Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2552331, 12 pages
https://doi.org/10.1155/2022/2552331

Research Article

Research on Application of Intelligent Corpus Annotation of Entity Extraction with Construction of Knowledge Graph

Xingli Liu,[1] **Junjie Fan**,[1] **and Haiqun Ma**[2]

[1] *School of Computer Science and Technology, Heilongjiang University of Science and Technology, Harbin 150020, Heilongjiang, China*
[2] *School of Information Management, Heilongjiang University, Harbin 150080, Heilongjiang, China*

Correspondence should be addressed to Haiqun Ma; mahaiqun@hlju.edu.cn

Received 10 August 2022; Accepted 8 September 2022; Published 11 October 2022

Academic Editor: Zaoli Yang

The purpose of this paper is to solve the problem of big data and small samples caused by the high manual annotation cost of a military corpus. The deep learning algorithm of entity extraction in the military field was organically combined with the method of bootstrapping loop iteration to complete a study on the application of intelligent corpus annotation of military field entities. With the experimental research showing that using a small number of military field entity corpus annotations for RoBERTa pretraining word vectors and BiLSTM-CRF models and based on the bootstrapping algorithm idea to complete 3 rounds of loop iterations and 10 rounds of cross-validation joint-voting model iterations, the best entity extraction model evaluation F value reached up to 91.5%. Finally, the 60M intelligent corpus annotation application testing was completed using the best model of iteration of this round, with a total of 178,177 sentences of military field corpus intelligently labeled, the number of entities that should be labeled reaching 417,734. Therefore, this is an efficient way of construction and evaluation of intelligent corpus annotation model in the military entity extraction field. The findings of this paper provide an effective way of how to complete the labeled corpus. The research serves as a first step for future research, for example, the construction of knowledge graphs and military intelligent Q&A.

1. Introduction

A corpus is a large-scale electronic text database that has gone through scientific sampling and processing. It is a cross-discipline combining linguistics, statistics, computer science, and other disciplines. The earliest definition of a corpus can trace back to 1982, when Professor Francis from Brown University believed that a corpus is a collection of texts used for language analysis, representative of a certain language, dialect, or a certain aspect of a language [1]. With the development and in-depth research of big data and artificial intelligence technology, corpus research has received more and more attention and focus. From the initial corpus collection for linguistic research to the current deeply labeled knowledge resources that support knowledge mining and discovery, corpora and related research have been fully explored in both depth and breadth [2]. Corpus annotation strategies are generally formulated according to the characteristics of corpus and the value of the content to be labeled. After a raw corpus is labeled, it is called a labeled corpus that can be further studied and used, based on which it can be seen that corpus annotation according to the established annotation specification is the core part of corpus construction [3], making the choice of a suitable annotation strategy particularly important. At present, there are three mainstream corpus annotation modes [4]: the first mode is annotation by experts in the field. This annotation mode is suitable for corpus annotation in professional fields, which can ensure annotation quality, but involves high annotation cost and a long cycle; the second mode is crowdsourcing-based annotation. This annotation mode can label corpus in large quantities at a low cost, but it is limited to simple annotation tasks, and the annotation process also needs to be carefully designed to ensure annotation quality; the third mode is group annotation. The corpus construction process under this annotation mode is similar to the construction of information retrieval evaluation set, with high-quality corpus able to be constructed without relying on experts, but it has high requirements for annotation groups. A military corpus is a labeled corpus consisting of monolingual or multilingual texts, whose content involves military (or military service) [5].

However, with the advent of the information age, the demand for intelligent information services from military corpora has increased. Although it is relatively easy to obtain big data-level military open-source literature information data, especially for annotation of special corpora of military scientific and technological intelligence, the dilemma of big data and small samples is still caused due to the high cost of manual annotation in the field. So the construction of an effective ecological corpus annotation method is particularly critical to the quality assurance of the construction of the field corpus.

Therefore, the purpose of this paper is to study the intelligent corpus annotation based on the research idea of organically combining the deep learning algorithm of entity extraction in the military field with bootstrapping loop iteration and K-fold joint voting-based evaluation of the iterative entity extraction model, with an intelligent corpus annotation strategy method of entity extraction and evaluation in the military field based on a loop iteration idea innovatively proposed, the research on the application of intelligent annotation of military entity corpus completed. This research provides a massive military corpus construction method for downstream intelligent information services based on the military field.

Although this paper only focuses on intelligent corpus annotation of the Chinese entity extraction, its methods will be applicable in other languages. The findings of this paper provide an effective way of how to intelligently complete the labeled corpus. The research serves as a first step for future research, for example, the construction of knowledge graphs and military intelligent Q&A.

Section 2 surveys this paper's relevant previous works. Section 3 introduces this paper's theoretical foundations; the research method is constructed in detail; and in Section 4, the experiment analysis and the application statistics of the model based on loop iteration and prototype are completed. Section 5 discusses the important findings and limitations of this paper. Finally, the conclusions, research and practical implications, and direction of future research are discussed in Section 6.

