Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 4, No. 2, 2022

Named Entity Recognition of Ancient Wine Texts Based on Deep Learning Models

Wei Zhang, Yadong Wu*, Weihan Zhang, Yuling Zhang, Xiang Ji

School of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin, China
* Corresponding author: Yadong Wu (Email:wyd028@163.com)

Abstract: Named entity recognition of ancient wine canonical books helps to excavate the history of wine culture, inherit ancient Chinese wine culture and help modern wine industry. In this paper, the collection of ancient wine data was carried out through the web and books, and after analysis and discussion, five types of entities were selected for manual entity annotation: wine name, person name, place name, event name and time, and the corpus was trained and compared using machine learning method CRF and deep learning method LTSM with BERT pre-training model. BERT pre-training model has the best effect among all models, and the reconciliation of entity recognition The average is higher than other models, up to 88.33%. The subsequent annotation process and BERT model can be optimized or the sample corpus can be expanded to improve the training effect, and the entity annotation system of ancient Chinese text can also provide a reference for subsequent research.

Keywords: Text mining, Natural language processing, Named entity recognition, Neural network; BERT.

1. Introduction

deposits. Wine culture refers to the material and spiritual culture produced during the production, sale and consumption of wine[1]. For the history and culture of wine, the ancients left many historical facts and stories related to wine in books. These documents from all dynasties and generations record the history of Chinese wine, the development direction and cultural elements of the history of Chinese wine, and the scale of the industry at that time, and the inheritance of Chinese wine culture.

For the study of wine history, there are currently no systematic historical materials available, often through experts who dig the history of wine from the vast amount of ancient documents plucked from the sand. If the required physical information in the ancient corpus is extracted with the help of certain algorithms, the manual workload can be greatly reduced. The emergence of natural language processing technology, applied to digital humanities research, can replace tedious manual analysis and provide a new approach to the excavation andanalysis of wine history and culture. Currently, natural language processing techniques can be used to identify information such as names of people, places, institutions and time in ancient texts in order to parse the semantic knowledge in ancient texts and thus improve the analysis and mining of the texts[2] .

This study takes ancient literature on the topic of wine as the object, builds a database of ancient wine literature, then establishes an entity annotation system for the text of the literature, builds a model for named entity recognition of ancient wine literature, and uses deep learning methods to build a deep data mining corpus in a proprietary domain to promote Chinese wine culture.

2. Research Background

The named entity recognition task is a subtask within the information extraction domain whose task goal is to find, identify and classify relevant entities, such as names of people, places and institutions, from sentences given a piece of unstructured text[3] . There are 3 main categories of mainstream approaches: lexicon and rule-based approaches, statistical machine learning-based approaches, and deep learning-based approaches.

2.1. Dictionary and rule-based approach

After the concept of named entities was introduced, early research mainly used the rule and dictionary approach[4]. This approach constructs a large number of rulesets or dictionaries, and then matches the Chinese characters to be recognized in the rulesets and dictionaries according to the requirements, and corrects them until they are successfully matched. This method has its limitation that it can only get high accuracy on small data sets, and it is only highly accurate on a specific corpus, and it is less applicable after adding other data.

As the number of entities to be recognized diversifies, more rule sets and lexicons need to be developed, and manual work becomes more laborious and complex. With the development of machine learning, named entity recognition tasks performed thereafter can also utilize statistical machine learning-based methods.

2.2. Statistical machine learning based approach

The methods based on statistical machine learning are: Hidden Markov model-based methods, maximum entropy-based methods, support vector machine based methods and conditional random field based methods[5] . There are many scholars using statistical machine learning based methods, especially the Conditional Random Fields (CRF) model for named entity recognition in modern and ancient Chinese. Huang Shuiqing et al. conducted named entity recognition of ancient Chinese place names from the Spring and Autumn Zuo's Biography corpus and obtained the conclusion that the CRF model outperformed the maximum entropy model[6] . Wang Dongbo et al. used the CRF model to study the automatic named entity recognition of human, place, and time names for five pre-Qin canonical texts, including the Biography of Gong Yang, and obtained a high accuracy

rate[7].

