Research on Named Entity Recognition Method of Chinese Classics Under the Supervision of Domain Knowledge

Wenjuan Zhao[1,a], Zhongbao Liu[2,b], Jian Lian[3,c]

[1] Library, Beijing Language and Culture University, Beijing 100083;

[2] Software College, Quanzhou University of Information Engineering, Quanzhou 362000;

[3] School of Information Science, Beijing Language and Culture University, Beijing 100083.

[a] zhaowenjuan1118@163.com, [b] 269495709@qq.com , [c] lj1422206819@163.com

Abstract. The current dominant named entity recognition methods of Chinese classics are classified as data-driven methods, which are limited by the data quality. The domain knowledge is introduced in this paper to supervise the process of the named entity recognition, so as to solve the poor performance problem because of the low-quality data. The experiments on the Historical Records corpus show that compared with the domain knowledge unsupervised case, the average accuracy, recall rate, and F1 value have respectively improved by 2.76%, 2.70%, and 2.75% under the supervision of domain knowledge. Domain knowledge plays an important role in improving the performance of the named entity recognition methods of Chinese classics.

Keywords: Domain knowledge; Chinese classics; Named entity recognition; BERT model; BiLSTM; CRF.

1. Introduction

For thousands of years, Chinese classics have been handed down from generation to generation, becoming an important carrier of excellent Chinese traditional culture. The deep accumulation of classical literature has become a solid foundation for the cultural confidence of the Chinese nation. In order to make the words written in Chinese classics come alive, researchers have conducted a series of effective explorations, especially in the context of the era of big data and artificial intelligence, using artificial intelligence technologies and methods to conduct deep excavation of Chinese classics, to generating a panoramic knowledge network and providing strong support for the learning and dissemination of Chinese classics. Named entity recognition is a key aspect of generating a panoramic knowledge network of Chinese classics and has received increasing attention from researchers. Named entity recognition of Chinese classic books refers to the automatic recognition of entities with specific semantic meanings such as names of people, places, institutions, time, official titles, etc., by using techniques and methods including rules, statistical models or deep learning models. The related research started relatively late in China. Some achievements have been made in the practice of named entity recognition of classics such as the Complete Library in Four Sections and Huangdi Neijing, but the research breadth and depth need to be further strengthened.

To summarize current researches, there are two main problems: On the one hand, the current dominant methods are mainly based on Long Short Term Memory (LSTM), Conditional Random Field (CRF), and Bidirectional Encoder Representations from Transformers (BERT); On the other hand, the current dominant methods are data-driven methods, whose performance depends on the data quality, and the actual data quality is not always high, which leads to the poor recognition results of these methods. Thus, this paper introduces domain knowledge in addition to the current Bi-LSTM+CRF model combined with the BERT model for named entity recognition and tries to guide the whole process of named entity recognition with domain knowledge in order to avoid the problem of poor recognition due to poor data quality. Through this paper, we aim to improve the performance of Chinese classics named entity recognition, further enrich the research results in this field, and open up new research ideas for researchers.

In this article, Section 2 introduces the related research on named entity recognition methods; Section 3 combines the research framework and proposes the recognition process of Chinese classic named entities guided by domain knowledge; Section 4 compares and analyzes the recognition effects of named entities with and without domain knowledge guidance on the basis of the corpus of the Historical Records; Section 5 concludes the whole article and indicates the next research assumptions.

2. Related Work

Currently, named entity recognition methods can be roughly divided into three categories: rule-based methods, statistical model-based methods, and deep learning-based methods.

Rule-based methods utilize features such as punctuation, keywords, positional words, directional words, etc., and manually construct finite rules for named entity recognition through pattern matching. Collins et al. gave firstly a seed rule set; then trained the seed rule set using unsupervised learning methods, and then obtained a large-scale rule set; and finally, utilized the rule set for named entity recognition on the corpus. The experimental results show that the accuracy of the above method can reach more than 91% [1]. Tan et al. used the context rules of the corpus to build up a rule base, and based on this, they recognized Chinese place names [2]. Wang Ning et al. obtained the structural features of company names and their contextual information through an in-depth analysis of Chinese financial news texts and proposed a recognition method based on two scanning processes on the foundation of establishing a rule base for company name recognition. The accuracy of the above method reaches 62.8% and the recall rate reaches 62.1% in the open test [3]. Zhou Kun first established a Chinese word segmentation model based on named entity recognition; then, in the process of word segmentation, a rule-based method was used to recognize named entities in the corpus; finally, the recognition results were analyzed to generate new rules, and update the rule base. The above method has a certain degree of self-learning ability and can obtain better named entity recognition performance [4]. Wang Hao proposed a named entity recognition method based on hierarchical pattern matching to recognize term abbreviations in academic papers [5]. In general, the rule-based method can obtain a better recognition effect on the specific corpus, but the method relies on manual rules. It is time-consuming and arduous, not flexible enough, has poor portability, and it is difficult to recognize named entities outside the coverage of the rules.

