ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324

Research on Ancient Weapons Named Entity Recognition From Chinese Historical Novels

Zhao Xiyuan[1,a]，Zhang Hao[2,b]

[1]School of Foreign Languages, Shandong University of Technology, China

[2]School of Computer Science And Technology, Shandong University of Technology, China

[a]zhaoxiyuan0402@163.com, [b]1095618608@qq.com

Abstract. As a basic task of natural language processing, named entity recognition has been applied in the field of Chinese literature. A method can be realized the ancient weapons named entity recognition and association analysis in Chinese historical novels,so this paper puts forward the recognition model and use knowledge mapping analysis. Taking ancient weapons in Romance of The Three Kingdoms as a characteristic entity example constructed a corpus,the BERT-BiGRU-CRF model is built to solve the named entity recognition of ancient weapons based on digital humanities.And the knowledge map is used to analyse the relationship of entities.[Result]The training effect of the combined model BERT-BiGRU-CRF is obviously better than the other four models, and finally the harmonic average F1-score of the ancient weapons' entities is improved to 88.44%.This method can solve the difficulty of recognizing the ancient weapons of Chinese historical novels, providing reference for the further entity correlation of Chinese historical literature.

Keywords: historical novels; ancient weapons; BERT-BiGRU-CRF model; text mining.

1. Introduction

China has a long history of weapons manufacturing, evolving from the Bronze Age to the Gunpowder Age, so ancient weapons are renowned worldwide for their sophistication and diversity. In Chinese historical novels, ancient weapons have been the subject of many literary figures to enrich the characters. It is evident that ancient weapons, as a distinctive object of depiction in Chinese historical fiction, symbolising the wisdom of the Chinese people and bear witness to the flourishing of Chinese power. While the study of ancient weapons has long been dominated by qualitative methods that explore their cultural implications. With the development of digital humanities technology, the inadequacy of analysing the distinctive areas has gradually emerged by traditional analytical models in classic Chinese historical novels.

Natural language processing techniques have been used as the primary method of analysis instead of manual textual research in many fields. Named Entity Recognition (NER) is one of the fundamental techniques used at the intersection of disciplines to recognize entities such as names, places, organizations, time and numerical expressions [1]. The application scope of NER has also been expanding, in the fields of economics, biology, medicine, law, military and literature [2, 6]. With the gradual increase in textual diversity, the traditional entity category recognition is short of novelty, leading to the need for distinctive domain mining. As an interdisciplinary research method, digital humanities is a product of the development of humanities domain knowledge, data collection and analysis techniques and algorithmic models [7]. As the digital humanities technology is developing, the deficiency of analyzing method by traditional modes has gradually emerged in the characteristic fields of classic Chinese historical novels. Digitisation of literature aims to use the corpus to further explore the textual logic at a deeper level, where the study of named entity identification is of great significance to the analysis of historical literary works as a fundamental task. This paper attempts to develop a new approach to the NER of ancient weapons in Chinese historical fiction texts, to analyze the correlation between the entities in conjunction with knowledge mapping.

ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324

2. Research Review

So far, named entity recognition methods have been broadly classified into rule-based, artificial feature-based and deep learning methods [8]. In early research of NER, rule-based methods were mainly influenced by the experience of linguists and annotated entities by rules, with high accuracy of recognition results but with limitations on the types. Subsequently, artificial feature-based methods trained better than before by statistical machine learning algorithms based on a large number of manually defined features. With significant advantages in automatically learning features, extracting deep semantic knowledge and alleviating data sparsity [8]. As a result, named entity recognition has also been applied to data mining and text analysis in various fields.

