International Journal of Knowledge and Language Processing
Volume 12, Number 2, Dec **2021**

AKLP International ©2021 ISSN 2191-2734
pp. 17–31

Review on Knowledge Structure Parsing of Ancient Chinese Medical Books for Single Texts

Huanqing, Li and Guanlin, Li

Peking University
No.5, Yiheyuan Road, Haidian District, Beijing, China
lihuanqing@stu.pku.edu.cn

Received xxx 2021; revised xxx 2021

ABSTRACT. Objective: Ancient Chinese medical books contain important academic knowledge of traditional Chinese medicine, in which the prescriptions, drug citation, and symptoms recorded still have immeasurable enlightenment and reference value for modern disease research. It is of great significance for knowledge management of professionals to structure text through knowledge discovery. Methods: This article conducts a retrospective study on the basic tasks in the analysis of ancient Chinese medicine books, including word segmentation, key entity concept recognition, relation extraction technology. Based on prior knowledge, the paper combined with the features of ancient Chinese medicine books, including its short length, lack of concepts, and the difficulty of discovering the relationship between concepts, and finally summed up the difficulties of ancient TCM (traditional Chinese medicine) books structure analysis technology. Results: The task of analyzing the structure of ancient Chinese medicine books faced great challenges: insufficient data annotation resources in the field of ancient Chinese medicine books, low overlap rate of single text content, lack of open data sets, insufficient statistical information, and difficulty in discovering hidden structures. Conclusion: To analyze the knowledge structure of ancient Chinese medicine texts, it is necessary to combine methods such as semi-supervised learning, self-supervised learning, and small-sample learning, in the absence of annotated data resources. In the future, more effors should be made to further structure and analyze the texts, and strengthen the discovery of the textual structure and knowledge structure in texts, thus to surpass previous studies.

Keywords: TCM ancient books knowledge mining, word segmentation, Named Entity Recognition (NER), relation extraction, supervised learning, single texts

1. Introduction

The acquisition and utilization of knowledge have become one of the most popular topics in natural language processing tasks. On the one hand, the effective utilization of world knowledge and text language knowledge can greatly improve the performance of semantic disambiguation, dialogue generation, intelligent retrieval, and other tasks. On the other hand, text knowledge mining has always been an important topic in the field of machine learning. With the rise of the digital era, a large number of books have been digitized, and a large number of electronic text resources have been produced. Text knowledge mining, which refers to the automatic acquisition of implicit knowledge from massive semi-structured and unstructured text data by natural language processing technology, can further transform the electronic unstructured text data into structured ones.

As the knowledge carrier of the practice and development of TCM since ancient times, ancient books collect the description of diseases, prescriptions, and other medical topics as well as the summary of clinical experience, in the long process traditional Chinese medicine development. They not only explain the medical theory but also contain a lot of valuable knowledge of medical history, etiquette system, social culture, language development, and so on. It is of great practical significance for the development of TCM and modern medicine and the study of medical history to sort out and annotate ancient books of TCM and dig out the knowledge contained therein.

Therefore, summarizing and reviewing the related technologies of knowledge mining in the field of ancient Chinese medicine books, and refining the obstacles currently faced in the analysis and research of them, will play a huge role in promoting the digitization of ancient books in the future, digging deep into the rich knowledge contained in ancient TCM books, and contributing to the overall human health.

2. Relevant Technologies of Text Knowledge Mining in the Field of Ancient Chinese Medicine Books

2.1 Word Segmentation

Unlike English, Chinese is ideographic with no obvious separators between words. However, in many natural language processing research and related engineering applications, an obvious word segmentation boundary is needed. For example, when building the index database of a search engine, it is often necessary to build an inverted index according to the word to search. In some natural language processing tasks, it is necessary to introduce part-of-speech information, such as shallow and deep syntactic analysis, semantic analysis, etc., and the premise of building part-of-speech is word segmentation. In some machine learning models, word-based algorithms are needed for feature extraction, such as document topic analysis, keyword clustering, and automatic keyword extraction. Therefore, the word segmentation task is usually the premise of the Chinese natural language processing task. Similar to modern Chinese, the texts of ancient Chinese medicine books also have no explicit word segmentation boundary, so automatic word segmentation is also needed to facilitate relevant research and resource processing.