The statistical model-based approach uses statistical models to learn from the manually labeled corpus and recognizes entities in the corpus on the basis of a given base of named entity types. The method transforms the named entity recognition problem into a named entity classification problem. The statistical models commonly used for named entity recognition are Decision Tree (DT), Maximum Entropy (ME), Conditional Random Field (CRF), Support Vector Machine (SVM), Hidden Markov Model (HMM), etc. Georgios et al. introduced C4.5 Decision Tree Model to recognize person names and organization names from the corpus [6]. Ning et al. introduced the maximum entropy model to recognize named entities from the corpus, in order to solve the problem of limited corpus size and sparse data faced by traditional statistical model-based methods. The experimental results show that the recall rate and F1 value of this model are improved by 1.17% and 0.41%, respectively, compared with the existing methods [7]. Sui Mingshuang et al. used to automatically identify chemical substances and disease entities from biomedical texts by constructing a CRF model that fuses multiple features [8]. Fajar et al. introduced an SVM model in an attempt to identify the narrator's name from the Indonesian translation of the Hadith Collection [9]. Indira et al. used an HMM model to identify usernames, organization names, and location information from short Twitter texts. Experimental results on real corpus showed that the model can reach F1 values of more than 64% [10]. Statistical model-based methods need to utilize a larger scale corpus for training, while the size of the corpus that can be used for named entity recognition in real applications is small, the above contradiction leads to the poor performance of this method in carrying out named entity recognition of a larger scale.

The deep learning-based method stems from the fact that deep learning models are widely used in the field of natural language processing. The biggest advantage of the deep learning model is that the model utilizes word vectors to represent the words in the corpus. On the one hand, this approach solves the data sparsity problem brought by the high-dimensional vector space, and on the other hand, the word vectors have a stronger semantic expression ability than the manually selected features. Therefore, this model is gradually introduced to named entity recognition. Inspired by the excellent performance of Long Short Term Memory (LSTM) network in Chinese word segmentation, Peng et al. proposed a named entity recognition method integrating LSTM and CRF models, which improved the F1 value by 5% compared with the traditional method [11]. To address the problem of over-reliance on manually labeled corpus faced by existing named entity methods, Lample et al. utilized a hybrid model of Bi-directional Long Short Term Memory (BiLSTM) and CRF for joint training of fusing a small amount of labeled corpus with a large amount of unlabeled corpus [12]. Bharadwaj et al. achieved a good effect of named entity recognition on languages with complex morphological changes such as Turkish by extracting morpheme features based on LSTM model [13]. Wu et al. used a hybrid model of LSTM and CRF with the introduction of an attention mechanism for prescription named entity recognition in response to the problems of scribbling and difficulty in recognizing Chinese prescriptions [14]. In addition, deep learning models such as Convolutional Neural Network (CNN) and Hybrid Neural Network (HNN) have been widely used in named entity recognition and achieved good recognition results.

