Named Entity Recognition Algorithm for Archaeological Site Corpus

Yong Xu [a], Yunke Peng [b], Hengna Wang [c, *], Xue'er Wang [d]

School of Management Science & Engineering, Anhui University of Finance & Economics, Bengbu 233000, China;

[a]xuyong@aufe.edu.cn, [b]1410939755@qq.com, [c]120081267@aufe.edu.cn, [d]2173600168@qq.com

*Corresponding author: Hengna Wang, female, master, her studying interesting include NLP, 120081267@aufe.edu.cn.

Abstract

Named entity recognition of archaeological site corpus is a key step in triple group extraction and question understanding. In view of the low accuracy of named entity recognition in archaeological site field, BERT_BiLSTM_CRF model is proposed, which dynamically generated a word vector containing text semantic information through BERT model, enhanced the semantic information contained in the word vector by learning the context information of BiLSTM model, and finally calculated the entity label using CRF model. Named entity recognition used BMESO sequence annotation to label the boundary of named entity and single named entity. BERT_BiLSTM_CRF Model was executed on the corpus on Ming Zhongdu Building, which was compared with classical models of BERT, BERT_BiLSTM. The performance of the model exceeds that of the classical models, where the value of P, R and F1 of the model our proposed are 99.79%, 99.22% and 99.53% respectively.

Keywords

Archaeological sites; Named entity recognition; BERT_ BiLSTM_ CRF.

1. Introduction

National Cultural Heritage Administration issued the "Fourteenth Five Year Plan" for the Protection and Utilization of Large Cultural Relics, proposing to use big data and artificial intelligence technology to promote the high-quality development of archaeological site parks, realize smart parks and smart cultural tourism, show the historical and cultural value of archaeological sites, and show the spiritual pursuit of the Chinese nation [1]。 When implementing smart parks and smart cultural tourism projects, the intelligent question answering system can be used as an auxiliary tool to meet the needs of users. Named Entity Recognition (NER) can extract named entities from text, and on this basis, it can extract triples and understand questions, which plays an important role in the subsequent steps of intelligent question answering system [2-4]。

Current BiLSTM_ The common deep learning NER model of CRF model [5]， Li[6]、Xiao[7] applied the BiLSTM_CRF model to the NER in the field of electronic medical records and traditional Chinese medicine, BiLSTM can use the bi-directional BiLSTM to learn the context information of the text sequence, The CRF random field infers the optimal labeling sequence according to the dependencies between sequences. Before inputting the sequence text into the BiLSTM_CRF model, the text needs to be vectorized. The traditional method uses the Word2Vec method to train word vectors or word vectors[8-10] but the word vectors trained by Word2Vec are static and are trained after using corpus training. The output vector cannot be changed, and the vector of a word or a word is fixed. For example, the word vector of "wo ai chi ping guo" and "ping guo" in "ping guo shou ji" is the same, but the semantics of the two are obviously different. the same. The BERT model[11] integrates the word embedding information, position information and sentence information of the text, and uses two tasks of MLM (language training model with mask) and NSP (next sentence prediction task) for training, and the final output word vector contains more Rich semantic information; and dynamically obtain word vectors according to the input text, which can solve the polysemy problem of Word2Vec. Therefore, using BERT to vectorize text and input it into a deep learning model can effectively improve the performance of NER. Guo [12] and Xu[13] used BERT and BiLSTM_CRF models to perform NER on Chinese resumes and biomedical fields, respectively, and achieved certain results. Improve. This paper uses the BERT_BiLSTM_CRF model[12] for entity recognition on a corpus of archaeological sites.

This paper collects the archaeological site corpus of the Ming Zhongdu Grand Site, and annotates the data. Then use the BERT_BiLSTM_CRF model for named entity recognition experiments. First, input the text into the BERT model to obtain the word vector with semantic information, then obtain the word vector of the text context information through BiLSTM, and finally calculate the optimal sequence annotation result through the CRF model.