2. Relevant Research

However, driven by the rapid development of big data and artificial intelligence in the past 30 years, the research centering on corpus construction and application is entering a new stage, and data annotation is the focus and difficulty of current corpus research. At present, data annotation strategies mainly involve manual annotation, machine-based annotation, and human-machine combination-based annotation, of which the research results of manual annotation methods include manual standards regularized by for the named entities and relationships of pediatric diseases, with consistency testing performed; in addition, researchers have improved annotation efficiency and organized and managed corpus conveniently using some existing annotation tools and software. For example, for the public security alert information corpus, the annotation tools are used to conduct multiple rounds of annotations of entities [6]; researchers are constantly exploring and trying to use computers to aid or even replace manual annotation, reduce labor cost, and improve annotation efficiency [2]. With the improvement of natural language processing and computer performance, a human-machine combination mode is mostly preferred to annotate corpus. For example, for the entity extraction of the TCM clinical case corpus, [7] proposed the use of the conditional random field named entity recognition method to interpret the automatic batch corpus annotation method; it proposed an automatic annotation method of text corpus for unexpected incidents, with the desired results achieved [8]; it put forward a design idea of a single-loop iterative intelligent corpus annotation system method, which can fully realize automatic annotation, with accuracy rate up to standard, theoretically after multiple rounds of iteration and optimization and expansion [9]. In recent years, research on named entity recognition methods based on statistical machine learning and deep neural networks has become one of the focuses of attention in the domestic academic and industrial circles. The CRF, BiLSTM, and extended neural network models have become the mainstream framework for this task [10, 11]. Pretraining models based on a large-scale unlabeled corpus, such as Google's BERT model [12] and text sequence annotation pattern recognition based on a high-quality labeled corpus, have become one of the important trends of this problem [13, 14] based on multiple rounds of manual annotation and consistency check, and have used deep learning named entity recognition to complete the construction of a corpus in the highway and bridge inspection field. With the increasing demand for informatization in the military field, research on text information extraction in the military field for intelligent information services has attracted the attention of some scholars. For example, [15] formulated a set of unified military term part-of-speech annotation specifications and military term corpus annotation specifications based on military term dictionaries and designed an automatically expanded military corpus entity feature extraction framework based on military term dictionaries. At the same time, relevant military data resources such as military fields and military services and arms have attracted attention. The research, construction, and application of military field corpora have just started in most countries, but western developed countries led by the United States have carried out the research and application of military corpora for a long time [16]. Many projects of the Defense Advanced Research Projects Agency, DARPA, run natural language technology on the basis of guaranteed corpus resources, including automatic translation, cross-language information detection, information extraction, and specific time tracking and search.

It is found through the above analysis that traditional supervised learning methods are based on the data drive, and involve strict requirements for the scale and quality of labeled training data. However, due to the particularity of the military field [17], the quantity of labeled corpus is not large, and there are few studies on training a high-quality military entity extraction method and intelligently constructing a corpus on the basis of a small number of annotations. The innovative contribution of this research lies in the following:

(1) An intelligent corpus annotation method based on bootstrapping loop iteration strategies for small samples in the military field was proposed: research on the intelligent corpus annotation strategy method was realized based on the organic combination of military entity extraction algorithm and K-fold joint-voting model's evaluation of iteration. The findings of this paper provide an effective way how to complete the labeled corpus.

(2) Intelligent corpus annotation system and application research of military entity extraction was completed, based on the best model selection obtained from the comparative experiment on the typical deep learning entity extraction algorithm; it is RoBERTa-BiLSTM-CRF and its loop iteration. The research serves as a first step for future research, for example, the construction of knowledge graphs and military intelligent Q&A.

3. Research Design and Method

This section makes a statement from three aspects, i.e., the design of intelligent corpus annotation strategies for military entities, the big data collection of corpora, and the manual annotation and quality evaluation of small sample entity corpus, of which the design of intelligent corpus annotation strategies for military entities is the overall plan design and key technology of this research, which can be called the basic framework of this research, while the big data collection of corpora is the cornerstone of annotation corpora, and the manual annotation and quality evaluation of small sample entity corpus are the beginning of the intelligent method of this research, the three together providing a necessary method guarantee for the intelligent corpus annotation experimental process.

3.1. Design of Intelligent Corpus Annotation Strategies for Military Entities

3.1.1. Overall Design Ideas. Due to the high cost of high-quality annotation, only a small amount of annotation data can be used for model training. Therefore, this paper uses RoBERTa-WWM-BiLSTM-CRF as the basic model, with bootstrapping algorithm ideas introduced to solve the intelligent annotation of entity corpus, with the joint-voting model of K-fold cross-validation used to evaluate the algorithm confidence index. On the one hand, the scale of the training set can be increased, and on the other hand, the expanded training set can also be used to iteratively train a named entity recognition model with strong generalization ability. The overall design of the research method is shown in Figure 1.

The specific steps of the intelligent annotation method of military entities based on loop iteration are as follows:

(1) Preannotation of small sample corpus: load the military domain dictionary and complete the preannotation;

(2) Manual small sample corpus annotation: complete the manually labeled corpus m_1 set according to the manual annotation system and specifications as the initial training corpus;

(3) Model training A: compare different entity extraction algorithms and select the best model A after evaluation as the best initial model B_1 for loop iteration;

(4) Loop iteration (B_1~B_n) and K-fold verification: apply the iteration idea and the model evaluation of K-fold joint voting to obtain the best intelligent annotation model C;

(5) Corpus intelligent annotation: use the best annotation model C iterated in this round of loop and intelligently tag the raw corpus in the military corpus to obtain the intelligently labeled corpus m_2;

(6) Enter the next round of model iteration: use the two parts $m_1 +\ m_2$ as the annotation training set of the second round and repeat steps (3)~(5) until corpus M (m_1+m_2,. . .,+m_n) is fully tagged.

3.1.2. Entity Extraction Algorithm: RoBERTa-WWM-BiLSTM-CRF. The main structure diagram of the RoBERTa-WWM-BiLSTM-CRF model is in this paper; this model is mainly divided into 3 parts, namely, the presentation layer, the sequence coding layer, and the prediction decoding layer. The presentation layer uses a large-scale pretraining language model RoBERTa-WWM to replace the traditional random initialization or word vector form extraction character-level representation, with each word mapped to a word vector through the embedding layer, followed by the use of the bidirectional transformer [18] structure for encoding based on the comprehensive context information, with the learned knowledge added to the token representation, the semantic information at the character level obtained, and then, the obtained word vector input into the BiLSTM layer for sequence encoding. The BiLSTM layer can combine the context information to extract high-dimensional features; finally, decoding of semantic information is conducted in the conditional random field (CRF) to eliminate wrong tags and predict the true tag sequence. This paper involves improvement of the classic model BiLSTM-CRF and introduction of the RoBERTa-WWM model. Through the masked learning of characters, the grammatical and semantic-level information between character contexts can be captured, and the semantic representation ability of character-level vectors can be enhanced.

