Hindawi
Scientific Programming
Volume 2018, Article ID 1890683, 10 pages
https://doi.org/10.1155/2018/1890683

Research Article

A Low-Cost Named Entity Recognition Research Based on Active Learning

Han Huang,[1] **Hongyu Wang**,[2] **and Dawei Jin**[1]

[1]*School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China*
[2]*School of Information Management, Wuhan University, Wuhan 430072, China*

Correspondence should be addressed to Dawei Jin; jdw@zuel.edu.cn

Received 10 August 2018; Accepted 28 November 2018; Published 18 December 2018

Guest Editor: Vicente García-Díaz

Named entity recognition (NER) is an indispensable and very important part of many natural language processing technologies, such as information extraction, information retrieval, and intelligent Q & A. This paper describes the development of the AL-CRF model, which is a NER approach based on active learning (AL). The algorithmic sequence of the processes performed by the AL-CRF model is the following: first, the samples are clustered using the *k*-means approach. Then, stratified sampling is performed on the produced clusters in order to obtain initial samples, which are used to train the basic conditional random field (CRF) classifier. The next step includes the initiation of the selection process which uses the criterion of entropy. More specifically, samples having the highest entropy values are added to the training set. Afterwards, the learning process is repeated, and the CRF classifier is retrained based on the obtained training set. The learning and the selection process of the AL is running iteratively until the harmonic mean F stabilizes and the final NER model is obtained. Several NER experiments are performed on legislative and medical cases in order to validate the AL-CRF performance. The testing data include Chinese judicial documents and Chinese electronic medical records (EMRs). Testing indicates that our proposed algorithm has better recognition accuracy and recall rate compared to the conventional CRF model. Moreover, the main advantage of our approach is that it requires fewer manually labelled training samples, and at the same time, it is more effective. This can result in a more cost effective and more reliable process.

1. Introduction

With the continuous popularization of Internet and mobile Internet and the continuous improvement of information infrastructure in various domains, the available digital resources have grown explosively in our metaindustrial societies [1]. On one hand, the sources and the volume of information have become more abundant. The density of useful data is decreasing, which makes it more difficult to mine valuable information. In the era of big data, it is difficult for people to manually analyze and filter information, due to their high volume and variety. Automatic or semiautomatic effective extraction from a large number of digital resources could lead to the mining of hidden knowledge. This could be achieved with the help of big data and artificial intelligence technologies like NER [2, 3]. NER is a process of recognizing and classifying words or phrases with special characteristics or meanings in a text. It belongs to the category of unsigned word recognition in lexical analysis, and it is an indispensable part of information extraction and retrieval, intelligent Q & A, and other natural language processing technologies [4].

The important role of NER in natural language processing has motivated a lot of research in the domains of library information and computer science, and it has resulted in the proposal of several methods. However, the categories and extensions of named entities significantly vary under different research scenarios and domains [5]. More specifically, the named entities (NAE) mainly refer to names of persons, places, and time. In the biomedical field, they can include medical terms such as protein, genome, or labels of diseases. In legislative domains, the NAE may include legal concepts, terms, or provisions. Obviously, the specification of NER depends on the field of study, and it is difficult to migrate directly [6]. Moreover, in the case of the domain-specific NER, such as legislative, ancient Chinese poetry and so on, the domain entities are relatively scarce, that is, the directly available training data are a minority. At

the same time, due to the high specialization of the data, the domain knowledge must be rich when annotating the domain texts manually. The necessary professional talents and the heavy workload require a lot of manpower and available resources. Therefore, the use of less annotated data to train an effective NER model has become an important goal of the domain NER research [7].

Active learning algorithms can be used to solve the problem of sparse data annotating, in machine learning. AL consists of the learning and the selection modules. The learning module is the process of training high-quality classifiers, while the selection module is the process of generating training sets from large amounts of data [8]. The core of AL is to use the learning algorithm in order to get the most useful data from the training set and then to add this data to the manually developed annotation set. In this way, we can get a classifier with a strong generalization ability, without requiring a high volume of annotated data [9]. This paper applies active learning to the field of NER. AL is the main algorithmic framework, and CRF is the corresponding classifier. We propose a new approach called AL-CRF, aiming not only to improve the efficiency recognition of the CRF model, but also to decrease the number of annotated training samples as well. The testing experiments on medical and legislative fields have proven that our proposed method can produce a more efficient NER model with fewer training samples, which can effectively cut the cost of manual annotation and improve the overall efficiency.

2. Background

2.1. Named Entity Recognition. NER has been defined as an important subtask of information extraction in the MUC-6 conference [10]. The most commonly used relative approaches can be divided into three categories: rule and dictionary-based methods, statistical, and mixed ones. Early NER mainly used rule and dictionary-based methodologies, which require the design and development of the rule sets by domain experts and the use of proper linguists. Nevertheless, the rules fail to cover all linguistic phenomena, the construction period is too long, and moreover, this approach has a lower likelihood of portability [7]. Statistical methods mainly include hidden Markov models (HMM) [11], maximum entropy (ME) [12], support vector machines (SVM) [13], and CRF [14]. These kinds of methods use the labelled corpus data to train the model combined with statistical probability. They are easily transplantable and they have comparatively short construction periods although they have more strict requirements for feature selection and much more dependence on the corpus. The mixed methods model is a combination of rules, dictionary-based approaches, and statistical ones, and they combine the advantages of both. They employ rules to filter the target text in advance, and they are reducing the state of the search space, based on statistical methods. Recently, some hybrid methods have been introduced, based on deep learning (DL), which combines DL with rules or statistical approaches [15]. At the same time, driven by the demand of natural language processing in various fields, the recognition objective of NER has also evolved from the initial person name, location name, and time to the words or phrases with special meanings in the recognition text. Researchers have also carried out NER research mainly for specific domain entities, such as fishery data [16], dietary data [17], and Chinese legal documents data [18]. CRF model is one of the most popular ones.

