Cluster Comput
DOI 10.1007/s10586-017-1146-3

Named entity recognition based on conditional random fields

Shengli Song[1] · **Nan Zhang**[2] · **Haitao Huang**[2]

Received: 8 June 2017 / Revised: 20 August 2017 / Accepted: 23 August 2017

Abstract Named entity recognition (NER) is one of the fundamental problems in many natural language processing applications and the study on NER has great significance. Combining words segmentation and parts of speech analysis, the paper proposes a new NER method based on conditional random fields considering the graininess of candidate entities. The recognition granularity can be divided into two levels: word-based and character-based. We use segmented text to extract characteristics according to the characteristic templates which had been trained in the training phase, and then calculate $P(y|x)$ to get the best result from the input sequence. The paper valuates the algorithm for different graininess on large-scale corpus experimentally, and the results show that this method has high research value and feasibility.

Keywords Named entity recognition · Conditional random fields · Graininess

1 Introduction

World Wide Web (WWW) brings great convenience to human life, and produces huge amount of data, which contains valuable information in different forms and structures. It is necessary to deal with the data in order to excavate valuable information. Because most of these data are semi-structured, unstructured data, no formal representation method, lacking in explicit semantic information, computer cannot achieve the automatic processing of these data. The emergence of semantic web is to solve the problems of information automatic processing and high precision retrieval, which cannot be solved at present. Semantic Web introduces the representation of semantic knowledge into the web, so that the semantic web will not only be limited to the page presentation form and the content of the page information, but to increase the support of semantic information to ensure that a variety of format data on the Web page can be understood by the machine to a certain extent, and can be automatically processed by the machine. Semantic annotation is the basis of semantic inference, it is guided by ontology, adding concept examples for multimedia data, the process of data attribute and object attribute, adding semantic information on data resources, making data resources from machine readable rise to machine understandable and processing, can effectively realize the integration and sharing of multi-source data resource of cross-domain, and provide support for semantic retrieval and management of upper-level data resources.

Named entity recognition [1] (NER) is a nontrivial NLP task for recognizing those entities which have specific meaning in texts. Broadly speaking, named entity is divided into entity class (person names, place names and institution names), digital class (currency, percentage, etc.) and time class (time and dates). In different areas of application it also has a different definition, such as in the biological, medical, pharmaceutical and other fields, it usually refers to the medical terms, drug nouns and other recognition. Entity recognition is a very important subtask of Information extraction and finds its applications in information retrieval, machine translation and other higher natural language processing (NLP) applications such as coreference resolution [2].

✉ Shengli Song
shlsong@xidian.edu.cn

[1] Software Engineering Institute, Xidian University, Xi'an 710071, China

[2] School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Published online: 08 September 2017

NER can be defined as identifying and categorizing certain types of data [3] and becomes a challenging learning problem. On the one hand, in most languages and domains, there is only a very small amount of supervised training data available. On the other hand, there are few constraints on these kinds of words that can be named, so generalizing from this small sample of data is difficult [4]. NER usually includes two tasks: one is to find the named entity, that is, to determine whether words or a combination of words is a named entity; the other task is to mark the type of named entity, that is, to determine the named entity marked as a type of entity recognition. As a fundamental technology of NLP [5], the study on NER has great significance. However, because of the characteristics of Chinese lexical and syntax, the development of Chinese NER is very slow and there are some key problems still need to be resolved.

Rules-based and statistic-based are the two main methods used in NER. The method which is based on the rules can represent knowledge visually and naturally, closest to the human's way of thinking. This method has proved conducive to reasoning and high accuracy. However, they rely heavily on the speech way, the domain, and the text formatting. From this perspective, it has a significant shortcoming in portability. Besides, the process of preparing rules costs a lot of human resources, but the results are not satisfying. The method which is based on statistical using existing mature mathematical models has good flexibility and robustness. Without too much human intervention, the results proved to be more objective. Due to the lack of data, using large amount corpus training is very important [6].

In this paper, we present a Chinese NER method that focused on person names, place names and institution names, based on CRFs and differentiated from different recognition granularities. The conditional random field (CRF) model can overcome the independence hypothesis of the hidden Markov model (HMM), also can solve the problem of inductive bias of the maximum entropy Markov model, and use a global optimization method, which is more and more widely used. In our proposed method, we divide the recognition granularities into two sizes: character-based and word-based. We use different characteristic templates and characteristic parameters for each granularity to verify the result of the method. Experiments show that the NER method has a good effect.

The paper consists of six sections, and is organized as follows: Sect. 1 introduces the background and the meaning of the paper. Section 2 discusses the related works. Section 3 presents the relevant knowledge of CRFs. Section 4 proposes a method for the NER based on CRFs. Section 5 is about the three experiments, in which the experimental results are analyzed. Finally, Sect. 6 analyzes existing deficiencies and prospects for future research.

2 Related works

Since the twenty-first century, the rapid development of the Internet information industry has made the amount of information increase exponentially. Due to a large amount of information on the Internet, there are so many severe tests in information extraction and information processing that automation technologies and tools became more and more urgent. Under such circumstances, many emerging technologies in information processing fields had emerged, such as information extraction, information retrieval, and machine translation, etc. NER was one of the most important branches of these NLP technologies.

NER plays an important role in information retrieval and extraction [7], question answer [8], machine translation [9] and so on. Lots of works have been done on NER. The approaches to NER can be classified into two sets, rule-based approaches and machine learning-based approaches [10]. Since rules were induced manually, such kind of approach is time-consuming and expensive. As the large NE tagged corpora are becoming available, machine learning based approach have been received more and more attentions [11].