The research on automatic Chinese word segmentation has been carried out since the last century and has developed relatively mature. The relevant tagging corpus is relatively rich, and the model algorithm is constantly updated and iterated, with high accuracy. In early Chinese word segmentation, the rule-based dictionary matching method based on maximum matching was mostly adopted. In the case of a relatively complete dictionary, its F value could reach 80% or even higher [1]. However, in the field of ancient Chinese, no dictionary can be applied to all dynasties. Ancient Chinese has been changing for thousands of years, and the words used in different times are inevitably different. In addition, ancient Chinese uses very concise words, often single-character words and many rare characters, which also limits the dictionary-based approach. Most importantly, the texts of ancient Chinese medicine books, as a kind of professional texts, contain a large number of professional terms only found in the field, which are often difficult to be comprehensively covered by existing dictionaries. Methods using statistical learning and supervised machine learning tend to have better generalization ability and higher accuracy and recall rate compared with dictionary-based methods, which usually include methods based on supervised learning, semi-supervised learning, and unsupervised learning.

At present, supervised learning is still the mainstream method of Chinese word segmentation, which has low training costs, high training efficiency, and can achieve good results. However, supervised learning relies too much on the training corpus and can only perform well in the field related to the training corpus in most cases. In addition, supervised learning cannot be carried out when there is little or no training corpus. Supervised learning of word segmentation algorithms is usually achieved through sequence tagging. Before a deep learning algorithm is applied to word segmentation, a representative algorithm is based on conditional random field (CRF, Conditional Random Field) [2] and Hidden Markov Model HMM [3]. Based on the homogeneous Markov hypothesis and observation independence hypothesis, the hidden Markov model models the observation sequence, that is, its hidden state. The hidden Markov model models the initial state probability, transition probability, and emission probability. In the word segmentation task, the observation sequence is the Chinese character, and its corresponding hidden state is the label of the word, representing the word head, word middle, word end, or word itself. In the HMM model, given the string C, the probability of obtaining W can be formalized as:

$$P_{HMM}(W|C) = \prod_{i=1}^{|c|} P_t(t_i|t_{i-1})P_e(c_i|t_i) \# \quad \text{(Equation 2.1)}$$

In this formula, $P_t$and $P_e$represent the transfer probability and emission probability at time $t_i$respectively. Based on the observation sequence, the Viterbi algorithm is usually used to solve the state sequence with the highest probability and get the most possible word segmentation path. Different from HMM model, CRF is an undirected graph model [4], which breaks the homogeneous Markov hypothesis and observation independence hypothesis, and has stronger modeling ability for sequences and hidden states. Therefore, CRF has become a more commonly used word segmentation algorithm for supervised

learning. Before the deep learning model was widely used, the CRF algorithm became the standard algorithm for word segmentation. In the learning settings based on string annotation, the optional types of word features mainly include N-gram features under Sliding Window. In ancient Chinese word segmentation based on CRF, Wang Xiaoyu et al. [5] added character classification and dictionary information based on the N-gram feature and obtained good word segmentation results. Based on N-gram features, the Chinese word segmentation model introduces more additional features, which further improves the model based on the feature template. In the study of word segmentation in ancient Chinese, the combination of word segmentation and part-of-speech tagging is trained and the template is made by CRF to obtain the tagging model of word segmentation and part-of-speech tagging, which can achieve better performance. Shi Min et al. [6] used the model of integration of word segmentation and POS tagging based on conditional random field in the word segmentation task of ancient prose. In the designation of feature template, detailed POS features were used, and additional features such as rhyme and radical were added, thus obtaining better word segmentation results. The limitation of traditional machine learning word segmentation based on feature engineering is obvious. Taking ancient Chinese as an example, to make the performance of the model more practical, it needs to carry out a large number of feature engineering, such as the tagging of part of speech, rhyme and radical, etc., which is highly dependent on domain expert knowledge. The model trained according to this method is only applicable to the field involved in the training data, and the generalization performance is not strong. With the development of computing power, deep learning methods are widely used in the field of natural language processing, especially in word segmentation, for it reduces the need for feature engineering and achieves good results. Cai et al. [7] used the LSTM model to directly model the sentence segmentation, abandoned the Sliding Window, and obtained similar performance to the traditional model requiring a lot of feature engineering. In the field of ancient Chinese word segmentation, LSTM and other depth models are also widely used. Liu Yutong et al. [8] used the model based on LSTM+CRF and achieved good results on the test set. However, compared with the study of word segmentation in modern Chinese, the development of word segmentation in ancient Chinese is relatively slow and often differs from the latest technology. In addition, both the traditional machine learning based on feature engineering and the word segmentation system based on deep learning rely on the high-quality annotated corpus to obtain a better part of speech performance. However, there are only a limited number of publicly available word segmentation data sets in the field of ancient Chinese, and most of these data are in the general field such as the Zuozhuan (Commentary of Zuo Qiuming), which are far from the data of ancient Chinese medicine books involved in this study in terms of dictions, content and times. Therefore, the supervised Chinese word segmentation method is not feasible.