2.2. Conditional Random Field. CRF is an undirected graph model proposed by Lafferty in 2001, which combines the characteristics of the ME and the HMM, and it considers the transition probabilities between contextual markers at the same time. The transition probability between tags is optimized and decoded in the serialization form, and the sequence data annotation is carried out by establishing the probability model [19]. CRF has strong reasoning ability and is widely used in sequential tagging tasks, such as part of speech tagging [20], significance testing [21], and new word discovery [22]. NER is also a special kind of sequence tagging problem, and the CRF has innate advantages in solving it. When applied to NER, the CRF has good stability, accuracy, and ease of use [23]. However, as a typical supervised model, it requires a lot of training data, and the convergence speed is slow. To solve these problems, many researchers combine other machine learning algorithms with CRF in an effort to improve its performance. Such efforts have been made by Deng et al. [24] who have proposed a short-term traffic flow forecasting model (MCRF) which is based on multiconditional random fields. It uses four kinds of feature functions to build multiple CRF feature subsets to reflect the multicumulative characteristics of traffic data. Xia et al. [25] have combined convolutional neural networks (CNN) with CRF, and they have proposed a hybrid classification approach for remote sensing images. These methods greatly improve the availability and effectiveness of the CRF model, but the training process is still inseparable from a large number of annotated data.

2.3. Active Learning. Active learning is a branch of machine learning (ML) that belongs to the area of artificial intelligence. It was originally proposed by Angluin of Yale University [26]. The learning module needs to continuously improve the classification accuracy and robustness of the classifier, and the purpose of the selection modules is to find out the most representative and extensive training data. Current research on active learning can be summarized in two aspects. On one hand, researchers have applied it to many fields. Wu et al. [27], Zhu et al. [28], Pohl et al. [29], have introduced the application of AL algorithm in social media data, spatial data annotation and image classification respectively. On the other hand, many researchers have put forward the idea of improving it. Wang et al. [30] have proposed a new multi-instance AL algorithm by combining diversity criterion with existing information measure. Patra et al. [31] have proposed the LAAL-ELM which is an online continuous learning method. Through the confidence meter of newly added data, this method selects the tagging set actively to update the classifier, and it reduces the

computational complexity. Active learning provides an algorithmic framework to solve the problem of sparse annotating data in the training process of ML. In practical research, machine learning algorithms are used as a classifier of active learning. Through the iteration of the learning and selection processes, the performance of the classifier keeps improving continuously.

To sum up, although NER has been developed and used for more than 20 years, the problem of this field has not been completely solved, due to the continuous diffusion of the named entity denotations in different scenarios and domains. In previous research efforts, CRF has been one of the most widely used approaches. However, its training requires a lot of annotated data. Though it is difficult to obtain annotated data in a specific field, AL can effectively solve this problem, as it is capable to find high-value data in order to train high-performance classifiers. Therefore, this paper combines active learning with the CRF model. It uses the CRF classifier, and it proposes a method to recognize the NAE which enriches the method of named entity recognition. At the same time, the training process requires a small amount of annotated data, which is very significant to the application areas where the annotated data is rare and the annotation is a hard task, e.g., in medical and legal cases.

3. Materials and Methods

3.1. CRF Model. The CRF model is a model that outputs the conditional probability of a random variable Y with a given random variable X. This model has various forms including the linear chain form, the matrix form, and so on. In the NER process, the CRF model is usually further simplified, that is, the random variables X, Y have the same graph structure, which is shown in Figure 1. X is the input text to be recognized, and $x_1, x_2, \ldots, x_{n-1}, x_n$ are sequences after word segmentation and feature tagging. The task of the CRF model is to predict the conditional probability of Y by training the model parameters, and the calculation method is shown in equation (1):

$$P(y \mid x) = \frac{1}{Z(x)} \exp \left(\sum_{i,k} \lambda_k t_k (y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l (y_i, x, i) \right), \quad (1)$$

where $Z(x)$ is a normalization factor, whereas t_k is an eigenfunction defined on the edge k, which is called transfer feature. It depends on the current position and on the previous position. Also, s_l is an eigenfunction defined on the node l which is called state feature, and it depends on the current position. The parameters λ_k and μ_l are the weights corresponding to t_k and s_l. The value of t_k and s_l is either 1 or 0. When the characteristic condition is satisfied, the value is 1, otherwise it is 0.