In recent years, the research on named entity recognition of Chinese classics has received attention. Tang Yafen, for the study of text mining and analysis, utilized the CRF model to automatically recognize ancient names on the Pre-Qin corpus [15]. Huang Shui-Qing et al. analyzed the effectiveness of CRF model and ME model in ancient place name recognition comparatively by constructing multi-feature templates on the Master Zuo's Spring and Autumn Annals corpus [16]. In terms of Chinese medicine classics named entity recognition, Wang Shikun, Zhang Wubei, and Meng Hongyu et al. recognized terms from Ming and Qing Dynasty ancient medical cases, Classified Case Records of Celebrated Physicians, and Shanghan Lun and other Chinese medicine classics, and their F1 values all reached more than 75% [17-19]. Gao Su et al. used the hybrid model of BiLSTM and CRF to identify the entities such as TCM cognitive methods, TCM physiology, TCM pathology, TCM nature, and treatment rules in the Huangdi Neijing [20]. Yan Chengxi et al. proposed HanNER, an automated extraction framework for named entities of ancient literature resources, and used an optimized BERT-CNN-BiLSTM-CRF model to achieve automatic extraction of online Chinese ancient literature entities [21]. Lin Litao and Wang Dongbo et al. recognized animal-named entities in the Historical Records with 91.6% accuracy using a SikuBERT pre-trained model constructed based on the training of the Complete Library in Four Sections corpus [22]. The above methods have greatly promoted the research on the recognition of named entities in Chinese classics, and made some breakthroughs in the recognition of named entities in ancient literature and intelligent processing of ancient literature, but there is still much improvement space in terms of the recognition effect and application scenarios.

3. Research Framework

Currently, researchers mainly utilize machine learning algorithms, especially deep learning models, to carry out research on Chinese classics named entity recognition. These researches rely on the data quality, while the actual data quality obtained is often unsatisfactory, and the above contradiction leads to the fact that the recognition effect of the existing research cannot be guaranteed. In order to solve the above problems, this paper proposes a named entity recognition method for Chinese classics under the supervision of domain knowledge, which makes full use of the advantages of domain knowledge and big data mining methods, and tries to guide the whole named entity recognition process by domain knowledge, in order to improve the effect of named entity recognition for Chinese classics. The research framework of this paper is shown in Figure 1. This paper takes the corpus "登丸山，及岱宗。" as an example, and gives the specific recognition process.

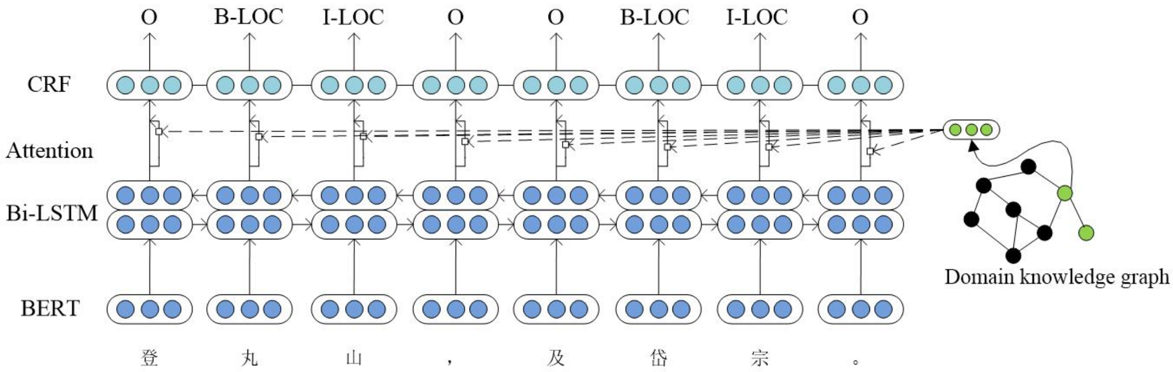


Fig. 1 Research framework

In the first step, the BERT (Bidirectional Encoder Representations from Transformers) model is able to utilize large-scale unlabeled corpus for training in order to obtain a rich-contextual representation of semantic information to adequately characterize the character-level, word-level, sentence-level, and even inter-sentence relational features of the corpus. Therefore, the model is utilized to generate vectors of shallow semantic features of the input corpus. First, Each character in the corpus "登丸山，及岱宗。" is vectorized to obtain the word vector, text vector, and the position vector of each character, and the vector sum of the word vectors, text vectors, and position vectors is input to the BERT model; Secondly, the Transformer structure of the model multiplies the input character vector sum by the Query matrix, Key matrix, and Value matrix to obtain the corresponding query vector q, key vector k, and value vector v; Finally, based on the query vector, key vector, and value vector, the semantic similarity between the input character vectors is computed respectively, and then the vectors of shallow semantic features of the input corpus are obtained.