2.3. Deep Learning Based Approach

As the development of deep learning techniques based on neural network models in the field of machine learning has become more advanced, more scholars are using deep learning-based methods for named entity recognition research. The common methods are CNN, RNN, LSTM, BiLSTM, and methods incorporating BERT models. For example, Wang Fan et al. conducted a study on medical named entity recognition using multilayer neural networks and compared it with the traditional CRF model, and the F-value reached 85.08%, which is 5.47% better than CRF[8]; Xie Tao et al. used LSTM combined with CRF model to recognize Song lyrics and the Historical Records as the ancient text corpus, and the F1 value improved by 14.16% compared with CRF[9] ; He Chunhui et al. introduced a bidirectional neural network Using a Bi-LSTM-CRF model, they extracted entities from electronic documents in the field of diabetes and achieved 89.14% accuracy on a dataset containing 15 entity classes[10] ; Cui Jingfeng et al. used a BERT pre-training model for the recognition of proper nouns in classical poetry of chrysanthemums and achieved an F-value of 91.60%[11] . In the biomedical field, new pre-trained models have also emerged. BioBERT, based on BERT, was pre-trained in the biological corpus PMC and PubMed, and the average F1 value was improved by 0.62 compared to BERT in nine publicly available BioNER datasets[12] ; Eklund et al. fused BioBERT with ELMo for feature extraction and contextual word embedding. reached the highest level on five BioNER datasets[13] . Sun Huan et al. construct a dynamic word vector-based fusion model BERT+BiLSTM+CRF for entity extraction of key information from traffic accident texts, with an average extraction accuracy of 92.40% for multiple samples[14].

3. Research Program

3.1. Research line

This section describes the data pre-processing and model selection in the study.

3.1.1. Original corpus selection and acquisition

Since there is no systematic database available for the corpus of the specialized fields involved in this study, the original corpus of this study mainly comes from the author's collection from two ways. The first is the crawling of web resources, by writing crawlers to collect articles on wine history and culture, etc. The second is the text extraction of books, by extracting the texts of e-books and physical books and saving them through ORC technology.

3.1.2. Corpus processing and database building

(1) Original corpus processing: The original corpus of this study is electronic, firstly, we download its pdf version on the Internet, extract the text using ORC technology, and proofread it with the help of the original book. The proofread text was obtained.

(2) Normalization: Using python language, we extract the document name, author name, dynasty, document content, annotation, etc. using regular expressions and so on.

(3) Data pre-processing: There are some problems with the processed data, such as duplication of poems, incorrect identification of ancient texts, missing article annotations, etc.; these problems are dealt with accordingly.

(4) Create a database: Use mysql to create a database of the wine canon.

3.1.3. Establishment of entity selection and labeling rules

After pre-processing to get the wine canonical data, several groups of entities were selected and some annotation rules were set after reviewing the data for analysis and discussion, followed by screening and comparison. The final entities and annotation rules were determined.

3.2. Deep learning model selection

Research on the technology of named entity recognition has been under development.

3.2.1. Rule- and lexicon-based approaches

The earliest approaches to the task of named entity recognition were rule- and lexicon-based approaches.

The rule method is generally manually customized, and the rule template is constructed by experts with rich linguistic knowledge to develop general rules applicable to general utterances to improve the accuracy of recognition. The rules usually include the conformational structure of the named entity, the lexical characteristics of the context of the named entity, relevant statistical information, etc. The main method is string regular matching.

The dictionary method requires a dictionary of entities in advance, and the process of entity identification is to find out whether the entity exists in the constructed dictionary. For example, we can construct a dictionary of personal names and geographical names corresponding to personal names and geographical names; in determining whether "Chang'an" is a named entity, if "Chang'an" exists in the dictionary of geographical names, it will be attributed as a geographical entity.

There are pros and cons to using a rule and lexicon based approach to named entity recognition. Pros: The correct rate of systems constructed using this method is often high. Disadvantages: It requires the accumulation of expertise, the determination of complex rules, and the construction of lexicon; the imperfection of rules and lexicon can lead to low recall; the most important point is that the rules and lexicon developed are often only applicable to the old corpus, and when a new corpus appears, the rules and lexicon need to be reconstructed. Therefore, the method has the disadvantages of high accuracy, but requires a lot of labor, poor portability, and low error tolerance.

3.2.2. Statistical-based approach

With the limitations of rule-based and dictionary-based methods, researchers needed to open up new approaches. Statistical-based methods emerged afterwards. The main idea of this method is to construct statistical models for parameter fitting on manually annotated corpus. Thus, the problem of poor portability of previous models was solved.