In China, named entity recognition has mainly been applied to modern texts, and the focus on Chinese historical fiction texts has only increased in recent years. In the context of "telling the Chinese story well", it should pay much attention to the external dissemination of classic Chinese historical fiction texts, making the study of literary texts with multiple perspectives and diversity. Relying on digital humanities technology to explore the semantic relationships of texts in the field of Chinese historical fiction has an important impact on the depth of textual analysis [9]. In a word, named entity identification methods have been applied in various classical texts flexibly, even more innovative in the types of entities. However, up to now, research on the NER of ancient weapon categories based on classic historical novels has not been published, and this area is worthy of excavation and analysis.

Named entity recognition is mainly about entity sequence annotation, extracting factual information with specific semantic meaning from unstructured text, carrying out entity relationship extraction, knowledge mapping and other related tasks to achieve text mining and analysis. Although the domestic named entity recognition research started late, the exploration of Chinese named entity recognition has been deeply concerned by the academic community. As a result, scholars have started to break the traditional modes and adopt a new interdisciplinary paradigm to further exploit canonical texts with good results. Currently, it focuses more on the model diversity in the field of historical texts, than models such as recurrent neural networks (RNN) and long short-term memory networks (LSTM) used for traditional deep learning. Li Na et al. performed automatic extraction of multiple types of named entities for local records based on a conditional random field model, with an accuracy of up to 98.28% [10]. Cui Dandan et al. proposed the Lattice LSTM model, which improved the F1 score by about 3.95% compared to the traditional BiLSTM-CRF model [11]. Subsequently, in October 2018, Google AI proposed a deep learning-based preprocessing model, BERT, which achieved good results in research in the field of natural language processing [12]. The above studies demonstrate that deep learning of relevant modeling methods using natural language processing techniques has been applied to text data analysis studies with good results.

At present, the following difficulties exist for NER of ancient weapons in historical novels: ① The incorrect splitting of longer weapon entities caused by over-sensitive word separation software under the interference of a single weapon entity, which affects the recognition accuracy. ② The number of weapon entities is relatively few compared to other names, and organization conventional entities. ③ The fuzzy boundary nature of weapon entities requires high model accuracy. In order to explore the NER techniques for feature entities in classic Chinese historical novels, this paper uses the text of Romance of the Three Kingdoms as the data to construct a corpus of ancient weapon feature entities, and then realise the extraction and association analysis of ancient weapons and other entities. As a classic Chinese historical novel, Romance of the Three Kingdoms has been famous all over the world, which includes detailed descriptions of ancient weapons.

To address the difficulty of recognising ancient weapon entities in novels in the featured domain, due to the significant advantages of pre-training models in acquiring semantic information, this study adds the pre-training model BERT which can extract higher quality contextual information compared to the pre-training model BERT. And the BiGRU network can capture global or local

ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324
features by inputting semantic vectors sequentially, improving the operational efficiency Then, CRF optimizes decoding and improves the accuracy of special entity recognition, so the BERT-BiGRU-CRF model is built for ancient weapons-like entity recognition in the field of classic novels, and the feasibility and advantages of the method are demonstrated through experiments. This study explores the methods and techniques of identifying and correlating ancient weapons naming entities in Chinese historical novels, which lays a foundation for the application of these methods and techniques in the field of Chinese literature, helping the process of overseas dissemination of Chinese classical culture.

3. Entities Corpus

3.1 Data source and entity definition

This study takes ancient weapons as the main identification object, and the requirements of data from historical fiction texts are rich variety and high frequency of use. Since the establishment of China, about 4,299 research papers have been written for Romance of the Three Kingdoms, which includes a wide variety of ancient weapons. Therefore, it was decided to use the books of The Romance of the Three Kingdoms as the source of data, with a total of over 850,000 words. Unlike modern military texts, ancient weapons in historical fiction with both the tools of war and the equipment of combat.

3.2 Entity annotation method and process

Before the pre-training of the model, the original text data were screened and cleaned according to the distribution of the four types of entities, and the data were first coarsely annotated using the YEDDA annotation software developed by the University of Science and Technology of Singapore, then manually corrected by secondary annotation, and then the above four types of entities were transformed into the BIO (B-begin,I-inside,O-out-side) annotation format for training the The BIO (B-begin,I-inside,O-out-side) annotation format was used to train the model. Examples of the annotation formats are shown in Table 1.