In the second step, the Bidirectional Long Short-Term Memory (Bi-LSTM) model is utilized to perform deep feature extraction on the vectors of shallow semantic features, and then obtain the vectors of deep semantic features. The model introduces gate structures such as the input gate, forgetting gate, and output gate, where the input gate and forgetting gate control the information that needs to be updated and forgotten by the hidden layer neurons, respectively, and the output gate determines the information that is output. The vector of shallow semantic feature is input into the model, and the deep semantic feature of the vector is extracted from the forward and backward directions respectively. The vector of deep semantic feature is obtained based on the fusion of the semantic features in both directions.

The third step is to utilize the jieba, Chinese word segmentation module, to split the corpus "登丸山，及岱宗。". The result is "登/丸山/，/及/岱宗/。", which consists of several feature words. Find the knowledge related to the feature words in the domain knowledge graph, and represent the knowledge as a triple (head entity, relationship, tail entity), such as (丸山, label, location); Vectorize the triple to obtain the knowledge vector by utilizing the TransE knowledge representation method.

In the fourth step, the vectors of deep semantic features and knowledge vectors are fused by using the attention mechanism to generate the fused feature vectors. The vectors of deep semantic features are from the deep learning model, while the knowledge vectors are derived from the domain knowledge graph, and the fusion of the two can make full use of the advantages of big data mining methods and domain knowledge to extract richer semantic information. The introduction of an attention mechanism can distinguish the important degree of different vectors. For example, if the knowledge vector (丸山, tag, location) has a low correlation with the vector of the deep semantic feature of "岱宗", the attention mechanism gives a larger weight to the vector of the deep semantic feature of "岱宗" and a smaller weight to the knowledge vector (丸山, tag, location). The fused feature vectors are more inclined to the vectors of deep semantic features, which effectively reduces the influence of the less relevant knowledge on the recognition results.

In the fifth step, the fused feature vectors are input into the Conditional Random Field (CRF) model, and the conditional probability is obtained by training with maximum likelihood estimation, which results in the output sequence with the largest conditional probability, i.e., the result of the named entity recognition: "[登, O] [丸, B-LOC] [山, I-LOC] [，, O] [及, O] [岱, B-LOC] [宗, I-LOC] [。, O], wherein the meanings of O, B-LOC, and I-LOC are given in Table 2.

4. Experimental results and analysis

4.1 Data sources

The manually labeled "The Basic Annals of the Five Emperors", "The Basic Annals of Xia", "The Basic Annals of Yin", "The Basic Annals of Zhou ", "The Basic Annals of Qin", "The Basic Annals of The First Emperor of the Qin", "The Basic Annals of Xiang Yu", " The Basic Annals of Emperor Han Gaozu", "The Basic Annals of Empress Dowager Lü", and "The Basic Annals of Emperor Xiao Wen" are used as the experimental corpus. The entity statistics of the experimental corpus are shown in Table 1. The partial knowledge graphs related to the Chinese classics are extracted from the encyclopedic Chinese knowledge graph CN-DBpedia as domain knowledge graphs.

Table 1. Table of statistical information about entities in the experimental corpus

experimental corpus	Number of character entities	Number of entities in location	Number of organization and post entities	Total number of entities
The Basic Annals of the Five Emperors	90	45	2	137
The Basic Annals of Xia	48	118	13	179
The Basic Annals of Yin	87	39	8	134
The Basic Annals of Zhou	241	102	16	359
The Basic Annals of Qin	411	182	34	627
The Basic Annals of The First Emperor of the Qin	278	261	47	586
The Basic Annals of Xiang Yu	283	192	29	504
The Basic Annals of Emperor Han Gaozu	325	252	42	619
The Basic Annals of Empress Dowager Lü	287	125	27	439
The Basic Annals of Emperor Xiao Wen	154	76	26	256
add up the total	2204	1392	244	3840

In this paper, we use the BIO format to annotate the experimental corpus with named entities. B denotes that the word is the beginning of the entity, I denotes the middle to the end of the entity, and O denotes other types of words, i.e., non-entities. Entities can be classified into three categories, person (Person, PER), location (Location, LOC), and organization (Organization, ORG). Given that the experimental corpus involves many post entities, ORG also denotes post entities in this paper. The IOB format annotation table is shown in Table 2.