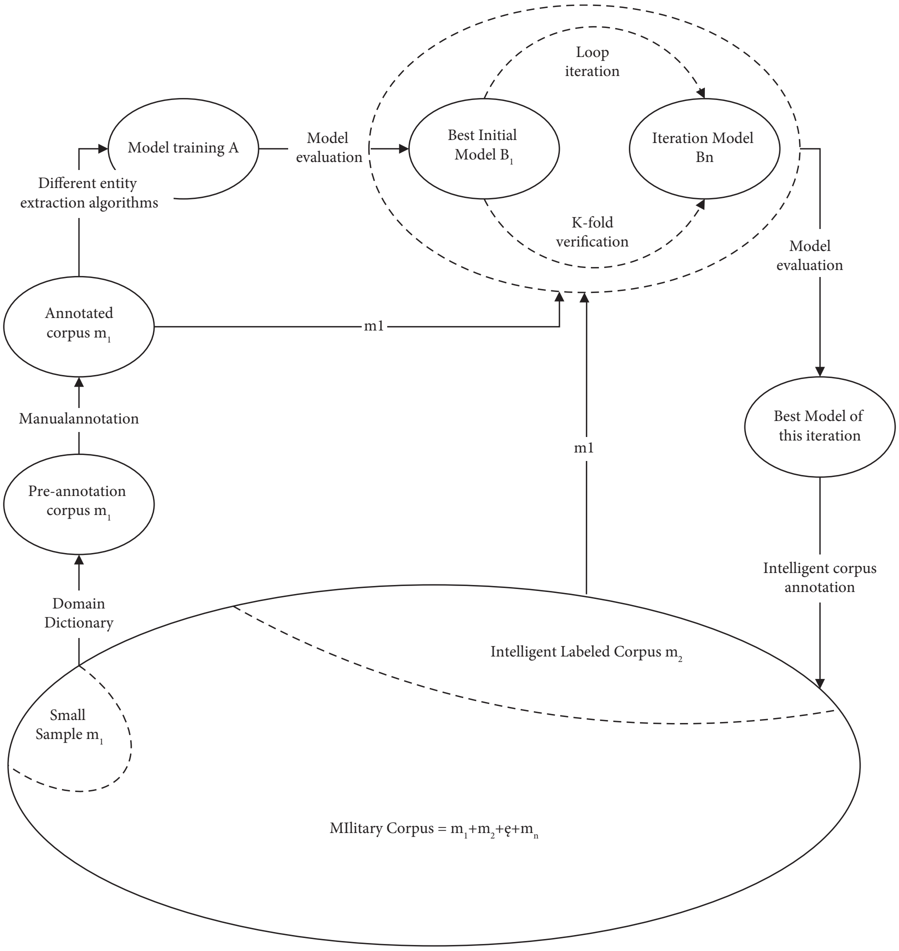

Figure 1: Overall design of the intelligent annotation method of military entities based on loop iteration.

(1) RoBERTa Layer. BERT belongs to a kind of pretraining language model and the whole of it is an autoencoder LM, which uses large-scale corpus to perform unsupervised pretraining for MLM [12] (masked language model) and NSP (next sentence prediction) tasks, while RoBERTa is a robust optimization-based BERT improvement scheme.

The RoBERTa model architecture has inherited the advantages of BERT and adopted a multilayer transformer structure. Each transformer has introduced a multihead attention mechanism. At the same time, RoBERTa has improved the pretraining task of BERT, eliminated the next sentence prediction (NSP) task that has little effect in BERT, and improved the efficiency of the model, with full sentences and doc sentences as input, supporting a maximum of 512 characters. In addition, it replaces the static masking mechanism in BERT with a dynamic masking mechanism. The original BERT masking mechanism is processed only once at the data preprocessing stage, with a certain percentage (15%) of words in the input sequence randomly masked, then most of these words (80%) covered with tags, a small part (10%) randomly replaced with words in the vocabulary, and the remaining part (10%) remained unchanged, followed by prediction of the masked words based on the context. RoBERTa has improved the masking mechanism of BERT, with dynamic masking adopted, masks dynamically generated while providing input to the model, a new mask pattern generated every time a sequence is input, and mask mark changes during training provided. In this way, in the process of continuously inputting a large amount of data, the model will gradually adapt to different masking strategies and learn different language representations, which can improve the semantic representation effects of downstream tasks.

On the other hand, BERT's text encoding mode uses character-level BPE vocabulary with a size of 30K, while RoBERTa uses byte-level BPE (byte pair encoding) vocabulary, including 50K subword units, with no any additional preprocessing of input, able to deal with common vocabulary in various NLP tasks.

The RoBERTa-WWM model adopts a whole word masking strategy (WWM) [18], which has adjusted the sample generation strategy at the pretraining stage for Chinese texts. Unlike BERT character granularity-based masking mode, word granularity masking has been adopted and this word-level masking mode can help improve the effect of Chinese named entity recognition tasks.

The word embedding layer, segment embedding layer, and position embedding layer in the figure above are all static word embedding layers, with the embedding matrix responsible for performing index-based table lookup. For the i-th token in the processed token sequence, its word vector is expressed as the following formula, where W_{token} is the token embedding matrix, W_{segment} is the segment embedding matrix, and W_{position} is the position embedding matrix.

$$e_i = W_{\text{token}}(t_i) + W_{\text{segment}}(s_i) + W_{\text{position}}(i). \tag{1}$$

First, the token embedding matrix is used to map the input to a 768-dimensional word vector; then, the segment embedding matrix is used to encode the sentence where the token is located. As this named entity recognition task can be uniformly deemed as a single-sentence input, the segment ID of the sentence where the input token is located is the same; then, the position of each token is encoded through the position embedding matrix to provide the position information of characters and solve the problem of the transformer network itself not having the ability to capture time-series information due to the attention mechanism, so the position embedding matrix is used to describe the features of data in terms of time series. The final word vector e_i is expressed as the sum of the above three 768-dimensional embedding vectors.

(2) BiLSTM Layer. The traditional feedforward neural network cannot process time-series data, while the recurrent neural network (RNN) uses loops to make data continuously circulate, remember past data, and update to the latest data. Among them, LSTM (long short-term memory) is a special recurrent neural network, which realizes the memory function in time through a gating mechanism to prevent the gradient disappearance and explosion existing in the RNN network after the length of the text sequence increases [19], and can solve the long-distance dependence problem.