In this research, the observation set x is the sequence set that comprises a text corpus after the word segmentation and feature automatic annotation, and y is the annotation type corresponding to the observation set x. In the feature

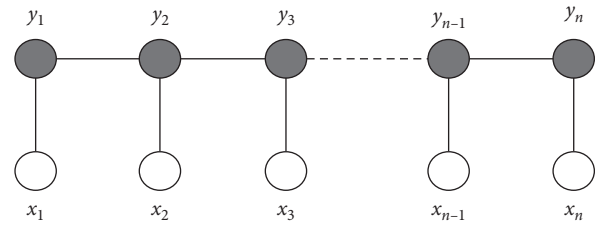

Figure 1: The structure of CRF model.

model construction, we use a 5-word tagging set that is expressed as $P = \{B, I, E, S, O\}$. Tag B is an entity starting word, and E represents the end of the entity. All of the entity is tagged as I except from the beginning and the final words. S represents the entity as a single word, and O is expressed as a word outside the entity.

3.2. Active Learning Algorithm. In this paper, the algorithmic process of the AL is shown in Figure 2. Firstly, the initial samples for training the basic CRF model are selected by following a certain strategy, and they are annotated by domain workers. Thereafter, according to the information, the CRF model is trained, the unlabelled samples are sorted according to certain ranking rules, and the top N samples are selected for manual annotating. Then the annotated data are added to the training set to retrain the CRF model. The learning process and selecting process are carried out iteratively until the exit condition is satisfied. Obviously, three key problems have to be solved in the AL process. First, the construction of the initial training set; second, the choice of the proper strategy for sample selection; and finally, the effective setting up of the iterative process and the quit condition.

The initial training set is used to train the benchmark classifier in the active learning algorithm. Therefore, selecting the representative initial training set can train the benchmark classifier with good recognition result, which would reduce the number of iterations and could accelerate the convergence process. Random Sampling is the basic algorithm for the construction of the initial training set. However, due to its limited size, the samples selected by this approach are considered less representative. On the other hand, the clustering method can aggregate samples with similar characteristics, so that the stratified sampling method based on the clustering results is more likely to choose the most representative samples.

In active learning, sample selection strategies (SSS) can be divided in two types, namely, the stream-based SSS and the pool-based one. The learning process of the stream-based SSS requires the processing of all unlabelled samples, which increases the query cost. In addition, since it requires presenting the sample annotation conditions in advance, it does not have good applicability [32]. The pool-based strategy is that selecting the sample with the highest contribution from the sample pool at a time, which reduces the query cost of the sample, so it is more widely used.

AL is a process which iteratively selects high-value samples for model training, in order to improve the efficiency of the classifiers. Although the increase in the number

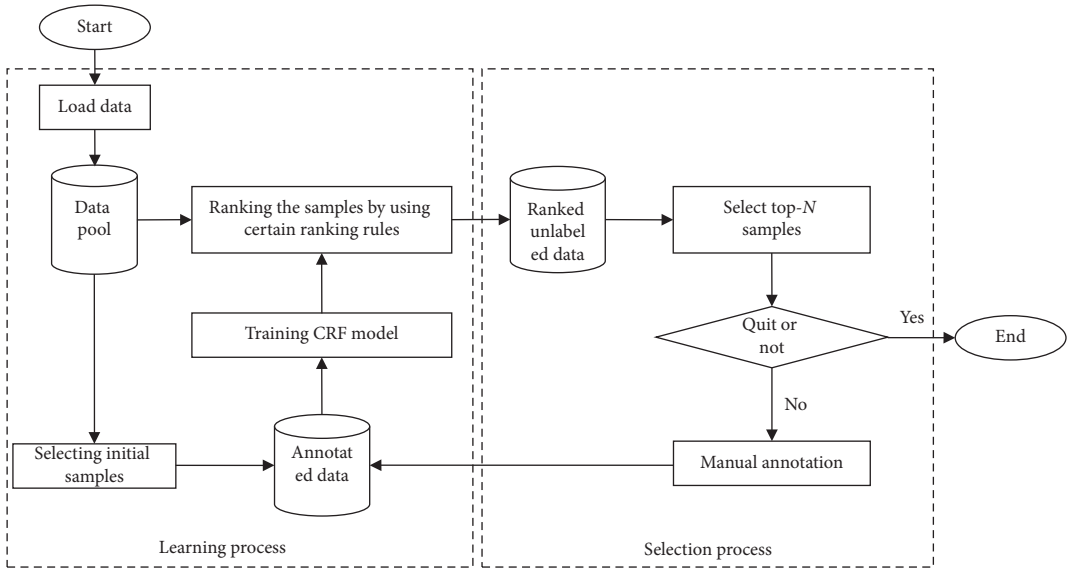

Figure 2: The process of active learning.

of iterations can improve the efficiency of the classifier, it also increases the workload of both the sample selection and the sample annotation. The training process strives to achieve the balance between sample labelling cost and classifier efficiency. Therefore, in general, the active learning algorithm will terminate the iteration when the model efficiency reaches the present precision or the number of samples reaches the given threshold value.

3.3. AL-CRF Model. The AL-CRF model takes active learning as the basic framework and the CRF as the basic classifier of AL. A hierarchical sampling method based on k-means clustering is used to select the training set for the initial learning. An SSS based on information entropy is adopted to select the samples for the iteration process. The quit condition of the iteration is based on a defined change rate of the F-value. The algorithm framework is shown in Figure 3.

In order to construct a more representative initial training set, the AL-CRF model uses the TF-IDF algorithm to vectorize the text data. It also employs the k-means algorithm to cluster the data, and it stratifies the data after clustering. The whole algorithmic process is described as follows:

(1) Vectorizing and normalizing data using TF-IDF to get the dataset X after loading the corpus.