Yarowsky et al. [12] used word alignment on parallel corpora to induce several text analysis tools from English to other languages for which such resources are scarce. An NE tagger was transferred from English to French and achieved good classification accuracy. However, Chinese NER is more difficult than French and word alignment between Chinese and English is also more complex because of the difference between the two languages.

In the early Chinese NER is based on heuristic rules methods. They have obtained the certain effect in smaller test set but not worked good on large-scale data set because it's not feasible to give the uniform rules of names recognition of flexible named entity on large-scale data set.

Some approaches have exploited Wikipedia as an external resource to generate NE tagged corpus. Kim et al. [13] build on prior work utilizing Wikipedia metadata and show how to effectively combine the weak annotations stemming from Wikipedia metadata with information obtained through English-foreign language parallel Wikipedia sentences. The combination is achieved using a novel semi-CRF model for foreign sentence tagging. The model outperforms both standard annotation projection methods and methods based solely on Wikipedia metadata. XLADA does not leverage Wikipedia because its content is poor in some languages like Chinese.

Fu et al. [14] presents an approach to generate large-scale Chinese NER training data from an English-Chinese discourse level aligned parallel corpus. It first employs a high-performance NER system on one side of a bilingual corpus. And then, it projects the NE labels to the other side according to the word level alignment. At last, it selects

labeled sentences using different strategies and generates an NER training corpus. This approach can be considered as passive domain adaptation while XLADA is active learning framework that filters out the auto-labeled data and selects the most informative training sentences.

Muslea et al. [15] introduced Co-Testing, a multi-view active learning framework, where two models are trained on independent and sufficient sets of features. The most informative sentences are the points of disagreement between the two models that could improve their performance and a human judge is asked for labeling them. On the other hand, XLADA looks for the most informative sentences for the target model and we don't have judges.

Jones et al. [16] adapted semi-supervised learning Co-EM to information ex-traction tasks to learn from both labeled and unlabeled data that makes use of two distinct feature sets (training document's noun phrases and context). It is interleaved in the supervised active learning framework Co-Testing. XLADA di ers in that cross-lingual label propagation on a parallel corpus is interleaved for automatic annotation instead of using Co-EM approach and that it adopts an unsupervised active learning strategy.

XLADA is more practical than the framework proposed by Li et al. [17] that depends on cross-lingual features extracted from the word-aligned sentence pair in training the target language CRF model. Hence, it isn't possible to extract named entities from a sentence in the target language unless it is aligned with a sentence in the source language.

Paliouras et al. use the C4.5 algorithm to implement a NER system that shows better performance than rule-based systems. Decision tree is a common learning method, which is based on the example of annotation corpus and constructs a tree to classify. Each of the nodes on the decision tree contains an attribute in the instance, which is then sorted by the property. The classification of an instance is equivalent to starting from the head node of the tree, judging by each attribute, and finally finding the process of the root node.

Different from most European languages, there is no space to mark word boundary between Chinese characters, so Chinese word segmentation (CWS) is the first step for Chinese language processing. From another point that there is no capitalization information to indicate entity boundary, which makes Chinese NER (NER) more difficult than European languages [18].

3 CRFs model

Models for many natural language tasks benefit from the flexibility to use overlapping, non-independent features. For example, the need for labeled data can be drastically reduced by taking advantage of domain knowledge in the form of word lists, part-of-speech tags, character n-grams, and capitalization patterns. While it is difficult to capture such inter-dependent features with a generative probabilistic model, conditionally-trained models, such as conditional maximum entropy models, handle them well. There has been significant work with such models for greedy sequence modeling in NLP [19].

The CRFs model [20], which is proposed by Lafferty in 2001, is a typical discriminant probability non-direction graph learning model based on the maximum entropy model and hidden Markov model, which focuses on the problem of serialization labeling. It is modeled on the target sequences based on the observational sequences, i.e. Enter the observation sequences to be labeled, calculate the joint probability distributions of the whole sequence, and obtain the best label sequences. The conditional-based model has the same advantages of the discriminant model while considering the transfer probability among the context tags. The characteristics of being in the form of serialization to optimize global parameter and to decode solve the problem of labeling bias [21] which is difficult to be avoided for other discriminant models (such as the maximum entropy Markov model).

CRFs are undirected graphical models, a special case of which correspond to conditionally-trained finite state machines. While based on the same exponential form as maximum entropy models, they have efficient procedures for complete, non-greedy finite-state inference and training. CRFs have shown empirical successes recently in POS tagging, noun phrase segmentation and CWS [19].

CRFs can be used in natural language processing tasks such as sequence marking, data segmentation, block analysis, and so on. In CWS, Chinese NR, ambiguity resolution, and other natural language processing tasks have good performance and widespread application.

3.1 Conditional random fields

The most popular class of probabilistic structured output methods are CRFs [22]. It is a kind of discriminant probability models which is well suited to sequence analysis. CRFs predict the probability of output sequence by giving an input sequence. With input sequence $X\,(X_1, X_2, \ldots, X_n)$, we define $Y\,(Y_1, Y_2, \ldots, Y_n)$ as the semantic annotation results sequence of random variables and $P\,(X\,|Y\,)$ as the probability distribution of annotation results. CRFs describes the undirected graphical model $G = (V, E)$, Where V is the set of nodes, E is the set of edges. Values of all nodes in G are in the set$X = \{X_v\,|v \in V\,\}$, and the annotation results are in the set$Y = \{Y_v\,|v \in V\,\}$.