The LSTM receives and transmits data through three gating units, i.e., the input gate, the output gate, and the forget gate [20]. Among them, the input gate is used to judge the value of new information, and conduct selection and weighted addition of such information; the output gate is used to learn when to let the information out of the storage unit; the forget gate is used to control the information from the storage unit of the last moment that can enter the storage unit at the next moment. In the hidden unit of LSTM, the forget gate is first calculated, with input being the information of the current unit and the output of the previous unit, with the calculation process as shown in the formula below [19], where σ is the activation function sigmoid, W is the weight matrix, and b is the bias.

$$f_t = \sigma\left(W_f * \left[h_{t-1}, x_t\right] + b_f\right). \tag{2}$$

Through the formula above, the output is mapped to between 0 and 1 to selectively allow the value of cell state C_{t-1} to pass. If there is an element in f_t that is 0, the corresponding element in C_{t-1} cannot pass, and the purpose to selectively forget information can be achieved.

In the LSTM unit, the next step is to calculate the input gate, with input being the information of the current unit and the output of the previous unit, the update value i_t determined through the sigmoid activation function, and the temporary cell state obtained using the tanh activation function, and then updated to obtain state, to achieve a round of update of information, the calculation formula is as follows [21]:

$$\begin{aligned} i_t &= \sigma\left(W_i * \left[h_{t-1}, x_t\right] + b_i\right), \\ \widetilde{C}_t &= \tanh\left(W_c * \left[h_{t-1}, x_t\right] + b_c\right), \\ C_t &= f_t * C_{t-1} + i_t * \widetilde{C}_t. \end{aligned} \tag{3}$$

Subsequently, the output gate is calculated in the LSTM unit, with input being the information of the current unit and the output of the previous unit, the calculation process similar to that of the forget gate, the state of the output cell obtained through the sigmoid activation function, and the final hidden state obtained by being multiplied by the neuron state processed by the tanh function, with the calculation formula as follows [21]:

$$\begin{aligned} o_t &= \sigma\left(W_o * \left[h_{t-1}, x_t\right] + b_o\right), \\ h_t &= o_t * \tanh\left(C_t\right). \end{aligned} \tag{4}$$

It can be seen from the structure and calculation process of the LSTM unit above that the unidirectional LSTM network can only use the forward information in the text sequence and cannot process the backward information. However, in the texts of the Chinese military news field, all contextual information is mutually related. Therefore, this paper uses a bidirectional BiLSTM network to combine the forward and backward LSTM networks, with the models of the contextual information of texts separately constructed through two independent LSTM networks, thereby obtaining the semantic input of texts with global characteristics.

(3) CRF Layer. Since the named entity recognition task can be regarded as a sequence annotation problem, there is also a sequence problem between tags. For example, I-PLA cannot independently appear, and it is sure to follow B-PLA; I-NAT and I-CIT cannot be adjacent. The hidden state output of the BiLSTM layer has only the characteristics of context information and does not involve the intricate dependence between different tags, which causes the output tag sequence information to be misplaced. Therefore, this paper uses conditional random field (CRF) [22] to obtain the best global sequence and improve the accuracy of prediction [23].

The CRF is an undirected graph model, which is an improvement of maximum entropy (MaxEnt) models and hidden Markov model (HMM) [24], and is a discriminative conditional probability distribution model. It has removed the conditional independence assumption of HMM and overcome the label bias problem of the generative oriented graph model. When used in named entity recognition tasks, it can select the optimal label sequence by adding constraints [25].

The input of the CRF layer is the output sequence x_t of the BiLSTM layer. Given a set of observation sequences, if the label of this set of observation sequences is obtained as y_t, the following calculation formula with the conditional probability being $p(y|x)$ can be obtained, where $Z(x)$ is the sum of all possible values for y as an normalization item, f_k is a characteristic function, and W_k is the weight of the characteristic function.

$$
\begin{aligned}
P(y|x) &= \frac{1}{Z(x)} \exp \sum_{k=1}^{k} W_k f_k(y, x), \\
Z(x) &= \sum_{y} \exp \sum_{k}^{K} W_k f_k(y, x).
\end{aligned}
\tag{5}
$$

Then, the CRF model is trained on the given dataset to determine the weight of the obtained feature function, followed by the use of the determined model to solve the sequence probability problem; in the final prediction, the Viterbi algorithm is used to search for the optimal path to obtain the final tag prediction sequence of military entity information.

3.1.3. Loop Iteration and Evaluation: Joint Voting Based on Bootstrapping Algorithm and K-Fold Cross-Validation.

Bootstrapping [26] is a classic algorithm of semisupervised learning (SSL) [27] also known as the self-learning algorithm. The snowball system is a sequence annotation system [10] designed with this method as the theory. The specific idea of the algorithm is as follows: training is started with a small amount of labeled data, and the dataset is continuously expanded through multiple rounds of iterations. It requires the correctness of the initial classifier classification and the reasonableness of the confidence calculation; otherwise, it is easy to cause the accumulation of errors and mistakes in the iteration and lead to the phenomenon of semantic drift.

The K-fold cross-validation is generally used to evaluate the performance of machine learning models, and it can also be used to make model selection. It refers to dividing the original data into K groups, making each subset data separately as a sequential verification set, the remaining $K-1$ group subsets as the training sets, and then obtaining K models. The final verification set accuracy rates of these K models are averaged to obtain the performance index under this classifier. It is generally used to reduce the variance of models. It is expected that the model will perform well on multiple sub-datasets than on a single dataset. For the model $F(\hat{f}, \theta)$, the unbiased estimator CV of the prediction error is obtained so as to select an optimal θ to minimize CV, where N is the total number of training samples, K is the number of subsets, and the size of each subset is represented by $m = N/K$.

$$
CV = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{m} \sum_{i=1}^{m} \left(\hat{f}^{k} - y_i \right)^2 . \tag{6}
$$

The algorithm flowchart used in the experiment in this paper is as shown in Algorithm1. The confidence is implicitly expressed through the voting consistency of multiple models. At the same time, because the dataset of the model is based on K-fold cutting, it can not only avoid the high classification consistency probability of the trained classifier due to bootstrapping's sampling with replacement, but also ensure the accuracy of the model based on the cross-validation errors and test errors calculated through cross-validation. Finally, through label obtained by the unanimous voting of the joint model, the addition of some labels with consistent but inappropriate prediction results is avoided by setting the threshold Th.

