Chinese texts has primarily focused on medicinal value from the perspective of traditional Chinese medicine. For example, Zou Li [**?**] examined information about Sophora tonkinensis in mainstream materia medica texts, discovering discrepancies between ancient records indicating non-toxicity and modern research findings, and identified its characteristic of being easily confused. Qu Baoquan [**?**] studied medication patterns in external hair-beautifying prescriptions from ancient Chinese medical texts by compiling records containing hair growth and blackening formulas from the Chinese Medical Classics Database. Yuan Daichang [**?**] conducted textual research on Lindera aggregata's name, morphology, and origin based on medical classics and materia medica texts, noting cognitive differences in quality, collection, processing, and efficacy.

Some scholars have also studied plant naming and imagery from literary perspectives. For instance, Yu Nana et al. [**?**] analyzed aquatic and hygrophytic plant imagery in the *Book of Songs*, exploring both practical value and underlying emotional-cultural significance. Tan Hongjiao [**?**] systematically studied ancient Chinese plant naming, summarizing characteristics and patterns. Wang Wei [**?**] examined object names in *Book of Rites*, conducting etymological research on eight monosyllabic object terms and classifying plant categories. Wang Lingyun [**?**] employed "close reading" methods to analyze plants in contemporary new poetry works based on four layers of meaning contained in "plant imagery," explaining meaning generation methods. Ma Kaiyan et al. [**?**] explored correlations between plants, vocabulary, and themes through topic modeling using

plant entries from typical literary works, mining correspondences between plant imagery and specific expressions.

Overall, existing research typically focuses on single classical texts with limited corpus scale, low automation in methodology, and insufficiently intuitive result presentation.

Named Entity Recognition Named entity recognition, as a fundamental natural language processing task, has evolved from rule-based methods to statistical machine learning approaches, and then to deep learning methods incorporating attention mechanisms and graph neural networks [**?**]. With advances in computing technology, named entity recognition techniques continue to update and optimize, with increasingly broad application prospects, becoming one of the important technical means in digital humanities research.

Research on named entity recognition for classical Chinese texts has yielded fruitful results, with machine learning and deep learning methods currently mainstream, eliminating feature engineering while achieving excellent recognition performance. For example, Li Na et al. [**?**] used digitized *Local Gazetteer Products* as corpus, annotated various name types including person names, place names, aliases, and citation names, achieving effective recognition using CRF. Xu Chenfei et al. [**?**] used the Yunnan volume of *Local Gazetteer Products* as corpus to recognize entities such as people, origins, citations, and product aliases, finding that Bi-LSTM-CRF performed better for citation recognition while BERT excelled at person name recognition. Wang Jingwei et al. [**?**] built an ALBERT-BiLSTM-CRF model for named entity recognition of diseases, prescriptions, medicines, syndromes, and symptoms in *Treatise on Cold Damage*, demonstrating superior performance compared to other models. Additionally, some scholars have innovated neural network structures, such as Y. Wang [**?**], who proposed a Polymorphic Graph Attention Network (PGAT) to capture dynamic correlations between characters and matched words from multiple dimensions to enhance character representation.

Overall, CRF, Bi-LSTM-CRF, and BERT are commonly used models for named entity recognition, with research objects including local gazetteer texts and traditional Chinese medicine classics. However, these studies employ general models for modern Chinese text processing, while classical Chinese differs significantly in grammar, morphology, syntax, genre, and stylistic features. Consequently, these approaches may suffer from low compatibility between model structure and text content, resulting in incomplete recognition.

Recently, classical Chinese pre-trained language models such as SikuBERT [**?**] have provided new options for intelligent classical Chinese information processing. Against this backdrop, this study constructs an effective classical Chinese plant entity recognition model based on pre-trained language models oriented toward classical Chinese text processing, using the constructed recognition model to assist knowledge graph construction.

