Research on Entity Recognition in Aerospace Engine Fields Based on Conditional Random Fields

Wenchao Gao¹*, Shanshan Zhao², Shiyi Zhu³, Shengbo Ren⁴
¹,²,³,⁴China University of Mining and Technology-Beijing, University, Beijing, Beijing, 100089, China
*Corresponding author’s e-mail: gaowc@cumtb.edu.cn

Abstract. With Entity recognition is an important basic tool for many natural language processing tasks such as information extraction, question answering systems, and syntactic analysis. Entity recognition is divided into general field and specific field. There were different recognition methods for different fields. At present, one of the more common entity recognition technologies is a rule-based method, and the other is a statistical method. Due to data sets scarcity in the aerospace engine field, there is a lack of corresponding entity recognition research in the aerospace engine field. This paper proposes a statistical-based entity recognition method to identify entities in a specific field based on conditional random fields. This method constructs a training set in the aerospace engine field through manual annotation, extracts entities in the aerospace engine field, and finally makes the accuracy rate, Recall rate and F value reach 92.7%, 87.1%, 89.8%.

1. Introduction

Entity recognition is the basic task of natural language processing and an important basic tool for many natural language processing tasks. The accuracy of entity recognition determines the effect of downstream tasks. Named entity recognition refers to recognizing entities with specific meanings in text. Entity recognition has two tasks, one is the identification of entity boundaries, and the other is the determination of entity types. The early entity recognition was mainly to identify entities such as person names, place names, organization names and time.

Due to the development of computer technology, only identifying people's names and other entities can no longer meet the task requirements, so researchers began to study domain entity recognition. Domain-specific entity recognition is to extract entities and categories that can describe the domain terms according to the characteristics of the domain data set. Domain-specific entity recognition is the basis for deep complex information extraction tasks, and has important applications in document classification, domain information retrieval and domain dictionary construction. At present, many fields have realized domain entity recognition. For example, in the medical field, medical entities such as disease names, imaging examinations, and diagnosis results are recognized from bioclinical medicine. Medical entity identification plays an important role in disease research, medical diagnosis, medical knowledge retrieval, etc. In the financial field, unknown financial entities are identified from the provided financial text, including financial platform names, enterprise names, project names, and products name. Banks, securities, insurance, funds and other institutions holding financial licenses, as
well as well-known Internet enterprises such as Tencent, Taobao, Jingdong, can better help financial work and related financial knowledge retrieval.

However, entity recognition itself is more difficult than English recognition. Chinese entities have no obvious signs and are interfered by word segmentation, while English word segmentation is not required. But, due to the various types and terms of entities in specific fields, boundary recognition is more difficult, which in turn increases the difficulty of entity recognition. At the same time, the recognition of named entities based on machine learning relies heavily on data annotations, and the lack of training data of annotation in specific domain also increases the difficulty of field entity recognition.

The complexity of domain entity recognition is mainly manifested in the following aspects:

1. Domain entity recognition lacks annotated data sets.
2. Due to the complexity and diversity of entity types, the combination of domain words does not have certain rules, which leads to the difficulty of domain words recognition.
3. The composition structure of the entity is relatively complex, and the length of the entity is not limited. There will also be multiple entity nesting, entity abbreviations, mixed Chinese and English, and other situations, which will bring difficulties to entity recognition.
4. The boundary between domain entity recognition and general entity recognition is not obvious. Domain entities include general entities, and it is difficult to label.

The main technical methods of domain entity recognition and general entity recognition are consistent. It is mainly divided into rules and dictionary-based methods and statistics-based methods.

Rule-based entity recognition methods mostly use linguistic knowledge and the lexical and syntactic characteristics of entities in the professional domain entities to construct a rule templates, and then match the corpus in the text with the rule templates. The rules and dictionary-based method is the earliest of the entity recognition method, as long as there are related templates, it is relatively simple to implement. However, depending on the quality of the rule templates, the transplant ability is poor and the cost is high.

Statistics-based methods mainly include Hidden Markov Model (HMM), Maximum Entropy (ME)[2], Support Vector Machine (SVM), Conditional Random Fields (CRFs)[3] and other methods based on statistics and Machine learning. The statistics-based method does not require the construction of rule templates and the participation of experts, and the method is more flexible. And compared with HMM, CRFs can achieve global optimization. Compared with ME, CRFs solves the problem of label bias. Although CRFs has these advantages, it still has the disadvantages of high training cost and high complexity. Therefore, it is necessary to select appropriate features to improve recognition performance and reduce or avoid the use of inefficient features[4].