3.2. Corpus Data Collection in the Military Field.

This research takes common sense knowledge of high-value targets in the intelligence information of military strategic reconnaissance as the research object, and recommended by experts in the field, data from relevant open-source intelligence information sources were collected. First of all, Python's BS4 was used to obtain and collect more than 50 intelligence information sources like open and authoritative military websites and public accounts to get an open-source military news corpus, with such sources including 12 public accounts like National Defense Science and Technology News, Military. China.Com, Global Military Magazine, and Military High-Tech Online and 40 news and encyclopedia websites including Weapons Encyclopedia, China Military's Authoritative Military News Website, ānd National Defense Science and Technology News—NetEase Subscription; the obtained military texts are preprocessed and used as research corpus, involving a total of 4,297,777 sentences, 644 MB, with 10 MB first selected for manual corpus annotation to extract the features of military entities.

3.3. Manual Military Corpus Annotation and Quality Evaluation

3.3.1. Entity Annotation System and Method.

On the basis of fully considering the needs of the military field concerned, referring to the distribution, type division, and observation range of various types of military targets, the named entity recognition in the corpus is defined into 15 categories, including nations (NAT), aircraft (PLA), regions (REG), arms (ARM), states (STA), cities (CIT), ports (POR), airports (AIP), coastlines (LIN), sea areas (WAT), islands (ISL), and other fine-grained military entities.

In addition, in the process of annotation of the military text corpus, the principle of "no overlapping, no nesting, and no pause punctuation" that is similar to that for literature should be followed. Since nonstandard Chinese word segmentation may cause error transmission problems for subsequent named entity recognition tasks, a BIO three-segment annotation method is used in the sentence-by-sentence annotation process of military corpus [28], in which the starting word for each entity is marked as "B-entity type," the subsequent mark is "I-entity type," and O means nonentity part. Therefore, there are a total of 31 types of tags in the labeled corpus, as shown in Table 1.

3.3.2. Process of Manually Labeled Entity Corpus.

In this research, a group annotation and domain expert combination-based annotation model was selected to complete the

```
Input: manual annotation training set $A$, a large number of preprocessed unlabeled corpus $B$, unlabeled data $B'$ extracted in each round, prediction result probability threshold Th, cross-validation fold number $K$.
Process:
(1) Divide training set $A$ into training set $A_1$ and test set $S$ based on the $9:1$ ratio;
(2) Divide training set $A_1$ into $d_1$, $d_2$, $d_3$,...,$d_K$ by means of K-fold cross-validation, with 1 copy reserved as a test set each time, the remaining $K$-1 copies used as training sets, with a model obtained through training $R_1$, $R_2$, $R_3$,...,$R_K$;
(3) Calculate the test errors $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$,...,$\varepsilon_K$ on the model $R_1$, $R_2$, $R_3$,...,$R_K$, respectively, and get the cross-validation error $\varepsilon = 1/K \sum_{i=1}^{K} \varepsilon_K$ after averaging, perform the final quality evaluation test on the test set S to obtain test errors $\varepsilon'$, and ensure the cross-validation error $\varepsilon$ and test error $\varepsilon'$ are smaller than the threshold by repeating $n$ rounds;
(4) Randomly extract a small part of data $B'$ from a large number of unlabeled corpus $B$;
(5) Integrate $R_1$, $R_2$, $R_3$,...,$R_K$ submodels into a joint-voting model with the same weight and tag the randomly selected data $B'$ by means of voting. When the consistency rate of the pseudolabel prediction results of $K$ models is greater than 80%, the tags are regarded as positive tags, with the labeling results added to the training set $A$ and deleted from $B$, the iterative learning of this round completed.
(6) Repeat steps (1)-(5) above and training is completed until the unlabeled set $B$ is empty or the maximum number of iterations $N$ is reached.
Output: the continuously expanding training set $A'$ and the named entity recognition model $R$ of joint voting
```

Algorithm 1: Loop iteration entity extraction model of bootstrapping and K-fold joint voting.

Table 1: The example of entity annotation in military corpora.

Word	Label	Word	Label
F	B-PLA	For	O
—	I-PLA	Rui	B-NAT
3	I-PLA	Shi	I-NAT
5	I-PLA	Air	O
Characteristics	O	Force	O
Advantage	O		

manually tagged part of the annotation model. A corpus annotation team consisting of 6 computer masters and 6 undergraduates in the cognitive intelligence laboratory spent a total of two months on the whole military field open-source data obtaining initial manual annotation and evaluation.

In order to improve the annotation efficiency, in this research, an annotation platform was established based on the text structured collaborative annotation tool (brat rapid annotation tool, https://brat.nlplab.org) of the B/S architecture and it supports part-of-speech annotation, named entity recognition and syntactic analysis, and other tasks and also supports such features as multiperson collaboration-based entities, relationships, attributes, and incident annotation and good drag-and-drop operation experience [29].

The corpus annotation research completion involves the following four stages.

Stage I is the formulation of the standards and methods for the annotation system. The characteristics of the collected corpus were analyzed in detail first, annotation specifications in the military field were referred to, the classification system of named entities and relations in this study was established, and the corpus annotation rules and methods were determined.

Stage II is the the preprocessing of text data. Semiautomatic data preprocessing was carried out by means of custom rules and manual review, with special characters such as HTML tags and zero-width spaces in the corpus removed through analysis of the characteristics of the collected corpus, the Unicode Chinese codes and English letters retained, traditional Chinese characters converted to simplified Chinese characters, and repeated values processed and missing fields supplemented to obtain noise-free pure text corpus.

Stage III is the preannotation based on the rough corpus. According to some encyclopedia-type websites such as Baidu Encyclopedia in the corpus collection process, Hanlp natural language processing tool was adopted to initially construct a military domain entity dictionary and carry out preliminary preannotation to minimize duplication of labor.

Stage IV is the manual annotation based on preannotated corpus. Two independent rounds of annotation were carried out with the AB grouping method. At the same time, the annotation rules were revised according to the feedback from the annotators, and the manually labeled corpus evaluation was conducted based on inter-annotator agreement (IAA).