Knowledge Graphs Knowledge graphs, also known as knowledge domain mapping maps, represent a novel form of knowledge representation that can visually organize and present concepts, concept attributes, and semantic relationships between different concepts in a particular field [**?**]. This technology has been applied to numerous domains including domain knowledge modeling [**?**], automatic question answering [**?**], and topic evolution analysis [**?**], particularly demonstrating extensive applications in organizing knowledge from traditional Chinese medicine classics. For instance, Zhang Jundong [**?**] constructed a knowledge ontology for infertility, presenting conceptual relationships within the field and using data mining methods to 完善本体语义关系, achieving semantic mapping and structured expression of traditional Chinese medicine clinical trial knowledge for infertility. Zhang Xiangxian et al. [**?**] used a top-down approach to construct an ontology model for Dunhuang and Turpan medical literature, building a knowledge graph for knowledge organization and visualization. Zhai Dongsheng et al. [**?**] extracted entities and relationships from traditional Chinese medicine patent texts through deep learning-based joint extraction models, completing knowledge graph construction based on traditional Chinese medicine knowledge graph ontology structures. Yang Yanling et al. [**?**] elaborated on construction methods for traditional Chinese medicine case knowledge graphs, conducting named entity recognition and extraction using entities such as diseases, symptoms, and medicines from cases to explore relationships. Li He et al. [**?**] studied bamboo and silk medical literature, constructing bibliographic and content ontologies to achieve visualization of bamboo and silk medical knowledge graphs.

Currently, knowledge graphs have become an important method in digital humanities research. For example, Cui Jingfeng et al. [**?**] used deep learning models to mine chrysanthemum-related knowledge and text associations in classical poetry. Zhang Yunzhong et al. [**?**] organized digital resources of historical figures to construct a red historical figure knowledge graph and build a question-answering platform. Liu Huan et al. [**?**] studied the *Zuo Zhuan* using SVM and BERT-LSTM-CRF models for question intent recognition and entity recognition, constructing a domain knowledge graph and building a question-answering system platform based on the Flask framework. Fan Qing et al. [**?**] constructed an intangible cultural heritage knowledge graph to form linked data and present hidden relationships. Zhong Yuanxin et al. [**?**] constructed an art image knowledge graph for the art image domain, demonstrating the advantages of knowledge graphs in knowledge organization applications compared to traditional databases.

These studies demonstrate that knowledge graphs offer advantages in knowledge organization, visualization, and correlation analysis. Meanwhile, continuous growth in domain knowledge makes general domain knowledge graphs require constant updates and supplementation, resulting in high construction and maintenance costs. However, knowledge graphs for digital humanities research are based on limited historical classics, granting them high stability and reducing later maintenance efforts. Stable domain knowledge graphs provide important

guarantees for knowledge retrieval and automatic question-answering applications. Therefore, this study selects knowledge graphs to organize and store plant-related knowledge and associations in classical texts.

Dataset Construction and Plant Word Distribution Statistics

This study selects the pre-Qin classics corpus constructed by Nanjing Normal University as the research object. This corpus contains 25 pre-Qin classics, which can be divided into four major categories according to the four-part classification system: "Classics, History, Masters, and Collections" [**?**], as shown in .

The corpus features rich content covering ancient military, cultural, and other aspects, comprehensively revealing the living conditions and social landscape of ancient people during the pre-Qin period while documenting numerous plants, thus possessing high research value.

Plant Word Annotation

Plant named entities refer to words denoting plants, hereinafter referred to as plant words. Among the 25 pre-Qin classics, some plant words refer to broad categories that cannot be clearly matched to specific plant varieties, such as "tree," "algae," "water grass," and "wild grass." Such plant words are excluded from subsequent annotation and statistical analysis. The annotation work employs manual annotation supplemented by dictionary matching, including three steps: dictionary-based pre-annotation, manual proofreading, and supplementary annotation.