2. Related work

Named Entity Recognition (NER) is to identify pre-defined categories and named entities with specific meanings from text, such as person names, place names, institution names, etc.[14],NER can be traced back to the 7th IEEE Artificial Intelligence Applications Conference in 1991. An article on identifying company names was published during[15] , and "Named Entity" was first proposed at MUC-6 (6th Information Understanding Conference) in 1996[16] . Named entity recognition is a basic task of natural language processing (NLP), an important technology in intelligent question answering systems, and the basis of information retrieval and answer generation[17] . Currently named entity recognition is mainly divided into three categories:

(1) Named entity recognition based on dictionary and rules. This method requires a lot of time to formulate dictionaries and rules. It has high accuracy and efficiency in small samples, but it relies too much on dictionaries and rules, and formulating dictionaries and rules will consume a lot of time and energy, and it will not perform well on a large number of data sets. not good. Bao[18] and Li[19] used the method of combining rules with dictionary to realize named entity recognition, and made dictionaries and rules by analyzing the root, affix and context features of named entities.

(2) Named entity recognition based on machine learning. Using a machine learning model to realize named entity recognition, that is, converting the problem into a classification problem, usually requires the use of vectors to represent data and features, and then input them into the machine learning model for training to determine whether a word or word is a named entity, which can be very good. The solution to the shortcomings of dictionary-based and rule-based methods. Commonly used models for machine learning named entity recognition are Hidden Markov Model (HMM)[20-23], Maximum Entropy Model (MEM)[24-26] , Support Vector Machine (SVM)[27-29] and Conditional Random Fields (CRF)[30, 31].

(3) Named entity recognition based on deep learning. Common deep learning models include Convolutional Neural Network (CNN)[32, 33] Recurrent Neural Network (RNN), Long Short-Term Memory Neural Network (LSTM)[34] , etc. There are also models that combine CNN, LSTM and other models together to form new Recognition models[35, 36] have also achieved better results.

With the application of pre-trained models in NLP, the use of pre-trained models Word2Vec[37-39] , GLOVE[40] , etc. has become the content of researchers. However, these pre-trained models cannot solve the problem of polysemy, and the emergence of the BERT[11] model solves this problem. The BERT model dynamically obtains the vector of the text in the field of the Mingzhongdu site, and then predicts the sequence label in the text through the BILSTM_CRF model to realize named entity recognition.

3. Model

The BERT_BiLSTM_CRF model receives the text that needs to be recognized by named entities, and calculates the sequence annotation results of the output text through the model. The model consists of 4 parts, input text, BERT module, BiLSTM module, and CRF module. Enter the text data "zhong lou yu gu lou ju li wu li", dynamically obtain the vector containing the text semantics through BERT, and enter the vector into the BiLSTM model to learn the text context information, and finally the optimal label sequence "B-building E-building O B-building E-building O O O O" is input through CRF, and it is judged that the bell tower and the drum tower belong to Building named entities.