(2) Choosing the number of K samples from the data set X randomly as C.

(3) Calculating the Euclidean distance between the remaining samples in X and C and classifying (assigning) the remaining samples in X, to the nearest cluster according to the distance.

(4) Calculating the mean value of each cluster and updating the original clustering center after all samples are divided.

(5) Comparing the new center with the previous clustering center. If there is no change, it will terminate; otherwise, go to step 2.

(6) Outputting the final clustering results.

(7) The initial sample set T was selected by stratified sampling according to the clustered results.

The AL-CRF model chooses a pool-based sample selection strategy, and it uses information entropy (IE) as a measure to evaluate the sample value based on uncertainty criteria. IE is a measurement unit used to measure the amount of information. The higher the value of the IE is, the more information is contained in the sample. This indicates that the classifier has not determined the proper category. Through the iterative process, the model predicts the sequence of IE values of the remaining samples, by employing the existing classifier. In this paper, the IE value of the sample is the sum of the IE value of each word in the sample, and the calculation method is shown in equations (2), (3), and (4):

$$H(x_j) = -\sum_i p(y_i \mid x_j) \log p(y_i \mid x_j), \quad (2)$$

$$H(s_t) = \sum_j H(x_j), \quad (3)$$

$$H(d_k) = \sum_t H(s_t). \quad (4)$$

where $H(x_j)$ represents the entropy value of word x_j, $p(y_i \mid x_j)$ indicates that the label belongs to the possibility of y_i under the given word x_j, $H(s_t)$ represents the entropy value of sentence s_t, and $H(d_k)$ represents the entropy value of the document d_k.

The AL-CRF model sets the change rate of F-value less than or equal than 0.1% as the iterative termination condition of the active learning. This means that $F_t - F_{t-1} \leq 0.1\%$, where F_t represents the F-value of the model in the t iteration, F_{t-1} represents the F-value of the model in the $t-1$ iteration, and F_0 represents the initial expression. The default value is zero. This is done in order to control the sample selection and to mark the cost of the training process.

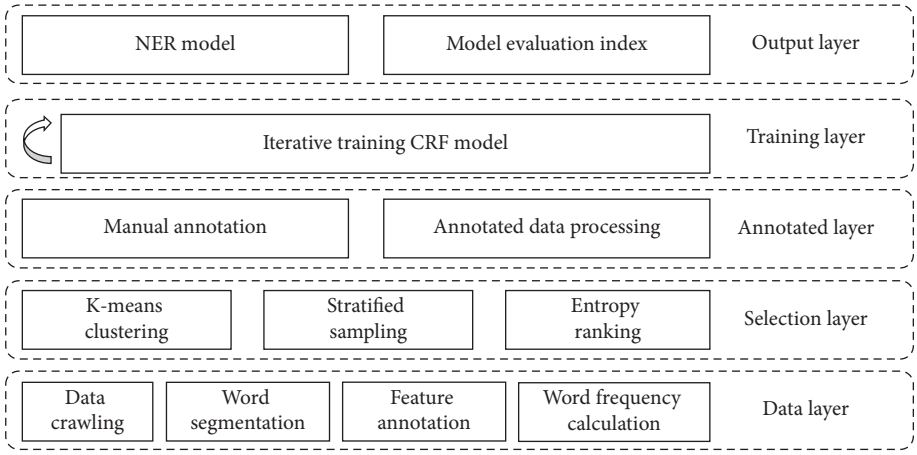

Figure 3: The algorithm framework of AL-CRF.

In summary, the AL-CRF model can be represented by the use of pseudocode as shown in the following Algorithm 1. The core of the algorithmic process is shown in Figure 4.

4. Experiment

In order to verify the effectiveness of the NER AL-CRF model and its performance in different domains, this paper selects two different datasets in the medical and the legislative fields. This completes the NER experiment.

4.1. Dataset. In the process of NER in the medical field, this paper uses the EMRs dataset which was released by the China Conference on Knowledge Graph and Semantic Computing (CCKS) 2017. To ensure the correctness of this experiment, we have selected only 300 EMRs that have been correctly annotated, whereas each EMR is divided into four parts, namely, "history characteristics," "hospital discharge," "general items," and "diagnosis and treatment." The total volume of the experimental data is 1,200 text numbers and 5 categories of entities. Overall, the total number is 23,719, and the distribution of categories is shown in Table 1.

In the legislative domain experiment, this paper obtained 61,515 copies of the judicial documents stored in the "Chinese justice document network." After clearing up the duplicates, blanks, and noncontent data, we got 59,788 valid data, containing 52,995 first-instance documents, 5,632 second-instance documents, 325 retrial documents, 37 penalty changes, and 799 documents in other categories. In this paper, 1,000 pieces of judicial documents have been extracted and manually annotated in the form of stratified sampling, which can be used as the corpus of the legislative NER experiment to reduce the cost of manual labelling. There are 73,217 legal entities in the corpus, including 5 categories of crime, penalty, legal principle, legal concept, and legal provision. The distribution is shown in Table 2.