First, we constructed a classical Chinese plant word dictionary from two data sources. The first source is the *Shi Cao* (Explaining Grasses) and *Shi Mu* (Explaining Trees) chapters from the *Erya*, a text compiled around the same period as pre-Qin classics that documents extensive plant-related content. The second source is *Ancient Chinese Plant Names with Illustrations*, an important achievement by Gao Mingqian after more than 30 years of research on ancient Chinese plant names, containing 4,394 ancient plant names and documenting a rich variety of ancient plants. We invited three graduate students with botanical research backgrounds to manually identify and organize plant words from these sources, forming a plant word collection. After merging and deduplication, we obtained the final classical Chinese plant dictionary for dictionary matching-based annotation.

Next, using a self-developed Python program with a maximum reverse matching strategy, we performed pre-annotation on content with part-of-speech "n" (noun) in the corpus based on the classical Chinese plant dictionary.

Finally, we conducted manual proofreading and supplementary annotation on the dictionary matching results to improve accuracy and comprehensiveness.

The authors and the three aforementioned graduate students with botanical backgrounds worked in groups (three per group) to complete this task. Each group first independently proofread and supplemented the pre-annotation results, then checked and confirmed another group's proofreading and supplementation. All personnel referred to the parallel corpus of classical and vernacular translations provided by Gushiwen website (https://www.gushiwen.cn/) to improve understanding accuracy of classical content.

Following these steps, an annotated corpus sample appears as: "[黍]/n 、/w [粱]/n 、/w [稻]/n 皆/d 二/m 行/n."

Plant Word Distribution Statistics

Across the 25 pre-Qin classics, we identified 4,576 plant word occurrences, comprising 364 unique plant words. During annotation, no plant words were found in the *Classic of Filial Piety*. In terms of total plant word count, the *Book of Rites and Ceremonies* contains the most with 635 plant words, followed by *Guanzi* (538), *Book of Songs* (472), and *Book of Rites* (457). Regarding unique plant word count, the *Book of Songs* ranks first with 134 unique plant words, followed by *Guanzi* (111) and *Book of Rites* (95). In terms of plant word proportion to total text length, the *Book of Songs* has the highest ratio at approximately 1.36%, being the only classic where plant words exceed one percent of the text. The total count, unique count, and text proportion of plant words in different classics are shown in .

As evident from , poetry and rhapsody classics contain relatively high proportions of plant words, while classics and historical texts show lower proportions. This suggests that using plants in poetry was a common phenomenon in ancient Chinese life. The *Book of Songs* ranking highly in total plant words, unique plant words, and text proportion is not coincidental. The *Book of Songs* frequently employs the "fu, bi, xing" writing techniques, where "bi" compares one thing to another and "xing" uses other things to introduce the main theme. In the *Book of Songs*, these "other things" are often plant words. For example, in "Wei Feng · Shuo Ren," "hands like tender reed shoots" uses the plant "tender reed shoots" to describe a beauty's soft hands; in "Jian Jia," "the reeds are lush, white dew turns to frost" uses two plants to express the poet's longing. Consequently, the *Book of Songs* leads in both variety and frequency of plant words.

Pre-Qin Classics Plant Knowledge Graph Construction

The technical route for plant knowledge graph construction is illustrated in [Figure 1: see original paper], comprising three main components: (1) Construction and application of a classical Chinese plant entity recognition model. Through comparative analysis, we input plant corpus into machine learning and various deep learning models to build an automatic plant named entity recogni-

tion model for classical texts, comparing model performance to determine the optimal approach. (2) Using the *Classic of Mountains and Seas* text as an application case, we employ the optimal model to identify plants, with manual verification of results to expand the knowledge graph's data sources. (3) Plant knowledge graph construction. Using the 25 pre-Qin classics, plant knowledge crawled from internet encyclopedia knowledge bases, and plant knowledge extracted from the *Classic of Mountains and Seas* as data sources, we construct the pre-Qin classics plant knowledge graph through entity linking and knowledge fusion methods.

Knowledge graphs represent entities and relationships in graph form, organizing entity data and relationship data through triples. Specifically, in knowledge graphs, the "node-edge-node" relationship can be viewed as a "subject-predicate-object" relationship, constituting one record. The representation model for classical Chinese plant words and associated knowledge is shown in [Figure 2: see original paper].

