Extracting Disease-Specific Clinical Experiences from Ancient Literature of Traditional Chinese Medicine with Deep Learning

Yongmei Lu

Sichuan University

Li Chen (✉ cl@scu.edu.cn)

Sichuan University

Tingting Zhang

Chengdu University of Traditional Chinese Medicine

Lingmei Bu

Sichuan University

Ying Ye

Chengdu University of Traditional Chinese Medicine

Zhonghua Yu

Sichuan University

Hao Wang

University of Western Australia

Research Article

Keywords: Traditional Chinese Medicine, Ancient literature, Disease-Specific Clinical Experiences, Deep learning

Posted Date: March 3rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-258949/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Extracting Disease-Specific Clinical Experiences from Ancient Literature of Traditional Chinese Medicine with Deep Learning

Yongmei Lu¹, Li Chen^{1*}, Tingting Zhang², Lingmei Bu¹, Ying Ye³, Zhonghua Yu¹ and Hao Wang⁴

Abstract

Background: Ancient literature of Traditional Chinese Medicine (TCM) contains massive clinical experiences which are important ingredient of TCM knowledge and valuable for TCM clinical practice of nowadays. However, it is difficult for TCM professionals to acquire such valuable experiences due to their massive volume and broad occurrence in the literature. Furthermore, different characteristics of ancient Chinese language from the modern one lead to additional challenges for analyzing the literature, regardless of how to perform the analyzing, manually or automatically with a software toolkit.

Methods: In order to overcome the aforementioned challenges, we formalize a novel information extraction task for ancient literature of TCM, and the entities to be extracted are Disease-Specific Clinical Experiences (DSCEs) occurring in the literature. For the purpose, we have collected two corpora from ancient literature of TCM and annotated them manually with DSCEs occurrence information for the diseases *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) and *jaundice* (黄疸) respectively. We further propose a deep learning and CRF-based algorithmic framework with character encoding of ancient Chinese, thus avoiding the special difficulty in word segmentation for ancient Chinese texts. We investigate the framework with different methods for contextual encoding of characters in a sentence, including CNN, Bi-LSTM and BERT, and diverse approaches to aggregate contextual information of characters into a sentence encoding, such as max-pooling and attention mechanism. After that all the encoded sentences in a section of the literature are passed through a Bi-LSTM-based sequence labelling model with CRF inference on its top to obtain an optimal label sequence for the sentences in the section.

Results: We conduct a series of experiments on the two corpora to verify the effectiveness of our framework for the task, and evaluate its effectiveness with different metrics in two granularities of labelling, namely accuracy/ F_1 -value in sentence-level labelling and precision/recall/ F_1 -value in correct recognition of the whole DSCEs.

Conclusion: The experimental results demonstrate that the deep learning and CRF-based framework with character encoding of ancient Chinese could achieve an accuracy of $80.40\% \pm 1.64\%$ and an F_1 -value of $76.73\% \pm 1.59\%$ for the sentence labelling, while for recognition of the whole DSCEs, it is able to obtain the recall of $44.97\% \pm 2.16\%$ and the precision of $51.13\% \pm 2.64\%$, meaning that the framework is a promising baseline for further development of the novel information extraction task for TCM.

Keywords: Traditional Chinese Medicine; Ancient literature; Disease-Specific Clinical Experiences; Deep learning

Background

Traditional Chinese Medicine (TCM) has been providing protection of Chinese people from different diseases for several thousands of years. In such a long

period of the practice massive experiences for diagnosis and treatment are distilled, which are recorded in ancient literature of TCM and have become an important ingredient of TCM knowledge. The experiences are also one of the most critical referenced resources for modern TCM clinical practice and research. Even the achievement of Youyou Tu for the 2015 Nobel

*Correspondence: cl@scu.edu.cn

¹Department of Computer Science, Sichuan University, Chengdu, China
Full list of author information is available at the end of the article

Prize of medicine, the therapy against Malaria with artemisinin, was inspired by an ancient clinical experience of that mentioned in *A Handbook of Prescriptions for Emergencies* (肘后备急方), namely “A handful of Qinghao immersed in two liters of water, wring out the juice and drink it all.” (青蒿一握,以水二升渍,绞取汁,尽服之)[1].

However, extracting manually the experiences from ancient literature of TCM is time-consuming and labor-intensive due to the massive volume of the literature. According to [2], there are more than 10,000 distinct TCM ancient books in China with more than 37,000 different versions, and the experiences are scattered across almost all of them. Furthermore, the literature is written in ancient Chinese language, a language that is different from the modern one, which leads to additional difficulties for analysis and understanding of the literature, regardless of the way of the analysis and understanding, manually or automatically with the off-the-shelf software toolkits such as information retrieval systems. An information retrieval system always returns a list of documents but the experiences needed for modern TCM research and clinical practice are hidden in several ancient books, where an experience constitutes just a text span in a section of the books. Moreover, the existing information retrieval systems for Chinese textual documents heavily rely on the word segmentation as a foundational operation, but in ancient Chinese language, there is a very vague bordering between word and character, which obviously hinders their direct applications in extraction of the experiences from ancient literature of TCM.

The importance of ancient literature for modern TCM research and clinical practice is realized early enough, but only recently some researchers have become to resort to the text mining and information extraction techniques in order to overcome the challenges mentioned above. These existing studies have focalized their attention mostly on semantic annotation[3], syndrome differentiation[4], clinical record classification[5, 6], knowledge graph construction[7], to name a few. However, to our best knowledge, there does not exist any study focusing specially on extracting clinical experiences from the literature of TCM, which are critical for TCM research and clinical practice, as mentioned above.

In order to bridge the gap in extraction of TCM clinical experiences, in this paper we propose a deep learning and CRF-based algorithmic framework to extract Disease-Specific Clinical Experiences (DSCEs)

from TCM ancient literature. The framework solves the task with a sentence sequence labelling approach[8] and incorporates character encoding of ancient Chinese into itself, thus avoiding the special difficulty in word segmentation for ancient Chinese texts. We investigate the framework with different methods for contextual encoding of characters in a sentence, as well as diverse approaches to aggregate contextual information of characters into a sentence encoding. All the sentences encoded with such a way in a section of the literature are passed through a Bi-LSTM-based sequence labelling model with Conditional Random Fields (CRF)[9] inference on its top to obtain an optimal label sequence for the sentences in the section.

In order to evaluate the performance of the framework for this novel information extraction task (i.e., extracting clinical experiences from ancient literature of TCM), we have collected two corpora from ancient literature of TCM and annotated them manually with DSCEs occurrence information for the diseases *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) and *jaundice* (黄疸) respectively. We perform a series of experiments on these corpora to verify the effectiveness of our framework, evaluating its effectiveness with different metrics in two granularities of labelling, namely accuracy/F₁-value in sentence-level labelling and precision/recall/F₁-value in correct recognition of the whole DSCEs. The experimental results show that the framework with character encoding of ancient Chinese could achieve an accuracy of 80.40%±1.64% and an F₁-value of 76.73%±1.59% for the sentence labelling, while for recognition of the whole DSCEs, it is able to obtain the recall of 44.97%±2.16% and the precision of 51.13%±2.64%, meaning that the framework is a promising baseline for further development of the novel information extraction task for TCM.