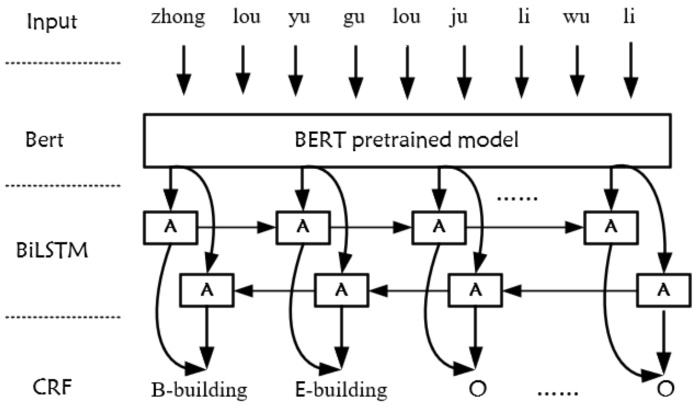


Figure 1 Model Architecture

4. Experiment and result analysis

4.1. Experimental data and evaluation indicators

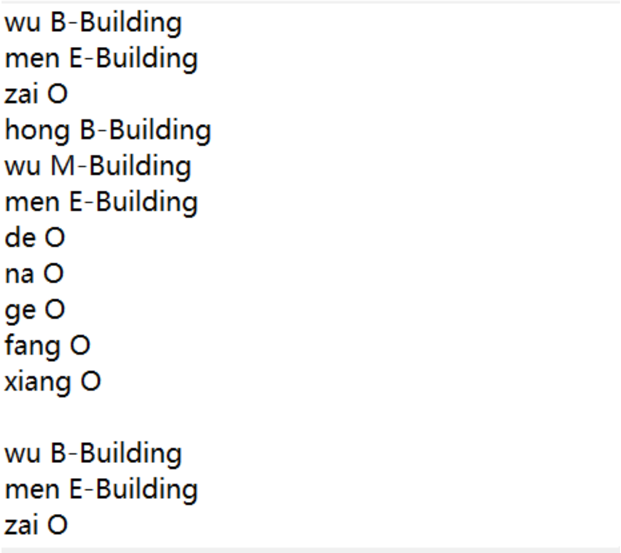
4.1.1. Experimental data

The experimental data in this paper comes from the hand-made Mingzhongdu question and answer pair, and the knowledge source is Mr. Wang Jianying's book "Mingzhongdu Research"[41] . There are a total of 4899 pieces of data, including five named entities of building, person, time, brick and character. The buildings are the various buildings in the central capital of the Ming Dynasty, such as Hongwumen, Bell Tower and so on. The persons are related in Ming Dynasty, such as Zhu Yuanzhang, Li Shanchang and so on. The time is the time when the relevant events occurred in the central capital of the Ming Dynasty, and the time is the time in the form of a year, such as the third year of Hongwu and the fourth year of Hongwu. Bricks are various types of bricks that designate the central capital, such as place name transfer, year bricks, carved bricks, etc. Characters refer to the characters engraved on the bricks. For example, the characters on the place name bricks include "Nanchang, Nanchang, Jiangxi Province" and "New Construction of Nanchang, Jiangxi Province".

This paper sorts out the named entities contained in the corpus, makes a named entity dictionary corresponding to the professional field, and uses the dictionary to sequence the named entities in the corpus. The sequence labeling method adopts the BMESO labeling method. B (Begin) represents the beginning of an entity, E (End) represents the end of the entity, M (Middle) represents the middle part of the entity except the beginning and the end, and S (Sigle) represents a single word as an entity , O(Outside) means all sequence annotations except the solid part are shown in Figure 2. There are a total of 7640 entities in the dataset, including 4252 building entities, 174 character entities, 882 time entities, 1402 brick entities, and 930 word entities.

Table 1 Number of dataset entities

Entity Type	Number of Entity
Building	4252
Person	174
Time	882
Brick	1402
Character	930
Total	7640



```
wu B-Building
men E-Building
zai O
hong B-Building
wu M-Building
men E-Building
de O
na O
ge O
fang O
xiang O

wu B-Building
men E-Building
zai O
```

Figure 2 Sequence annotation example

4.1.2. Evaluation indicators

There are five types of named entities identified in this paper. P (precision), R (recall) and F1 value (F1-score) are used to evaluate the performance of the model, and complete matching is performed during calculation. When all labels of an entity are correctly labeled time is correctly identified. In the confusion matrix, TP means that the predicted value and the actual value are both positive, FP means that the predicted value is positive and the real value is negative, FN means that the predicted value is negative and the real value is positive, and TN means that the predicted value and the real value are both negative. Confusion matrix and formula (a) calculate P, R, F1.

Table 2 Confusion matrix

Predicted value	True value	
	Positive example	Negative example
Positive example	TP	FP
Negative example	FN	TN

$$\begin{cases} P = \dfrac{TP}{TP+FP} * 100\% \\ R = \dfrac{TP}{TP+FN} * 100\% \\ F1 = \dfrac{2*p*R}{P+R} * 100\% \end{cases} \tag{a}$$

4.2. Experimental results and analysis

4.2.1. Analysis of experimental results of different models

The BERT_BiLSTM_CRF model proposed in this paper was tested on the corpus of the Mingzhongdu archaeological site, and compared with several other common models to verify the performance of the model in this paper. The results of the comparison test are shown in Figure 3and Table 3 Comparison of the recognition results of each model.