Due to the lack of data sets in the engine field, there is also the problem of domain entity recognition. Therefore, this paper uses a conditional random fields model based on statistical methods, uses the CRF++0.58 toolkit to recognize entities in the aerospace engine field, and uses features such as words and word frequency to train and learn to realize the entity recognition of data in the aerospace engine field and obtain a high accuracy rate.

2. Conditional Random Fields

Conditional random fields (CRFs) was proposed by Lafferty[5] et al. in 2001. It is an undirected probability graph model, which combines the characteristics of hidden Markov model and maximum entropy model. CRFs is a special case of Markov random field, which solves the problem of tag bias caused by hidden Markov model. In addition, features of context can be taken into account to make better feature selection. CRFs is to calculate the conditional probability distribution density of another set of output random variables given a set of input random variables. A general CRFs model is shown in Figure 1.
In the above model, the observation sequence is represented by \(x_1, x_2, \ldots, x_T\), and the implied state sequence is represented by \(y_1, y_2, \ldots, y_T\), so the figure above represents the observation value \(x_i\) and \(y_{i-1}, y_i, y_{i+1}\) are related.

The design of the CRFs model mainly resolves the setting of parameters and the selection of features, and the model focuses on solving three problems:

1. **Probability problem**: Conditional probability is calculated on the premise of given model parameters, observed values and state values.
2. **Prediction problem**: Given model parameters and observations, calculate the state value of the maximum conditional probability.
3. **Training problem**: Given training data, the model parameters are trained with the goal of maximizing conditional probability.

CRFs can be applied in many fields. Yang Yan et al.\([6]\) proposed an image classification and recognition method based on CRFs in the field of image, which realized accurate classification and recognition of images. In the field of natural language processing, Che J et al.\([7]\) proposed an Chinese word segmentation method based on a CRFs, which realized automatic word segmentation of Chinese. Kaixin Liu et al.\([8]\) proposed the named entity recognition method of Traditional Chinese Medicine (TCM) clinical medical case symptoms based on CRFs, which realized the entity recognition of TCM clinical complex symptoms. C. Janarish Saju et al.\([9]\) proposed an entity recognition based on CRFs in bank big data, which realized entity recognition in a new field.

Entity recognition in engine field based on CRFs is proposed in this paper, which realizes the recognition of specialized words in engine field. Entity recognition in engine field is mainly to realize the prediction problem of CRFs. In other words, set the random field \(p(x|y)\) of the linear chain component, Under the condition that the input observation sequence is \(x\), calculate the conditional probability of the value of the output sequence, and the annotation of the observation sequence is finally obtained, which has the form of formulas (1) and (2).

\[
p(y|x) = \frac{1}{z(x)} \exp \left( \sum_{ik} \lambda_k t_k (y_{i-1}, y_i, x, i) + \sum_{il} u_l s_l (y_j, x, i) \right)
\]

among them:

\[
z(x) = \sum_y \exp \left( \sum_{ik} \lambda_k t_k + \sum_{il} u_l s_l (y_j, x, i) \right)
\]

(1)

(2)

t_k and s_l are eigenfunctions also called transition feature and state feature, \(\lambda_k\) and \(u_l\) are the corresponding weight and the normalization factor. And the sum \(z(x)\) is done over all possible output sequences.
The prediction algorithm of CRFs is Viterbi algorithm. It can be obtained from Equation (3).

\[ y^* = \arg \max_y p_y(y \mid x) \]

\[ = \arg \max_y \frac{\exp(w^* F(y, x))}{z_y(x)} \]

\[ = \arg \max_y \exp(w^* F(y, x)) \]

\[ = \arg \max_y w^* F(y, x) \tag{3} \]

According to Equation (3), the prediction problem of CRFs is the problem of maximizing the probability of demoralization to obtain the optimal sequence marker. The maximum unnormalized probability decomposition calculation can be obtained by formulas (4) (5).

\[ \max_y w^* F(y, x) \tag{4} \]

Among them:

\[ w = (w_1, w_2, \ldots, w_k)^T \]

\[ F(y, x) = (f_1(y, x), f_2(y, x), \ldots, f_k(y, x)) \]

\[ f_k(y, x) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, x, i), k = 1, 2, \ldots \tag{5} \]

Due to the wide application of CRFs, many CRFs tools have appeared. The experiment of this article also selected the CRF++ toolkit. The current common CRFs versions are CRF++0.53, CRF++0.58, this article selects the CRF++0.58 to identify the engine field entities. CRF++ is a well-known open source tool for CRFs, written in C++ language. Its most important function is the use of feature templates. So we can automatically generate a set of feature functions, instead of generating our own feature functions, what we have to do is to find features, such as part of speech.