Summarily, the major contributions of this work are four-fold: (1) a novel information extraction task is proposed for extracting DSCEs from ancient literature of TCM; (2) a deep learning and CRF-based algorithmic framework with character encoding of ancient Chinese is devised to solve the task; (3) a series of experiments is conducted to investigate the performance of the framework with alternative techniques in its different modules; (4) the two benchmark corpora and the source code of this paper are released for further development in solution of the novel information extraction task^[1].

^[1]https://github.com/153lym/TCM_DSCEs

Related work

In recent years, many disciplines witness fast growth in exploiting machine learning and text mining technologies to discover knowledge hidden in a massive volume of data. Similarly, some efforts have raised in TCM which utilize machine learning and text mining technologies for discovering knowledge from prescriptions and clinical records, such as treatment rules mining[10, 11], medical term extraction[12–14], syndrome differentiation[15], knowledge graph construction[16] and fine-grained entity corpus construction [17]. However, majority efforts in these studies are devoted to structured data or unstructured textual data written in modern Chinese language, in spite of the importance of ancient literature for modern TCM research and clinical practice, as mentioned in Section *Background*.

Fortunately, a couple of pioneering works on knowledge discovery from ancient literature of TCM appear now. Weng et al.[3] proposed to employ a HMM (Hidden Markov Model) for recognizing medical phrases related to spleen and stomach from ancient TCM literature, and on the basis develop a knowledge-mining system to support the spleen-related research in TCM. There are various academic schools in TCM formed at different regions of China. Nie et al.[4] presented a data mining approach to investigate the characteristics of the *Bashu* (巴蜀) school in exposing the relationships among diseases, symptoms and herbs. The source data used for the mining are the ancient literature published by the school of TCM. Yao et al.[5] investigated the issue of automatic classification of clinical records from *Classified Medical Records of Distinguished Physicians Continued Two* (二续名医类案), a famous book with many clinical records provided by distinguished TCM physicians from different periods of the history. In order to solve the issue, they examined a variety of shallow classification algorithms (such as SVM, MaxEnt and Random Forest) with various kinds of textual features from the literature, including document embedding powered with TCM MeSH. In their subsequent work Yao et al.[6] further proposed a BERT-based method for the issue, and with the pre-trained model a much better performance is gained. Zhou et al.[7] outlined a knowledge graph construction based on the key concepts related to diseases, symptoms, prescriptions and herbs extracted from ancient books of TCM, demonstrating its possible applications for TCM research and clinical practice.

Knowledge discovery from ancient TCM literature is still in its infancy and achieves only limited results due to the unique characteristics of ancient Chinese

language relative to the modern one, as mentioned in Section *Background*, which always lead to a failure when applying the existing Natural Language Processing (NLP) tools on these ancient Chinese texts. Therefore, some authors turn to the fundamental tasks for ancient Chinese literature processing of TCM such as word segmentation and Part-of-Speech (POS) tagging in [18].

Put simply, the studies for automatic processing of TCM ancient literature are limited, and to the best of our knowledge, there is no research on extraction of DSCEs from ancient literature of TCM, in spite of their importance in TCM academic research and clinical practice. In order to bridge the gap in mining of TCM ancient literature, in this paper we formalize a novel information extraction task for TCM ancient literature, where the entities to be extracted are DSCEs mentioned in the literature, and further propose a deep-learning-based framework to solve the task. We detail the task and the framework in the following section.

Task definition and method

In this section we elaborate on the task definition of extracting DSCEs from ancient literature of TCM and introduce the framework for solving the task, detailing different methods in it. The solution in general consist in a sequence labelling process with the sequence-to-sequence deep learning architecture as its core component coupled with a CRF inference level on its top to obtain an optimal label sequence for all the sentences in a section of an ancient book of TCM. The technical details of the framework are described in *Sequence labelling framework* subsection .

Task definition

The new information extraction task we formalize here is to identify text spans in ancient TCM literature which describe experiences for a particular disease summarized by ancient TCM doctors and described in the literature. We call such text spans DSCEs for exposition brevity. Figure 1 exemplifies some segments from an ancient TCM book titled with *Qixiao Liangfang* (奇效良方), where the text spans underlined are describing some experiences for *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血). It is obvious from the underlined contents that an experience description for a particular disease almost always contains its pathogenesis, syndrome differentiation, clinical symptoms, etc., which are critical references in modern TCM research and clinical practice.

…… 治胎冷腹胀，痛引两胁，小便频数，大便虚滑。诃子（面裹煨，去核）白术（各二钱）陈皮（去白）高良姜（炒）木香（不见火）白芍药陈米（炒）甘草（炙，各一钱）上作一服，水二盅，生姜五片，煎至一盅，不拘时服，忌食生冷之物。治妊娠怔悸，睡里多惊，腹胁膨胀，坐卧不宁。白茯苓（去皮）麦门冬（去心）…… 生姜三片，竹茹弹子大一块，煎至一盅，食前服。治妊娠恶阻，胃寒吐逆，反胃吐食，及心腹刺痛。人参（四钱）丁香（一钱）藿香（二钱半）上作一服，水二盅，煎至一盅，不拘时服。…… 治妊娠伤寒。阿胶（炒）人参桑寄生白茯苓（去皮）白术（炙，各等分）上为细末，每服二钱，不拘时糯米饮调下。治妊娠宿有冷，胎痿不长，或失于将理，伤胎多坠，此药补荣卫，养胎气。白术川芎（各一两）吴茱萸（汤泡，半两）甘草（炙，一两半）上为细末，每服二钱，食前温酒调下，忌生冷果实之物。治妊娠伤寒，头疼壮热，肢节烦疼。前胡白术（各一钱）黄芩（二钱，炒）梔子仁知母木香（各二钱）上作一服，水二盅，竹茹一块，葱白三寸，煎一盅，不拘时服 …… ……

Figure 1 DSCEs examples in an ancient TCM book

Considering that in most situations an experience description consists of several successive sentences in a section of an ancient TCM book, and hardly traverses borders between sections, we thus propose to regard the task as a sequence labelling issue for the sentence sequence in a section of an ancient TCM book, and employ the traditional BIO tagging scheme on a sentence level. Performing sequence labelling directly on a section rather than on a book could greatly shorten the sequence length to be labelled and thus raise the efficiency of the labelling. After the sequence labelling with the BIO scheme for a section, any text span in this section, starting at a sentence with tag ‘B’ and continuing until a first encounter of an another tag ‘B’ or ‘O’ after the starting sentence, is formed into a DSCE. Figure 2 shows the top flowchart of the DSCEs extraction process.

Sequence labelling framework

For sequence labelling on a section of an ancient TCM book, we propose a deep learning and CRF-based algorithmic framework with character encoding of ancient Chinese in order to avoid the special difficulty in word segmentation for ancient Chinese texts. The framework is depicted in Figure 3.