Because of the CRFs' characteristic of non-direction, it is difficult to guarantee that the conditional probability of each node obtained from its adjacent point is consistent with the one obtained from other nodes. Therefore, the representation of the joint probability needs to find out the product of a series

of local functions from a set of principles of conditional Independence. The simplest local function is a potential function defined on the maximal connected subgraphs and is a function of strictly positive real values.

For each input sequence $X\,(X_1, X_2, \ldots, X_n)$ of length n, the probability of output sequence $Y\,(Y_1, Y_2, \ldots, Y_n)$ is defined by the Formula (1).

$$p\,(y\,|x\,) = \frac{1}{Z\,(x)} \exp\left(\sum_{e\in E}\sum_{i}\lambda_i t_i\,(e, y\,|_e\,, x) + \sum_{v\in V}\sum_{k}\mu_k s_k\,(v, y\,|_v, x\,)\right) \tag{1}$$

where $Z\,(x)$ is a normalized distribution function, causing all of the output sequence probabilities to be 1. $Z\,(x)$ is the Formula (2).

$$Z\,(x) = \sum_{y}\left(\exp\left(\sum_{e\in E}\sum_{i}\lambda_i t_i\,(e, y|_e, x) + \sum_{v\in V}\sum_{k}\mu_k s_k\,(v, y|_v, x)\right)\right) \tag{2}$$

where $y|_e$ and $y|_v$ represent the edges and nodes of the undirected graph consisted by annotation sequence, respectively. t_i represents the transfer characteristic function of the edge e, s_k represents the state characteristic function of the nodes v, μ_k and λ_i represent the weights of a characteristic of nodes and edges.

The characteristic functions of edges can be defined as the characteristics between two consecutive nodes in CRFs so that the Formula (1) can be rewritten as:

$$p\,(y|x) = \frac{1}{Z\,(x)} \exp\left(\sum_{i=1}^{n}\left(\sum_{j}\lambda_j t_j\,(e, y_{i-1}, y_i, x) + \sum_{k}\mu_k s_k\,(v, y_i, x)\right)\right) \tag{3}$$

where n represents the length of sequence X, y_i represents the annotation result of the ith element in X.

3.2 Characteristic function

Before defining the characteristic function, we need to construct a characteristic collection of real numeric, $b\,(x, i)$, on an observation sequence to describe the empirical distribution characteristics of the training data. It is defined as:

$$b\,(x, i) = \begin{cases} 1 & \text{if the observed value x at the position i is the last name word} \\ 0 & \end{cases} \tag{4}$$

When each characteristic function is represented as an element in a feature sequence collection $b\,(x, i)$, if the current state (State characteristic function) or the previous state with the current state (the transfer characteristic function) have a specific value, all characteristic functions are real values. The value of the state characteristic function and the transfer characteristic function can be expressed as:

$$t\,(y_{i-1}, y_i, x, i) = \begin{cases} b\,(x, i) & y_{i-1} = B, y_i = M \\ 0 & otherwise \end{cases} \tag{5}$$

$$s\,(y_i, x, i) = \begin{cases} b\,(x, i) & y_i = B \\ 0 & otherwise \end{cases} \tag{6}$$

To unify the state characteristic function and the transfer characteristic function, we can rewrite the state characteristic function to the following form:

$$s_k\,(y_i, x, i) = s_k\,(y_{i-1}, y_i, x, i) \tag{7}$$

It can be either a transfer characteristic function or a state characteristic function, which is uniformly represented by two characteristic functions by $f_k\,(y_{i-1}, y_i, x, i)$

$$F_k\,(y, x) = \sum_{i}^{f_k(y_{i-1}, y_i, x, i)} \tag{8}$$

Thus, under the condition of observing sequences, the conditional probability of the corresponding label sequences can be rewritten as follows:

$$p\,(y\,|x\,) = \frac{1}{Z\,(x)} \exp\left(\sum_{k}\lambda_k F_k\,(y, x)\right) \tag{9}$$

where is $Z\,(x)$ the normalized factor:

$$Z\,(x) = \sum_{y}\exp\left(\sum_{k}\lambda_k F_k\,(y, x)\right) \tag{10}$$

The characteristic of CRFs is defined throughout the observation sequence, but in practices, the observation value in a suitable window around the current position is sufficient to be the condition.

3.3 Parameter estimation

For the CRFs model, the main task of modeling is to estimate the weight λ of the characteristic from the training data, and

the logarithmic maximum likelihood parameters estimate the value of $\lambda = (\lambda_i, \lambda_2, \cdots, \lambda_n)$ from the independent training data, the parameters λ_i are estimated using L-BFGS method.

Assuming the given training set $D = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_\Gamma, Y_\Gamma)\}$, the maximum likelihood method is used to estimate the parameters according to the maximum entropy model. For the conditional probability model $p(y|x, \lambda)$, the logarithmic likelihood function of the training set D is:

$$L(\lambda) = \log \prod_{x,y} p(y|x, \lambda)^{\tilde{p}(x,y)} = \sum_{x,y} \tilde{p}(x, y) \log p(y|x, \lambda) \tag{11}$$

$\tilde{p}(x, y)$ is the Probability of empirical distribution for training samples. The formal formulas for the conditional probability are:

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp\left(\sum_k \lambda_k F_k(y, x)\right) \tag{12}$$

where $Z(x)$ is the normalized factor. Thus, the probability of empirical distribution and the mathematical expectation for conditional probabilities obtained by the CRFS model can be expressed as follows:

$$E_{\tilde{p}}[f_k] \stackrel{def}{=} \sum_{x,y} \tilde{p}(x, y) \sum_{i=1}^{n+1} f_k(y_{i-1}, y_i, x, i) = \sum_{x,y} \tilde{p}(x, y) F_k(x, y) = E_{\tilde{p}}[F_k] \tag{13}$$

$$E_p[f_k] \stackrel{def}{=} \sum_{x,y} \tilde{p}(x) p(y|x, \lambda) \sum_{i=1}^{n+1} f_k(y_{i-1}, y_i, x, i) = \sum_{x,y} \tilde{p}(x) p(y|x, \lambda) F_k(x, y) = E_p[F_k] \tag{14}$$