Since there is no delimiter in Chinese itself, Chinese word segmentation is the basis of the data analysis. In order to improve the accuracy of the word segmentation, this paper attempts to construct the professional dictionaries in both medical and legislative fields by using disease symptoms, treatment technologies, crimes, legal institutions, and legal words obtained from the Internet. Then, we import them into the NLPIR segmentation tool of the Chinese Academy of Sciences and cut the words of EMRs and of the judicial documents, respectively.

After the completion of the manual annotation of the corpus, the program has been used to tag the POS and the length of each word automatically and to record whether it is left or right boundary word. According to the annotation method of 5-word tagging sets, the format of the corpus is shown in Table 3.

4.2. Experimental Design. In this paper, CRF and AL-CRF models have been used to recognize entities in medical and legislative domain for EMRs and judicial documents, respectively. Ten crossover trials have been conducted in the specific experiment, where the training and the testing sets have been determined based on a ratio of 9 : 1. The CRF++ has been used as a tool for training and evaluating CRF models and the Spark platform has been selected for text quantization and text clustering.

In the AL-CRF experiment, the initial corpus of the active learning model has been set to 100 copies, the sample size has been increased to 50 per each round of iteration, and the growth rate of the harmonic mean F has a value less than 0.1% as the iterative termination condition. In the CRF experiment, random sampling has been used to select the equivalent of the AL-CRF documents of the training set. Then the test data have been used to evaluate the effects of the two models respectively.

5. Results and Discussion

5.1. The Evaluating Indicator. According to the common indicator system of NER, we selected the following evaluation indicator, involving precision rate (P), recall rate (R), and F-measure (F) [18]. The calculation method is shown in equations (5), (6), and (7):

```
Initialization: unlabelled dataset U, F_0 = 0, i = 0, t = 0, initialization data number n, additional number N in iteration
// k-means clustering
select cluster centers randomly as C_i
do
  for u in U do
    for c in C_i do
      if dis(u, c) is min_u then
        the cluster of u is c
      end
    end
  end
  update C_i to C_{i+1}
while C_i! = C_{i+1}
output the clustered dataset S
select n samples from S by stratified sampling
annotate n samples into T
train CRFs by T
t⟵t + 1
calculate the F-value of CRFs as F_t
while F_t − F_{t−1} > 0.1% do
  calculate entropy in {S − T}
  rank the {S − T} according to the entropy
  annotate top N samples into T
  train CRFs by T
  t⟵t + 1
  calculate the F-value of CRFs as F_t
end
```

Algorithm 1: The pseudocode of AL-CRF.

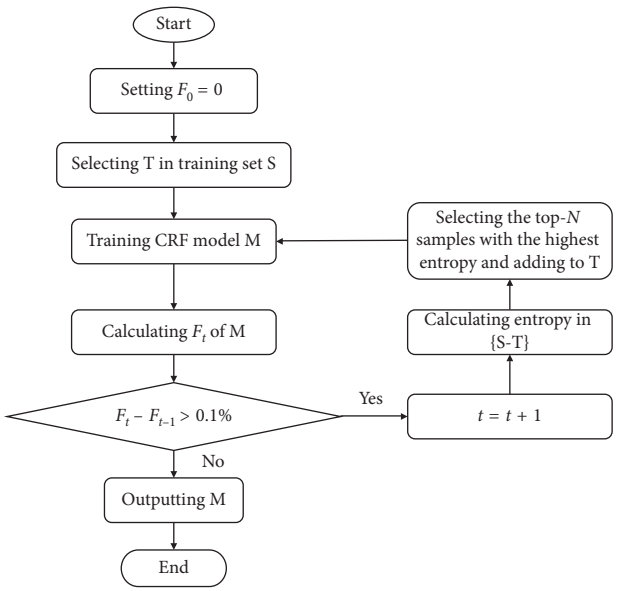

Figure 4: The flowchart of AL-CRF.

$$P = \frac{A}{A + W} * 100\%, \tag{5}$$

$$R = \frac{A}{A + U} * 100\%, \tag{6}$$

$$F = \frac{2 * P * R}{p + R} * 100\%, \tag{7}$$

where A represents the number of entities recognized correctly, W indicates the number of entities that are recognized by mistake, and U represents the number of entities that are not recognized at all.

5.2. Clustering Number Selection. In order to determine the number of k-means clustering, we have conducted 5 rounds of experiments on the EMRs data and on the judicial documents, respectively. Each experiment divided the training and testing sets according to the ratio of 4 : 1 to conduct clustering experiments.

Since the number of clusters is not too large, each round of experiments contains 14 clustering experiments, with the number of clusters ranging from 2 to 15. The sum of squared errors (SSE) of each round in the EMR experiments is shown in Figure 5.

According to the general principle of the elbow method, when the number of clusters is 5 and 8, the SSE value is lower. In order to further determine the number of clusters, we select the initial sample set which is selected by stratified sampling with 5 and 8 clusters to train the initial recognition model. Additionally, we use three alternatives as control groups. In the first one, we select the samples randomly; in the second mode, we employ stratified sampling with 2

Table 1: The entity distribution of the EMRs.

Category	Body(B)	Symptoms and signs (SS)	Examination and inspection (EI)	Disease and diagnosis (DD)	Treatment (T)
Number	8,282	6,941	6,903	657	936

Table 2: The entity distribution of the judicial documents.

Category	Charge(C)	Penalty(P)	Legal principle (LP)	Legal concept (LC)	Law(L)
Number	1,745	4,732	209	63,820	2,711

Table 3: The format of the corpus.