Table 2. BIO format annotation table

tab	meaning
B-PER	The starting point of the character entity
I-PER	The middle part of the character entity
B-LOC	The starting point of the location entity
I-LOC	The middle part of the location entity
B-ORG	Starting points for organization and post entities
I-ORG	Middle parts for organization and post entities
O	non-entities

The experimental corpus set is divided into training sets and test sets according to the ratio of 8:2. The evaluation metrics used in the experiments are precision P, recall R, and F1 value, defined as follows:

$$P=\frac{TP}{TP+FP}$$

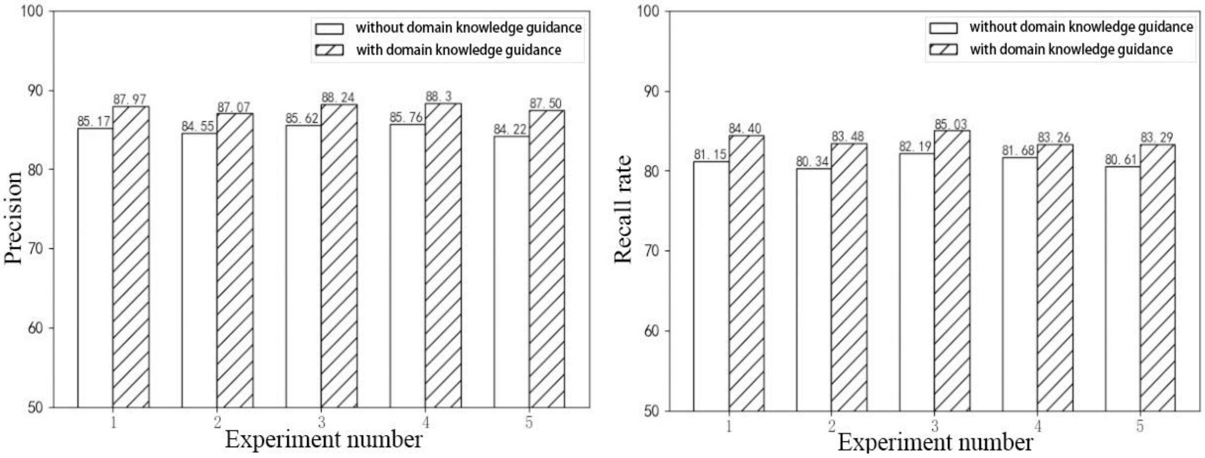
$$R=\frac{TP}{TP+FN}$$

$$F1=\frac{2P\cdot R}{P+R}$$

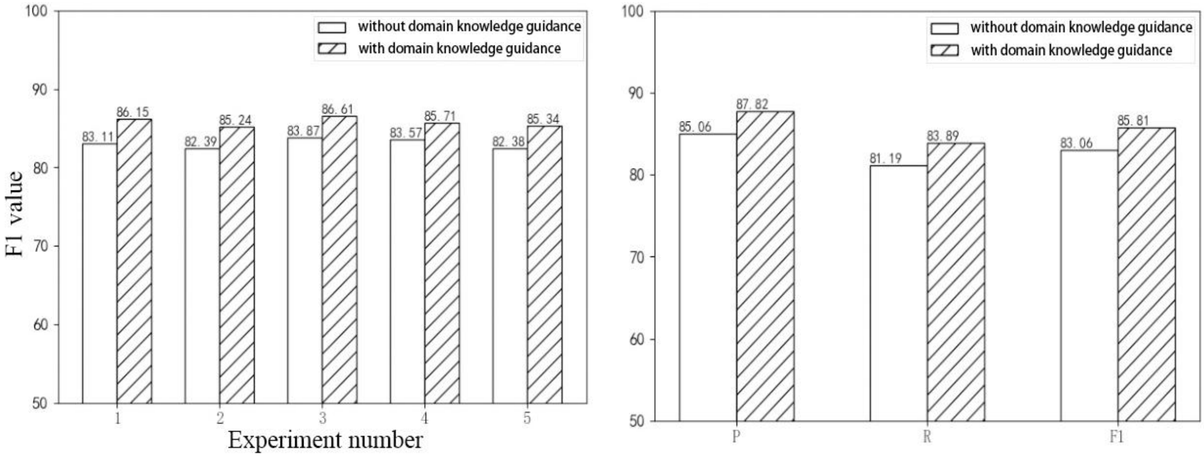
where TP(true positive), FP(false positive), and FN denote the number of correct entities recognized, wrong entities identified, and unidentifiable entities, respectively.