According to the logarithmic likelihood function, the corresponding parameter is obtained by the first derivative (the specific deduction procedure is omitted) can be obtained, L-BFGS is significantly more efficient than traditional iterative scaling and even conjugate gradient. This method approximates the second-derivative of the likelihood by keeping a running, finite-sized window of previous first-derivatives. L-BFGS can simply be treated as a black-box optimization procedure, requiring only that one provide the first-derivative of the function to be optimized. Assuming that the training labels on instance j make its state path unambiguous, let s(j) denote that path, and then the first-derivative of the log-likelihood is:

$$\frac{\partial L(\lambda)}{\partial \lambda_k} = E_{\tilde{p}}[F_k] - E_p[F_k] \tag{15}$$

According to the maximum entropy principle [23], the expectation of the distribution characteristic for the conditional probability model's equals the expectation of empirical distribution, and the problem of parameter estimation can be solved by the optimal method.

In the above introduction, the calculation expression of logarithmic likelihood function $L(\lambda)$ gradient is given, i.e., the mathematical expectation of empirical distribution $\tilde{p}(x, y)$ minus the mathematical expectation of conditional probability $p(y|x, \lambda)$ obtained from the model. The mathematical expectation of empirical distribution is the number of random variables (x, y) in the training data set to satisfy the characteristic constraint. The mathematical expectation calculation of conditional probability is essentially the calculating conditional probability $p(y|x, \lambda)$.

4 CRFs based NER framework

CRFs model is essentially a conditional probability model based on statistics. Given an input sequence, the CRFs model can predict the probability of the output sequence by establishing a consistent exponential model for joint probability of input sequence. In this way, eigenvalues in various states can compromise with each other, not only the advantage of the maximum entropy models can remain, but also the problem of labeling bias can be solved.

4.1 NER framework

The process of CRFs based NER can be divided into two phases. The first phase is training phase and the second is recognizing phase, as shown in Fig. 1. In training phase, we use characteristics extracted from training corpus to train characteristic templates. According to training result, we can get the weight of each characteristic. In recognizing phase, recognition is the process of using feature templates, combining the trained weight parameters, doing NER, and finally obtaining recognition results. We use segmented Chinese test text to extract characteristics according to the characteristic templates which had been trained in the first phase, then calculate $P(y|x)$ and get the best result from the input sequence. At last, we do the NE tag for the entire sentence to obtain the result.

4.2 Recognition granularity

The recognition granularity is usually divided into small granularity and large granularity, in this paper, we divide recognition granularity into two levels: word-based and character-based. The recognition granularity is a kind of named entity tagging strategy, which is composed of complex and full name or abbreviation for named entity in

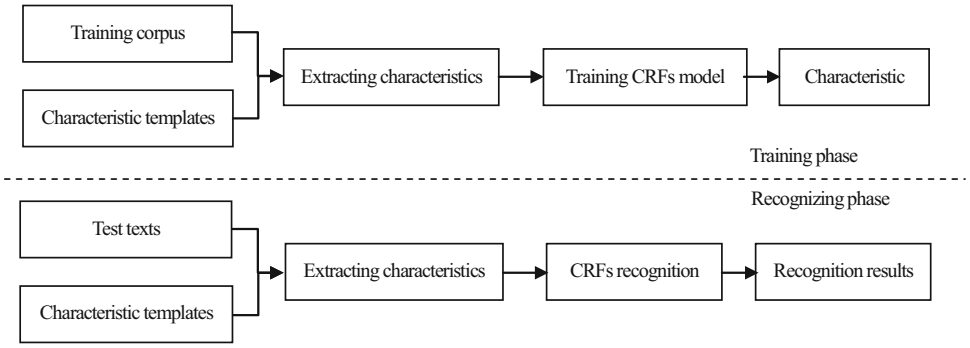

Fig. 1 NER framework based on CRFs

the process of naming entity annotation. The content of a character-based partitioning strategy is to label only the name entities that cannot be separated. It refers to a single character, and there is no obvious correlation between characters, such as the person's name: "Hu Wenbo", according to the small size of the strategy, respectively labeled "Hu" "Wen" and "Bo", which are not separated. The small granularity strategy can effectively reduce the effect of the difference of named entity composition on the recognition of named entity. With this strategy, you can get a complete named entity by combining the successive small-grained named entities in the recognition results and using the callout of the last entity as a callout for the merged entity. Large-grained strategies are mainly for entities composed of multiple words, an entity may contain multiple words, each of which has an independent meaning, such as the name of the institution: "Xi'an Local Taxation Bureau", which can be regarded as "Xi ' an + Local + Taxation Bureau", the word-based division is more accurate recognition of the entity.

One important problem in NER is which granularity is to be chosen: character-based or word-based. The accuracy of the recognition results will be largely affected because of the different recognition granularity. Now we will determine which granularity will be used in recognition according to the differences between person names, place names and institution names.

Due to that the Chinese name is many open collections, there is a concurrent phenomenon between the name and surname. If there is not special treatment, it is likely to be ambiguous with the context of the environment. The structure of Chinese names could be generalized as "family name + first name," which could be subdivided into "family name + character + character" corresponding to the names has three characters and "family name + character" corresponding to the names has two characters. The characteristics of Chinese names are the distribution of using characters loosely and widely.