Unsupervised word segmentation usually uses statistical information from large-scale unstructured original corpus, so it has good adaptability to any domain and does not need to annotate corpus. Unsupervised word segmentation methods can be roughly divided into

two categories: discriminant method and generative method. In the discriminant method, goodness measures are usually designed for candidate word strings to determine whether they are formed based on the measures. Commonly used metrics include mutual information [9], nVBE [10], minimum description length [11], etc. Generative methods include HMM model [12], HDP (Hierarchical Dirichlet Process) based methods [13], and Nested Pitman-Yor Process [14]. Among them, Chen et al. [12] combined the word-based unsupervised BHMM and word-based unsupervised HDP model through probability multiplication and normalization. Suppose C is the string to be sliced and W is the word string after sliced, then given string C, the conditional probability of sliced W is:

$$P_J(W|C) = \frac{1}{Z(C)} P_D(W|C) P_M(W|C) \#(\text{Equation 2.2})$$

$P_D(W|C)$ is the conditional probability of W given by the HDP model, $P_M(W|C)$ is the conditional probability of W given by HMM, and $Z(C)$ is the normalized factor to ensure that the result is a probability distribution. In the process of model inference, nVBE is introduced to initialize Gibbs sampling, to accelerate the convergence of sampling and get better performance.

Yu Jingsong et al. [15] combined the generative model HDP and the large-scale pre-training language model BERT to conduct iterative training under the unified Bootstrap framework and achieved good performance in ancient Chinese word segmentation. In the HDP model, C is set as the string to be segmented and W is the word string after segmentation, then given string C, the conditional probability of segmented W is:

$$P_{HDP}(W|C) = \prod_{i=0}^{|w|} P_{HDP}(w_i|w_{i-1}) \#(\text{Equation 2.3})$$

w_i is the ith word in the spliced string. HDP word segmentation model is a non-parametric 2-gram model, and the binary grammar model assumes that each different word has a different distribution with all possible adjacent words. Joint modeling of these distributions is carried out through the Diriclet process, as shown in Equations 2.4 and 2.5:

$$w_i|w_{i-1} \sim DP(\alpha_1, G_0) \#(\text{Equation 2.4})$$

$$G_0 \sim DP(\alpha, H) \#(\text{Equation 2.5})$$

DP stands for Dirichlet process, α for concentration parameter, H for the prior distribution of words, and is the measurement result observed. In the case of the observed participle sequence h, the posterior probability $P_D(w_i|w_{i-1} = l, h)$ as follows:

$$\frac{n_{<w_{i-1},w_i>} + \alpha_1 P_D(w_i|h)}{n_{<w_{i-1},*>} + \alpha_1} \#(\text{Equation 2.6})$$

$n_{<w_{i-1},w_i>}$ represents the number of co-occurrences of the binary grammar $n_{<w_{i-1},w_i>}$ in the observed segmentation sequence. Similarly, $P_D(w_i|h)$ can be calculated as:

$$\frac{t_{wi} + \alpha H(w_i)}{t + \alpha} \#(\text{Equation 2.7})$$

Chinese Restaurant Process [16] is used to calculate the Dirichlet Process, t_{wi} represents the total number of tables related to wi, T is the total number of tables, and $H(w_i)$ is the prior distribution of words. In the Dirichlet process, the commonly used model inference algorithm is Gibbs Sampling. In each sampling, Gibbs Sampling fixed other dimensions and selected one dimension for sampling. Through iterative training, the model converges under the condition that the probability of string segmentation is constant, to carry out word segmentation.

Based on the HDP model, Yu et al used MSIT (Multi-stage Iterative Training) [15], a method based on iterative-training algorithm, to introduce the high-parameter pre-training model, aiming to improve the accuracy of unsupervised word segmentation with the knowledge implicit in the pre-training model. Using three stages of iterative training, MSIT can perform high-quality unsupervised word segmentation on target domain text by building a domain language model, an unsupervised model based on statistical information, and iterative training of both models. In the open test using the manually annotated Chinese word segmentation data set, the F1 value of the unsupervised learning word segmentation model in the annotation set of Zuozhuan reaches 90%, and it has achieved good performance in the ancient Chinese corpus of multiple periods, making it the optimal word segmentation model in the current ancient Chinese word segmentation.

2.2 Named Entity Recognition

Automated knowledge extraction and parsing usually start with entity extraction. Named Entity Recognition (NER) is the key step of text knowledge mining and knowledge structure analysis. It automatically identifies and classifies concepts from texts, and the recognized objects are generally proper nouns such as human name, place name, and organization name [17]. In the field of medical research, NER has a stronger professional, and the recognition objects also include drug names, symptoms, and other domain-specific concepts. NER is mainly based on the rule method of dictionary and language features, machine learning method of statistical learning, and deep learning method of neural network. Compared with the previous two methods, deep learning-based methods are widely used in NER tasks because they do not require complex feature engineering to construct text features and can achieve better performance in the case of sufficient data.