4.2 Experimental results and analysis

The purpose of the experiment is to test the effect of domain knowledge graph on the recognition results of named entities. In the experiment, the hidden layer of BiLSTM contains 200 neurons, the learning rate is selected as 0.005, and the dropout value is 0.5. In order to ensure the validity of the experiment, the experiment is designed by using the five-fold cross-validation method. The five-fold cross-validation method divides the experimental corpus set into five equal parts, four parts are used as the training set, the remaining one is used as the test, and the average of the five experimental results is taken as the final experimental results. The results are shown in Figure 2, where the unit is %.



(a) Comparison of precision with and without domain knowledge guidance (b) Comparison of recall rate with and without domain knowledge guidance



(c) Comparison of F1 value with and without domain knowledge guidance (d) Mean value for results of five experiments

Fig. 2 Comparison of experimental results with and without domain knowledge guidance

From Fig. 2(a)-(c), it can be seen that the named entity recognition precision, recall rate, and F1 value under the guidance of domain knowledge are better than the case without domain knowledge guidance. From Fig. 2(d), it can be seen that the mean values of named entity recognition precision, recall rate, and F1 value under domain knowledge guidance reach 87.82%, 83.89%, and 85.81%, respectively. Compared to the case without domain knowledge guidance, the above metrics are improved by 2.76%, 2.70%, and 2.75%, respectively.

The results of recognizing three types of entities such as Person PER, Location LOC, Organization, and Post ORG are shown in Figure 3.

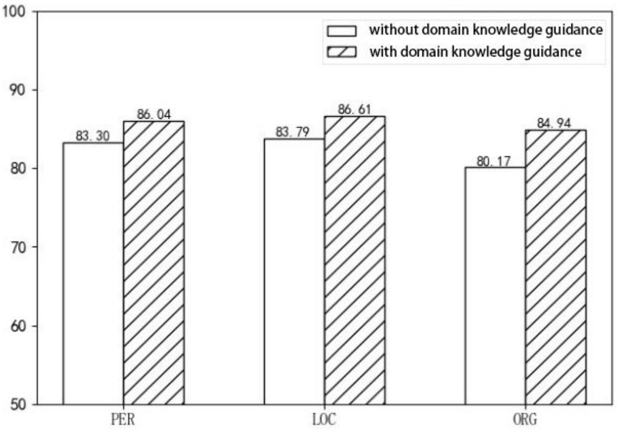


Fig. 3 Results of identification of three types of entities (F1 values)

As can be seen in Figure 3, the F1 values of three types of entity with domain knowledge guidance for PER, LOC, and ORG reach 86.04%, 86.61%, and 84.94%, respectively, while the F1 values of three types of entity without domain knowledge guidance are 83.30%, 83.79%, and 80.17%, respectively. With or without domain knowledge guidance, the ORG entity recognition results are worse than PER and LOC, which is mainly due to the smaller scale of the organization and post entities in the training corpus set, resulting in an insufficient training model and poorer entity recognition results. Compared with the case without knowledge guidance, the three types of entity recognition results of PER, LOC, and ORG under domain knowledge guidance are improved by 2.74%, 2.82%, and 4.77% respectively, among which the ORG entity recognition effect is most obviously improved. It can be seen that the introduction of domain knowledge on a smaller training corpus set can better improve the named entity recognition effect.

5. Experimental results and analysis

The current dominant named entity recognition methods for Chinese classics belong to the data-driven method, and the performance of this method is limited by the quality of the data. Therefore, this paper introduces domain knowledge graph and proposes a named entity recognition method for Chinese classics under the supervision of domain knowledge graph. The method attempts to guide the named entity recognition process from domain knowledge in the whole process, in order to solve the problem of poor recognition effect caused by poor data quality. The experimental results on the Historical Records corpus show that the named entity recognition precision, recall rate, and F1 value under the guidance of domain knowledge are all improved to different degrees compared with the case without domain knowledge guidance. In this paper, the experimental corpus set is labeled manually, which is time-consuming and arduous. The next step is to introduce automated labeling techniques to enlarge the corpus labeling scale, with a view to improving the training quality of the model and obtaining better named entity recognition results.