The entire knowledge graph consists of such triples. The same subject typically contains multiple relationships, and as knowledge accumulates, the entity relationship network continuously expands, eventually containing massive data and knowledge. [Figure 3: see original paper] shows an example: "Plant Tang has the efficacy of treating kidney deficiency and lower back pain," where Plant Tang is the subject, "has the efficacy" is the predicate, and "treating kidney deficiency and lower back pain" is the object.

Knowledge Extraction

Knowledge extraction aims to identify and extract plant named entities from classical texts. This study employs sequence labeling models for knowledge extraction, specifically CRF, Bi-LSTM-CRF, and various pre-trained language models, briefly introduced below.

(1) **CRF**
Conditional Random Field (CRF) is a discriminative undirected graphical model. When applied to labeling problems, it becomes a discriminative model that predicts output sequences based on input sequences. Its learning method obtains conditional probability models through maximum likelihood estimation on given training datasets; for prediction, it seeks the output sequence with maximum conditional probability given an input sequence.

(2) **Bi-LSTM-CRF**
Bi-LSTM-CRF consists of Bi-LSTM and CRF components. The CRF layer can learn transition probabilities between labels in the dataset to correct Bi-LSTM layer outputs, improving model prediction accuracy.

(3) **Deep Pre-trained Language Models**
Deep pre-trained language models are semantic representation models trained through self-supervised methods on large-scale unsupervised cor-

pora, containing lexical, syntactic, and contextual information from the corpora. Domain-specific pre-trained language models can further improve downstream task performance on corresponding corpora. Therefore, this study specifically selects SikuBERT and SikuRoBERTa [**?**], classical Chinese pre-trained models oriented toward digital humanities. SikuBERT and SikuRoBERTa were trained by Nanjing Agricultural University based on the *Siku Quanshu* corpus. During pre-training, both models used traditional Chinese character vocabularies without punctuation, with sentence segmentation at the character level. SikuBERT was obtained by continuing training BERT-base-Chinese on *Siku Quanshu* corpus, removing the next sentence prediction task that provides limited performance improvement. SikuRoBERTa was obtained by continuing training Chinese RoBERTa-Chinese (with whole word masking) on *Siku Quanshu* corpus. The *Siku Quanshu* corpus used for continued training comprises the main text (excluding annotations) of the Wenyuan Pavilion edition in traditional Chinese, totaling approximately 530 million characters.

To comprehensively compare and select the optimal model, this study also includes guwenBERT (https://github.com/Ethan-yt/guwenbert), BERT-base-Chinese [**?**], and Chinese-roberta-wwm-ext [**?**] for comparative experiments. guwenBERT was obtained by Beijing Institute of Technology through continuing training RoBERTa-Chinese (with whole word masking) on the Daizhige ancient literature corpus, which contains 15,694 ancient books totaling approximately 1.7 billion characters. BERT-base-Chinese, trained by Google on Chinese Wikipedia data, demonstrates good versatility for Chinese natural language processing tasks. Chinese-roberta-wwm-ext is a Chinese pre-trained language model developed by the HIT-SCIR and iFLYTEK joint laboratory using whole word masking technology based on general Chinese corpora.

Corpus Preprocessing

Before constructing the plant named entity automatic recognition model, corpus preprocessing is required. Through statistical analysis of all plant word lengths in the corpus, we determined to use a 5-tag labeling set as the annotation specification for preprocessing. The 5-tag set can be represented as R={B-P, E-P, M-P, S-P, O}, where "B-P" indicates the starting character of a plant word, "M-P" indicates middle characters, "E-P" indicates the ending character, "S-P" indicates single-character plant words, and "O" indicates all other characters not part of plant words. Preprocessed corpus samples are shown in .

The computer configuration for this experiment is as follows: operating system CentOS 3.10.0, CPU of 4 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz, 256GB memory; GPU of 6 NVIDIA Tesla P40 with 24GB memory. CRF and Bi-LSTM-CRF use default training parameters. Since deep pre-trained models including SikuBERT, SikuRoBERTa, and BERT-base-Chinese share the same neural network architecture, we set identical training parameters as shown in .