In theory, all geographical names can constitute a huge limited set. However, from the practical point of view, it is very impractical to traverse all the names through an exhaustive method. What can be found through analysis is that the structure of place names could be generalized as "name + key word." The "name" refers to the name of the place, "key words" are a description of the characteristics of a place, such as "Mount Taibai" which "Taibai" is the name of the place and "Mount" is the key word. The characteristics of place names era the same as the characteristics of Chinese names. Because of the distribution of using characters loosely and widely, recognition on Chinese names and place names will be supposed to base on characters.

Institution names can be nested person names, place names or sub-institution names with a complex composition and rich structure. Each component of an institute name is not a character but the word which has its meaning, such as "The People's Republic of China". Because of the granularity of institution names is large, a word-based approach can achieve better results [24].

4.3 Characteristic templates

The quality of characteristics selection is very important in NER based on CRFs. It will largely affect the final recognition results. When recognition granularity has been selected, the characteristic templates can be defined. Characteristic template format as used herein refers to the settings from CRFs++ toolkit. The characteristic templates could be divided into unitary templates and binary templates (shown in Tables 1 and 2, the window size of templates is three by default and it will be further selected in subsequent experiments). The templates will be used to examine the characteristics between and among characters, words, and part-of-speeches.

NER uses the character as a unit and each character in the text will be identified individually. There is no part-of-speech for a single character so that only the characteristics between and among characters will be considered. Using word as a unit, except for the characteristics between and among characters and words, the characteristics between and among part-of-speeches will also be added into characteristic templates.

Table 1 Unitary template

No	Characteristic template	Meaning
1	U01:%x[−1,0]	The first word (character) on the left
2	U02:%x[0,0]	The current word (character)
3	U03:%x[1,0]	The first word (character) on the right
4	U04:%x[−1,1]	The first part-of-speech on the left
5	U05:%x[0,1]	The current part-of-speech
6	U06:%x[1,1]	The first part-of-speech on the right

As shown in Table 1, taking the word-based recognition for example, we can find that: "%x[0,0]" means the current word which is being analyzed, "%x[−1,0]" means the first word on the left of the current word, "U05:%x[0,1]" means the current part-of-speech which is being analyzed, and "%x[−1,1]" means the first part-of-speech on the left of the current.

As shown in Table 2, taking the word-based recognition for example, we can find that: "%x[−1,0]/%x[0,0]" means the relationship between the current word and the previous word, "%x[−1,1]/%x[0,1]" means the relationship between the current part-of-speech and the previous part-of-speech.

4.4 NER process

Preprocessing Preprocessing does word segmentation and POS tagging from the original text, so that the original text implied words, parts of speech and context and other characteristics can be expressed explicitly. We use the LTP platform to achieve the text of the word segmentation and POS tagging, its meaning as shown in Table 3.

Training set tagging Training set tagging is a manual way to mark the named entity of the training set to provide a model for the training of CRFs models. We take the word as the text corpus segmentation granularity, use "bio" annotation method, adopt the big granularity strategy annotation training set, and get the training set entity annotation sequence, where b (begin) denotes the beginning of the entity (left boundary), I (internal) denotes the interior and end of the entity (right boundary), and O (other) denotes words, words, and punctuation other than entities. According to the text feature of the 1998 People's Daily Corpus, select the annotation 10 class named entity, its annotation method is shown in the Table 4.

Table 2 Binary template

No	Characteristic template	Meaning
1	U10:%x[−1,0]/%x[0,0]	The first word (character) on the left/the current word (character)
2	U11:%x[1,0]/%x[0,0]	The first word (character) on the right/the current word (character)
3	U12:%x[−1,1]/%x[0,1]	The first part-of-speech on the left/the current part-of-speech
4	U13:%x[1,1]/%x[0,1]	The first part-of-speech on the right/the current part-of-speech

Table 3 Examples of the meanings of LTP POS tagging

Tag	Description	Example	Tag	Description	Example
a	Adjective	美丽	ni	Organization name	华北电管局
b	Other noun-modifier	大型, 西式	nl	Location noun	首都
c	Conjunction	和, 虽然	ns	Geographical name	北京
d	Adverb	很	nt	Temporal noun	近日, 明代
e	Exclamation	哎	nz	Other proper noun	诺贝尔奖
g	Morpheme	茨, 甥	o	Onomatopoeia	哗啦
h	Prefix	阿, 伪	p	Preposition	在, 把
i	Idiom	百花齐放	q	Quantity	个
j	Abbreviation	公检法	r	Pronoun	我们
k	Suffix	界, 率	u	Auxiliary	的, 地
m	Number	一, 第一	v	Verb	跑, 学习
n	General noun	苹果	wp	Punctuation	,。!
nd	Direction noun	右侧	ws	Foreign words	CPU
nh	Person name	李鹏, 汤姆	x	Non-lexeme	萄, 翱

Table 4 Naming entity categories and annotation methods

Named entity	Entity begining	Entity internal and entity other
工厂	B-Fac	I-Fac
地名	B-Pla	I-Pla
人名	B-Person	I-Person
坐标	B-Co	I-Co
方位趋向	B-Az	I-Az
任务	B-Task	I-Task
时间	B-Time	I-Time
日期	B-Date	I-Date
数量	B-Num	I-Num
机构	B-Institution	I-Institution

Table 5 Word segmentation, POS tagging, entity tagging

Word	POS	NE tagging
李鹏	nh	B-Person
华北电管局	ns	I-Pla
电厂	n	I-Fac
询问	v	I-Az
今年	nt	B-Time
华北电网	ns	B-Institution
首都	n	I-Pla
北京	ns	I-Pla
供电	v	I-Az

The example sentence is "李鹏向华北电管局电厂负责人详细询问了今年华北电网向首都北京供电情况".