Choose appropriate parameters and design a feature template according to your needs to train and test the data. There are four main parameters that can be adjusted:

- a CRFs-L2 or CRFs-L1

Normalized algorithm selection. The default is CRFs-L2. Generally speaking, the effect of the L2 algorithm is slightly better than that of the L1 algorithm, although the value of the non-zero features in the L1 algorithm is substantially smaller than that in the L2 algorithm.

- c float

This parameter sets the hyper-parameter of CRFs. The higher the value of C, the higher the degree of CRFs fitting training data. This parameter can adjust the balance between overfitting and unfitting. It also can be found by cross-validation and other methods to better parameters.

- f NUM

This parameter sets the cut-off threshold of the feature. CRF++ uses features that appear at least NUM times in the training data. The default value is 1. This option is useful when using CRF++ to large scale data, where there may be millions of features that occur only once.

- p NUM

If the computer has multiple CPUs, then the training speed can be increased through multithreading. NUM is the number of threads.

3. Entity recognition in the aerospace engine field

3.1. Entity in the aerospace engine field

Using the CRFs of machine learning to identify entities in the engine field is supervised learning. It requires a certain scale of labeled training data. However, the engine field lacks domain data corpus we
need to build a certain scale of engine training corpus. This article is based on Jieba word segmentation tools segmented the initial text and annotated part of speech, and then manually annotated the training data set and corrected the results of Jieba word segmentation. This article uses the book "Encyclopedia of World Missiles and aerospace Engines" to study and analyze the engine-related entities in it. The provided data set is 1.74 MB, 70% is used for training corpus, 30% is used for test corpus.

We formulate certain labeling standards for the obtained training data of aerospace engines, and then obtain training data through manual labeling based on Jieba word segmentation. The marking standards are as follows:

1. Words that directly describe or indicate an engine, such as "rocket engine", "aspirated engine".
2. Words representing the structure of aerospace engine systems and related aerospace terms, such as: "space vehicle", "propulsion system", "check valve". Although these words do not denote aerospace engines, they also describe aerospace-related vocabulary and are marked as entities in the training set of this article.
3. The identification of entities should be as accurate and complete as possible, such as: "third-level engine".
4. The aerospace engine vocabulary composed of a mixture of Chinese and English is labeled as a vocabulary. Such as: "YF-40 Liquid Rocket Engine".
5. With national aerospace engine vocabulary, we should mark the country and engine words as a vocabulary. Such as: "China Long March 4A (LM-4A) carrier rocket".

Through the analysis of the aerospace engine data, it is found that the aerospace engine entity has the following characteristics:

1. The entity type is highly specialized, and entities in the aerospace engine field rarely appear in other fields.
2. There is no clear definition of the boundary of related entities, and it is easy to blur the scope of the boundary when marking entities.
3. The length of the entity is uncertain, ranging from two characters to more than a dozen characters. The part of speech composed of words does not have a certain rule, and there will be a mixture of Chinese and English, and unified characteristics cannot be found according to the context.
4. The term nesting phenomenon exists in the aerospace engine entities.
5. Data sparseness exists in entities of the aerospace engine field, and many aerospace entities have only appeared once or twice in the text.

3.2. aerospace engine entity recognition based on CRF++

The entity recognition of aerospace engines based on CRFs, using the CRF++ toolkit for experiments, the main process of using CRF++ for entity recognition in the aerospace engine field is:

1. Collect experimental corpus and perform preprocessing operations such as word segmentation and part-of-speech tagging on the corpus.
2. Select appropriate parameters and feature training model.
3. Model establishment and prediction.
4. Analysis results.