As shown in Figure 3, our framework consists of four layers, namely an embedding layer, a sentence encoding layer, a sentence contextual encoding layer and a CRF layer. The embedding layer is responsible for initializing every character in a sentence with its corresponding dense real-valued vector for further processing in the next layers. The sentence encoding layer, taking the dense real-valued vectors of all the characters in a sentence, aims to capture the contextual embeddings of the Basic Feature Units (BFUs) in the

sentence, and on the basic aggregate them into an encoding of the sentence. A BFU for a sentence could be its character, character n-gram or even the $[CLS]$ returned by the BERT model, depending on the model used for capturing the contextual embeddings. All the sentences in a section encoded in the sentence encoding layer are passed through the sentence contextual encoding layer to further incorporate contextual information in the section into all its sentence encodings. After that, these sentence encodings are fed into a multilayer perceptron respectively, which returns a BIO-tag probability distribution for every sentence. Finally, the CRF layer takes the BIO-tag probability distributions for all the sentences in the section and generates an optimal BIO-tag sequence for the sentence sequence of the section. We examine diverse techniques in the layers of the framework, and detail them in the subsequent subsections.

Embedding Layer

In recent years, the deep learning approaches with word embedding are so popular in processing textual data that almost all the state-of-the-art algorithms for such data processing are exclusively deep-learning-based. However, such approaches always suffer from the issue of out-of-vocabulary (OOV) words due to the skewed distribution of words in textual data. In order to overcome the OOV challenge, some approaches to embed sub-words (characters for Chinese) rather than whole words are proposed, which demonstrate an improved performance in many text processing applications[6, 19, 20]. Contrary to processing texts of modern Chinese language, processing ancient Chinese texts has to be confronted with another special difficulty, namely the difficulty in word segmentation

due to the vague difference between a word and a character. Therefore, embedding individual characters, instead of whole words, is more reasonable for ancient Chinese texts. Therefore, in our framework a sentence is regarded as a sequence of characters, and characters are our minimal processing unit rather than words. In the embedding layer the characters are embedded with the dense real-valued vectors pre-trained specially for ancient Chinese in TCM. Concretely, we have collected 693 ancient TCM books as the corpus, regarding the characters as the words to be embedded and train-

ing the character embedding vectors with Google's word2vec tool^[2] on the corpus. We call the character embedding obtained in such a way *TCM-char* in the subsequent part of the paper for exposition brevity. Besides, we also investigate the character embeddings trained in [21] with the BERT model (we call such character embeddings *Bert-base* from now on), and additionally the character embeddings formed with con-

^[2]<https://code.google.com/archive/p/word2vec/>

catenation of *TCM-char* and *Bert-base*, and call such embeddings *TCM-char-Base*.

Sentence Encoding Layer

The task of the sentence encoding layer is to capture the contextual representations of the features in a sentence and further aggregate the representations of the features into the presentation of the sentence for the subsequent processing. The features could be single characters or n-grams of characters depending on the methods employed to capture the contextual representations. We call all these diverse features Basic Feature Units (BFUs) for brevity. Formally, suppose $S=(c_1, c_2, \dots, c_N)$ is the sentence to be encoded, and it consists of N Chinese characters $c_1, c_2, \dots$, and c_N , where c_i is the embedding of its i -th character initialized with *TCM-char*, *Bert-base* or *TCM-char-Base*. The sentence encoding layer learns the representation Z (a real-valued vector) of S through two steps, namely BFU contextual encoding step and aggregation step. We detail them below.

BFU Contextual Encoding We examine three alternative approaches for BFU contextual encoding, namely CNN, Bi-LSTM and BERT, and accordingly resulting in different particular BFUs, i.e., character n-grams for CNN, characters for Bi-LSTM and BERT, $[CLS]$ for BERT. We elaborate on learning contextual encodings for these BFUs with their corresponding approaches in more details below.

- **CNN-based BFU contextual encoding**
The capability of CNN in capturing local linguistic dependencies motivates us to choose it for our character n-gram BFU contextual encoding. Particularly, let g be a convolution kernel with a width a , and $c_{i:i+j}$ be the successive concatenation of the character embeddings $c_i, c_{i+1}, \dots$, and c_{i+j} in S , the BFU contextual embedding generated at the position i in S , denote with BFU_i^g , is calculated with

$$BFU_i^g = \tanh(g \cdot e_{i:i+a-1} + b), \quad (1)$$

where b is the bias. With the padding method named *SAME*, for a convolution kernel g and a sentence of length N , we can obtain a sequence of N distinct BFUs encoded contextually in real-valued vectors, i.e.

$$BFU^g = (BFU_1^g, BFU_2^g, \dots, BFU_N^g). \quad (2)$$

They are aggregated into the sentence representation in the subsequent, *aggregation* step.

- **Bi-LSTM-based BFU contextual encoding**
Bi-LSTM is characterized with its long-short memory, which could benefit learning long-distance dependencies in textual data. We thereby examine it in our framework for BFU contextual encoding. For a position i in S , the Bi-LSTM generates a BFU contextual encoding being the concatenation of the forward and the backward hidden outputs, i.e.

$$BFU_i = [\overrightarrow{LSTM}(c_i); \overleftarrow{LSTM}(c_i)], i \in [1, N]. \quad (3)$$

- **BERT-based BFU contextual encoding**
In recent years we can always find BERT[22] in state-of-the-art algorithms for NLP tasks. Indeed, due to its multi-layer bidirectional transformer and the self-attention mechanism[23] employed, BERT has been applied successfully in solutions of almost all the NLP tasks. We therefore incorporate BERT into our framework and investigate its performance for BFU contextual embedding in the task of DSCEs extraction from ancient TCM literature. The BERT for BFU contextual encoding in our framework is depicted in Figure 4, where $[CLS]$ is added as a special character to denote the entire sentence, and $c_{[CLS]}$ is its initial embedding. Through its multi-layer bidirectional transformers and the self-attention, the BERT incorporates contextual information inside the sentence and generates a contextually embedded vector for $[CLS]$ as well as for every character in the sentence. We thereby obtain two groups of BFUs from the BERT for a sentence. The first group consists of a single contextual embedding for $[CLS]$, denoted with $BFU^{(2)}$ in Figure 4, while the second group contains all the character contextual embeddings in the sentence, i.e., $BFU_1^{(1)}$, $BFU_2^{(1)}$, $\dots$, and $BFU_N^{(1)}$ in Figure 4.

Aggregation After capturing the BFU contextual embeddings for the sentence S with any alternative approach for BFU contextual embedding, the aggregation step composes them into a representation of the entire sentence S necessary for the subsequent processing and the final sentence labelling. We examine two diverse ways of the aggregation, namely max-pooling[24] and attention[25].