Table 5 shows the use of LTP tools for Word segmentation and POS tagging. The resulting tagging sequence is [B-person, I-pla, I-Fac, I-az, O, B-time, b-institution, I-pla, I-az, O].

Feature template selection and feature extraction The greatest advantage of CRFs model is the ability to use characters, words, parts of speech and contextual information synthetically. When using CRFs model for naming entity recognition under the large granularity strategy, the feature selection will affect the recognition effect, and the features of the optional feature include lexical features, lexical features and contextual features. A feature template is a predefined by combination of identity features that are used to train and identify named entities in the CRFs model. After word segmentation, pos tagging and manual annotation entities, the named entity annotation is shown in the Table 5, and the words and parts of speech are chosen as the distinguishing features. If the feature%x[0, 0] represent the word 询问 in the sentence, the feature template represents the characteristics and its contents as shown in the Table 6.

Table 6 Characteristics and content of representative in feature template

Feature	Word
%x[−3,0]	李鹏
%x[−2,0]	华北电管局
%x[−1,0]	电厂
%x[0,0]	询问
%x[1,0]	今年
%x[−1,1]	华北电网
%x[−2,1]	北京
%x[−3,1]	供电

5 Experiment and evaluation

The training and recognizing phases of the CRFs based NER are completed in aid of the CRFs toolkit called CRFs++: Yet another DRF toolkit 0.53. We select the People's daily corpus (PFR) which takes the content of People's Daily in 1998 as the corpus. To cope with the using of CRFs++, we need to shape the format of the corpus.

5.1 Evaluation method

The essence of the recognition model training is to obtain the optimization parameters of the CRFs model. After getting the trained CRFs model, the test set is used to evaluate the performance of the model to judge the identification method. The performance of the CRFs model after training is evaluated by the accuracy rate (p), recall rate (R) and F1-measure (FL).

Where the accuracy rate indicates the proportion of all recognized named entities that are correctly identified:

precision = correctly recognized NE/recognized NE

$$p = \frac{A}{A + B} \tag{16}$$

The recall rate indicates that the correct recognition of the named entity is the proportion of all named entities. All named entities refer to standard named entities that are artificially standard in the test corpus.

recall = correctly recognized NE/golden NE

$$R = \frac{A}{A + C} \tag{17}$$

Table 7 Results for recognition

	Correct target	Incorrect target
Recognized	A	B
Unrecognized	C	D

For NER tasks, if the recall rate is higher, it indicates that the correct recognition of the named entity is more, thus causing all the identified naming entities to increase, it is possible to lead to a decline in accuracy. Therefore, the F-1 standard is used to balance accuracy and recall.

F-1 measure = 2 × Precison × Recall/(Precision+Recall)

$$F1 = \frac{P \times R \times 2}{P + R} \tag{18}$$

where A, B, C and D represent the number of corresponding samples, their specific meanings are shown in Table 7.

5.2 Phase I

This experiment is designed to measure which window size of the characteristic templates will lead to a better performance of the recognition results, complete the training of characteristic templates, and obtain the weight of each characteristic function. To objectively reflect the recognition effect, we decompose the corpus in Jan and Feb 1998 into eight pieces. The four pieces will be used for training the characteristic templates with different window sizes, and the others will be used for selecting the window size. The experiment results are shown in Tables 4 and 5 by using accuracy, recall, and F-measure as the evaluation index.

The final experimental results are shown in Tables 8 and 9.

As shown in Table 4, when the window size of the character-based characteristic template is 7, which the range is $(-3, -2, -1, 0, 1, 2, 3)$, the recognition results will be best. As shown in Table 5, when the window size of the word-based characteristic template is 5, which the range is $(-2, -1, 0, 1, 2)$, the recognition results will be best.

5.3 Phase II

We decompose the corpus in Mar 1998 into four pieces used for testing the recognition performance for recognizing person names, place names and institution names. To verify the effect of the granularities, we will use the character-based characteristic template (size 7) and the word-based characteristic template (size 5) for recognizing person names, place names and institution names. The experimental results are shown in Table 10 using accuracy, recall, and F-measure as the evaluation index.

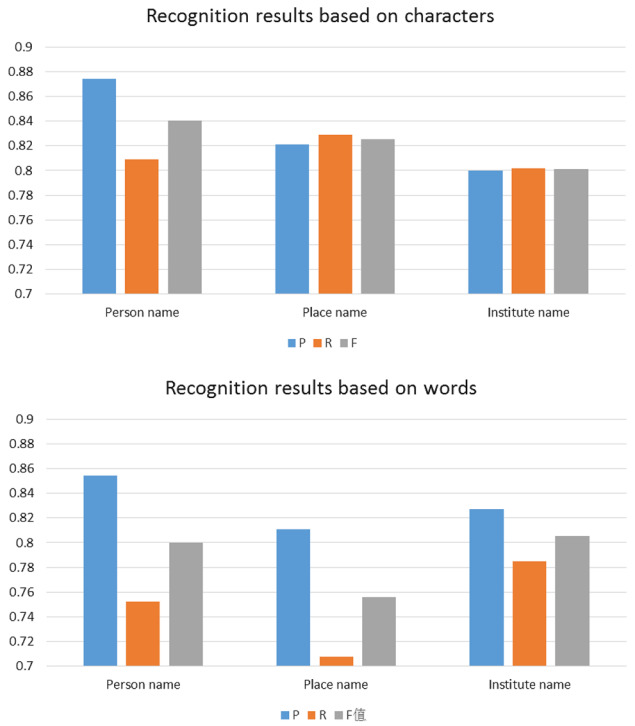

Fig. 2 The comparison of recognition results among different recognition granularities