3.3.3. Statistics of Artificial Entity Annotated Corpora.

The first artificial entity annotated corpus involves a total of 2,779,277 characters, with 60,657 annotated entities, the statistics of entities of various kinds as shown in Table 2. In the statistics of annotated entity corpus, countries, aircraft, and bases come top and ports come bottom in the order of entities.

3.3.4. Quality Evaluation of Manually Annotated Corpus.

The inter-annotator agreement (IAA) refers to the degree of agreement between two independent annotators [30]. There has been extensive research on the reliability of manually annotated datasets based on IAA indicator evaluation. Normally, in the research of entity manual annotation corpora, the F value is used to calculate IAA [31]. The specific practice is to use the final entity annotation result B as the standard answer, calculate the accuracy (P) and recall rate

Table 2: The entity distribution of manually annotated corpora of entities in the military field.

Entity type	Number of entities (PCs)	Entity type	Number of entities (PCs)
NAT	34483	SAT	684
PLA	7961	WAT	994
VES	4083	REG	621
MIS	2423	RAD	408
CIT	1978	ISL	290
ARM	1702	AIP	121
BAS	4706	LIN	109
POR	94	Total	60657

(R) of the first annotation result A, and then calculate the F value, with the calculation formulas as shown in equations (1)~(3).

$$P = \frac{\text{number of consistent } A \text{ and } B}{\text{total of } B},$$

$$R = \frac{\text{number of consistent } A \text{ and } B}{\text{total of } A}, \tag{7}$$

$$F = \frac{2 * P * R}{P + R}.$$

In the case of determining entity consistency, only when the entity text, entity type label, and start and end are the same, the entity annotations can be considered as consistent [2]. According to statistics, the consistency of corpus entities constructed in this paper is labeled as 89.8%, and this manually annotated corpus is reliable [30].

4. Experimental Results and Analysis

This section states from the three aspects, i.e., experimental environment and parameter setting, experimental evaluation indicators, and experimental results and analysis of the experimental design part of the intelligent corpus annotation strategies for entity extraction in the military field.

4.1. Experimental Environment and Parameter Setting. The experimental environment selected in this research is as shown in Table 3.

The hyperparameter setting used in this research is as shown in Table 4.

4.2. Experimental Evaluation Indicators. In this paper, the BIO annotation mode was adopted and the most commonly used evaluation indicators in the field of named entity recognition, including precision rate P (precision), recall rate R (recall), and $F1$ (F-measure) values, were used to judge the accuracy of entity recognition, with the specific formulas, where TP represents samples with positive actual prediction, FP represents samples that are negative but predicted to be positive, and FN represents samples that are positive but predicted to be negative.

Table 3: The experimental environment.

Operating system	Ubuntu
CPU	Intel Xeon E5-2678 v3
GPU	NVIDIA GeForce RTX 2080 Ti
Python	3.6.0
TensorFlow	2.2.0
RAM	62G
Video memory	11G
Hard disk	200G

Table 4: The hyperparameter setting.

Parameter name	Parameter values
Batch_size	128
Seq_max_len	256
Dropout	0.4
Learning rate	$8e-3$
GRU unit	128
LSTM unit	128
Epoch	5
Optimizer	RAdam
kernel_size	5
Random embedding size	300
Word2vec embedding size	300
Word2vec window	5
Word2vec iter	5

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

$$R = \frac{TP}{TP + FN}, \tag{8}$$

$$F1 = \frac{2PR}{P + R}.$$

P, in relation to the prediction result, refers to the proportion of the number of samples that are positive and predicted to be positive in the total number of the samples that are predicted to be positive; R refers to the number of samples that are positive and predicted to be positive in all annotated samples, i.e., how many have been recalled from the perspective of annotation; $F1$ value is the evaluation indicator of comprehensive accuracy rate and recall rate and is the harmonic average of the recall rate and the accuracy rate.

4.3. Experimental Results and Analysis

4.3.1. Entity Extraction Model Results and Evaluation Analysis. For testing the effect of sample corpus, the 1.5 M annotated data which have been manually checked are divided into the training set, the validation set, and the testing set in a 8 : 1 : 1 ratio, with the effects of different word vectors and different downstream structure models on the military field entity recognition tasks, compared with the experimental results as shown in Table 5.

First, as can be seen from the table above, the experimental analysis of different randomly initialized word vectors is as follows:

(1) Using the word2vec mode, the accuracy rate increased by 5.5%, and the $F1$ value increased by 4%;

(2) After the BERT pretraining language model was introduced as a word vector, compared with the word2vec mode, the accuracy rate increased by 6.6%, the recall rate increased by 6.9%, and the $F1$ value increased by 6.2% overall, indicating that the pretraining language model obtains prior semantic information from large-scale unlabeled corpus, greatly improving the effect of downstream tasks;

(3) However, for Chinese corpus, due to the fact that there are no separators between Chinese words, BERT can only provide semantic information at the character level. After the use of the pretraining language model of RoBERTa-WWM, word-level masking was conducted at the MLM stage of its pretraining task, with word-level semantic information provided. Therefore, it can be seen that the RoBERTa-WWM model achieved better experimental results than BERT. The P value increased by 1.6%, the R value increased by 0.2%, and the $F1$ value increased by 0.9% on the whole.

Second, as can be seen from the table above, the experimental analysis of different downstream models is as follows:

(1) Compared with RoBERTa-WWM-CRF, the P, R, and F values of RoBERTa-WWM-LSTM-CRF model increased by 3.4%, 2.2%, and 2.9%, respectively, indicating that the addition of a variant recurrent neural network with a gating mechanism made it possible to capture better sequence relationships and make a significant improvement in sequence feature extraction tasks;

(2) After the use of mutually independent bidirectional networks, the P, R, and F values of the experimental model increased by 7.7%, 3.2%, and 6.5%, respectively, indicating that in the military field named entity recognition tasks, the contextual information is also very important.

(3) Compared with the RoBERTa-WWM-CRF model, the P, R, and F values of the RoBERTa-WWM-BiGRU-CRF model increased by 6.4%, 6.8%, and 6.6%, respectively, indicating that the addition of gate RNN made it possible to capture sequence relationships and make a significant improvement in sequence feature extraction tasks. As GRU is a simplified version of LSTM [24], it has fewer parameters and a simpler gating mechanism. Therefore, it can be seen that compared with the RoBERTa-WWM-BiGRU-CRF model, the P value and $F1$ value of this experimental model increased by 6% and 3.6%, respectively.