Word	POS	Length	Is left	Is right	Tag
无	V	1	N	N	O
发热	Vi	2	Y	Y	SS-S
,	Wd	1	N	N	O
时	Ng	1	N	N	O
有	Vyou	1	N	N	O
咳嗽	Vi	2	Y	Y	SS-S
、	Wn	1	N	N	O
咳	V	1	Y	N	SS-B
痰	n	1	N	Y	SS-E
,	Wd	1	N	N	O
无	V	1	N	N	O
胸	Ng	1	Y	N	SS-B
闷	V	1	N	Y	SS-E
、	Wn	1	N	N	O
气	N	1	Y	N	SS-B
短	a	1	N	Y	SS-E

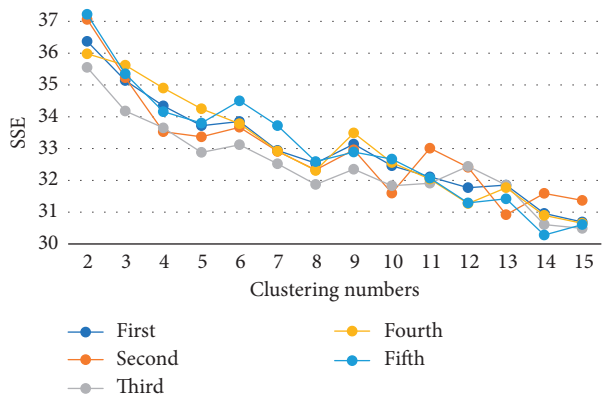

Figure 5: The SSE of each round in EMR experiments.

clusters; and in the third, we adopt again stratified sampling with 14 clusters. This is done in order to compare the influence of the number of clusters on the selection of the initial training set. The average recognition effect of the above experiments is shown in Table 4.

By analysing the results in Table 4, it can be seen that selecting the initial sample set after clustering the initial sample into the appropriate number of categories can improve the accuracy and the recall rate of the initial model. When the number of clusters is 8, the recognition efficiency of the model is the best. Therefore, for EMRs, the optimal number of clusters is 8. According to the same experimental procedure, the optimal number of clusters is 10 for the legislative judicial documents.

Table 4: The average recognition effect of initial model in EMRs.

Selection method	P	R	F
Random	0.8100	0.7658	0.7873
2 clusters	0.8115	0.7534	0.7813
5 clusters	0.8053	0.7707	0.7876
8 clusters	0.8174	0.7767	0.7965
14 clusters	0.8039	0.7682	0.7856

5.3. The Evaluation of AL-CRF and CRF

5.3.1. The Dataset of EMRs. In this paper, 1,200 EMRs have been divided to 10 equal parts. The training and the testing sets have been divided according to the ratio 9 : 1, and 10 comparative experiments have been carried out. The experimental results are shown in Table 5.

According to the results of Table 5, the recognition efficiency of the AL-CRF model tends to be stable when the number of iterations is 10, using 600 training samples. It is found that the CRF and the AL-CRF models have good recognition efficiency in the Chinese EMRs, with the recognition accuracy reaching over 90%, and most of the entities have been recognized. However, the recognition efficiency of the AL-CRF model is obviously better than the one of the CRF and the recognition accuracy can reach up to 95%. The F-value of the model increases almost by 3.65%. Specifically, the recognition effect of the five categories of entities in the medical field is shown in Table 6.

It can be seen that the AL-CRF model is superior to the CRF model in the recognition effect of various entities. In both models, the recognition effect is the best for symptoms and signs, while the entity recognition effect of treatment, examination, and inspection is not good, which may be related to the mixing of drug information into the entity of treatment and the unclear boundary between the entity of examination and inspection. Through analysis of experimental data, it is found that due to the imported custom dictionary, the word segmentation of symptoms and signs entity is more accurate, and its good recognition effect is related to the result of word segmentation.

5.3.2. The Dataset of Chinese Judicial Documents. At the same time, 10 comparative experiments have been conducted on the judicial documents. The experimental results are shown in Table 7.

Table 5: Comparison table of experimental results in EMRs.

Number of iterations		AL-CRF			CRF		
		P	*R*	*F*	*P*	*R*	*F*
1	11	0.9603	0.9314	0.9456	0.9223	0.9012	0.9116
2	10	0.9552	0.9328	0.9439	0.9252	0.8992	0.912
3	9	0.9516	0.9273	0.9393	0.8993	0.8632	0.8809
4	10	0.9501	0.9316	0.9408	0.9233	0.9014	0.9122
5	11	0.9627	0.9462	0.9543	0.9236	0.8979	0.9106
6	10	0.9513	0.932	0.9416	0.9124	0.9005	0.9064
7	10	0.9498	0.9297	0.9396	0.9157	0.8999	0.9077
8	11	0.9544	0.9406	0.9475	0.9208	0.9025	0.9136
9	10	0.9531	0.9201	0.9363	0.9082	0.8963	0.9022
10	10	0.9602	0.9224	0.9409	0.9144	0.9015	0.9079
Mean		0.9549	0.9314	0.9430	0.9165	0.8964	0.9065

Table 6: The recognition effect of 5 categories in medical field.