Table 1 BIO Format

word	label	word	label	word	label
关	p	月	K	交	O
羽	p	刀	K	战	O
持	O	欲	O	。	O
青	B-k	与	O	陈	O
龙	I-k	黄	B-p	县	B-o
偃	I-k	忠	I-p	令	I-o

4. BERT-BiGRU-CRF model

The overall model structure of the BERT-BiGRU-CRF model for the recognition of named entities of weapons in the canonical texts mentioned in this study is shown in Figure 1. First, the text is input to the pre-trained BERT model, which generates a word vector for each character and converts it into a low-dimensional vector form, combining its linguistic regularity and lexical features. Subsequently, the word vector is input to the BiGRU layer, which performs bi-directional semantic encoding based on pre and post-textual information features, and extracts semantic temporal features by processing the serialised data through directed loops. Finally, the relevant semantic vectors are fed into the CRF layer for decoding, and the tag sequence with the highest probability is output.

ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324

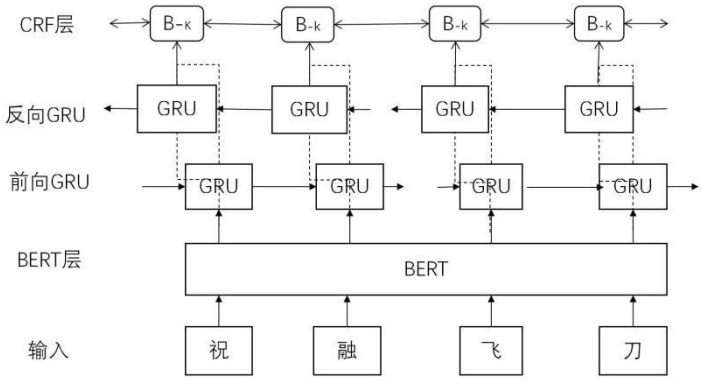

Fig1 BERT-BiGRU-CRF structure

4.1 BERT pre-training model

In October 2018, the BERT pre-training language model was proposed by Google AI and the structure is shown in Fig. 2. To better fuse the information of the upper and lower features of the word vector, the model uses a bidirectional Transformer as the encoder and two main sub-layers placed in a single coding unit, namely a multi-headed self-attentive layer and a feed-forward neural network layer [12].

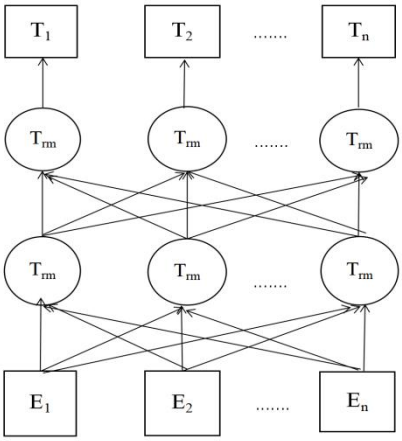

Fig.2 BERT Structure

The Self-Attention section is the core module of the Transformer. The main task of this module is to calculate the interrelationships between the words in each sentence and the whole, and to indirectly reflect the relevance and importance of each word according to their interrelationships. The calculation equations are as follows.

The BERT model uses two target tasks to complete learning during pre-training, including the masked language model (MLM) and adjacent sentence prediction (NSP). During training, the MLM randomly masks 15% of the word sequence, replacing 80% of the masked words with [Mask] masks, leaving 10% of the words unchanged, and replacing 10% of the words with random codes. The masked words are predicted during training based on the unmasked words. Adjacent sentence prediction, on the other hand, is based on the correlation between two sentence features, thus determining the exact position of two sentences. The input text is pre-trained with the BERT model and the final output is calculated after a multi-layer Transfomer extraction, which contains multi-level features and more complete semantic information. Therefore, this study introduces the BERT pre-trained language model, which is able to better grasp the word vector features and sequence patterns through in-depth learning by double-layer decoding, and the model is more efficient and high quality in completing the task of training applications for small-scale corpora compared to traditional shallow learning models.

ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324

4.2 BiGRU layer

A Gated Recurrent Unit (GRU) is a specific recurrent neural network (RNN) model that performs machine learning tasks related to memory and clustering through the connection of a series of nodes, with the GRU carrying information over multiple time periods to influence subsequent time periods. While the LSTM structure contains forgetting gates, input gates and output gates, which often result in phenomena such as gradient explosion in traditional recurrent neural network models, the LSTM layer only partially solves the gradient problem and is computationally time-consuming, the GRU can be considered a variant on the LSTM as both have similarities in design, where the presence of gated recursive unit helps to adjust the input weights of the neural network in order to solve the vanishing gradient problem. And structurally, GRU combines the forgetting and input gates into one, with only update and reset gates present. Therefore, the GRU layer used in this paper not only has the advantages associated with LSTM, but also has a relatively simple structure, saving computation time [4].

In terms of how the GRU unit works, it can retain only useful information and has a simple structure, reducing the complexity of the computation. However, a simple GRU cannot meet the application requirements of canonical texts. Therefore, this study introduces a reverse GRU to learn backward semantics, which can extract key features of the word vectors of text data backward to ensure the depth and comprehensiveness of model learning, and combines the forward GRU with the reverse GRU, i.e. BiGRU network model. The structure of BiGRU network is shown in Figure 3.

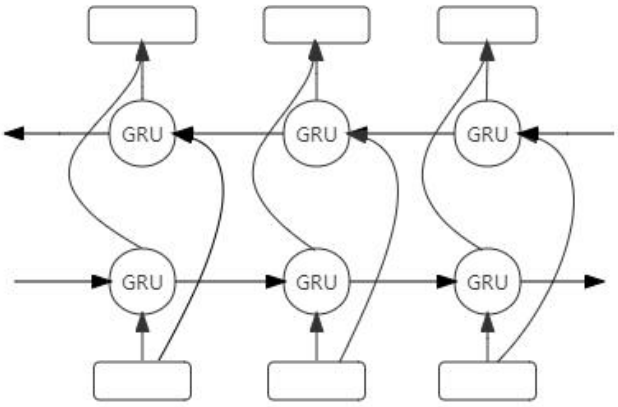

Fig.3 BiGRU Network

4.3 CRF layer

After the global feature extraction and local semantic capture by the BERT pre-trained language model and BiGRU layer, the CRF layer is used as the output layer for deep learning, and the CRF layer is used for decoding. the contextual feature vector in the hidden state output of the BiGRU layer does not reflect the dependencies between the labels. The CRF layer is called a linear chain conditional random field model, and compared with the traditional softmax, which only focuses on the optimal solution of local features, CRF decoding, as a discriminative model, can learn to extract all word-related features and obtain the optimal solution of sequence annotation on the global whole, improving the final prediction of label accuracy.

5. Experiments and analysis of results

5.1 Experimental environment and model parameters

In this paper, the experiments were conducted on Ubuntu 18.04, based on Python 3.6 and TensorFlow 1.14.0 framework, and accelerated by Nvidia GeForce RTX 2070 SUPER (16G) video memory.

ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324

In the BERT-BiGRU-CRF model, the BERT-Base was 12 layers, the batch_size value was set to 4, the max_seq_len was 202, and the initial learning rate was 0.00001. To prevent overfitting problems during training, the dropout was increased at both ends of BiGRU, and the corresponding value was 0.5.