Based on the analysis of the above experimental results, the best model combination of RoBERTa-WWM-BiLSTM-CRF was selected as the base model to enter the iterative loop model training.

4.3.2. Iteration and Evaluation Analysis of the Best Entity Extraction Model. In this paper, let $K = 10$, and the experiment was made with the 10-fold cross-validation mode. After the best hyperparameters were determined and the model was fixed, iteration was conducted by constructing a joint-voting model based on bootstrapping algorithm and 10-fold cross-validation. The results of the joint model after three rounds of iterations are as shown in Table 6. The P, R, and F values of the model in each round are the data of the joint-voting model on the testing set S.

As shown in the table above, base is the joint-voting model obtained from the best hyperparameters under 10-fold cross-validation. The first, second, and third rounds are iterative effects. It can be seen that the best results were achieved in the second round, and the values of P, R, and $F1$ in the third round were all slightly reduced, with the specific causes analyzed as follows.

In the second round of model evaluation, the best effect was achieved because of the method proposed in this paper. First, cross-validation was used for the evaluation of the initial model, avoiding the problem of some data being unused that was caused during the division of the initial training set. Second, the accuracy of the model was ensured through the model's cross-validation error and testing error on the testing set S. Next, the expansion of the quality of the training set was guaranteed by utilizing voting for the implicit representation of the confidence and through the prediction result probability threshold Th, indicating the idea of making full use of unlabeled data was effective and that the noise introduced by iteration was offset by the large quantity of benefits brought by labeled data.

In the third round of iteration, the $F1$ value slightly decreased, showing a situation of falling instead of rising. Although the recall rate increased by 1.3% compared to the second round of iteration, the accuracy rate decreased by 2.3%, indicating the related noise generated by the use of unlabeled data in this round was greater than the effectiveness of iterative labeled data, so the model needs to be used with caution. Returning to the second-round model to continue iterating can be considered or the data with lower confidence judged by the joint-voting model can be

Table 5: The evaluation of different deep learning entity extraction algorithms.

Experiment content	Model	*P*	*R*	*F*1
Different word vectors	BiLSTM-CRF	0.810	0.806	0.807
	Word2vec-BiLSTM-CRF	0.819	0.829	0.823
	Bert-BiLSTM-CRF	0.873	0.898	0.885
Different downstream models	RoBERTa-WWM-CRF	0.778	0.846	0.810
	RoBERTa-WWM-LSTM-CRF	0.812	0.868	0.839
	RoBERTa-WWM-BiGRU-CRF	0.829	0.904	0.865
This experimental model	RoBERTa-WWM-BiLSTM-CRF	0.889	0.900	0.894

Table 6: The evaluation of the iterative joint model.

Model	*P*	*R*	*F*1
Base	0.889	0.900	0.894
1	0.901	0.915	0.908
2	0.909	0.921	0.915
3	0.886	0.934	0.909

Table 7: The entity distribution of intelligent annotated corpora of entities in the military field.

Type of entities	Number of entities (PCs)	Type of entities	Number of entities (PCs)
NAT	331471	SAT	2449
PLA	30106	WAT	8329
VES	6366	REG	2169
MIS	3864	ARM	2117
CIT	7073	BAS	23790
Total of intelligently annotated entities	417734		

extracted for manual annotation and correction, and at the same time, iteration can be terminated, with the second-round iteration model with the best performance selected to extract named entities.

It is found through experiments that the method used in this paper has effectively improved the effects of the named entity recognition model for military tasks in the small sample scenario of domain tasks after multiple rounds of iterations, with each round of iteration increasing the F value by about 1%, which proves the effectiveness of the iteration method in named entity recognition tasks in the military field.

4.3.3. Intelligent Corpus Annotation Results. Based on the existing experimental conditions, in order to verify the application effects of the intelligent corpus annotation strategy, in this research, 60 MB in the corpus was chosen to test the intelligent annotation system for corpus in the applied military corpus, with a total of 178,177-sentence physical corpus containing research tasks in the military field finally completed, with the numbers of characters and entities being 10,502,307 and 417,734, respectively, with the statistics of labeled entities of various kinds as shown in Table 7.

5. Discussion

The goal of this paper was to provide a general method of how to complete the corpus annotation intelligently in the field of the military for entity abstract in Chinese. The method has a better effect and can improve the precision and recall rate. Different from previous studies [3, 9, 15], on one hand, this paper integrates the pretraining language model RoBERTa-WWM, which preprocesses the data in the way of a whole word masking strategy, so it is more suitable for entity recognition of military news text in Chinese; on the other hand, this paper solves the contradiction between large data and small samples caused by the high annotation cost in the deep learning algorithm of entity extraction by bootstrapping loop iteration strategies; and it not only includes the design method and experiment but also application statistics and prototype.

This paper has some limitations. First, the effect of entity recognition still needs to be improved in the labeled small samples. It can be combined with the external multiple data augmentation method. The next research is to combine the effective data augmentation with loop iteration in named entity recognition. It is significant in the field of the sparse corpus or difficult annotation.

Second, the accuracy in the training set is further improved as a seed entity corpus. Although this paper's iterated training set also has better accuracy, it is manually corrected and checked, as stated in Section 4.3. The next step is to try to fuse the feature enhancement of the domain knowledge in the pretraining model. Finally, the optimization of the whole process of intelligent corpus annotation is acquired.

Third, the corpus of this paper only focuses on automatic annotation of military news entity extraction. In general, relational intelligent annotation is also of great research value. Therefore, the next step is to expand the research on automatic annotation of corpus in breadth, including not only the static information corpus tagging of entities and relationships, but also the automatic tagging research of dynamic event corpus.