For the sentence S with its BFU contextual embeddings $BFU_1, BFU_2, \dots$, and BFU_N , their max-pooling aggregation results in the representation Z of S that $Z = \max\{BFU_1, BFU_2, \dots, BFU_N\}$. If the

Figure 4 The BERT for BFU contextual encoding

BFU context embeddings are obtained with the $[CLS]$ of the BERT (i.e., the first group of the BFU contextual embeddings), $Z = BFU^{(2)}$. For the CNN-based BFU contextual encoding, we employ several convolution kernels in order to capture different semantical aspects of a n-gram, thus obtaining multiple max-pooling aggregations for a sentence. If we employ k convolution kernels, and for the j -th convolution kernel g_j the BFU contextual embeddings for the sentence S are $BFU_1^{g_j}, BFU_2^{g_j}, \dots$ and $BFU_N^{g_j}$, then the sentence representation for g_j is $z_j = \max \{BFU_1^{g_j}, BFU_2^{g_j}, \dots, BFU_N^{g_j}\}$, and the sentence representation for all the convolution kernels are $Z = [z_1; z_2; \dots; z_k]$, where the punctuation mark semi-colon denotes vector concatenation.

With the attention mechanism, BFU contextual embeddings for a sentence are aggregated according to their respective contributions for the sentence, and the contribution strengths are learned automatically. The process of the attention-based aggregation is shown in Figure 5, which demonstrates that, for the sentence S with its BFU contextual embeddings $BFU_1, BFU_2, \dots$, and BFU_N , their attention weights are learned with the query vector u_c at first, after that $BFU_1, BFU_2, \dots$, and BFU_N are weighted-summed to generate the representation Z of S . The attention weights are calculated through the formulas (4) and (5), and the weighted-summed operation in the formula (6), where W_c is the weight matrix, and b_c is the bias. If the BFU context embeddings are obtained with the single $[CLS]$ of the BERT, the attention aggregation leads to $Z = BFU^{(2)}$. For the CNN-based BFU contextual encoding, the attention aggregation is performed for every convolution kernel to obtain the sentence representation towards the convolution kernel, and all the representations resulted towards k different convolution kernels employed are concatenated to form the final sentence representation Z .

Figure 5 The process of the attention-based aggregation

$$u_i = \tanh(W_c BFU_i + b_c), \quad (4)$$

$$\alpha_i = \frac{\exp(u_i^T u_c)}{\sum_i \exp(u_i^T u_c)}, \quad (5)$$

$$Z = \sum_i \alpha_i BFU_i. \quad (6)$$

Except the max-pooling and attention, we also examine a simple aggregation method for the Bi-LSTM-based BFU contextual encoding, where the last BFU encoding is employed to represent the sentence, i.e., $Z = BFU_N$.

Sentence Contextual Encoding Layer

In order to further capture contextual information in a section, the sentence contextual encoding layer takes the representations of its sentences generated in the

sentence encoding layer, and updates their representations with a Bi-LSTM operating on the sentence sequence of the section. Suppose the section D to be sequentially labelled contain successive sentences $S_1, S_2, \dots$, and S_L , their sentence representations are $Z_1, Z_2, \dots$, and Z_L respectively. The contextual encoding h_i for the sentence S_i in D generated with the Bi-LSTM is

$$h_i = [\overrightarrow{LSTM}(Z_i); \overleftarrow{LSTM}(Z_i)]. \quad (7)$$

The resulted h_i is then fed into a MLP to transform h_i into a probability distribution r_i over the elements of the BIO label set $K = \{ 'B', 'I', 'O' \}$, and such label distributions for all the sentences in a section are passed through the CRF layer in order to determine an optimal label sequence for the section. The CRF layer is presented in the next subsection.

CRF Layer

For a section D containing successive sentences $S_1, S_2, \dots$, and S_L , with their corresponding label distributions $r_1, r_2, \dots$, and r_L , the CRF layer aims to determine an optimal label sequence $y^* = (y_1^*, y_2^*, \dots, y_L^*)$ for D through scoring all its possible label sequences. For a label sequence $y = (y_1, y_2, \dots, y_L)$, the CRF scores it with

$$score(y) = \sum_{i=1}^L r_i[y_i] + \sum_{i=1}^{L-1} A_{y_i y_{i+1}}, \quad (8)$$

where $r_i[y_i]$ denotes the probability of S_i labelled with y_i , and $A_{y_i y_{i+1}}$ is the probability of y_{i+1} following y_i learned in the training phase. The scores for all possible label sequences for D are then transformed into a probability distribution over these label sequences as in the formula (9),

$$P(y) = \frac{e^{score(y)}}{\sum_{\hat{y} \in Y} e^{score(\hat{y})}}, \quad (9)$$

where Y are the set of all possible label sequences for D . We use the traditional Viterbi algorithm[26] to choose an optimal label sequence for D , i.e., the label sequence with the maximal probability of label sequence, as expressed in the formula (10).

$$y^* = \underset{\hat{y} \in Y}{\operatorname{argmax}} score(\hat{y}), \quad (10)$$

All the model parameters in the layers of embedding, sentence encoding, sentence contextual encoding and CRF are trained jointly using the negative

log-likelihood loss with the L_2 norm regularization, as shown in the formula (11).

$$\mathcal{L} = \log \left(\sum_{\hat{y} \in Y} e^{score(\hat{y})} \right) - score(y) + \lambda \|\omega\|_2, \quad (11)$$

where ω denotes all the model parameters, and λ is a hyperparameter trading off the norm regularization term against the loss.

Experiments

In order to evaluate the performances of our framework with the alternative approaches applied in its different layers and modules for the DSCEs extraction task, we have collected two corpora from ancient literature of TCM and annotated them manually with DSCEs occurrence information for the diseases *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) and *jaundice* (黄疸) respectively. In this section, we introduce these two corpora and their annotations, report the experimental results of our framework on the corpora, and present the implications which could be drawn from the empirical results.

Datasets and evaluation metrics

The first corpus^[3], named PAC in the following exposition, is oriented towards extracting DSCEs of *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) from ancient TCM literature. When constructing PAC, the two co-authors of this paper from TCM domain had chosen a list of query keywords according their own experiences, including “pregnant (妊)”, “pregnant (娠)”, “pregnant (孕)” and “fetus (胎)”, and retrieved such ancient TCM books from *Chinese Medical Classics* (中华医典)^[4] that match at least one character from the list. With the retrieval process, they have determined 114 books mentioning *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) in *Chinese Medical Classics* (中华医典). After that, all the sections in these books that contain one or more keywords from the list are given to two TCM graduate students which are asked to annotate DSCEs of *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) occurring in these sections. All the sections containing at least once an DSCE occurrence of *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) are composed into the corpus PAC, ignoring the other sections. After that the corpus PAC annotated with such manual way in

^[3]The two terms, namely corpus and dataset, are used interchangeably in this paper for ease of its exposition.

^[4]*Chinese Medical Classics* (中华医典) is a large-scale electronic collection of ancient TCM book with its special content organization and search functionality.

transformed into the BIO tagging form, such as the example in Figure 6. Another corpus, named JAUNDICE in the following exposition, is constructed and annotated with the similar way, except that it is oriented towards extracting DSCEs of *jaundice* (黄疸) rather than *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血), and accordingly the query keywords for this disease are “jaundice (黄疸)”, “jaundice (疸)” and “jaundice (瘰)”. The two co-authors finally have determined 70 books mentioning *jaundice* (黄疸). Table 1 summarizes statistics of the two annotated datasets.