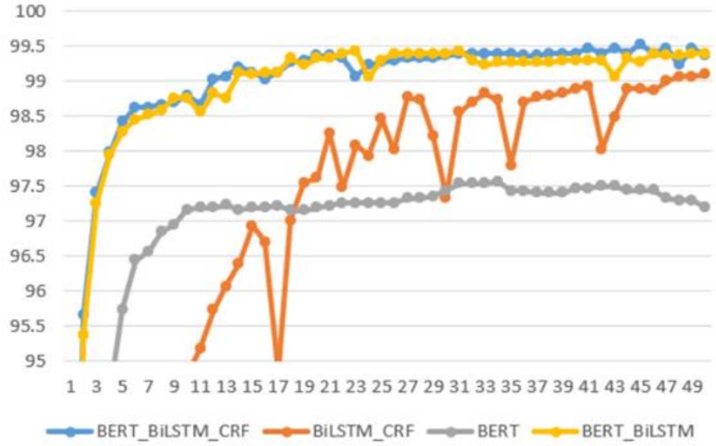


Figure 3 The F1 value of each model changes with the number of training times

Table 3 Comparison of the recognition results of each model

Model	P/%	R/%	F1/%
BERT_BiLSTM_CRF	99.79	99.22	99.53
BiLSTM_CRF	99.79	98.28	99.07
BERT	99.03	97.35	97.57
BERT_BiLSTM	99.79	99.09	99.4

According to Table 3, it can be seen that the BERT_BiLTM_CRF model has achieved good results in the Mingzhongdu archaeological site corpus, and the precision, recall, and F1 are 97.79%, 99.22%, and 99.53%, respectively. Compared with the BiLSTM_CRF model F1, the F1 is improved by 0.46%. The input of BiLSTM_CRF uses Word2Vec to convert the input text into a vector representation. The experimental results show that the BERT model has a better effect in text vector representation than Word2Vec. Compared with the BERT model, the F1 value is improved by 1.96%, indicating that the BiLSTM and CRF layers can better learn the contextual information of the text. Comparing the BERT_BiLSTM model, it is found from the training curve that the training curves of the two are very close. The recognition results show that the difference in F1 is 0.13%, which is very small. The reason may be that the input corpus is a manually formulated question-and-answer pair text, which is structural and named entity meeting. Appears in a fixed position, and can be more accurately predicted without adding the CRF layer.

4.2.2. Analysis of experimental results of different named entities

In order to learn more about the recognition effect of the model in this paper on different named entities, this paper conducts a detailed comparative experiment on the recognition results of five named entities, and the recognition results are shown in Table 4. It can be seen from Table 4 that the recognition effect of person named entities is the best, mainly because there are fewer person named entities and the person name is relatively simple. The recognition effect of character named entities is the worst, because the lengths of character named entities vary, and the composition of place name bricks, five-element bricks, and guardian bricks is complex.

Table 4 The recognition effect of the model in this paper on different named entities

named entity class	R/%	P/%	F1/%
Building	99.76	99.88	99.82
Person	100	100	100
Time	100	98.86	99.43
Brick	100	98.91	99.45
Character	100	96.63	98.29

5. Conclusions

This paper proposes a named entity recognition model BERT_BiLSTM_CRF in the field of large ruins. The model achieves a F1 of 99.53% on the experimental data, and has achieved good results. With the development of deep learning technology, the research on named entity recognition will become more and more in-depth, and the model in this paper still has shortcomings. The follow-up work is mainly carried out from the following aspects:

(1) The dataset is small, and the text has certain rules. After that, the dataset will be expanded and free text will be added.

(2) Try different data labeling methods to verify the impact of data labeling methods on model training.

(3) Apply the model to the intelligent question answering system to realize the Mingzhongdu intelligent question answering system.