In Fig. 2, the two images are based on the result of word and word-based entity recognition, which can be seen that the recall rate of the method based on word segmentation is significantly higher than that based on the word segmentation. In recognition of person names and place names, the accuracy and f-values of the method are all higher than the NER method based on the words, but the accuracy and f-values of the organization name are higher than those of the unit. This indicates that the greater the recognition granularity, the higher the accuracy of the recognition of the mechanism name, but the larger the recognition granularity will result in the decrease of recall rate.

As apparent from the experiment results, when recognizing person names and place names, using the character-based method is better in accuracy, recall and F-measure, which reach up to 0.8745, 0.8089, 0.8404 and 0.8212, 0.8291, 0.8251. This shows that the character-based recognition is more suitable for the recognition of person names and place names.

When recognition institution names using the word-based method, the accuracy, recall and F-measure have reached up to 0.8272, 0.7850 and 0.8055, which is higher in accuracy and F-measure compared to character-based method but is slightly lower in recall rate.

In summary, the word-based method is better and more suitable for recognizing institution names. There are some shorthands for complex organization names. For example,

Table 8 The selection for window sizes of the character-based characteristic template

	Size: 5 characters			7			9		
	P	R	F	P	R	F	P	R	F
1	0.8366	0.7654	0.7954	0.8734	0.8076	0.8392	0.8233	0.8058	0.8145
2	0.8145	0.7956	0.8049	0.8252	0.8280	0.8266	0.7960	0.8031	0.7995
3	0.7977	0.7662	0.7816	0.8401	0.8118	0.8257	0.7800	0.7989	0.7893
4	0.8234	0.7847	0.8036	0.8325	0.8254	0.8289	0.8005	0.8153	0.8078
Avg	0.8181	0.7780	0.7964	0.8428	0.8182	0.8301	0.7999	0.8058	0.8028

Table 9 The selection for window sizes of the word-based characteristic template

	Size: 3 words			5			7		
	P	R	F	P	R	F	P	R	F
1	0.8135	0.7368	0.7945	0.8559	0.7534	0.8014	0.8025	0.7075	0.7520
2	0.8071	0.7017	0.7507	0.8218	0.7368	0.7770	0.7946	0.6526	0.7166
3	0.7923	0.6913	0.7384	0.8272	0.7850	0.8055	0.8104	0.6501	0.7215
4	0.8234	0.7120	0.7637	0.8364	0.7553	0.7938	0.8023	0.6859	0.7395
Avg	0.8090	0.7105	0.7618	0.8353	0.7576	0.7944	0.8024	0.6740	0.7324

Table 10 Recognition results

	The character-based NER			The word-based NER		
	P	R	F	P	R	F
Person name	0.8745	0.8089	0.8404	0.8545	0.7524	0.8002
Place name	0.8212	0.8291	0.8251	0.8109	0.7079	0.7559
Institute name	0.8000	0.8018	0.8009	0.8272	0.7850	0.8055

"NDRC" is short for "National Development and Reform Commission," it is no longer a complex structure consisting of many words. In this case, using word-based method will cause recognition frailer.

5.4 Phase III

The superiority of CRFs phase compared to the HMM and maximum entropy Markov model [25] is illustrated in the previous theory, and the conclusion needs to be validated in the third phase experiment. This experiment adopts the data set used in the second stage, and the feature template is kept unchanged, and the HMMs [26] uses the first-order HMM. For the names of persons and places, the character-based recognition granularity is adopted, while for the institution names, the word-based recognition granularity is adopted. The testing results are shown in Table 11.

In Table 11, compared with other two methods, CRF is better in accuracy, recall and F-measure. As can be seen, more intuitively, Fig. 3 shows the comparison of NER efficiency among three probabilistic models, especially for the part of accuracy. The CRFs reaches almost 0.85, which is the highest among these three methods. For HMM, the average value of accuracy, recall, and F-measure is respectively 0.7046, 0.6701, and 0.6895. For MEM, the average value of accuracy, recall, and F-measure is respectively 0.7884, 0.7509 and 0.7691. As for CRFs, the average value of accuracy, recall, and F-measure is respectively 0.8410, 0.8077 and 0.8237. According to the experimental results, CRFs is the best one to solve the problem.

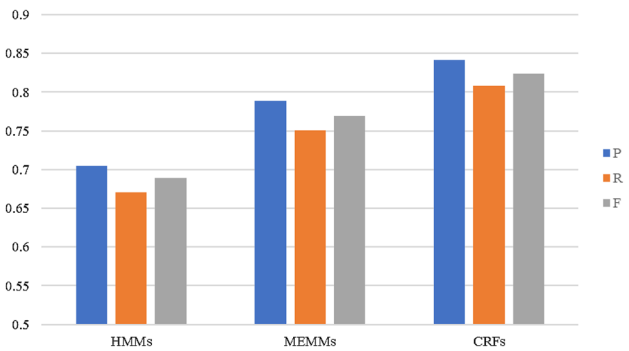

Fig. 3 The comparison of NER efficiency among three probabilistic models

HMMs is the representative of the generative model, because it has strict independence hypothesis when modeling, and cannot effectively integrate many kinds of information. Especially it is powerless for long-distance information, making it unsuitable to deal with the problem of sequence labeling. MEMMs is a conditional probability model, but it cannot solve the tag bias problem, making its recognition effect less than CRFs.