Table 1 Statistics of the two datasets

Dataset	#DSCEs	#Section	#Sentence	Tags		
				#B	#I	#O
PAC	2521	1189	25586	2521	11395	11670
JAUNDICE	296	177	3267	296	816	2155

We evaluate our framework on the annotated datasets towards two granularities of labelling, namely sentence-level labelling and DSCE-level labelling. For the sentence-level labelling we employ the traditional metrics for the evaluation, including precision, recall, F_1 -value and accuracy. Specifically, for a tag i from $\{B', I', O'\}$, the precision P_i , recall R_i , and F_1 -value F_{1_i} are defined respectively in the formulas (12), (13) and (14).

$$P_i = \frac{\text{\#sentences with actual tag } i \text{ labelled correctly by our framework}}{\text{\#sentences labelled by our framework with tag } i}, \quad (12)$$

$$R_i = \frac{\text{\#sentences with actual tag } i \text{ labelled correctly by our framework}}{\text{\#sentences with actual tag } i}, \quad (13)$$

$$F_{1_i} = \frac{2 \times P_i \times R_i}{P_i + R_i}. \quad (14)$$

We further aggregate the precision, recall and F_1 -value for a single tag into their corresponding metrics for all the tags from $\{B', I', O'\}$ with the macro-average method. Hereafter all the mentions of precision, recall and F_1 -value mean their corresponding macro-averaged aggregation over all the tags in $\{B', I', O'\}$.

Except precision, recall and F_1 -value, we also evaluate the framework towards the sentence-level labelling with its accuracy Acc , which calculated with the formula (15).

$$Acc = \frac{\text{\#sentences labelled correctly by our framework}}{\text{\#sentences}}. \quad (15)$$

For the DSCE-level labelling we employ precision, recall and F_1 -value to assess our framework. They are defined respectively in the formulas (16), (17) and (18).

$$P_{DSCE} = \frac{\text{\#actual DSCEs identified by our framework as DSCEs}}{\text{\#spans identified by our framework as DSCEs}}, \quad (16)$$

$$R_{DSCE} = \frac{\text{\#actual DSCEs identified by our framework as DSCEs}}{\text{\#actual DSCEs}}, \quad (17)$$

$$F_{1_{DSCE}} = \frac{2 \times P_{DSCE} \times R_{DSCE}}{P_{DSCE} + R_{DSCE}}. \quad (18)$$

Experimental Setup

Considering different characteristics of the two annotated datasets, in particular their difference in size, we treat them separately in our experiments. Concretely, the dataset PAC is divided randomly into training, validation and testing sets according in the ratio of 8:1:1, while the dataset JAUNDICE is split only into two sets, namely training set and testing set in the ratio of 9:1 due to its relatively small size. All the hyperparameters of our framework are tuned on the validation set from PAC and the resulted hyperparameter settings are applied to the experiments on JAUNDICE. The parameter settings are given in Table 2.

We implement our framework in PyTorch, taking advantage of the Adam[27] optimizer in the training phase to learn its parameters and incorporating the dropout operation into its sentence encoding and sentence contextual encoding layers to avoid its overfitting. We evaluate our framework based on the 10-fold cross-validation schema, and report not only the mean of a metric over all the 10 folds but also its corresponding 95%-confidence interval.

Table 2 Hyper-parameter settings

Hyper-parameter Name	Value
dimensionality of <i>TCM-Char</i>	64
dimensionality of <i>Bert-base</i>	768
widths of convolution kernels	3, 4 and 5
#convolution kernels for a particular width	300
dimensionality of the Bi-LSTM-based BFU encoding	300
dimensionality of the sentence contextual encoding	300
dropout probability	0.5
coefficient of the L2 regularization term λ	0.0001
learning rate for the CNN-based and Bi-LSTM-based BFU encodings	0.0001
learning rate for the BERT-based BFU encoding	5e-5

(《大》) 妇人怀胎，有七月八月而产者，有至九月十月而产者，有经一年二年乃至四年而后产者，各根据后法治。/O 知母丸 治十月未足，而痛如欲产，兼治产难及子烦。/B 用知母一味为细末，炼蜜丸如鸡豆大。/I 温酒嚼下，日三服。/I 丸如桐子大，饮下二十丸。/I 治妊娠月数未足，而似欲产腹痛者。/B 槐子 蒲黄 (等分) 上为末，蜜丸如桐子大。/I 温酒下二十丸。/I 以痛止为度。/I 又方 蒲黄如枣核大，筛过，以井花水调服。/B (《千》) 日月未足而欲产者，捣菖蒲根汁一二升灌喉中。/B (本草云：菖蒲治胎动不安下血。) /I (罗) 治妇过月不产方。/O 四物汤 香附 桃仁 枳壳 缩砂 紫苏 水煎服，即生。/O

Figure 6 An example of the BIO-tagging for a section from ancient TCM literature

Experimental results on the dataset PAC

We first report the experimental results of our framework on the dataset PAC with the alternative techniques described in Subsection *Sequence labelling framework* of Section *Task definition and method*. In order to investigate the effects of these alternative techniques, we conduct a series of experiments under different combinations of the alternatives for the three layers, namely the character embedding, the BFU contextual encoding and the aggregation. For the character embedding we have three alternatives, i.e., *TCM-char*, *Bert-base* and their concatenation *TCM-char-Base*, while for the BFU contextual encoding we have CNN-based character n-gram BFUs, Bi-LSTM-based character BFUs, BERT-based character BFUs and BERT-based single *[CLS]* BFU. For the aggregation we have the max-pooling, the attention, and the simple aggregation choosing the last BFU from the Bi-LSTM-based character BFUs as the aggregated sentence representation. For ease of the demonstration, we call the last aggregation method *Simple*. Furthermore, we also examine a hybrid approach for the two groups of the Bi-LSTM-based BFU contextual encodings (i.e., the single *[CLS]* BFU and the character BFUs) and their respective aggregations (i.e., max-pooling and attention). Specifically, the BFU contextual encoding module generates all the two groups of BFUs, namely $BFU^{(2)}$ (the encoding of the *[CLS]*) and $\{BFU_1^{(1)}, BFU_2^{(1)}, \dots, BFU_N^{(1)}\}$, and then the aggregation module, performing the max-pooling aggregation on $BFU^{(2)}$ and attention aggregation on $\{BFU_1^{(1)}, BFU_2^{(1)}, \dots, BFU_N^{(1)}\}$ at first, concatenates the two aggregation results as the aggregated sentence representation. We call the hybrid approach for the BFU contextual encoding as well as for the aggregation *Hybrid*. The experimental results are given in Table 3.