Model	Category	*P*	*R*	*F*
AL-CRF	Body (B)	0.9523	0.9197	0.9357
	Symptoms and signs (SS)	0.9791	0.9556	0.9672
	Examination and inspection (EI)	0.8042	0.8396	0.8215
	Disease and diagnosis (DD)	0.9381	0.9187	0.9283
	Treatment (T)	0.7938	0.7607	0.7769
CRF	Body (B)	0.9312	0.8899	0.9101
	Symptoms and signs (SS)	0.9746	0.9306	0.9521
	Examination and inspection (EI)	0.7659	0.8065	0.7857
	Disease and diagnosis (DD)	0.9133	0.8919	0.9025
	Treatment (T)	0.7624	0.7123	0.7365

Table 7: Comparison table of experimental results in judgement documents.

Number of iterations		AL-CRF			CRF		
		P	*R*	*F*	*P*	*R*	*F*
1	9	0.9314	0.9603	0.9456	0.8468	0.8824	0.8642
2	10	0.9328	0.9552	0.9439	0.8854	0.8992	0.8922
3	9	0.9273	0.9516	0.9393	0.8993	0.9132	0.9062
4	8	0.9316	0.9501	0.9408	0.8333	0.8714	0.8519
5	9	0.9462	0.9627	0.9543	0.8836	0.8979	0.8907
6	9	0.932	0.9513	0.9416	0.9024	0.9205	0.9114
7	10	0.9297	0.9498	0.9396	0.9057	0.9299	0.9176
8	8	0.9406	0.9544	0.9475	0.8708	0.9025	0.8864
9	9	0.9201	0.9531	0.9363	0.8982	0.9363	0.9169
10	9	0.9224	0.9502	0.9361	0.8844	0.9215	0.9026
Mean		0.9314	0.9539	0.9425	0.881	0.9075	0.894

According to the results of Table 7, the recognition efficiency of AL-CRF model tends to be stable when the number of iterations equals to 9, and 550 training samples are considered. Meanwhile, the AL-CRF model has also achieved promising results for the entity recognition in the case of the legislative domain. The recognition accuracy and recall have been improved, and the *F*-value has been increased by 4.85% compared with the *F*-value of the CRF model. Specifically, the effect of the 5 categories of entities in the recognition of the legislative field is shown in Table 8.

Trying to assess the effect of various entities to the recognition, in the case of the legislative data, we conclude that the AL-CRF model has a reliable performance and an obvious superiority, compared with the CRF, especially in legislative conceptual entities. Although, due to the wide range of legislative conceptual entities (e.g., "plaintiff" and "legal person)" and to the existence of relational concepts (e.g., "obligation of delivery" and "liability for compensation)" and finally due to the differences in the description of various legal documents, the overall recognition efficiency is still not as high as it should be. In addition, the large number of long entities in legislative principle entities has a negative effect.

6. Conclusions

In this paper, the active learning algorithm is applied to the domain of NER, and the hybrid AL-CRF model, which

Table 8: The recognition effect of 5 categories in legal field.

Model	Category	P	R	F
AL-CRF	Charge (C)	0.9713	0.9801	0.9757
	Penalty (P)	0.8727	0.8943	0.8834
	Legal principle (LP)	0.8062	0.8331	0.8194
	Legal concept (LC)	0.8898	0.9187	0.9041
	Law (L)	0.9765	0.9784	0.9774
CRF	Charge (C)	0.9672	0.9706	0.9689
	Penalty (P)	0.8583	0.8754	0.8668
	Legal principle (LP)	0.7926	0.8118	0.8021
	Legal concept (LC)	0.8547	0.8774	0.8659
	Law (L)	0.9718	0.9735	0.9726

employs the CRF classifier, is proposed. Through the recognition experiments of medical entities and legal entities, it is found that this method can train a good NER model with less annotated data, and it has a certain improvement over the CRF model in accuracy and recall. It can significantly reduce the labour cost for a large number of annotated data of the traditional methods, and it can speed up the convergence rate of the model. Thus, it can be more suitable for the recognition of domain entities with high annotation cost, and it lays the foundation for other natural language processing tasks in the specific domain.

Though this research effort is very promising, there are certain shortcomings. In terms of experimental data, although two datasets from different domains were adopted, their size was not big enough. Regarding the model itself, the adopted k-means approach can improve the initial sample quality; however, it is sensitive to noise and the outliers in the initial training set may affect the recognition efficiency of the model.

The extraction of relations between entities and knowledge fusion will be the key point in the next step of our research, in an effort to resolve the existing deficiencies.

Data Availability

This research has two different datasets. One of them is the open documents which is derived from "Chinese judgement document network" (http://wenshu.court.gov.cn/), and the other is the electronic medical records released by the China Conference on Knowledge Graph and Semantic Computing 2017.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the Innovative Education Program for Graduate Students of Zhongnan University of Economics and Law (no. 201811409).

References

[1] Z. Liu and Q. Zhang, "Research overview of big data technology," *Journal of Zhejiang University (Engineering Science)*, vol. 48, no. 6, pp. 957–972, 2014.

[2] R. Grishman and B. Sundheim, "Message understanding conference-6: a brief history," in *Proceedings of 16th International Conference on Computational Linguistics COLING 1996 Volume 1*, Copenhagen, Denmark, August 1996.

[3] W. Zhao, L. Gao, and A. Liu, "Programming foundations for scientific big data analytics," *Scientific Programming*, vol. 2018, Article ID 2707604, 2 pages, 2018.