When constructing an ontology or knowledge graph, concepts are generally obtained as nodes through entity extraction from text. The performance of a NER system is affected by many factors, including language factor, text type factor, and entity type. For Chinese NER, language is generally regarded as a continuous sequence of words, and the method of sequence annotation is used for entity recognition. Among them, the text field for NER has a great impact on the performance of the entity recognition system. Generally speaking, the NER performance of the medical field is lower than that of the general field [17]. Entity type also has a great influence on NER systems. Entity types in general domains, such as names and dates, are difficult to identify, while entity types in specialized domains, such as drug names and chemical names, rely on domain text resources, are difficult to generalize, and often include nested entity types, so are difficult to identify. In the field of ancient TCM books, named entities include entities in the field of TCM, such as prescriptions and medicinal materials, as well as entities commonly seen in ancient texts, such as dynasties

and names. These entities have rich semantics, some of them are highly specialized, and there are a lot of nested entities (such as nested medicinal materials in prescriptions), so it is difficult to identify them.

The existing entity recognition systems are generally based on rules, supervised learning methods, semi-supervised and self-supervised methods. Early entity recognition systems were generally based on rules. These systems rely on prior knowledge to formulate domain-related entity type dictionaries, formulate linguistic-based grammar rules based on texts, and conduct entity discovery based on rules and dictionaries. For example, Rau et al. [18] proposed a system for identifying entity types of company names based on heuristic algorithms and manual rules. This type of entity recognition system can achieve better performance in its target domain because it relies on a large amount of domain expertise and therefore is costly and difficult to generalize. With the development of machine learning, systems based on statistical learning methods and supervised learning have been widely adopted. These methods regard named entity recognition as a classification task and have better generalization performance and lower annotation requirements compared with rule-based systems but still rely on domain knowledge. These include the methods for modeling sequences mentioned in the previous section, such as HMM, CRF, and so on. The first HMM-based entity recognition system was proposed by Bikel et al. [19] in the last century, which recognized named entities such as names and dates and achieved high accuracy and recall rate. Szarvas et al. [20] proposed a multi-language NER system based on the decision tree correlation algorithm and obtained the final NER result by training multiple decision tree classifiers and using the integration algorithm. Settles[21] builds a named entity recognition system in the medical field based on CRF and rich feature templates and predicts entities such as proteins and cell types, whose F1 value reaches 70% in general. The NER system based on supervised learning and statistics can achieve good entity recognition performance in the general domain, but its performance is highly dependent on the developed feature template and annotation data.

With the development of big data and deep learning, the deep learning method based on supervised learning has been widely used in entity recognition systems and achieved good results. Through self-supervised tasks such as autoregression, deep learning systems can extract features from text without relying on feature engineering, and deep learning models can fit complex data with nonlinear distribution well, so they quickly become the main method used in NER. Language model based on neural network is the core of system based on deep learning. The language model is a kind of model used to describe the probability of sequence generation [24]. For a given sequence of characters $(t_1, t_2, t_3 \ldots t_n)$ and the current character tk, the forward language model passes the history sequence of the current character $(t_1, \ldots, t_{k-1})$ gives the conditional probability of the current character:

$$p(t_1, t_2, \ldots, t_n) = \sum_{k=1}^{n} p(t_k | t_1, t_2, \ldots, t_{k-1}) \#(\text{Equation 2.8})$$

Similar to the forward language model, the backward language model is based on the reverse order of the sequence, giving the conditional probability of the occurrence of the current character through the future sequence of the current character:

$$p(t_1, t_2, \ldots, t_n) = \sum_{k=1}^{n} p(t_k | t_{k+1}, t_{k+2}, \ldots, t_n) \#(\text{Equation 2.9})$$

Representative models of deep language models are the RNN model and its variant model [22], which can effectively model character sequences and word sequences to obtain context-related feature representation, and achieve good performance and generalization results in NER tasks, and are less dependent on domain knowledge and feature engineering. Such models include the LM-LSTM-CRF model [23], in which language model and sequence model share character-level text representation, and feature representation of language model and word-level feature representation of sequence model is superimposed for named entity prediction, achieving good performance in a variety of NER tasks.