The current mainstream statistical models are maximum entropy model, support vector machine, hidden Markov model (HMM) and conditional random field (CRF). Among them, the conditional random field is the most effective model for entity recognition tasks based on statistical models, which solves the long distance dependencies between sequences and complex contextual relationship features that cannot be handled by HMM models.

The idea of statistical modeling is mainly a serialized annotation method, where each word in the text is labeled with a particular tag, commonly known as BIO annotation method.

The BIO labeling method labels each element into three

categories, the beginning of an entity X is "B-X", the end of entity X is "I-X", and non-entities use "O". For example, Du Fu's poem "Li Bai's poems are a hundred pieces of poetry, and he sleeps in a restaurant in Chang'an City." In the poem, "Li Bai" is a person's name, so the entity "Li Bai" is labeled as "B-PER", "I-PER "; similarly, "Chang'an" is labeled as "B-LOC" and "I-LOC", and other non-entities are labeled with Other non-entities are labeled with "O". This basically determines the left and right boundaries of the named entity.

3.2.3. Deep learning based approach

Statistical-based methods have been able to achieve high accuracy and better portability, but there are still some problems. First, manual feature selection is not the optimal choice, so that important details may be lost in the text-to-feature step, often leading to a subsequent recognition model with high accuracy but not reaching a high overall performance; the most critical drawback is the need for input features, which takes a lot of time when performing complex feature engineering.

Deep learning has been increasingly researched in natural language processing in recent years, improving efficiency in various tasks in comparison to previous models. Deep learning methods mainly use word vector techniques to convert text into vector representations, which are subsequently trained using neural networks. The word vector model word2vec is proposed on the basis of One-Hot discrete representation of text, which can be used to obtain vectorized representations of words by context.

Subsequently, the most common network structures in text tasks are RNN, LSTM, BiLSTM, and BiLSTM+CRF structures. Using a bi-directional LSTM can solve the problem of gradient disappearance in one-way networks by keeping the important weight information in the memory unit and finally using the activation function for the output of recognition probabilities.

3.2.4. Ancient text named entity recognition model with BERT model

The current mainstream framework for named entity recognition is the word vector obtained from word2vec training plus the model of BiLSTM+CRF. In Chinese and even in ancient languages, the structure needs to be adapted.

First of all, Chinese datasets need to be word-separated compared to English, especially for ancient Chinese word-separation, it is often better to use character-level input or a combination of word-level and word-level input.

When extracting contextual features, using the Transformer architecture can improve the effectiveness of RNN-like models. A BERT pre-trained model with Transformer architecture can be used to mask some words in a given sentence and let the BERT model predict these words so that Transformer will refer to the contextual information when encoding a word. The output text undergoes word-level vector transformation, which mainly contains three aspects: position vector, text vector, and word vector, and then is passed into the bidirectional Transformer layer, which after learning can output the label corresponding to each word and identify the entities in the text.

At present, the pre-training model of BERT is widely used in natural language processing tasks, but there is still much room for improvement for ancient texts. In this paper, we choose to use BERT pre-training model to compare with BiLSTM+CRF model, and use statistical model CRF and bidirectional recurrent neural network model BiLSTM alone to analyze the most suitable named entity recognition model, and apply it to the same type of text to verify.

4. Research Content

4.1. Corpus acquisition and processing

There is no large collection of classical poetry literature on wine that has been compiled. The corpus in this study is mainly from the author's collection.

4.1.1. Original corpus collection

The main sources of the original corpus are websites and books. Book data mainly include books mentioned in the database of wine culture features, such as "History of Chinese Wine", "Chinese Geography of Wine", "Chinese Wine Dictionary", etc. Using ORC recognition and python language operations, the required textual information is extracted. Websites mainly include ancient wine related websites such as Ancient Poetry and Literature, Baijiu Says Chinese; using python language, the crawler is used to obtain information on ancient book names, authors, dynasties, texts and analysis and appreciation by professionals related to ancient wine from each website, and subsequently the book text information is integrated with the website text information and stored in a formatted form in the established database for the specialty area.