The training and test corpus requirements of CRF++ must follow a certain format, as shown in Table 1.

| Table 1. CRF++ Corpus Format |
|-----------------------------|
| Missile                     | n  | 2  | B    |
| and                         | p  | 1  | M    |
| Aerospace                   | n  | 2  | M    |
| Engine                      | n  | 3  | E    |
| Cover                       | v  | 2  | O    |
From Table 1, each text contains multiple tokens, and each line represents a token. A token contains multiple columns of data, which are displayed in each column. The first column is the word itself, the last column is the word label, and the middle is the selected feature.

The specific process of aerospace engine entity recognition is shown in Figure 2.

3.2.1. Pre-treatment
Preprocessing the data after collecting the corpus used in the experiment is the foundation of the model building. The pure text in the book "Encyclopedia of World Missiles and aerospace Engines" is extracted, and operations of removing spaces and empty lines are performed as preliminary preprocessing. Then the text obtains the initial word segmentation through the Jieba word segmentation tool, and obtains the word itself, part of speech, and word length after the word segmentation. After obtaining the basic features of the words, the text is labeled with lexical positions by manual methods. The lemma tagging is to obtain the boundary information of each word. The current general lemma tagging methods mainly include 2-tag tagging method, 4-tag tagging method and 6-tag tagging method.

The tag set of 2-tag is \{B,E\}, B represents the first word of an entity, and E represents the middle or end of the entity. B and E form an entity. 4-tag is more complex than 2-tag, but the recognition of the result will be more accurate. 4-tag tag set is \{B,M,E,S\}, B represents the first word of the entity words, and M represents the middle of the entity Part, E represents the tail of the entity, S represents a separate entity. 6-tag implements a more detailed entity tagging than 4-tag. The label set of 6-tag is \{B,M,E,W,O\}, B represents the first word of the entity words, M represents the middle part of the entity, E represents the tail of the entity, and W represents a separate entity in the field. In the aerospace engine field expressed as a separate engine word into an entity, O represents other components of the sentence except the engine entity. The more complex the annotation set is, the training speed of CRF model will be affected of the CRFs model, but it can also improve the result of entity recognition.

This paper uses the 6-tag label set to label entities. The definition of the 6-tag label set is shown in Table 2.
Table 2. The definition of 6-tag label set

| Label | Definition |
|-------|------------|
| B     | B represents the first word of the entity words |
| M     | M represents the middle part of the entity |
| E     | E represents the tail of the entity |
| W     | W represents a separate entity in the field |
| O     | O means other components of the sentence except the engine entity |

3.2.2. Feature selection

CRFs have two main problems to be solved: one is the parameter setting, and the other is the feature selection. The selection of features is critical to the results of the model. When Liu Bao et al. used rule-based and statistical methods to extract scientific and technological terms, they used the word itself and part of speech as features. When Huang Yi et al. used conditional random fields to acquire the context of domain terms, they used words and parts of speech, combined with feature word dictionaries and punctuation as their features, and trained the model. This article selects commonly used words and part of speech features, and adds word length as a new feature. As shown in Table 1, the first column is the characteristics of the word itself after jieba word segmentation, the second column is the part-of-speech characteristics of the words, the third column is the word length characteristics of the current words, and the fourth column is the lexeme characteristics of the word. The three features selected in this article are as follows:

1. The word itself.
   Our task is to obtain the entity in the field of aerospace engines. According to the characteristics of the entity in the field of engine, most words only circulate in this field, so the word itself covers the largest information of the entity, and the word itself is the most important feature.

2. Part of speech
   Certain rules can be found according to the compositional characteristics of words and phrases. Chinese word formation has certain characteristics. By observing entities in the field of aerospace engines, most of the entities’ part-of-speech combinations are "n+n", "v+n" and other formats, so the part of speech Feature is also one of the very important features.

3. Word length
   Engine words generally have more than one word. This feature can be used to determine whether the word is part of the entity by the word length.

3.2.3. The Model of establishment and prediction

The training of CRFs model needs a feature template, which is used to provide a unified template for the feature functions in the CRFs model, so as to propose the relevant feature functions from the training text, and then train to obtain the weight of each feature function. The definition of the template directly determines the recognition effect of the final model.

Commonly used feature templates are divided into unigram template and bigram template. The unigram template starts with "# Unigram" and the binary feature template starts with "big gram". Compared with the unigram feature template, the binary feature template can improve the recognition efficiency of the model, but when the number of features is relatively large, the binary feature template will greatly increase the training time of the CRFs model. Therefore, this article uses unigram features template in the aerospace engine entity recognition.