For long string sequences, the neural language model based on RNN and its variants are difficult to capture word dependence between long distances, and its architecture is difficult to carry out large-scale parallel computing, which reduces computational efficiency to a certain extent. On the contrary, based on the attention model USES the connection of feed-forward neural network architecture, to be able to capture the relationship between any two words, and through the weights in dynamic decision model of the mechanism of attention, and variable-length sequence modeling, compared with the model based on RNN or CNN, better able to capture the distance dependence, therefore, Neural network models based on self-attention are the mainstream models of NER at present, including Transformer model [24], BERT[25] and a large parameter pre-training model based on Transformer, etc. These models have been applied to NER tasks in various fields of text, and have achieved the current optimal results.

The models and algorithms used in NER tasks in ancient Chinese medicine books lag behind those used in modern Chinese. Most of them adopt LSTM-based models or statistical learning models based on feature engineering. For example, Meng Hongyu et al. [26] automatically extracted TCM terms such as disease name, pulse condition, and prescription based on CRF. Ye Hui et al. [27] developed the feature template of CRF based on diverse features, and extracted symptoms and drug entities. Zhang Yipin et al. [28] extracted diseases, formulae, medicinal materials, and other entities based on the BiLSTM+CRF model. When the training samples were about 30,000, the F value reached more than 90%. Gao Su et al. [29] used the BiLSTM+CRF model to identify TCM cognition methods, TCM physiology, and other entities in Huangdi Neijing, and the F value reached more than 85% when the training samples were about 10,000. The method based on supervised learning requires carefully designed feature templates and a certain scale of annotation data, which rely on expert domain knowledge and require different annotation data for different entities. To achieve higher model performance, a large amount of human and material resources are required to annotate samples.

2.3 Relationship Extraction

Relationships usually refer to meaningful relationships between two or more entities [30]. In unstructured natural language text tasks, relation extraction refers to extracting semantic relationships from text, usually by defining naming relationships between named entities. Entity relation extraction is the core step of knowledge graph construction and ontology learning engineering, as well as the foundation of deep text parsing and structuring.

Relationship extraction methods are mainly divided into three categories: supervised extraction based on syntactic features, such as constructing syntactic dependency maps [31]; Semi-supervised extraction methods include Bootstrapping method, of which SnowBall

method is a representative one [32]; Active Learning method [33]; Method of Label Propagation [34]; Methods based on unsupervised learning, mainly including clustering methods. In recent years, with the development of deep learning, supervised relationship based on the deep model has been developed, which is greatly improved compared with traditional methods. Zhou[35] et al. used a bidirectional circulating neural network to extract features from sentences in combination with attention mechanism and adopted a classification-based method to extract features, which achieved good results. Wei et al. [36] unified the tasks of relationship extraction and entity extraction in a hierarchical annotation framework, avoiding error transmission in the pipeline of entity extraction and relationship extraction.

In the field of TCM ancient books, most of the relationship extraction models for TCM text use feature extraction of recurrent neural network + supervised learning classifier for relationship extraction. Luo Jigen et al. [37] proposed a relationship recognition algorithm based on a bidirectional LSTM network and gradient ascension tree, which marked more than 20,000 sentences and 11 categories of data sets. Word2vec was used to vector represent TCM texts, and gradient ascension tree was used as a classifier. Using supervised learning method to fit the existing feature relationship to find the relationship of TCM texts, the average F value is more than 80%; Zhao Lipeng [38] constructed GRU+CNN+ATTN model based on cyclic gate neural network + convolutional neural network + attention, used cyclic neural network and convolutional neural network to extract global feature and local feature input classifiers of text, and used supervised learning method to train relationship recognition model. The accuracy rate of relation classification is about 80% on 5000 sentences corpus.

Based on the investigation of knowledge mining-related literature in the TCM field, it is found that there are still many problems in TCM text parsing and knowledge discovery at present, among which the main problems are more manual annotation data and training data, and poor performance in a complex context. Most of the existing researches focus on the extraction of inter-sentence relationships, and the performance of inter-sentence extraction is poor. Most of the existing studies focus on supervised learning and the open relation extraction based on weakly supervised and unsupervised has poor performance.

3. Difficulties Faced by Knowledge Mining of Ancient Chinese Medicine Books.

3.1 Limitations on the Implied Text Structure of Single Texts

The single text refers to short text without explicit text structure and with limited implicit knowledge structure. Different from single texts, non-single texts are usually longer, have explicit text structure division, and contain rich implicit knowledge structure. Compared with single texts, non-single texts are easier to identify and extract. A single text is usually derived from a short article, a paragraph, or even a few sentences in the field of expertise, rather than a single text, which can be an essay, a monograph, a collection of articles, or an article with a clear structure and rich themes.