Table 11 The comparison of NER results among HMM, MEMM and CRFs models

	HMM			MEM			CRFs		
	P	R	F	P	R	F	P	R	F
Person name	0.7139	0.6531	0.6903	0.8221	0.7712	0.7958	0.8745	0.8089	0.8404
Place name	0.7096	0.6846	0.6969	0.7783	0.7575	0.7678	0.8212	0.8291	0.8251
Institute name	0.6904	0.6725	0.6813	0.7648	0.7239	0.7438	0.8272	0.7850	0.8055
Avg	0.7046	0.6701	0.6895	0.7884	0.7509	0.7691	0.8410	0.8077	0.8237

6 Conclusion

NER has always been a very important and relatively difficult subject in the fields of NLP. The common NER methods are based on the low-level information such as morphological, part of speech and do not involve the syntactic and semantic information of the high-level, thus introducing the complete information into NER is the inevitable trend for the development of the NER technology. This paper proposes a CRFs model considering the graininess in NER, experiments show that CRFs has high precision and recall rate. The improvement of NER precision will help improve the precision of information extraction and relation inference. This paper has analyzed different recognition granularities, designed a recognition model based on CRFs, and has achieved good results. However, there is still a long road in NER and many areas need to be improved. The vision for the future works is introducing more grammar and semantic information, removing dependencies in data training so that the model has better universality.

References

1. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
2. Bhargava, R., Vamsi, B., Sharma, Y.: Named entity recognition for code mixing in indian languages using hybrid approach. Facilities **23**, 10 (2016)
3. Şeker, G.A., Eryiğit, G.: Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content. Semant. Web **8**(5), 625–642 (2017)
4. Lample, G., Ballesteros, M., Subramanian, S., et al.: Neural architectures for named entity recognition (2016). arXiv:1603.01360
5. Chowdhury, G.G.: Natural language processing. Ann. Rev. Inf. Sci. Technol. **37**(1), 51–89 (2003)
6. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**(6), 1554–1563 (1966)
7. Müller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. **2**(11), e309 (2004)
8. Lehnert, W.G.: The Process of Question Answering: A Computer Simulation of Cognition. Lawrence Erlbaum Associates, Hillsdale (1978)
9. Cho, K., Van Merriënboer, B., Gulcehre, C., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation (2014). arXiv:1406.1078
10. Goldberg, D.E., Holland, J.H.: Genetic algorithms and machine learning. Mach. Learn. **3**(2), 95–99 (1988)
11. Suxiang, Z.: Based cascaded conditional random fields model for Chinese Named Entity recognition In: Signal Processing. ICSP 2008. 9th International Conference on. IEEE, pp. 1573–1577 (2008)
12. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Human Language Technology Conference, pp. 109–116 (2001)
13. Kim, S., Toutanova, K., Yu, H.: Multilingual named entity recognition using parallel data and metadata from Wikipedia. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (2012)
14. Fu, R., Qin, B., Liu, T.: Generating Chinese named entity data from a parallel corpus. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, pp. 264–272 (2011)
15. Muslea, I., Minton, S., Knoblock, C.A.: Active learning with multiple views. J. Artif. Intell. Res. **27**, 203–233 (2006)
16. Jones, R., Ghani, R., Mitchell, T., Rilo, E.: Active learning for information extraction with multiple view. In: Proceedings of the European Conference in Machine Learning (ECML 2003), vol. 77, pp. 257–286 (2003)
17. Li, Q., Li, H., Ji, H.: Joint bilingual name tagging for parallel corpora. In: Proceedings of CIKM 2012 (2012)
18. Mao, X., Dong, Y., He, S., et al.: Chinese word segmentation and named entity recognition based on conditional random fields. In: IJCNLP, pp. 90–93 (2008)
19. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL -Volume 4. Association for Computational Linguistics, vol. 2003, pp. 188–191 (2003)
20. Zhao, H., Huang, C.N., Li, M.: An improved Chinese word segmentation system with conditional random field. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: July, 1082117 (2006)
21. Joseph, K.: Bradley and Carlos Guestrin. Learning tree conditional random Felds. In: International Conference on Machine Learning (ICML 2010) (2010)
22. Tran, T., Phung, D., Bui, H., et al.: Hierarchical semi-Markov conditional random fields for deep recursive sequential data. Artif. Intell. **246**, 53–85 (2017)
23. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Comput. Linguist. **22**(1), 39–71 (1996)
24. Sutton, C., McCallum, A.: An introduction to conditional random fields, pp. 21–23 (2010). arXiv:1011.4088v1 [stat.ML]
25. McCallum, A., Freitag, D., Pereira, F.: Maximum entropy Markov models for information extraction and segmentation. In: Proceedings of the 17th International Conference on Machine Learning (ICML' 2000), pp. 591–598 (2000)
26. Seymore, K., McCallum, A., Rosenfeld, R.: Learning hidden Markov model structure for information extraction. In: Proceedings of AAAI'1999 Workshop on Machine Learning for Information Extraction (1999)

Shengli Song is currently an associate professor with Xidian University, China. He received his Ph.D. degree in Computer Science and Technology from Xidian University, Xi'an, China, in 2011. His current research interests include semantic computing, text analytics and intelligent system.

Nan Zhang is currently a master student majoring in Software Engineering in Xidian University. Her research interest is text analytics and knowledge reasoning.

Haitao Huang is currently a master student majoring in Software Engineering in Xidian University. His research interest is mobile data analysis.