The feature template format is: %x[row, col]. x can be U or B, corresponding to two types. The number in square brackets is used to calibrate the feature source, row represents the row relative to the current position, 0 is the current row; col corresponds to the column in the training file. Only the first column (number 0) is used here, which is the text.
For the aerospace engine entity recognition, when the selected features are the word itself, part of speech and word length, the established feature template is shown in Figure 3.

![Feature template](image)

Figure 3. Feature template Figure.

U00 to U09 correspond to the feature template of "word itself", U10 to U14 correspond to the feature template of "part of speech", and U15 to U17 are feature templates of "word length".

After the feature template is specified, the CRF++ toolkit can be used to train the model on the training set while testing the model.

Model training mainly includes two aspects: input and output respectively. The input is a preprocessed data set, and the output is a model composed of feature functions and parameters.

The model test uses the trained template, the data is processed into the result that can be recognized by the model as input, and the output is the recognized aerospace engine entity.

4. Experimental process and result analysis

4.1. Experimental data

In the field of Chinese word segmentation, SIGHAN-2nd International Chinese Word Segmentation Bakeoff is the most authoritative evaluation. This article uses the data set provided by Microsoft Research Asia for SIGHAN bakeoff 2005 as a general field data set, all of which are "People's Daily". The content in January 1998, with a total of 1.7 million words, is a news language with a wide range of sentences and standardized data formats. The number of training data labeled entities is shown in Table 3.

| BME | O  |
|-----|----|
| 1255055 | 1113336 |

As shown in Table 3, the number of entities and non-entities of People's Daily are 1,255,055 and 1,113,336 respectively.

This article proposes an entity recognition algorithm for the aerospace engine field. The bake off 2005 data set lacks aerospace engine field data. Therefore, this experiment uses the aerospace engine field data from the "World Missiles and aerospace Engines". The provided data set is 1.74 MB. 70% is used for training corpus and 30% is used for test corpus. The number of training data labeled entities is shown in Table 4.

| BME | W  | O  |
|-----|----|----|
| 25932 | 16085 | 161380 |
As shown in Table 4, the aerospace engine field entities include 42017 words in combination and individual words, and a total of 161,380 words in the general field.

4.2. Experimental environment

| Classification       | Parameter name | Configuration                      |
|----------------------|----------------|------------------------------------|
| Hardware environment | CPU            | Intel® Core™ i5-5200U CPU @ 2.2GHZ |
|                      | RAM            | 8GB                                |
|                      | Operating      | Windows10                          |
|                      | language       | python3.6                          |
| Software environment | Development tools | pycharm                          |

4.3. Experimental results

4.3.1. Parameter selection

There are mainly four parameters that can be adjusted: normalization parameters, data fitting degree parameters, cut-off thresholds for setting features, and the number of threads.

The training parameters selected in this article are: CRF-L2, -c=4.0, -f=3, p=1.

4.3.2. Evaluation Criteria

The performance of a method is mainly measured by the accuracy of the method's recognition. The commonly used indicators for evaluating the accuracy are accuracy (P), recall (R) and F1 value. The calculation method is as follows:

\[
P = \frac{TP}{TP + FP} \quad (6)
\]

\[
R = \frac{TP}{TP + FN} \quad (7)
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \quad (8)
\]

4.3.3. Result analysis

This article first compares the entity recognition of CRF++ tool in general field and aerospace engine field. The results are shown in Table 5.

| Data set         | accuracy | recall | F1 value |
|------------------|----------|--------|----------|
| People's Daily   | 96.5%    | 96.4%  | 96.5%    |
| aerospace engine | 92.7%    | 87.1%  | 89.8%    |

It can be seen from Table 5 that after using CRF++ for entity recognition, the recognition results in the general field will be significantly better than the recognition results in the aerospace engine field. This is mainly due to the relatively complex entities in the aerospace engine field, fewer training data sets, and manual annotation errors.

The analysis of the results generated by using different entity recognition algorithms on the aerospace engine data is shown in Table 6.

| Entity recognition algorithm | accuracy | recall | F1 value |
|------------------------------|----------|--------|----------|
| CRFs general field template  | 63.2%    | 63.4%  | 63.3%    |
The general domain model generated after training on the People’s Daily data set using CRFs tool, and the test set of the aerospace engine using the general domain template. It can be seen that the entity recognition result is not good, because the entity types in the general domain are inconsistent with those in the aerospace engine domain. The inconsistent types result in poor entity recognition. The effect of using jieba word segmentation is better than the general domain model, but the recognition result is still not ideal. The CRFs tool was used to train the space engine field data and generate the template with the three features, and the test was conducted with the template. The accuracy rate of 92.7% was obtained, which was higher than the other two methods.