[4] Z. Sun and H. Wang, "Overview on the advance of the research on named entity recognition," *New Technology of Library and Information Service*, vol. 26, no. 6, pp. 42–47, 2010.

[5] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: a systematic review," *Computer Science Review*, vol. 29, pp. 21–43, 2018.

[6] J. Zhao, "A survey on named entity recognition, disambiguation and cross-lingual coreference resolution," *Journal of Chinese Information Processing*, vol. 23, no. 2, pp. 3–17, 2009.

[7] S. Huang, X. Zheng, and D. Chen, "A semi-supervised learning method for product named entity recognition," *Journal of Beijing University of Posts and Telecommunications*, vol. 36, no. 2, pp. 20–23, 2013.

[8] W. Yang, X. Tian, S. Wang, and X. Zhang, "Recent advances in active learning algorithms," *Journal of Hebei University (Natural Science Edition)*, vol. 37, no. 2, pp. 216–224, 2017.

[9] M. A. Carbonneau, E. Granger, and G. Gagnon, "Bag-level aggregation for multiple-instance active learning in instance classification problems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, pp. 1–11, 2018.

[10] N. Chinchor, "MUC-6 named entity task definition (version 2.1)," in *Proceedings of 6th Conference on Message Under-Standing*, Columbia, Maryland, November 1995.

[11] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-per-formance learning name-finder," in *Proceedings of Fifth Conference on Applied Natural Language Processing*, pp. 194–201, Washington, DC, USA, April 1997.

[12] A. E. Borthwick, "A maximum entropy approach to named entity recognition," Master's Dissertation, New York University, New York City, NY, USA, 1999.

[13] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," in *Proceedings of 19th International Conference on Computational Linguistics*, pp. 1–7, Taipei, Taiwan, September 2002.

[14] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of Seventh Conference on Natural Language Learning at HLT-NAACL*, vol. 4, pp. 188–191, Edmonton, Canada, May 2003.

[15] Y. Liu, "Named entity recognition in Chinese Micro-blog based on deep learning," *Advanced Engineering Sciences*, vol. 48, no. 2, pp. 142–146, 2016.

[16] J. Sun, H. Yu, and Y. Feng, "Recognition of nominated fishery domain entity based on deep learning architectures," *Journal of Dalian Ocean University*, vol. 33, no. 2, pp. 265–269, 2018.
[17] Q. Wei, T. Chen, R. Xu, Y. He, and L. Gui, "Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks," *Database*, vol. 2016, article baw140, 2016.
[18] L. Zhang, C. Qin, and W. Ye, "Research on legal field entity automatic recognition model based on conditional random fields," *New Technology of Library and Information Service*, vol. 11, pp. 46–52, 2017.
[19] J. Sun, J. Li, and G. Zhou, "An unsupervised Chinese part-of-speech tagging approach using conditional random fields," *Computer Applications and Software*, vol. 28, no. 4, pp. 21–23, 2011.
[20] S. Paisitkriangkrai, J. Sherrah, P. Janney, and V. D. Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36–43, Boston, MA, USA, June 2015.
[21] S. Qian, Z. H. Chen, M. Q. Lin, and C. B. Zhang, "Saliency detection based on conditional random field and image segmentation," *Acta Automatica Sinica*, vol. 41, no. 4, pp. 711–724, 2015.
[22] F. Chen, Y. Liu, C. Wei, Y. Zhang, M. Zhang, and S. Ma, "Open field neologism discovery based on conditional random field," *Journal of Software*, vol. 5, pp. 1051–1060, 2013.
[23] H. Guo, "Accelerated continuous conditional random fields for load forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 8, pp. 2023–2033, 2015.
[24] Z. Deng, J. Ren, and L. B. Liu, "Short-term traffic flow prediction algorithm based on multiple CRF model," *Computer Engineering and Design*, vol. 38, no. 10, pp. 2887–2891, 2017.
[25] M. Xia, G. Cao, G. Wang, and Y. Shang, "Remote sensing image classification based on deep learning and conditional random fields," *Journal of Image and Graphics*, vol. 22, no. 9, pp. 1289–1301, 2017.
[26] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, no. 4, pp. 319–342, 1988.
[27] J. Wu, A. Guo, V. S. Sheng, P. Zhao, and Z. Cui, "An active learning approach for multi-label image classification with sample noise," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 3, article 1850005, 2018.
[28] J. Zhu, H. Wang, B. K. Tsou, and M. Y. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 6, pp. 1323–1331, 2010.
[29] D. Pohl, A. Bouchachia, and H. Hellwagner, "Batch-based active learning: application to social media data for crisis management," *Expert Systems with Applications*, vol. 93, pp. 232–244, 2018.
[30] R. Wang, X. Z. Wang, S. Kwong, and C. Xu, "Incorporating diversity and informativeness in multiple-instance active learning," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, pp. 1460–1475, 2017.
[31] S. Patra, K. Bhardwaj, and L. Bruzzone, "A spectral-spatial multicriteria active learning technique for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 12, pp. 5213–5227, 2017.
[32] J. Long, J. Yin, E. Zhu, and W. Zhao, "A summary of active learning research," *Journal of Computer Research and Development*, vol. 45, no. 1, pp. 300–304, 2008.