4.1.2. Data pre-processing

While processing the text data, the following problems were identified: (1) Duplication of canonical texts. Duplicate use of canonical texts found in the website and those recorded in books occurred, and the completely duplicated canonical texts were filtered and deleted by comparison, keeping as complete versions as possible. (2) Handling of rare characters and misspellings. There were many misidentified rare words in the ORC identification books, so we collected various materials for comparison and found the correct words; those that could not be found were deleted as appropriate. (3) Processing of shorter or longer sentences. Some verses or field interception content is incomplete, as far as possible to consult the data to complete, if it can not be found, it will be deleted to ensure that the entity recognition of the sentence for more than two sentences or length of more than 8 words; for longer sentences or paragraphs, according to the actual situation of the sentence, we will split it into several short sentences, so as not to affect the effect of recognition. (4) Retention of data. In order to make the annotation smooth and the experiment smooth, we retained the data of the following dimensions: document name, author name, dynasty, document content, and comment. After preliminary data processing, the preliminary established database of wine ancient books was obtained, which contains a total of 3251 poetry literature data, involving dynasties including Wei and Jin, Tang and Song, Yuan, Ming, and Qing dynasties.

4.2. Entity labeling system establishment

The definition of named entity identification presented in the Sixth Message Understanding Conference contains three major categories and seven subcategories of entities; the three major categories include named entities, temporal expressions and quantitative expressions, while named entities are divided into time, place and institution. To carry out the entity identification work, we have to carry out the establishment of the wine antiquity entity system, so we have to combine the definition of named entity identification with the target analysis and establish the entities for this study.

4.2.1. Entity Establishment

We decided to use the defined entity categories (person name, institution name, place name, time, date, currency and percentage, time category, number category) first, and then add or remove unnecessary entities through targeted analysis to improve the efficiency of identification.

(1) Firstly, for the textual information in this paper, the corpus used is Chinese literary texts, so some entities in English texts and modern Chinese texts can be excluded: currency, numerical categories, and percentages. Second, for this study, only dynasties and chronologies are retained for the time-class entities, while specific times, such as seasons and living times, are not included. Therefore, the date and time categories are combined into dynastic entities. Finally, the institution name entity was excluded because of the low frequency of institution names. Three valid entities were extracted: person name, place name, and dynasty name.

(2) In the contents of numerous poetic and ancient literature related to wine collected by the author, most of them mention aliases or types of wine, so the entity of wine name was added to this study. Since wine literature contains a large number of activities related to wine, such as rituals, songs and dances, etc., the entity of activity names was added.

(3) This gives us five types of entities: person name, place name, dynasty name, wine name, and activity name. After establishing the entities, you need to refine the naming rules for each entity.

Person's name: specific to the person's name, excluding official names, title names, nickname names, etc.

Geographical names: including provincial, city and county-level place names (ancient place names may be different at the local level), the word appearing in the county, county and other locations combined with the word marked together, not marked with other specific scene names such as field, marketplace and other specific place names.

The name of the dynasty: including the name of the dynasty and the alias of the dynasty, not marked with the specific time, such as "eight years of spring" in the eight years of the Dali spring is not marked.

Wine names: The following rules are defined according to the characteristics of wine names; the use of color and turbidity to define the name of the wine is not labeled, such as "fine wine" and other adjectives to describe the entity of wine is not labeled, the word wine appears together with the word wine label.

Event name: Include the name of the event related to wine.

4.2.2. Entity annotation and statistics

(1) Corpus proofreading

We performed an initial error correction of the text during pre-processing, and then proofread it again in the part of the entity to be annotated; by proofreading it twice, the error rate was greatly improved and a more correct version was retained.

(2) Markup rules and tool selection

The annotation involves three rounds, the first round first uses the annotations of each data in the database to annotate with the literal meaning of the text and documents the unclear entities; the second round allows the annotators to exchange annotations, conduct sampling tests and discuss the annotation of unclear entities, and get the final annotation results by searching for information in depth; the third round invites experts to check and control the annotated text and check the overall In the third round, experts were invited to check and control the annotated text, and to check and correct the whole text, and finally to obtain a corpus of ancient wine texts for this study.

The annotation tool used in this study is YEDDA, which is a lightweight collaborative text span annotation tool open-sourced on Github and developed using the tkinter package in python; it can be used to annotate entities or events on text (Chinese, English and almost any other languages), symbols and even emoticons, which is very effective for manual annotation of text; the operation is simple and easy to use. The user only needs to select the text span and use the shortcut keys, and the text span is automatically annotated; it also supports exporting the annotated text to sequence text, which greatly simplifies the operation of transferring formats; it can automatically generate ann files for model training; in the updated version, it is possible to use administrator status to evaluate and analyze the annotation quality among multiple annotators through a simple visual interface, and to generate statistical reports.