Analysis of the results generated by entity recognition using different features on the aerospace engine data is shown in Table 7.

| Selected features                                      | accuracy | recall | F1 value |
|--------------------------------------------------------|----------|--------|----------|
| The word itself                                        | 90.7%    | 84.2%  | 87.3%    |
| The word itself, part of speech                         | 92.3%    | 86.9%  | 89.5%    |
| The word itself, part of speech, word length            | 92.7%    | 87.1%  | 89.8%    |

It can be seen from Table 7 that the recognition of the model obtained by selecting the three features for training is relatively accurate. At the same time, adding part-of-speech features on the basis of the word itself has a greater impact on the results, while the word length has a small impact on the recognition results, but it is also improved accuracy to a small extent.

4.4. Experimental analysis

Analyzing the experimental results, it was found that the errors were mainly concentrated in the following aspects:

1. The recognition words are not complete, such as "CIAOTRJ verification engine" is identified as "CIAOTRJ"/" verification "/" engine ".

2. The recognized word is partly more than the correct entity. Except for the case of incorrect word segmentation, there are cases where, for example, "fuel supply pipeline" is recognized as "regulated fuel supply pipeline".

3. Because there is no uniform standard, there are some ambiguities in the labeling. For example, according to the labeling rules, "J33-A-18A turbojet engine" is judged as an entity, but the recognition result is "turbojet engine".

4. Some words are not considered as entities in the aerospace engine field, but they may also be recognized due to their own characteristics or similar context and terminology, such as "regulation system".

5. The recognition result is incorrect due to inevitable errors in manual labeling.

From the results, although there are still errors in the results, the recognition effect is relatively good.

5. Conclusion

This paper proposes an entity recognition method in the field of aerospace engines based on CRFs, using the famous open source tools CRF++ over the airport, using artificially labeled data sets, and constructing a model based on the word itself, part of speech, and word length. By comparing experiments, the effectiveness of CRF++ and the necessity of three characteristics are obtained.

From the experimental point of view, this paper still has certain defects. Due to the sparse aerospace engine data, the comparison of experimental results using different methods is relatively high; there are certain errors in the process of manually labeling the training set, which will also have a great impact
on the results. Next, we will consider training with fewer data sets to obtain better results and reduce errors in manual data annotation.

Acknowledgments
This work had supported by the Special Funds for Fundamental Research Funds for Central Universities (2020YQJD15) and National University Student Innovation Training Project Funding (C202004828).

References
[1] Liu Z., Zhu C., Zhao T. Chinese Named Entity Recognition with a Sequence Labeling Approach: Based on Characters, or Based on Words?[M]// Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. Springer Berlin Heidelberg, 2010.
[2] RATNAPAPKHI A.A Maximum Entropy Model for Part-of-speech Tagging[C]//Conference on Empirical Method in Natural Language Processing. 1996: 133-142.
[3] AFFERTY J, MCCALLUM A, PEREIRA FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence data[C]//Proceedings of the 18th International Conference on Machine Learning 2004:109-112.
[4] Maosong S, Dayang S, Tsou B K. Chinese Word Segmentation Without Using Lexicon and Hand-Crafted Training Data[C]//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2, 1998: 1265-1271.
[5] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[M]. Scholarly Commons, 2001: 282-289.
[6] Yang Yan, Wen Bo Huang, Yun Ji Wang. Image Classification Based on Conditional Random Fields. 2014, 3207:4901-4905.
[7] Che J, Tang L, Deng S, et al. Chinese Word Segmentation Based on Conditional Random Field[J]. Personal and Ubiquitous Computing: 1-12.
[8] Kaixin Liu, Qingcheng Hu, Jianwei Liu, Chunxiao Xing. Named Entity Recognition in Chinese Electronic Medical Records Based on CRF[C]//Web Information Systems and Applications Conference. 2017.
[9] C. Janarish Saju, S. Ravimaran. A Tool to Extract Name Entity Recognition From Big Data in Banking Sectors. 2020, 17(2): 18-39.