Since the implicit structure of a single text is not obvious, and the explicit semantic and textual information is not rich, it is difficult to directly analyze the knowledge structure of a single text and to carry out the task of analyzing individual texts of ancient Chinese medicine books through traditional supervised learning. Supervised learning relies on a manually annotated large-scale corpus, and the model can infer the categories of new data

by fitting the probability distribution of labels of the existing annotated data, thus indexing concepts and relationships. However, this method relies too much on the manually annotated data: the analysis of target resources depends on the artificially annotated labels in the annotated data, and the model training based on these data may not be able to be well generalized to the fields not involved in the annotated data. However, for the single text proposed in this paper, due to its short length and few concepts, the implied knowledge structure may not coincide with the concepts involved in the annotation field, making it difficult to parse. In addition, it also faces the problem of missing annotation data mentioned in the previous article: it is difficult to annotate data in the field of ancient Chinese medicine books, with high requirements for domain knowledge and natural processing knowledge, and there is no available public data set at present.

Therefore, knowledge structure analysis of a single text can only be carried out through small samples or unsupervised learning. However, knowledge structure analysis in this paper relies on the full indexing of knowledge concepts. Theoretically, the maximum number of concept relations is C_n^2, where N is the number of categories of all concepts. Assuming n=10, the maximum number of relationship categories is 45. Coupled with the development and test sets of test model generalization performance, a certain number of annotations are still required. Although the number of labeling has decreased significantly compared with traditional supervised learning, some labeling work is still needed. In addition, from the perspective of engineering, small sample learning is still a kind of supervised learning. If the parsing task changes, the data set needs to be adjusted manually without the construction of engineering pipelines.

It is also difficult to parse a single text traditionally based on unsupervised learning. Unsupervised relationship extraction is usually based on statistical information of rules and articles. For non-single text, the implicit concepts in the paper are rich, which can effectively extract unsupervised relations based on word statistics, such as Snowball [32] algorithm which is commonly used based on large-scale data iteration rules, Stanford OpenIE[39] algorithm which extracts relations based on syntactic analysis of part of speech, etc. However, for a single text, the above word-based statistical information algorithm cannot be carried out, because a single text is short and does not contain rich explicit grammatical information. Compared with non-single text, the single text is shorter and there is no explicit text structure to extract relationships. However, for multiple texts with explicit structure norms, relationships can be extracted by making rules. In addition, a single text contains relatively fewer concepts, statistical information is not rich, and hidden structure is difficult to discover; Non-single text has more dominant grammatical information, such as word frequency and co-occurrence, which can be used to better mine knowledge structure.

3.2 The Existing Knowledge Structure Analysis Algorithm Does Not Match

In ancient Chinese medicine books, based on machine learning and with the help of natural language processing, the biggest limitation is the absence of annotated data. In terms of professional fields, there are relatively few automated digital researches on ancient Chinese medicine books, and even fewer public data sets are available. Based on literature research, it is found that most of the natural language processing tasks for Chinese medicine texts are carried out based on their manually annotated non-public data sets. Therefore, although commonly used knowledge extraction and text knowledge structure

construction algorithms are mature and have good performance, they are difficult to be carried out. In the field of ancient Chinese medicine books, there is no public data set that can be used as a test set for method evaluation, and there is no annotated data. All data need to be constructed by ourselves. Technically, it is impossible to use more mature machine learning methods such as supervised learning. Traditional machine learning methods include SVM, CRF, HMM, and so on. Based on an annotated corpus, the generation probability of natural language tag sequence is modeled by constructing features, to predict new tags. These methods can achieve high accuracy and recognition speed but rely on the artificial construction of features.

Supervised learning relies on a manually annotated large-scale corpus, and the model can infer the categories of new data by fitting the probability distribution of labels of the existing annotated data, thus indexing concepts and relationships. However, this method relies too much on the manually annotated data: the analysis of target resources depends on the artificially annotated labels in the annotated data, and the model training based on these data may not be able to be well generalized to the fields not involved in the annotated data.

Extracting conceptual relationships from a single text of ancient Chinese medicine texts cannot be performed based on commonly used unsupervised relationship discovery algorithms. Existing unsupervised knowledge structure analysis algorithms are usually based on the syntactic analysis of the text. Through word segmentation, part of speech tagging, dependency parsing, and semantic role tagging, the algorithm matches chunks that meet certain grammar rules from a large number of texts and outputs the matched contents as duals or triples to obtain the corresponding knowledge structure. Common tools include TextRunner, Stanford OpenIE, OLLIE, DeepDive, etc. Combined with relevant reviews [40], the more popular or representative structure analysis tools are summarized, as shown in Table 1:

Table 1 Summary of some unsupervised relationship discovery tools

Name	Principle
Snowball	Bootstrap Template Iteration
TextRunner	Dependency Analysis, POS Features
Stanford OpenIE	Clause Construction, NER, POS Features, Dependency Analysis
PredPatt	Universal Dependencies (POS, Formal Grammar, etc.)
DeepDive	POS Features, Entity Recognition, Dependency Features, Rule Reasoning

Analysis of its principles shows that most of the relationship extraction operations of these tools are performed through syntactic analysis. The syntactic analyzer is learned on pre-labeled language resources through supervised learning. In the inference process, based on pre-trained part-of-speech tagging tools and syntactic analysis. The device extracts the shallow syntactic relationship between different language blocks from the sentence and finds frequent sets from a large number of texts through a predetermined small rule set or

Bootstrap Iteration, to perform knowledge discovery. Therefore, for single texts of ancient Chinese medicine books, none of the above extraction models can be performed for the following reasons:

1. These tools still rely on part of speech analysis and syntax analyzer at the bottom. These parts of speech analysis and syntactic parsers are obtained by supervised training of annotated data from specific resources, and cannot be generalized for texts without corresponding annotated resources.

2. These tools rely on large-scale corpus iteration to discover association rules of the relationships between language blocks, based on which to predict new relationships. When there is no large-scale corpus and implicit semantic knowledge in the text is insufficient, that is, for a single text, it is impossible to discover relationships.

3. All knowledge structure extraction tools are word-oriented, which cannot be applied to ancient Chinese texts with many words or ancient Chinese medicine books that cannot express entities with words.

The study further selected Stanford OpenIE, the most commonly used relational extraction tool, to test the above views on a single text of ancient Chinese medicine books. The results showed that OpenIE could not extract any relationship on a single text, as shown in Figure 1:

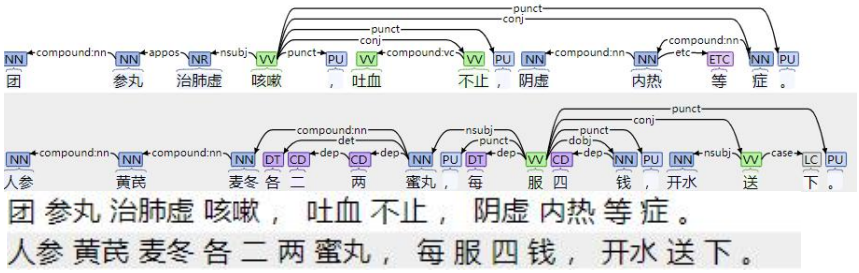

Figure 1 OpenIE recognition results of the single text (unable to extract relation)

Based on the above analysis, it is impossible to use the explicit knowledge structure and grammatical information in the non-single text to analyze the knowledge structure in the single text of ancient Chinese medicine books.

4. Conclusions and Prospects.

The numerous ancient books of TCM cost a lot of manpower and material resources to sort out. Traditional supervised learning and unsupervised learning methods are not suitable for the single knowledge analysis of ancient Chinese medicine books. They are mainly manifested in the following points: there is no labeled data can be used to train supervised learning models in the field of ancient Chinese medicine books. It requires certain domain knowledge to collate and proofread it, which increases the difficulty of manual proofreading of ancient texts; the implicit concepts in a single text are not rich and difficult to extract, and there is no explicit discourse structure in ancient Chinese medicine books. Moreover, due to the short length and lack of rich concepts, it is usually difficult to discover the relationship between concepts, and to construct the implicit discourse structure based on it.

Based on the research of predecessors, this paper systematically reviewed the basic

task of the Chinese medical knowledge mining and its related technology, showing that using supervised learning and unsupervised learning of traditional methods to analyze the knowledge structure does not adapt to the characteristics of TCM ancient books, namely, its hidden discourse structure. Therefore, it is unable to realize the full of ancient text parsing. In the future, research on knowledge mining of ancient Chinese medicine books should focus more on in-depth analysis of text knowledge structure and knowledge-centered text processing, to obtain more refined knowledge resources of ancient Chinese medicine books.

REFERENCES

[1] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 647-657.

[2] Wallach H M. Conditional random fields: An introduction[J]. Technical Reports (CIS), 2004: 22.

[3] Rabiner L, Juang B. An introduction to hidden Markov models[J]. ieee assp magazine, 1986, 3(1): 4-16.

[4] Wallach H M. Conditional random fields: An introduction[J]. Technical Reports (CIS), 2004: 22.

[5] Wang Xiaoyu, Li Bin.Automatically Segmenting Middle Ancient Chinese Words with CRFs [J]. Data Analysis and Knowledge Discovery, 2017, 1(05): 62-70.