5.2 Experimental evaluation criteria

This study uses the most common evaluation criteria used in named entity recognition tasks, namely accuracy (P), recall (R) and the harmonic mean F1 value, to measure the performance of the model applied to the domain of the canonical weapons category in a comprehensive manner. p is the recognition rate of correctly identified entities, R is the recognition rate of correctly identified entities in the test set and F1 is the harmonic mean of P and R. Theoretically, the higher the P and R values, the higher the accuracy and recall, but this is not the case and there is a possibility of an inverse relationship between the two in the experiments. The F1 value is therefore a comprehensive evaluation indicator of the performance of the test model.

5.3 Experimental procedure

In order to demonstrate the recognition effectiveness of the BERT-BiGRU-CRF model adopted in this paper in recognising weapons-like entities in the canon, four sets of comparison experiments will be set up in this paper, using the BiLSTM, CRF, BiLSTM-CRF and BERT-BiGRU-CRF models respectively, to compare the experimental results according to the evaluation metrics mentioned above.

5.4 Experimental results

The results of BERT-BiGRU-CRF for four types of entities recognition are shown in Table 2.

Table2 The Results of BERT-BiGRU-CRF

Types	P/%	R/%	F1/%
names	89.03	98.00	93.30
localization	91.32	87.68	89.46
weapon	86.71	90.25	88.44
official title	91.97	87.60	89.73

The experimental results show that the BiLSTM model and the CRF model are not ideal in terms of overall recognition effect for the four classes of entities, and for the ancient weapons class entities in the canonical text with high complexity, the single model training is not effective in obtaining more complete semantic information. With the addition of the CRF model to the BiLSTM, the recall values significantly improved, and the preliminary F1 values increased by 3.94% compared to the single BiLSTM model for the weapon class entities, demonstrating that the CRF layer can comprehensively extract and discriminate sequences of entity-related features and improve the model recognition effect. The overall effect of using the BERT-BiGRU-CRF model is better than the other models, and the best F1 value of the weapon class entity reaches 88.44% at one time, indicating that after the introduction of the BERT pre-training model, the contextual feature information can be obtained through training, and a more complete semantic information can be obtained through feature extraction by the 12-layer bi-directional Transformer. The BiGRU layer replaces the BiLSTM layer, mainly because the former structure is simpler compared to the latter, and the former has superiority in time under the same amount of computational tasks. Under the features of small corpus size, complex entity names and short length, for the recognition of named entities of ancient weapons, the BERT-BiGRU-CRF model embodies, compared to other models, a significant advantages, see Figure 4.

ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324

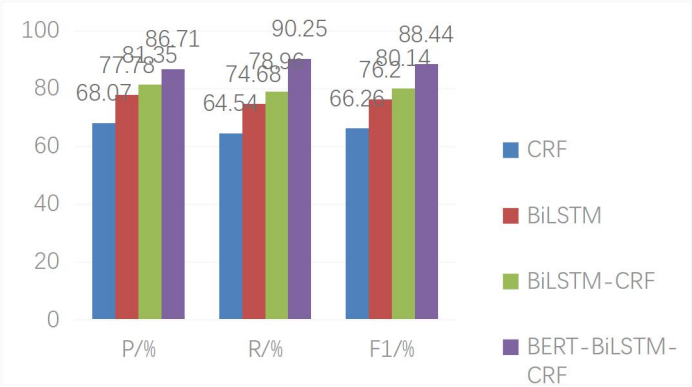

Fig4. The Results of Four Models

5.5 Knowledge Graph Presentation

Based on the correspondence between ancient weapon class entities and other entities such as other names, place names and official positions in the experimental data set, a basis for further mining and analysis of canonical data by combining feature entities is laid. Ancient weapon category named entity identification study, with the help of association relationship visualization, can be explored through the weapon category to analyse the character of relevant characters in the canon, social and in material information changes, reasons for victory and defeat in battle, etc. In summary, the feature entities are closely associated with each entity, providing clues and technical means for further mining to explore the characteristic canon.