We can see in Table 3 that, among all the combinations of alternatives in the character embedding,

BFU contextual encoding and aggregation, the combination of *Bert-base* character embedding, CNN-based character n-gram BFUs and max-pooling aggregation are best in all the performance measure metrics, e.g., the accuracy of $80.40\% \pm 1.64\%$ and the F_1 -value of $76.73\% \pm 1.59\%$. Furthermore, the results also demonstrate that, for a fixed BFU contextual encoding approach and a fixed aggregation approach, the *Bert-base* consistently outperforms the other two character embedding alternatives, e.g., the average accuracy of the *Bert-base* for the Bi-LSTM-based character BFUs and the attention aggregation is 74.80%, about 4.5% higher than the *TCM-char* and 3.5% higher than the *TCM-char-Base* for the same BFU contextual encoding approach and the same aggregation approach.

With regard to the BFU contextual encoding, the CNN-based approach performs almost equally well as the BERT-based approach. Both of them outperform the Bi-LSTM-based BFU contextual encoding, regardless of the approach employed for the subsequent aggregation. The superiority of CNN over Bi-LSTM in the BFU contextual encoding explains a possibility that, although there is a very vague bordering between word and character in ancient Chinese language and thus it is difficult to perform word segmentation in its text, the character n-grams captured by CNN could operate as proxies of words. BERT also has similar ability to capture character n-grams for the BFU contextual encoding with its transformer mechanism and character position embedding. For the alternative aggregation techniques, the attention is useful only for the Bi-LSTM-based BFU contextual encoding, regardless of the character embedding method employed.

Due to its better performance, we employ the *Bert-base* as the only character embedding approach in the subsequent experiments. We first report experimental results of our framework on the dataset PAC for recognition of the whole DSCEs, which are given in Table 4.

Table 3 The results of sentence-level labelling on the dataset PAC

Character embedding	BFU contextual encoding	Aggregation	P(%)	R(%)	F_1 (%)	Acc(%)
TCM-char	CNN-based character n-gram BFUs	Max-pooling	73.93±1.86	70.73±1.38	71.23±1.56	75.00±1.82
		Attention	74.60±0.68	73.17±1.05	73.50±0.69	76.80±0.95
	Bi-LSTM-based character BFUs	Simple	67.70±1.17	69.33±1.66	63.50±2.51	67.20±2.64
		Attention	70.23±2.24	71.07±1.66	67.20±3.02	70.30±3.36
Bert-base	CNN-based character n-gram BFUs	Max-pooling	78.60 ± 1.43	75.67 ± 1.79	76.73 ± 1.59	80.40 ± 1.64
		Attention	74.80±1.62	71.97±1.65	73.07±1.65	76.60±1.95
	Bi-LSTM-based character BFUs	Simple	73.17±1.51	70.20±1.82	71.37±1.64	75.40±1.80
		Attention	72.47±1.59	69.77±1.83	70.73±1.78	74.80±1.88
	BERT-based single [CLS] BFU	Max-pooling	77.28 ± 1.51	75.34 ± 1.41	75.92 ± 1.51	78.87 ± 1.96
	Hybrid	Hybrid	76.85±1.90	75.59±1.63	72.65±1.65	73.81±1.68
TCM-char-Base	CNN-based character n-gram BFUs	Max-pooling	75.23±1.91	73.80±2.23	73.67±2.32	76.50±2.62
		Attention	75.53±1.43	73.50±1.23	74.00±1.25	77.30±1.33
	Bi-LSTM-based character BFUs	Simple	69.20±1.77	71.53±1.84	66.77±2.74	69.80±3.05
		Attention	70.67±1.75	72.43±1.61	68.60±2.40	71.30±2.62

Table 4 The results of recognizing the whole DSCEs on the dataset PAC (*strict*)

Character embedding	BFU contextual encoding	Aggregation	P_{DSCE} (%)	R_{DSCE} (%)	F_{1DSCE} (%)
Bert-base	CNN-based character n-gram BFUs	Max-pooling	51.13 ± 2.64	42.80±2.03	46.49±1.91
		Attention	44.85±2.75	37.90±2.15	40.99±2.15
	Bi-LSTM-based character BFUs	Simple	44.17±2.65	36.66±2.45	39.90±1.95
		Attention	38.74±2.42	33.09±2.52	35.62±2.35
	BERT-based single [CLS] BFU	Max-pooling	49.90 ± 1.74	44.97 ± 2.16	47.22 ± 1.52
	Hybrid	Hybrid	50.78±2.20	42.87±2.54	46.41±2.16

The experimental results in Table 4 demonstrate that, as for the sentence-level labelling, for recognizing the whole DSCEs both of the CNN-based and BERT-based BFU contextual encodings perform almost equally well, outperforming the Bi-LSTM-based BFU contextual encoding. Comparing Table 4 with Table 3, we can see much degraded performances in recognizing the whole DSCEs, which exemplify special difficulties in our novel information extraction task. Different from entities to be extracted in other information extractions tasks, the DSCEs to be extracted for a particular disease often have a larger span in an ancient TCM book. Furthermore, in such a book DSCEs always occur sparsely. All these phenomena make our framework have to face three challenges unsolved in machine learning, namely skew data distribution, data sparsity and long-distance dependency in sequence labelling. Therefore, our information extraction task deserves further and continuous studies.

Of course, the results in Table 4 are discouraging. However, if we weaken our goal from exact recognition of the DSCEs to their approximate appearing positions in an ancient TCM book, the results become encouraging and such positions are also useful for TCM professionals to acquire clinical experiences from ancient TCM literature. Table 5 shows the results of recognizing the whole DSCEs in the weakened situation, where a recognized DSCE is regarded as being precise if its span is overlapped with a manually annotated

DSCE span and a manually annotated DSCE is regarded as being recalled if its span is overlapped with a recognized DSCE span, and the P_{DSCE} , R_{DSCE} and F_{1DSCE} are calculated according to the formulas (16), (17) and (18) respectively under the weakened definitions of a precise test instance and a recalled test instance. We call such assessment for recognizing the whole DSCEs *loose*, while the original exact assessment *strict*.

The experimental data in Table 5 indicates that our framework performs very well in determining mention positions of the DSCEs in ancient TCM literature, although it is faced with the difficulty in identifying their exact spans (i.e., their beginning and ending positions). The data also supports the conclusion above that both of the CNN-based and BERT-based BFU contextual encodings perform almost equally well, outperforming the Bi-LSTM-based BFU contextual encoding.