[6] Shi Min, Li Bin, Chen Xiaohe. CRF Based Research on a Unified Approach to World Segmentation and POS Tagging for Pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010, 24(2): 39-45.

[7] Cai D, Zhao H, Zhao H, et al. Neural word segmentation learning for Chinese[J]. arXiv preprint arXiv:1606.04300, 2016.

[8] Liu Yutong, Wu Bin, Xie Tao, Wang Bai.New Word Detection in Ancient Chinese Corpus[J]. Journal of Chinese Information Processing,2019,33(01):46-55.

[9] Chang J S, Lin T. Unsupervised word segmentation without dictionary[C]//ROCLING 2003 Poster Papers. 2003: 355-359.

[10] Magistry P, Sagot B. Unsupervised word segmentation: the case for mandarin chinese[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012: 383-387.

[11] Magistry P, Sagot B. Can MDL Improve Unsupervised Chinese Word Segmentation? [C]//Sixth International Joint Conference on Natural Language Processing: Sighan workshop. 2013: 2.

[12] Chen M, Chang B, Pei W. A joint model for unsupervised Chinese word segmentation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 854-863.

[13] Goldwater S, Griffiths T L, Johnson M. A Bayesian framework for word segmentation: Exploring the effects of context[J]. Cognition, 2009, 112(1): 21-54.

[14] Mochihashi D, Yamada T, Ueda N. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009: 100-108.

[15] Yu Jingsong, Wei Yi, Zhang Yongwei, Yang Hao. Word Segmentation for Ancient Chinese Texts

Based on Nonparametric Bayesian Models and Deep Learning [J]. Journal of Chinese Information Processing,2020,39(6).

[16] Blei D M, Griffiths T L, Jordan M I. The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies[J]. Journal of the ACM (JACM), 2010, 57(2): 1-30.

[17] Goyal A, Gupta V, Kumar M. Recent named entity recognition and classification techniques: a systematic review[J]. Computer Science Review, 2018, 29: 21-43.

[18] Rau L F. Extracting company names from text[C]//Proceedings the Seventh IEEE Conference on Artificial Intelligence Application. IEEE Computer Society, 1991: 29, 30, 31, 32-29, 30, 31, 32.

[19] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]//Proceedings of the fifth conference on Applied natural language processing. 1997:194-201.

[20] Szarvas G, Farkas R, Kocsor A. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms[C]//International Conference on Discovery Science. Springer, Berlin, Heidelberg, 2006: 267-278.

[21] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]//Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP). 2004: 107-110.

[22] Jing K, Xu J. A survey on neural network language models[J]. arXiv preprint arXiv:1906.03591, 2019.

[23] Liu L, Ren X, Shang J, et al. Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1215-1225.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[25] Devlin J, Chang M W, Lee K, et al. Bert: Research on deep Bidirectional Transformers [J]. arXiv preprint arXiv:1810.04805, 2018.

[26] Meng Hongyu, Meng Qinggang. TCM Terminology Extraction Method and Application Based on Conditional Random Field[J].CHINESE ARCHIVES OF TRADITIONAL CHINESE MEDICINE 2014, 032(010):2334-2337.

[27] Ye Hui, Ji Donghong. Research on Symptom and Medicine Information Abstraction of TCM Book Jin Gui Yao Lue Based on Conditional Random Field [J]. Chinese Journal of Library and Information Science for Traditional Chinese Medicine, 2016, 040(005):14-17.

[28] Zhang Yipin, Guan Bei, Lv Yirun, et al. Study on the Entity Extraction Method of Traditional Chinese Medicine on the Basis of Deep Learning[J]. Chinese Journal of Library and Information Science for Traditional Chinese Medicine, 2019, 40(02):62-67.

[29] Gao Su, Jin Pei, Zhang Dezheng. Study on the Entity Extraction Method of Traditional Chinese Medicine on the Basis of Deep Learning[J]. Journal of Medical Informatics, 2019, 005(001):113-123.

[30] Pawar S, Palshikar G K, Bhattacharyya P. Relation Extraction: A Survey [J]. arXiv preprint arXiv:1712.05191, 2017.

[31] Kuffner R, Josik K, Kupchak K, et al. Relation extraction from dependency parse trees: a review [J]. Bioinformatics, 2007, 23 (3) : 365-371.

[32] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections[C]//Proceedings of the fifth ACM conference on Digital libraries. 2000:85-94.

[33] Angeli G, Tibshirani J, Wu J, et al. Combining distant and partial supervision for relation extraction[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1556-1567.

[34] Chen J, Ji D, Tan C L, et al. Relation extraction using label propagation-based semi-supervised learning[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006: 129-136.

[35] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.