For the annotation of this study, YEDDA is an extremely suitable manual annotation tool. Firstly, we need to perform joint annotation of multiple people in 3.2.2, and the administrator status of this tool can perform a visual display of the joint annotation results to help optimize the annotation. Secondly, the annotation can modify the entity name by itself, and finally it will generate ann file which can be used directly in the model, saving a lot of manpower and time to study how to convert the annotation result format. Ultimately this study uses custom categories to customize the five entity categories selected in the previous section and exports them as ann files for use.

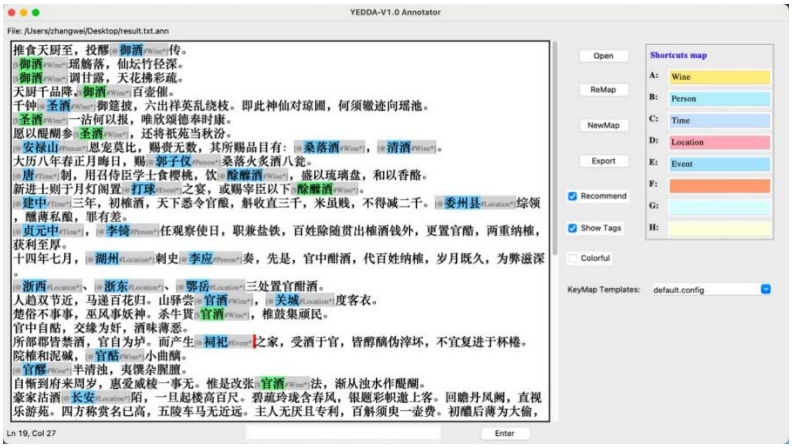


Figure 1. YEDDA tool interface

(3) Entity Statistics

After counting, 5 types of named entities are involved; these 5 types of entities are present in 2870 out of all 3251 sentences. Among them, there are a total of 382 types of wine names, which appear 1984 times; there are a total of 581 occurrences of human names; there are a total of 737 occurrences of place names; there are fewer types of events and dynasties, which appear relatively more often, 503 and 489 times respectively. Some representative high-frequency entities are listed in Table 1.

Table 1. High-frequency entities for wine names, dynasties, and events

Wine Name	Location	Events
官酒	长安	祭祀
烧酒	洞庭	酿酒
清酒	巴蜀	宴乐
黄酒	江南	歌舞
官醪	蓬莱	上贡
浊醪	南山	客宴

4.3. Environment construction and parameter setting

In this paper, the model uses the two main categories mentioned before: the most common CRF model in statistical learning and the deep learning model for comparison experiments, the named entity recognition experiments use a MacOS with 16GB of memory, the programming language uses python, and the cross-validation approach is adopted to divide the corpus into training and test sets in a 9:1 ratio for experiments.

4.3.1. Experiments on statistical-based machine learning models

The statistical-based machine learning model experiments are conducted using the CRF+0.72 tool, and the model training and testing are performed according to the established data features and the established feature templates. If we want to get better model training effect, we have to choose a suitable feature template. In the process of named entity recognition of this corpus, the word length feature is mainly involved, and the entity length of this corpus is mainly concentrated in 1 and 2. Based on this feature, the feature window of the CRF model in this paper is chosen to be 3, and added to the feature template for model training and testing.

4.3.2. Deep learning model experiments

(1) BiLSTM

The BiLSTM model for this experiment is built using the open source Torch framework from Facebook. The model parameters before the experiment are set as follows: the dimension of the training word vector is set to 300, the dropout value is set to 0.2 to prevent the model from overfitting; the batch_size of the training set is set to 64, the dimension of the forward and backward hidden states is 300; the model uses a learning rate of 0.01 and is optimized using the Adam optimizer to reduce the model loss; the number of iterations is set to 100 and save the results of each iteration.

(2) BiLSTM-CRF

The BiLSTM-CRF model for this experiment uses the same Torch framework as the above BiLSTM model with the same parameter settings, and finally the CRF layer is used to obtain iterative results by statistical knowledge for the final classification.

(3) BERT

This experiment uses Google's pre-trained model BERT-Base-chinese, which has 12-layer, 768-hidden, 12-heads, and 110M parameters. parameters). The hyperparameters are set as follows: the maximum sequence length is 256, the batch_size is 8, the dropout value is set to 0.1, the learning rate of the model is set to 0.01, and the number of training is set to 500, and the training is stopped when the model is not improved for 10 consecutive times.