Experimental results on the dataset JAUNDICE

We further perform the experiments of our framework on the dataset JAUNDICE in order to investigate its generalization ability for DSCEs of a particular disease with fewer annotated training data and adaptability of its hyper-parameters from one disease to another disease. As introduced in subsection *Datasets and evaluation metrics*, the dataset JAUNDICE is much smaller than PAC, and composes of sections with annotated

Table 5 The results of recognizing the whole DSCEs on the dataset PAC (*loose*)

Character embedding	BFU contextual encoding	Aggregation	$P_{DSCE}(\%)$	$R_{DSCE}(\%)$	$F1_{DSCE}(\%)$
Bert-base	CNN-based character n-gram BFUs	Max-pooling	88.73 ± 1.34	89.79 ± 1.46	89.21 ± 0.76
		Attention	86.37 ± 1.44	86.93 ± 1.62	86.60 ± 0.80
	Bi-LSTM-based character BFUs	Simple	84.37 ± 2.40	82.44 ± 3.11	83.23 ± 1.70
		Attention	83.19 ± 2.47	85.22 ± 2.25	84.11 ± 1.78
	BBERT-based single [CLS] BFU	Max-pooling	88.71 ± 1.25	88.89 ± 1.54	88.76 ± 0.71
	Hybrid	Hybrid	88.97 ± 1.33	84.15 ± 2.99	86.42 ± 1.86

DSCEs for *jaundice* (黄疸), other than *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血), the annotated disease of PAC. We split the dataset JAUNDICE into training set and testing set in the ratio of 9:1, and train our framework with the parameter settings in Table 2, which are tuned on the dataset PAC. Similarly, we evaluate our framework on JAUNDICE based on the 10-fold cross-validation schema, and report the mean of a metric over all the 10 folds and its corresponding 95%-confidence interval. The experimental results of the sentence-level labelling are given in Table 6.

It is obvious from Table 6 that, due to the small size of the dataset JAUNDICE and the hyper-parameters tuned outside of the dataset, the performance of our framework in sentence-level labelling on JAUNDICE is much poorer than on PAC, but the same tendency is hold, i.e., both of the CNN-based and BERT-based BFU contextual encodings perform almost equally well, outperforming the Bi-LSTM-based BFU contextual encoding. The corresponding experimental results for recognizing the whole DSCEs on JAUNDICE are shown in Table 7, where we can see the similar phenomena that our framework performs well in determining mention positions of the DSCEs in ancient TCM literature, but is difficult to identify their exact beginning and ending positions.

Discussion

In this paper we define a novel information extraction task for recognizing DSCEs in ancient TCM literature, and propose a framework with alternative techniques in its modules to solve the task. The experimental results shown in Section *Experiments* demonstrate that the framework is effective in helping TCM professionals finding disease-specific clinical experiences in ancient TCM literature. Furthermore, we examine different combinations of the alternative techniques in our framework, and find that the *Bert-base* character embedding is appropriate than *TCM-char* for our task, and the CNN-based and BERT-based approaches are almost equally powerful in capturing the features than the Bi-LSTM-based approach.

Although the results are encouraging, the specificities of our task with regard to the long span of the DSCEs and their sparse occurrences in ancient TCM literature have made the task much difficult than other information extraction task, which deserves further deep research.

Conclusion

We present a novel extraction task where the entities to be extracted are DSCEs in TCM ancient literature in order to assist TCM professionals in obtaining valuable disease-specific clinical experiences in massive volume of ancient TCM literature. For the purpose we have collected two corpora from ancient literature of TCM and annotated them manually with DSCEs occurrence information for the diseases *pregnant abdominalgia and colporrhagia* (妊娠腹痛及下血) and *jaundice* (黄疸) respectively. Furthermore, we propose a framework for solve the task, examining different alternative techniques in the framework to overcome the challenges specific to ancient TCM text processing, such as the character embeddings without segmenting the words and the attention mechanism on character n-gram encodings. The experiments performed verify the usefulness of our framework, but also demonstrating the difficulties specific to the task, especially in determining the exact span of a DSCE and solving the issue of the occurrence sparsity, which are our subsequent studying direction for this new information extraction task.

Abbreviations

DSCEs: Disease-Specific Clinical Experiences; TCM: Traditional Chinese medicine; CRF: Conditional Random Fields; CNN: Convolutional Neural Network; Bi-LSTM: Bi-directional Long Short-Term Memory; BERT: Bidirectional Encoder Representation from Transformers; BFUs: Basic Feature Units; PAC: Pregnant Abdominalgia and Colporrhagia

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials

The datasets used and analyzed during the current study are available from <https://github.com/153lym/TCM.DSCEs>.

Table 6 The experimental results of sentence-level labelling on the dataset JAUNDICE

Character embedding	BFU contextual encoding	Aggregation	P(%)	R(%)	F_1 (%)	Acc(%)
Bert-base	CNN-based character n-gram BFUs	Max-pooling	69.40 ± 4.64	62.17 ± 4.41	63.03 ± 5.35	72.70 ± 4.40
		Attention	65.90 ± 4.57	64.20 ± 3.51	64.03 ± 3.94	72.40 ± 3.94
	Bi-LSTM-based character BFUs	Simple	62.13 ± 3.99	53.37 ± 3.59	54.33 ± 4.22	67.60 ± 4.27
		Attention	61.20 ± 3.56	53.57 ± 3.49	54.33 ± 3.85	66.70 ± 4.05
	BERT-based single [CLS] BFU	Max-pooling	68.68 ± 3.46	65.68 ± 4.77	66.13 ± 4.48	73.35 ± 3.81
	Hybrid	Hybrid	68.06 ± 3.69	59.95 ± 7.19	60.16 ± 8.37	70.83 ± 5.11

Table 7 The results of recognizing the whole DSCEs on the dataset JAUNDICE

Character embedding	BFU contextual encoding	Aggregation	strict			loose		
			P_{DSCE} (%)	R_{DSCE} (%)	F_1_{DSCE} (%)	P_{DSCE} (%)	R_{DSCE} (%)	F_1_{DSCE} (%)
Bert-base	CNN-based character n-gram BFUs	Max-pooling	26.78 ± 8.44	23.13 ± 6.80	24.79 ± 7.54	86.88 ± 4.80	76.35 ± 7.42	80.72 ± 5.05
		Attention	22.86 ± 6.08	23.01 ± 4.41	22.74 ± 5.18	80.86 ± 7.25	85.26 ± 4.34	82.21 ± 3.86
	Bi-LSTM-based character BFUs	Simple	22.55 ± 5.62	16.33 ± 5.26	18.66 ± 5.16	79.21 ± 4.08	61.22 ± 7.81	68.30 ± 5.19
		Attention	20.43 ± 5.24	13.86 ± 3.66	16.33 ± 4.16	82.61 ± 4.61	63.36 ± 8.65	70.28 ± 5.31
	BERT-based single [CLS] BFU	Max-pooling	37.88 ± 7.13	35.23 ± 7.23	36.24 ± 7.03	82.71 ± 4.52	75.24 ± 7.71	78.16 ± 5.27
	Hybrid	Hybrid	46.55 ± 6.43	30.96 ± 6.81	34.96 ± 5.94	86.68 ± 4.82	61.51 ± 13.98	68.47 ± 11.56

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was obtained for this study.

Author's contributions

Conceptualization and Methodology, CL and YH; Review and Editing, LM, CL and YH; Dataset tagging, ZT, YY and WH; Experiment, LM, BM and WH; all authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Author details

¹Department of Computer Science, Sichuan University, Chengdu, China.

²College of Medical Information Engineering, Chengdu University of TCM, Chengdu, China. ³Basic Medical college, Chengdu University of TCM, Chengdu, China. ⁴Faculty of Engineering and Mathematical Sciences, University of Western Australia, Australia.