This study selects precision (P), recall (R), and F1-score [**?**] to evaluate model

performance. presents the test results for each model, showing that SikuBERT and SikuRoBERTa demonstrate outstanding performance, with SikuRoBERTa achieving the best results at 85.44% F1-score.

The *Classic of Mountains and Seas*, also compiled during the pre-Qin period, records extensive content on ancient geography, history, animals, plants, and medicine. We applied the optimal model to this text to supplement the classical Chinese plant dictionary. The optimal model's identification result for one sentence reads: "其葉如【榆】葉而方，其實如【赤菽】，食之已聾。" The term "赤菽" was not an annotated entity in the training corpus, yet the model accurately recognized it, demonstrating good practical effectiveness. In total, we identified 366 plant word occurrences comprising 121 plant types from the *Classic of Mountains and Seas*, including both previously annotated plant words and newly recognized ones such as "棪" and "橿."

Knowledge Fusion

Although knowledge extraction yields entities, relationships, and attributes, different knowledge sources introduce noise and duplicate data. Cleaning and integrating such data benefits the improvement and optimization of the pre-Qin classics plant knowledge graph. The knowledge fusion process consists of two main steps: entity linking and knowledge merging.

Entity linking primarily fuses duplicate knowledge from different data sources during data collection, merging semantically identical entities into one. Due to limited data sources, this study employs manual judgment to determine whether entities share the same meaning. For example, "Domestic Distribution" on the "Zhiwutong" website (https://www.zhiwutong.com/) and "Distribution Location" on the "Zhiwuzhi" website (https://www.iplant.cn/) both describe plant locations within China, representing the same entity that should be merged.

Knowledge merging primarily integrates attributes of the same plant entity from different source websites based on entity linking. For instance, "rice" has alias knowledge recorded on "Zhiwutong" but not on "Zhiwuzhi," while "Zhiwuzhi" records functional value and ecological habits not present on "Zhiwutong." Manual selection and merging of attribute content from different sources for the same entity can comprehensively absorb plant knowledge, enriching the pre-Qin classics plant knowledge graph.

This study employs two methods for plant knowledge extraction: first, manually annotating plant words in pre-Qin classics, then using the SikuRoBERTa model trained on annotated data to automatically identify plant words in the 25 pre-Qin classics and *Classic of Mountains and Seas* with manual proofreading; second, crawling plant-related data from "Zhiwutong" and "Zhiwuzhi" websites using Python, obtaining final structured data for knowledge graph construction through data processing and integration.

Regarding data presentation, "Zhiwutong" and "Zhiwuzhi" mostly store data in semi-structured forms. In terms of content, "Zhiwutong" stores attributes

and relationships including family name, genus name, flora, aliases, sources, properties, efficacy, domestic distribution, foreign distribution, altitude, habits, medicinal parts, medicinal functions, medicinal indications, textual research, and chemical components. "Zhiwuzhi" stores scientific names, common names, synonyms, distribution locations (Chinese), morphological characteristics, ecological habits, images, specimens, specimen distribution, flora, protection levels, protection value, protection measures, and cultivation points. Comparing and integrating these knowledge contents ensures comprehensiveness of relationships and thoroughness of attributes in the pre-Qin classics knowledge graph.

This study selects the graph database Neo4j to construct the pre-Qin classics plant knowledge graph. Neo4j supports semantic queries of entities and relationships using the Cypher query language, offers faster query speeds on highly connected data, and provides visualization functions.

Integrating plant entities recognized by the pre-Qin classics plant named entity recognition model with multi-dimensional plant knowledge crawled from external databases, we store all obtained plant knowledge in the Neo4j graph database according to the knowledge representation model's data structure. This achieves multi-source knowledge fusion, with specific entity relationships and attribute knowledge shown in and visualized in [Figure 4: see original paper].