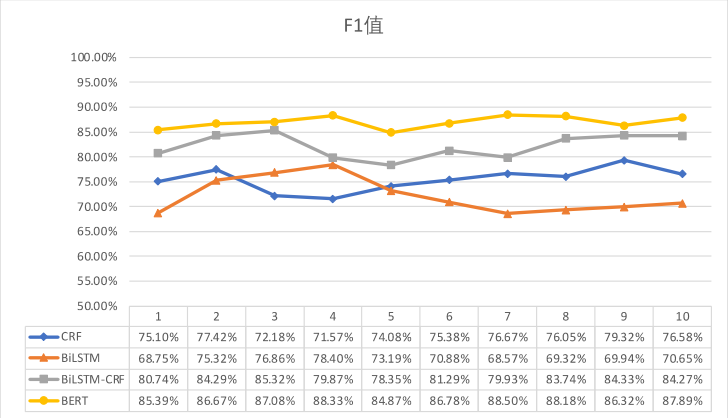
4.4. Evaluation indexes and experimental results

4.4.1. Evaluation indicators

This study uses the classical validation methods in the field of named entity recognition: accuracy, precision, recall, and F1 values. Accuracy A refers to the proportion of all correctly identified samples to the total samples; precision P is the proportion of correctly identified samples to the identified entity samples; recall R refers to the proportion of correctly identified samples to the actual entity samples; F1 value refers to a function of precision P and recall R. F1 value is equal to 2 times the value of P multiplied by the value of R and divided by the sum of P plus R. It takes into account the contradiction between P and R The F1 value is the weighted summed average of the precision and recall rates.

4.4.2. Experimental results

This experiment uses four models for comparison experiments: CRF, BiLSTM, BiLSTM+CRF, and BERT four models, and the number of experiments epoch is taken 10 times. The F1 value line graph in Figure 2 is obtained statistically.



	1	2	3	4	5	6	7	8	9	10
CRF	75.10%	77.42%	72.18%	71.57%	74.08%	75.38%	76.67%	76.05%	79.32%	76.58%
BiLSTM	68.75%	75.32%	76.86%	78.40%	73.19%	70.88%	68.57%	69.32%	69.94%	70.65%
BiLSTM-CRF	80.74%	84.29%	85.32%	79.87%	78.35%	81.29%	79.93%	83.74%	84.33%	84.27%
BERT	85.39%	86.67%	87.08%	88.33%	84.87%	86.78%	88.50%	88.18%	86.32%	87.89%

Figure 2. Results of 4 models to identify the F1

That is, the results of this experiment show that for the named entity recognition of ancient wine and ancient poetry texts, the BERT-based model is the best application, surpassing the BiLSTM and CRF models.

5. Summary

5.1. Problem Analysis

In this experiment, the highest F1 value using the BERT model was 88.33%, which did not exceed 90%. After conducting a specific analysis of each step, the following reasons for the identification errors were found.

5.1.1. Total sample size is too small

Generally the number of samples used for machine learning and deep learning should be as many as possible so that the features of the text can be learned, while the data in this study is small; on the one hand, because similar text data is difficult to collect, and on the other hand, manual labeling requires a lot of manpower and time. A small number of samples will make the recognition accuracy lower.

5.1.2. Difficulty in recognizing single words

Due to the difference between ancient Chinese and modern Chinese, the characters in ancient Chinese have the characteristics of refinement, which makes most entities are less than two words, and there will be a large number of single-word entities, such as the word "mash" in the name of wine is also labeled as a kind of wine name, or some people's names are shortened to one word, which will also affect the training effect and reduce the accuracy. This will also affect the training effect and reduce the accuracy of recognition.

5.1.3. Low frequency word recognition error

There are some entities that appear particularly few times, and entities that appear with low frequency also affect the recognition effect, resulting in the machine not being able to recognize this low frequency entity.

5.1.4. Manual labeling errors

The use of cross-validation reduces the manual annotation errors, but the manual annotation errors are still unavoidable.

5.2. Exploration of improvement

For the reason that the accuracy rate of recognition of 5 types of entities is not high enough, improvements can be made in the following directions.

(1) Expanding the corpus. Expanding the size of the corpus can increase the frequency of entity occurrences, make model training more effective, learn more text features, and thus improve the effectiveness of recognition.