6. Conclusion

Currently, research related to canonical literature in the field of digital humanities is expanding in depth, and named entity recognition and association analysis through feature entities is another active exploration of text mining in the canonical domain. Due to the representativeness and complexity of ancient weapons-like entities in the canonical literature, the recognition process leads to poor recognition results of generic models. In this study, taking The Romance of the Three Kingdoms as an example, ancient weapons are selected as feature entities, a relevant corpus is constructed, deep learning and knowledge graph methods are fused, a naming recognition method based on the BERT-BiGRU-CRF model is built, and the corpus data is divided into proportions for multiple experiments, and the final F1 value of weapons class entities reaches 88.44%, with higher recognition accuracy than other models and shorter computing time shorter, and the use of knowledge graph to correlate feature entities for further interpretation and analysis of canonical content, proving the feasibility and superiority of the application of the method. In comparison with the traditional models, the BERT-BiGRU-CRF model obtains word vectors combined with contextual information through the BERT layer, and uses BiGRU-CRF to obtain the optimal annotation sequence, demonstrating that the model approach not only ensures fast training speed under simple structures, but also can solve the contextual information to a certain extent The ambiguity problem is solved, and the difficulties such as unclear boundary and semantic complexity of entity recognition of feature entities in the canonical text are solved, and the accuracy of entity recognition is effectively improved. In the subsequent research, it is necessary to further expand the application scope of canonical texts, increase the number of corpus annotations in each feature entity domain, broaden the application scope of the discipline, and make a pavement for the exploration of canonical texts with Chinese characteristics related to mining.

References

[1] Nadeau D , Sekine S . A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30 (1) :págs. 3-26.

ISSN:2790-167X DOI: 10.56028/aehssr.3.1.324
[2] Yi Liu,Jiahuan Lu,Jie Yang,Feng Mao.Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax[J].Mathematical Biosciences and Engineering,2020,17 (6) :7819-7837.

[3] Ren N, Bao T,Shen G,Guo T.Fine-Grained Named Entity Recognition Based on Deep Learning：A Case Study of Tomato Diseases and Pests[J].Information Science,2021,39 (11) :96-102.

[4] Qin Q,Zhao S,Liu C.A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic MedicalRecords[J].Complexity,2021,https://doi.org/10.1155/2021/6631837

[5] Feng Yuntian, Zhang Hongjun, Hao Wenning. Named entity recognition for military text [J]. Computer Science, 2015,42 (07): 15-18 + 47.

[6] Yin Xuezhen, Zhao Hui, Zhao Junbao, Yao Wanwei, Huang Zelin. Military domain named entity recognition for multi-neural network collaboration [J]. Journal of Tsinghua University (Natural Science Edition), 2020,60 (08): 648-655.

[7] Ouyang Jian. Digital Humanities research of the Humanities from the perspective of Big Data [J]. Library Magazine, 2018,37 (10): 61-69.

[8] Li J,Sun A,Han J,et al.A Survey on Deep Learning for Named Entity Recognition[J].IEEE.2018,34 (1) :50-70.

[9] Eddy S R .Hidden Markov models.[J]. Current Opinion in Structural Biology.1996, 6 (3) :361-365.

[10] Zhu Haodong, Yang Lizhi, Ding Wenxue, et al. Chinese microblog named entity recognition based on topic tags and CRF [J]. Journal of Central China Normal University: Natural Science Edition, 2018,52 (3): 316-321.

[11] Li N. Construction of Joint Automatic Recognition Model of Multi-Type Named Entities for Local Records[J]. Library Tribune.2021,41 (12) :113-123.

[12] Cui J, Zheng D, Wang D, Li T. Named Entity Recognition of Chrysanthemum Poetry Based on Deep Learning Models [J]. Information Studies:Theory & Application.2020,43 (11) :150-155

[13] Devlin J,Chang M W,Lee K,et al.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J].arXiv preprint arXiv:1810.04805,2018.