Acknowledgements

The work was funded by the Key Research and Development Project of Anhui Province (Grant No. 2022O07020001).

References

[1] Notice of the State Administration of Cultural Heritage on Printing and Distributing the "14th Five-Year Plan for the Protection and Utilization of Great Sites" [Z]. 2021

[2] Tao Yongqin. Design and Implementation of Intelligent Question Answering System in Professional Field [J]. Computer Application and Software, 2018, 35(05): 95-101.

[3] Yang Wei, Sun Deyan, Zhang Xiaohui, et al. Named Entity Recognition Algorithm for Power Intelligent Question Answering System [J]. Computer Engineering and Design, 2019, 40(12): 3625-30.

[4] Zhang Fangrong, Yang Qing. Research on Entity Relationship Extraction in Knowledge Base Question Answering System [J]. Computer Engineering and Applications, 2020, 56(11): 219-24.

[5] Cui Dandan, Liu Xiulei, Chen Ruoyu, et al. Ancient Chinese Named Entity Recognition Based on Lattice LSTM [J]. Computer Science, 2020, 47(S2): 18-22.

[6] Li Gang, Pan Rongqing, Mao Jin, et al. Entity Recognition of Chinese Electronic Medical Records Integrating BiLSTM-CRF Network and Dictionary Resources [J]. Modern Intelligence, 2020, 40(04): 3-12+58.

[7] Xiao Rui, Hu Fengju, Pei Wei. Named Entity Recognition of Traditional Chinese Medicine Text Based on BiLSTM-CRF [J]. World Science and Technology-Modernization of Traditional Chinese Medicine, 2020, 22(07): 2504-10.

[8] Jiang Xiang, Ma Jianxia, Yuan Hui. Named Entity Recognition in Ecological Governance Technology Field Based on BiLSTM-IDCNN-CRF Model [J]. Computer Applications and Software, 2021, 38(03): 134-41.

[9] Zeng Qingxia, Xiong Wangping, Du Jianqiang, et al. Named Entity Recognition of Electronic Medical Records Combined with Self-Attention BiLSTM-CRF [J]. Computer Applications and Software, 2021, 38(03): 159-62+242.

[10] Ding Shengchun, Fang Zhen, Wang Nan. Named Entity Recognition in Business Domain Based on Bi-LSTM-CRF [J]. Modern Intelligence, 2020, 40(03): 103-10.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. 2018.

[12] Guo Juncheng, Wan Gang, Hu Xinjie, et al. Named Entity Recognition in Chinese Resume Based on BERT [J]. Computer Applications, 2021, 41(S1): 15-9.

[13] Xu Li, Li Jianhua. Biomedical Named Entity Recognition Based on BERT and BiLSTM-CRF [J]. Computer Engineering and Science, 2021, 43(10): 1873-9.

[14] Li Dongmei, Ross, Zhang Xiaoping, et al. A Review of Named Entity Recognition Methods [J]. Computer Science and Exploration: 1-18.

[15] F R L. Extracting company names from text; proceedings of the IEEE Conference on Artificial Intelligence Applications, F, 1991 [C].

[16] B G R S. Message understanding conference 6: a brief history; proceedings of the Proceedings of the 16th Conference on Computational Linguistics-Volume 1, F, 1996 [C].

[17] Zhou Botong, Sun Chengjie, Lin Lei, et al. Automatic question answering of large-scale knowledge bases based on LSTM [J]. Journal of Peking University (Natural Science Edition), 2018, 54(02): 286-92.

[18] Bao Minna, S. Lauglau. Research on Mongolian Named Entity Recognition Based on Dictionary Matching [J]. Journal of Minzu University of China (Philosophy and Social Sciences Edition), 2017, 44(03): 165-9.

[19] Li Nan, Zheng Rongting, Ji Jiuming, et al. Research on Chinese chemical substance naming recognition based on heuristic rules [J]. Modern Library and Information Technology, 2010, (5): 5.

[20] Yu Hongkui, Zhang Huaping, Liu Qun, et al. Chinese Named Entity Recognition Based on Cascaded Hidden Markov Models [J]. Journal of Communications, 2006, (02): 87-94.