References

1. The Nobel Prize in Physiology or Medicine 2015. <https://www.nobelprize.org/prizes/medicine/2015/summary/>
2. Liu, Y., Zhou, Y.: Discussion on the standard of word marking of chinese medical literature corpus. Chinese Journal of Information on Traditional Chinese Medicine **9**, 85–87 (2002)
3. Weng, H., He, W., Ou, A., Deng, L., He, C., Li, H., Yan, S.: Ancient medical literature semantic annotation using hidden markov models. In: 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 37–40 (2014)
4. Nie, J., Ren, Y., Jiang, R., Xu, X.: Construction and application of data mining system of bashu ancient tcm book records. Chinese Journal of Library and Information Science for Traditional Chinese Medicine, 13–15 (2015)
5. Yao, L., Zhang, Y., Wei, B., Li, Z., Huang, X.: Traditional chinese medicine clinical records classification using knowledge-powered document embedding. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1926–1928 (2016)
6. Yao, L., Mao, Z.J.C., Zhang, Y., Luo, Y.: Traditional chinese medicine clinical records classification with bert and domain specific corpora. Journal of the American Medical Informatics Association **26**, 1632–1636 (2019)
7. Zhou, Y., Qi, X., Huang, Y., Ju, F.: Research on construction and application of TCM knowledge graph based on ancient chinese texts. In: IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume, pp. 144–147 (2019)
8. Akhundov, A., Trautmann, D., Groh, G.: Sequence labeling: A practical approach. arXiv:1808.03926 [cs] (2018)
9. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning (2001)
10. Lin, F., Xiahou, J., Xu, Z.: TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model. Multimedia Tools and Applications **75**(22), 14203–14232 (2016)
11. Ye, Y., Xu, B., Ma, L., Zhu, J., Shi, H., Cai, X.: Research on treatment and medication rule of insomnia treated by TCM based on data mining. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2503–2508 (2019)
12. Jiang, Q., Li, H., Liang, J.: Free text mining of TCM medical records based on conditional random fields. In: 2016 International Conference on Progress in Informatics and Computing (PIC), pp. 242–247 (2016)
13. Jiang, Q., Li, H., Liang, J., Wang, Q., Luo, X., Liu, H.: Multi-combined features text mining of TCM medical cases with CRF. In: 2016 8th International Conference on Information Technology in Medicine and Education (ITME), pp. 621–626 (2016)
14. Liu, L., Wu, X., Liu, H., Cao, X., Wang, H., Zhou, H., Xie, Q.: A semi-supervised approach for extracting TCM clinical terms based on feature words. BMC Medical Informatics and Decision Making **20**, 118 (2020)
15. Xu, Q., Tang, W., Teng, F., Peng, W., Zhang, Y., Li, W., Wen, C., Guo, J.: Intelligent syndrome differentiation of traditional chinese medicine by ANN: A case study of chronic obstructive pulmonary disease. IEEE Access **7**, 76167–76175 (2019)
16. Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., Sun, Z., Tang, B., Chang, T.-H., Wang, S., Liu, Y.: Real-world data medical knowledge graph: construction and applications. Artificial Intelligence in Medicine **103**, 101817 (2020)
17. Zhang, T., Wang, Y., Wang, X., Yang, Y., Ye, Y.: Constructing fine-grained entity recognition corpora based on clinical records of traditional chinese medicine. BMC Medical Informatics and Decision Making **20** (2020)
18. Fu, X., Yuan, T., Li, X., Wang, Z., Zhou, Y., Ju, F., Li, J., Chen, X., Xiaoming, S.: Research on the method and system of word segmentation and POS tagging for ancient chinese medicine literature. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2493–2498 (2019)
19. Lee, J., Cho, K., Hofmann, T.: Fully character-level neural machine translation without explicit segmentation. Transactions of the Association for Computational Linguistics **5**, 365–378 (2017)
20. Li, X., Meng, Y., Sun, X., Han, Q., Yuan, A., Li, J.: Is word segmentation necessary for deep learning of Chinese representations? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3242–3252 (2019)
21. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A.,

- Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
22. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs] (2019)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
24. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *Journal of Machine Learning Research* **12** (2011)
25. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)
26. Forney, G.D.: The viterbi algorithm. *Proceedings of the IEEE* **61**(3), 268–278 (1973)
27. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. *Computer Science* (2014)

Figures

..... 治胎冷腹胀，痛引两胁，小便频数，大便虚滑。诃子（面裹煨，去核）白术（各二钱）陈皮（去白）高良姜（炒）木香（不见火）白芍药陈米（炒）甘草（炙，各一钱）上作一服，水二盅，生姜五片，煎至一盅，不拘时服，忌食生冷之物。治妊娠怔悸，睡里多惊，腹胁膨胀，坐卧不宁。白茯苓（去皮）麦门冬（去心）..... 生姜三片，竹茹弹子大一块，煎至一盅，食前服。治妊娠恶阻，胃寒吐逆，反胃吐食，及心腹刺痛。人参（四钱）丁香（一钱）藿香（二钱半）上作一服，水二盅，煎至一盅，不拘时服。..... 治妊娠伤寒。阿胶（炒）人参桑寄生白茯苓（去皮）白术（炙，各等分）上为细末，每服二钱，不拘时糯米饮调下。治妊娠宿有冷，胎痿不长，或失于将理，伤胎多坠，此药补荣卫，养胎气。白术川芎（各一两）吴茱萸（汤泡，半两）甘草（炙，一两半）上为细末，每服二钱，食前温酒调下，忌生冷果实之物。治妊娠伤寒，头疼壮热，肢节烦疼。前胡白术（各一钱）黄芩（二钱，炒）梔子仁知母木香（各二钱）上作一服，水二盅，竹茹一块，葱白三寸，煎一盅，不拘时服.....

Figure 1

DSCEs examples in an ancient TCM book

Figure 2

The flowchart of the DSCEs extraction process

Figure 3

The architecture of the framework

Figure 4

The BERT for BFU contextual encoding

Figure 5

The process of the attention-based aggregation

〔《大》〕妇人怀胎，有七月八月而产者，有至九月十月而产者，有经一年二年乃至四年而后产者，各根据后法治。/O 知母丸 治十月未足，而痛如欲产，兼治产难及子烦。/B 用知母一味为细末，炼蜜丸如鸡豆大。/I 温酒嚼下，日三服。/I 丸如桐子大，饮下二十丸。/I 治妊娠月数未足，而似欲产腹痛者。/B 槐子 蒲黄（等分）上为末，蜜丸如桐子大。/I 温酒下二十丸。/I 以痛止为度。/I 又方 蒲黄如枣核大，筛过，以井花水调服。/B 〔《千》〕日月未足而欲产者，捣菖蒲根汁一二升灌喉中。/B （本草云：菖蒲治胎动不安下血。）/I （罗）治妇过月不产方。/O 四物汤 香附 桃仁 枳壳 缩砂 紫苏 水煎服，即生。/O

Figure 6

An example of the BIO-tagging for a section from ancient TCM literature