6. Acknowledgements

This research project is supported by Science Foundation of Beijing Language and Culture University (supported by “ the Fundamental Research Funds for the Central Universities ”) (Approval number: 22YJ200005)

This research project is supported by Fujian Provincial Social Science Planning Project "Research on Government Data Governance from the Perspective of Data Ecology" (Approval number: FJ2021B126).

References

[1] Collins M, Singer Y. Unsupervised models for named entity classification[C]. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Singapore, 1999: 100-110

[2] Tan H Y, Zheng J H, Liu K Y. Research on method of automatic recognition of Chinese place name based on transformation. Journal of Software,2001,12(11):1608-1613

[3] Wang Ning, Ge Ruifang, Yuan Chunfa, et al. Company name identification in Chinese financial domain [J]. Journal of Chinese Information Processing, 2002,16 (2):1-6

[4] Zhou Kun. Research on named entity recognition based on rules [D]. Hefei University of Technology, 2010

[5] Wang Hao. Named entity extraction based on hierarchical pattern matching [J]. New Technology of Library and Information Service, 2007, 5: 62-68

[6] Georgios P, Vangelis K, Georgios P, et al. Learning decision trees for named-entity recognition and classification [C]. Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, Germany, 2000: 1-6

[7] Ning H, Yang H, Tan Y Z. A method of Chinese named entity recognition based on maximum entropy model [C]. Proceedings of the 2009 International Conference on Mechatronics and Automation, Changchun, China, 2009: 95-106

[8] Sui Mingshuang, Cui Lei. Extracting chemical and disease named entities with multiple-feature CRF model [J]. New Technology of Library and Information Service, 2016, 10: 91-97

[9] Fajar A Y, Moch A B, Arief F H. Narrator’s name recognition with support vector machine for indexing Indonesian Hadith translations [J]. Procedia Computer Science, 2019, 157: 191-198

[10] Indira S A, Moch A B, Ibnu A. Named entity recognition on Indonesian tweets using hidden markov model [C]. Proceedings of the 7th International Conference on Information and Communication Technology, Kuala Lumpur, Malaysia, 2019: 1-5

[11] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning [C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 149-155

[12] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C]. Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, USA, 2016: 260-270

[13] Bharadwaj A, Mortensen D, Dyer C, et al. Phonologically aware neural model for named entity recognition in low resource transfer settings [C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Texas, USA, 2016: 1462-1472

[14] Wu G H, Tang G E, Wang Z R, et al. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition [J]. IEEE Access, 2019, 7: 113942-113949

[15] Tang Yafen. Research of a Automatically recognizing name in Pre-Qin ancient Chinese classics [J]. New Technology of Library and Information Service, 2013, 7: 63-68

[16] Huang Shuiqing, Wang Dongbo, He Lin. Research on constructing automatic recognition model for ancient Chinese place names based on Pre-Qin corpus [J]. Library and Information Service, 2015, 59(12): 135-140

[17] Wang Shikun, Li Shaozi, Chen Tongsheng. Recognition of Chinese medicine named entity based on condition random field [J]. Journal of Xiamen University (Nature Science), 2009, 48(3): 359-364

[18] Zhang Wubei, Bai Yu, Wang Peiyan ,et al. An automatic domain terms extraction method on traditional Chinese medicine books [J]. Journal of Shenyang Institute of Aeronautical Engineering, 2011, 28(1): 72-75

[19] Meng Hongyu, Xie Qingyu, Chang Hong, et al. Automatic identification of TCM terminology in Shanghan Lun based on conditional random field [J]. Journal of Beijing University of Traditional Chinese Medicine, 2015, 38(9): 587-590

[20] Gao Su, Jin Pei, Zhang Dezheng. Research on named entity recognition of TCM classics based on deep learning [J]. Technology Intelligence Engineering, 2019, 5(1): 113-123

[21] Yan Chengxi, Tang Xuemei, Yang hao, et al. HanNER: A General Framework for the Automatic Extraction of Named Entities in Ancient Chinese Corpora [J]. Journal of the China Society for Scientific and Technical Information, 2023,42(02):203-216

[22] Lin Litao, Wang Dongbo, Liu Jiangfeng, et al. Animal Named Entity Recognition in Ancient Chinese Classics from the Perspective of Digital Humanities ：Based on SikuBERT Pre-training Model [J]. Library Tribune, 2022,42（10）: 42-50