6. Conclusions

This paper proposes research on the application of an intelligent corpus annotation strategy for military entity extraction. In this research, based on the idea of bootstrapping semisupervised learning, a loop iteration-based annotation model was constructed, with a military entity extraction model with the best RoBERTa-BiLSTM-CRF evaluation indicators selected and an iterative annotation model completed, 10-fold joint-voting evaluation conducted, and annotation model with the best indicators obtained, with the application research results showing that the iterative model obtained with this method can complete intelligent annotation of military entity corpus involving a large data volume. It effectively solves the problem between subjectivity and high cost of manual annotation and is the first step to provide corpus guarantee for the construction of knowledge maps in military domain and downstream intelligent Q & A and other intelligence services.

Future research should proceed in two directions. First, a comprehensive method in depthly intelligent corpus annotation should be thought to improve the precision of corpus annotation. Second, the scope of intelligent corpus annotation should be broadened to event corpus annotation and relationship corpus annotation.

Data Availability

The data supporting the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This article is an outcome of the major project "Research on Public Data Open Utilization System and Capacity Building Orienting Digital Development" (No. 21&ZD336) supported by the National Social Science Foundation of China and "Research on the Institutional Innovation of National Intelligence Work under the Overall National Security View" (No. 20ATQ004) supported by the National Social Science Foundation of China, and this project is undertaken by Professor Mahaiqun.

References

[1] W. N. Francis, "Problems of assembling and computerizing large corpora," *Computer Corpora in English Language Research. Ed. Stig Johansson. Bergen*, pp. 7–24, Norwegian Computing Centre for the Humanities, 1982.

[2] S. Huang and D. Wang, "Review of corpus research in China," *Journal of Information Recording Materials*, vol. 11, no. 3, pp. 4–17, 2021.

[3] H. Zan, T. Liu, C. Niu, Y. Zhao, K. Zhang, and Z. Sui, "Construction and application of named entity and entity relations corpus for pediatric diseases," *Journal of Chinese Information Processing*, vol. 34, no. 5, pp. 19–26, 2020.

[4] F. Xia and M. Yetisgen-Yildiz, "Clinical corpus annotation: challenges and strategies," in *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in Conjunction with the International Conference on Language Resources and Evaluation (LREC)*, vol. 67, Istanbul, Turkey, 2012.

[5] L. Ma and X. Wang, "The building and application of a parallel corpus of military texts in military. Translation," *National Defense Science & Technology*, vol. 30, no. 1, pp. 38–41, 2009.

[6] R. Cao and W. Du, "Construction of the corpus oriented to entity annotation in public security area," no. 3, p. 5, Telecommunications information, 2021.

[7] L. Feng, *Automatic Approaches to Develop Large-Scale TCM Electronic Medical Record Corpus for. Named Entity Recognition Tasks*, Beijing Jiaotong University, 2015.

[8] W. Liu, X. Wang, Y. Zhang, and Z. Liu, "An Automatic-Annotation Method for Emergency Text Corpus," *International Journal of Speech Technology*, vol. 3, no. 2, pp. 76–85, 2017.

[9] Y. Liu and X. Lu, *A Cyclic and Iterative Intelligent Corpus Annotation System*, Guangdong. Communication Technology, no. 2, pp. 76–79, 2021.

[10] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence annotation," *arXiv preprint*, 2015, https://arxiv.1508.01991.

[11] X. Zhenyu, S. Jiang, L. Zhang, and R. Bao, "Research on name recognition of film critics. Based on multi-feature Bi-LSTM-CRF," *Journal of Chinese Information Processing*, vol. 33, no. 3, pp. 94–101, 2019.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional. Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[13] L. Liu and D. Wang, "A review on named entity recognition," *Journal of the China Society for. Scientific and Technical Information*, vol. 37, no. 3, pp. 329–340, 2018.

[14] T. Mo, L. Ren, J. Yang, L. Tong, S. Jiang, and L. Dong, "Construction of named entity. Recognition corpora in the field of regular inspection of highways and bridges," *Journal of Computer Applications*, vol. 40, no. S1, pp. 103–108, 2020.

[15] B. Zhou, H. Zhang, R. Zhang, Y. Feng, and Y. Xu, "Construction of military corpus for entity annotation," *Computer Science*, vol. 46, no. 6, pp. 540–519, 2019.

[16] H. Wang and M. Zhou, "Research on the Practicality of Naval Military Science and Technology. English from the Perspective of Internationalization," vol. S1, pp. 1103-1104, China Extracurricular Education (Issued Every Ten Days), 2014.

[17] L. Zhang, J. Li, L. Tang, and M. Yi, "Deep learning recognition method for target entity in military field. Based on pre-trained BERT," *Journal of Information Engineering University*, vol. 22, no. 3, pp. 331–337, 2021.

[18] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.

[19] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic. speed prediction using remote microwave sensor data," vol. 54, pp. 187–197, Transportation Research Part C Emerging Technologies, 2015.

[20] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook. of brain theory and neural networks*, vol. 3361, no. 10, 1995.

[21] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: a search space. Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[22] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: probabilistic models for. segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289, Morgan Kaufmann Publishers, USA, June 2001.

[23] M. Ma, Q. Yang, and T.·T. Askar·Hamdulla, "Chinese named entity classification. Based on word vector and conditional

random field," *Computer Engineering and Design*, vol. 41, no. 9, pp. 2515–2522, 2020.

[24] K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," arXiv preprint arXiv: 1406.1078, 2014.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector," *Space*, arXiv preprint arXiv:1301.3781, 2013.

[26] N. Abe and H. Mamitsuka, "Query Learning Strategies Using Boosting and Bagging," in *Proceedings of the Fifteenth International Conference On Machine Learning*, pp. 1–9, DBLP, 1998.

[27] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "*D*istant supervision for relation extraction without labeled data," in *Proceedings of the Joint, Conference Of the 47th Annual Meeting Of the ACL and the 4th International Joint Conference On Natural Language Processing Of the AFNLP*, pp. 1003–1011, 2009.

[28] J. Su, B. He, H. Wu et al., "Annotation. Scheme and corpus construction for cardiovascular diseases risk factors from Chinese electronic medical records," *Acta Automatica Sinica*, vol. 45, no. 2, pp. 420–426, 2019.

[29] H. He, *Introduction to Natural Language Processing*, Vol. 21, Posts &Telecom Press, Beijing, 2019.

[30] G. Hripcsak and A. S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, 2005.

[31] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.