Visualization of Pre-Qin Classics Plant Knowledge

Plants and Classics After importing all classic title entities and plant entities with their relationships into the graph database, we can intuitively observe plants recorded in two or more classics. Taking the *Book of Songs* and *Chu Ci* as examples, [Figure 5: see original paper] clearly shows 26 common plants recorded in both texts, including "mulberry," "kudzu," and "mugwort." Furthermore, the number of directed edges between specific plants and different classics can reflect a plant's importance during the pre-Qin period. For example, the plant "chestnut" frequently appears in pre-Qin classics including the *Book of Songs* and *Chu Ci*, and indeed held important social status as a major grain crop at that time. Simultaneously, the number of directed edges pointing to classics can roughly estimate a classic's richness in plant documentation. Therefore, the association between classics and plants provides both a visual knowledge network for plant research and a new meaningful perspective for literary history studies.

Plants and Attributes After importing all entity relationships and attributes for classical plants into the graph database, the generated knowledge graph is shown in [Figure 6: see original paper]. This randomly selects the plant "rice" from classics and visualizes its knowledge including Chinese name, scientific name, aliases, and attribute knowledge from morphological characteristics such as life form, branches, roots, stems, leaves, fruits, flowers,

phenology, habitat, altitude, and foreign distribution. The visualization reveals that "rice" has aliases including "grain," "he," "jing," and "paddy," is an annual plant with loose, hairless leaf sheaths, and serves as a major grain crop. This presentation and organization method broadens approaches to learning plant knowledge and provides plant researchers with semantic-based query pathways, allowing them to limit or narrow search ranges based on specific plant attribute information for further research. Beyond single plant attributes, the pre-Qin classics plant knowledge graph can also explore associations between different plant attributes, such as directly querying plant groups with specific attributes for comparative studies to uncover more valuable information.

Plants and Efficacy During the pre-Qin period, besides serving as grain crops, plants' medicinal value was also significant. For example, the *Yellow Emperor's Inner Canon* compiled during the Warring States to Qin-Han period and the subsequent *Shennong's Herbal Classic* documented extensive plant medicinal properties. The *Classic of Herbal Medicine · Preface* records: "One hundred and twenty superior medicines serve as sovereigns, nourishing life in accordance with heaven, non-toxic, and safe for long-term consumption," reflecting the medicinal value of plants such as ginseng, licorice, rehmannia, coptis, and jujube. Thus, some plant efficacies were discovered and utilized during the pre-Qin period. Therefore, this study separates plant entities and efficacy entities during knowledge graph construction to facilitate exploration of associations between different plants and their medicinal values, with partial efficacy attribute visualization shown in [Figure 7: see original paper].

[Figure 7: see original paper] not only displays which efficacies plants possess but also associates plants with the same efficacy, such as "Hu" and "Chun" for "regulating qi and relieving pain," and "Rongshu" and "Bao" for "dispelling wind and dampness." Particularly in traditional Chinese medicine, this presentation clarifies relationships between different plants and efficacies, not only helping researchers discover plants with similar efficacies but also enabling accurate querying and understanding of complex relationships between plants and efficacies for better research on plant characteristics and applications, thereby facilitating traditional Chinese medicine research.

Conclusion

This study meticulously annotated plant words in 25 pre-Qin classics, constructing a pre-Qin classics plant entity corpus. Based on CRF, Bi-LSTM-CRF, and various deep pre-trained language models, we constructed an automatic classical Chinese plant entity recognition model for classical texts, providing an effective method for plant knowledge mining in classics. By associating plant entities recognized from classics with external encyclopedia knowledge bases such as "Zhiwutong" and "Zhiwuzhi," we built a pre-Qin classics plant knowledge graph and visualized it. This knowledge graph holds potential application value for

plant knowledge discovery and can provide data support for plant knowledge retrieval and automatic question answering.