(2) Adding split-word features. The problem of difficult single-word recognition can be optimized by adding split-word features to the model, which can distinguish word words from multi-word words and thus improve the recall rate of the model.

(3) Optimize the manual annotation process. For custom samples, the manual labeling results greatly affect the training results, and more standard labeling rules need to be developed, and setting a more rigorous labeling process can lead to improved training results.

(4) Using new training models. After the BERT model researchers have proposed more new models, whose recognition effects on different corpus are yet to be tested, and new models can be used for training in subsequent studies, and the perspective of entity relations can also be added for entity recognition exploration.

5.3. Outlook

Named entity recognition is gradually developing, and the current research on recognition in ancient Chinese is just beginning. Research on different corpora can help the development of named entity recognition technology, and the mining of texts such as ancient wine and ancient poems can also promote the deepening of the association between natural language processing technology and ancient Chinese corpus, opening up new areas of research, so as to achieve, for example, the discovery of ancient culture by computer technology. In this study, we take ancient wine and ancient poems as corpus, establish rules for their named entity recognition annotation, use cross-validation method, use four models, based on CRF model in statistical machine learning based method, combined with LSTM network in deep learning, use BERT pre-training model, and make comparison to get the advantages and promotion value of BERT model in ancient text entity recognition. At present, new variants of the BERT model also appear in the subsequent research, for its recognition accuracy on different corpus still needs to be studied, and the subsequent hope is to explore the application performance of the new model on different corpus through more model comparisons, using more ancient text data for verification.

References

[1] Chen Yu Hou. Visual analysis of knowledge map of Sichuan wine culture research[J]. Journal of Sichuan Institute of Technology (Social Science Edition),2017,32(06):10-25.

[2] Li Na. Construction of automatic extraction model for aliases of ancient books of Fangzhi based on conditional random fields[J]. Journal of Chinese Information,2018,32(11):41-48+61.

[3] Zhao SH, Luo R, Cai ZP. A review of Chinese named entity recognition[J]. Computer Science and Exploration,2022,16(02):296-304.

[4] Sui Chen. Research on Chinese named entity recognition based on deep learning [D]. Zhejiang University,2017.

[5] Li, J., Wang, P.. A review of Chinese named entity recognition research methods[J]. Computer Age,2021(04):18-21.DOI:10.16644/j.cnki.cn33-1094/tp.2021.04.005.

[6] Huang Shuiqing, Wang Dongbo, He Lin. Research on the construction of automatic recognition model of ancient Chinese place names based on pre-Qin corpus[J]. Library Intelligence Work,2015,59(12):135-140.DOI:10.13266/j.issn.0252-3116.2015.12.020.

[7] Wang D.B., Gao R.Q., Shen S., Li B. Research on automatic identification of basic entity components of historical events for pre-Qin canonical texts[J]. National Library Journal,2018,27(01):65-77.DOI:10.13666/j.cnki.jnlc.2018.01.009.

[8] Zhang Fan,Wang Min. Medical named entity recognition based on deep learning[J]. Computing Technology and Automation,2017,36(01):123-127.

[9] Xie Tao. Research and implementation of named entity recognition based on ancient literature [D]. Beijing University of Posts and Telecommunications,2018.

[10] He, Chun-Hui, Wang, Meng-Xian, He, Xiao-Bo. Named entity identification in diabetes domain based on two-layer Bi-

LSTM-CRF model[J]. Journal of Shaoyang College (Natural Science Edition),2020,17(01):21-26.

[11] Cui Jingfeng,Zheng Dejun,Wang Dongbo,Li Tingting. Named entity recognition of chrysanthemum classical poems based on deep learning model[J]. Intelligence Theory and Practice,2020,43(11):150-155.DOI:10.16353/j.cnki.1000-7490.2020.11.024.

[12] Jinhyuk Lee,Wonjin Yoon,Sungdong Kim,Donghyeon Kim,Sunkyu Kim,Chan Ho So,Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining.[J]. CoRR,2019,abs/1901.08746.

[13] Naseem U, Musial K, Eklund P, et al. Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding [C]//2020 International joint conference on neural networks (IJCNN). IEEE, 2020: 1-8.

[14] Sun Huan. Traffic accident text analysis based on BERT+BiLSTM+CRF model and improved Apriori algorithm [D]. Chang'an University, 2021. DOI:10.26976/d.cnki.gchau.2021.000506.