[21] Le Juan, Zhao Xi. A Named Entity Recognition Algorithm for Peking Opera Institutions Based on HMM [J]. Computer Engineering, 2013, 39(06): 266-71+86.

[22] ZHOU G, JIAN S. Named entity recognition using an HMM-based chunk tagger [J]. proc acl, 2002.

[23] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An Algorithm that Learns What\"s in a Name [J]. 1999, 34(1-3): 211-31.

[24] SAHA SK, SARKAR S, MITRA P. Feature selection techniques for maximum entropy based biomedical named entity recognition [J]. Journal of Biomedical Informatics, 2008, 42(5).

[25] BORTHWICK A, STERLING J, AGICHTEIN E, et al. NYU: Description of the MENE Named Entity System as Used in MUC7 [J]. 1998.

[26] BENDER O, OCH F J, NEY H. Maximum Entropy Models for Named Entity Recognition [J]. Proceedings of Conll –, 2003.

[27] ISOZAKI H, KAZAWA H. Efficient Support Vector Classifiers for Named Entity Recognition; proceedings of the International Conference on Computational Linguistics-volume, F, 2002 [C].

[28] Li Lishuang, Huang Degen, Chen Chunrong, et al. Recognition of place names in Chinese text based on support vector machine [J]. Journal of Dalian University of Technology, 2007, (03): 433-8.

[29] Chen Xiao, Liu Hui, Chen Yuquan. Recognition of Chinese Organization Names Based on Support Vector Machine Method [J]. Computer Application Research, 2008, (02): 362-4+7.

[30] Wang Shikun, Li Shaozi, Chen Tongsheng. TCM Named Entity Recognition Based on Conditional Random Fields [J]. Journal of Xiamen University (Natural Science Edition), 2009, 48(03): 359-64.

[31] Zheng Rongting, Li Nan, Ji Jiuming, et al. Research on the identification of chemical substance names in Chinese [J]. Modern Library and Information Technology, 2010, (06): 48-52.

[32] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural Language Processing (almost) from Scratch [J]. Journal of Machine Learning Research, 2011, 12(1): 2493-537.

[33] LIN Y, HONG L, YI L, et al. Biomedical Named Entity Recognition based on Deep Neutral Network [J]. International Journal of Hybrid Information Technology, 2015, 8(8): 279-88.

[34] HUANG Z, WEI X, KAI Y. Bidirectional LSTM-CRF Models for Sequence Tagging [J]. Computer Science, 2015.

[35] Chen Dexin, Zhan Yuanyuan, Yang Bing, et al. Research on online medical entity extraction based on CNN-BiLSTM model [J]. Library and Information Work, 2019, 63(12): 105-13.

[36] Liu Yupeng, Li Dongdong. Chinese Named Entity Recognition Method Based on BLSTM-CNN-CRF [J]. Journal of Harbin University of Science and Technology, 2020, 25(01): 115-20.

[37] Cao Yiyi, Zhou Yinghua, Shen Fahai, et al. Research on Named Entity Recognition of Chinese Electronic Medical Records Based on CNN-CRF [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019, 31(06): 869-75.

[38] Han Pu, Liu Yizhuo, Li Xiaoyan. Research on entity recognition of Chinese electronic medical records based on deep learning and multi-feature fusion [J]. Journal of Nanjing University (Natural Science), 2019, 55(06): 942-51.

[39] Wang Huan, Zhu Wenqiu, Wu Yuezhong, et al. Named Entity Recognition Based on the Fault Field of CNC Machine Tool Equipment [J]. Journal of Engineering Science, 2020, 42(04): 476-82.

[40] NING G, BAI Y. Biomedical named entity recognition based on Glove-BLSTM-CRF model [J]. Journal of Computational Methods in Sciences and Engineering, 2020, 21(1).

[41] Wang Jianying. Research on the Central Capital of the Ming Dynasty [M]. Research on the Central Capital of the Ming Dynasty, 2005.