This study has several limitations. First, the performance of the classical Chinese plant entity automatic recognition model has room for improvement. For example, low-frequency plant words in the training corpus such as “migu” and “panmu” were not successfully recognized. Second, entities, relationships, and attributes in the pre-Qin classics plant knowledge graph require further expansion. During knowledge completion, some plant-related knowledge such as endangered categories, protection levels, and economic value have not been fully integrated. Future research will consider increasing training corpus scale and exploring more advanced entity recognition methods. Additionally, while expanding and optimizing the pre-Qin classics plant knowledge graph, we will explore knowledge graph applications in automatic question answering and knowledge retrieval.

References

[1] Shi Weiwei, Zhou Changxing, Liu Kai, et al. Study on phytoremediation of heavy metal Cd contaminated soil[J]. China Resources Comprehensive Utilization, 2022, 40(09): 93-95.
[2] Liu Bing, Xiang Xiaomei, Tan Lu, et al. Functional trait diversity of seed plants in Dehang Canyon, Hunan Province[J]. Acta Botanica Boreali-Occidentalia Sinica, 2022, 42(9): 1591-1599.
[3] Mo Weijun. Research on traditional Chinese medicine health care of medicinal plants[J]. Journal of Nuclear Agricultural Sciences, 2021, 35(03): 768.
[4] Meng Yingjun. Study on the noun words in Erya · Shicao[D]. Guilin: Guangxi Normal University, 2010.
[5] Zou Li, Zhao Huanjun, Li Na, et al. Textual research and toxicity analysis of Sophora tonkinensis[J]. Modern Traditional Chinese Medicine, 2021, 41(5): 19-
[6] Qu Baoquan, Liu Dun, Hou Wenbin, et al. Analysis of medication patterns in external hair-beautifying prescriptions from ancient traditional Chinese medicine classics based on data mining[J]. China Medical Herald, 2021, 18(21): 126-129+149.
[7] Yuan Daichang, Yuan Ling, Yuan Panpan, et al. Textual research on Lindera aggregata[J]. Shanxi Traditional Chinese Medicine, 2021, 37(07): 55-58.
[8] Yu Nana, Wang Yingying, Yu Jingyi. Analysis of aquatic and hygrophytic plant imagery in the historical classic Book of Songs[J]. Wetland Science and Management, 2022, 18(05): 54-57+61.
[9] Tan Hongjiao. Study on ancient Chinese plant naming[D]. Hangzhou: Zhejiang University, 2004.
[10] Wang Wei. Study on object words in Book of Rites[D]. Changchun: Northeast Normal University, 2005.

[11] Wang Lingyun. Study on plant imagery in contemporary Chinese new poetry[D]. Kunming: Yunnan University, 2017.
[12] Ma Kaiyan, Xiao Yao, Chen Qian, et al. Study on plant imagery in contemporary Chinese literary works from the perspective of digital humanities[J]. Digital Humanities Research, 2022, 2(02): 35-45.
[13] Song Xuhui, Yu Hongtao, Li Shaomei. Chinese named entity recognition based on graph attention network with character-word fusion[J]. Computer Engineering, 2022, 48(10): 298-305.
[14] Li Na, Bai Zhentian, Bao Ping. Analysis of knowledge organization path for ancient books based on Local Gazetteer Products[J]. Ancient and Modern Agriculture, 2016(01): 105-113.
[15] Xu Chenfei, Ye Haiying, Bao Ping. Research on automatic entity recognition model for local gazetteer product materials based on deep learning[J]. Data Analysis and Knowledge Discovery, 2020, 4(08): 86-97.
[16] Wang Jingwei, Xiao Li, Luo Jiawei, et al. Named entity recognition research based on Treatise on Cold Damage[J]. Computer and Digital Engineering, 2021, 49(08): 1584-1587.
[17] WANG Y, LU L, WU Y, et al. Polymorphic graph attention network for Chinese NER[J]. Expert Systems with Applications, 2022, 203: 117467.
[18] Liu Chang, Wang Dongbo, Hu Haotian, et al. Research on automatic word segmentation for classics integrating external features for digital humanities—taking SikuBERT pre-trained model as an example[J]. Library Forum, 2022, 42(06): 44-54.
[19] WANG Q, MAO Z, WANG B, et al. Knowledge Graph Embedding: A Survey of Approaches and Applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.
[20] Zhou Lina, Hong Liang, Gao Ziyang. Construction of Tang poetry knowledge graph and design of intelligent knowledge services[J]. Library and Information Service, 2019, 63(2): 24-33.
[21] Zhou Yi, Liu Zheng, Su Xiaoqing, et al. Construction of question-answering knowledge graph ontology model integrating multi-level data[J]. Library and Information Service, 2022, 66(5): 125-132.
[22] Huang Wei, Lu Guoqiang, Zhao Xu. Research on microblog topic evolution path based on knowledge graph[J]. Information Studies: Theory and Application, 2022, 45(3): 173-181.
[23] Zhang Jundong. Research on knowledge ontology construction for infertility traditional Chinese medicine clinical trials[D]. Beijing: China Academy of Chinese Medical Sciences, 2022.
[24] Zhang Xiangxian, Li Shiyu, Shen Wang, et al. Research on knowledge organization of Dunhuang and Turpan medical literature from digital humanities perspective[J]. Library and Information Service: 1-16.
[25] Zhai Dongsheng, Lou Ying, Kan Huimin, et al. Research on construction and application of traditional Chinese medicine knowledge graph based on multi-source heterogeneous data[J]. Data Analysis and Knowledge Discovery: 1-23.
[26] Yang Yanling, Li Yan, Shuai Yaqi, et al. Construction of knowledge graph

based on traditional Chinese medicine cases[J]. Journal of Medical Informatics, 2022, 43(10): 50-54.
[27] Li He, Zhu Linlin, Liu Jiayu, et al. Research on knowledge organization of bamboo and silk medical literature based on ontology[J]. Library and Information Service: 1-12.
[28] Cui Jingfeng, Zheng Dejun, Wang Dongbo, et al. Named entity recognition for chrysanthemum classical poetry based on deep learning models[J]. Information Studies: Theory and Application, 2020, 43(11): 150-155.
[29] Zhang Yunzhong, Guo Dong, Wang Yage, et al. Research on red historical figure knowledge question-answering service framework based on knowledge graph[J]. Library and Information Service, 2021, 65(16): 108-117.
[30] Liu Huan, Liu Liu, Wang Dongbo. Research on domain knowledge graph automatic question answering from digital humanities perspective[J]. Science and Technology Information Research, 2022, 4(1): 46-59.
[31] Fan Qing, Shi Zhongchao, Tan Guoxin. Construction of intangible cultural heritage knowledge graph[J]. Library Forum, 2021, 41(10):
[32] Zhong Yuanxin, Xia Cuijuan. Preliminary study on art image knowledge graph construction[J]. Library Forum, 2022, 42(02): 109-118.
[33] Zhang Qi, Jiang Chuan, Ji Youshu, et al. Construction of integrated automatic annotation model for word segmentation and part-of-speech for multi-domain pre-Qin classics[J]. Data Analysis and Knowledge Discovery, 2021, 5(03): 2-11.
[34] Wang Dongbo, Liu Chang, Zhu Zihe, et al. SikuBERT and SikuRoBERTa: Construction and application research of pre-trained models for Siku Quanshu oriented to digital humanities[J]. Library Forum, 2022, 42(06): 31-43.
[35] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.
[36] CUI Y, CHE W, LIU T, et al. Pre-Training With Whole Word Masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29:
[37] ATTERER M, SCHÜTZE H. Prepositional Phrase Attachment without Oracles[J]. Computational Linguistics, 2007, 33(4): 469-476.

Author Contributions:
Wu Mengcheng: Conceptualized research framework, performed data annotation and analysis, wrote the manuscript;
Lin Litao: Designed the study, constructed models, revised the manuscript;
Qi Yue: Revised the manuscript;
Huang Shuiqing: Revised the manuscript;
Wang Dongbo: Determined research topic, reviewed and revised the manuscript;
Liu Liu: Revised the manuscript.

Note: Figure translations are in progress. See original paper for figures.