Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 480879, 6 pages
http://dx.doi.org/10.1155/2015/480879

Research Article

Automatic Recognition of Chinese Personal Name Using Conditional Random Fields and Knowledge Base

Chuan Gu,[1] Xi-ping Tian,[1] and Jiang-de Yu[2]

[1]*School of Software Engineering, Anyang Normal University, Anyang, Henan 455000, China*
[2]*School of Computer and Information Engineering, Anyang Normal University, Anyang, Henan 455000, China*

Correspondence should be addressed to Xi-ping Tian; txp_696@163.com

Received 24 February 2015; Revised 10 May 2015; Accepted 18 May 2015

Academic Editor: Chih-Cheng Hung

According to the features of Chinese personal name, we present an approach for Chinese personal name recognition based on conditional random fields (CRF) and knowledge base in this paper. The method builds multiple features of CRF model by adopting Chinese character as processing unit, selects useful features based on selection algorithm of knowledge base and incremental feature template, and finally implements the automatic recognition of Chinese personal name from Chinese document. The experimental results on open real corpus demonstrated the effectiveness of our method and obtained high accuracy rate and high recall rate of recognition.

1. Introduction

Named Entity refers to identified entity using name, such as person name, place name, and organization name [1]. Named Entity Recognition is the process of identifying name or symbol of specified events from natural language documents. At present, there have been a lot of works on Named Entity Recognition, especially on the person name, place name, and organization name. Chinese personal names account for a large proportion in the named entity, so Chinese personal name recognition is a subquestion and also is the spot of Chinese Named Entity Recognition. There are different forms of Chinese names, and the choice of words forming name is very optimal, so automatic Chinese personal name recognition is one of the difficulties for Named Entity Recognition.

In the past several decades, many effective Chinese names recognition algorithms have been proposed, which can be classified into two categories: methods based on rules and methods of statistics. The method based on rules commonly match artificial rules to recognize Chinese names [2]. The algorithm integrating boundary templates and local statistics has been developed to identify Chinese names, which identified possible names using boundary templates extracted from corpus with frequency and then corrected the recognition results by local statistics of context and several heuristic rules [3]. The experimental result proved effectiveness of the proposed algorithm in literature and high speed for small-scale corpus, but there were drawbacks, such as the fact that the coverage of rules was limited, the fact that the design was difficult, and the fact that transportability was poor. The methods of statistics usually are used to recognize Chinese names by statistical language model based on counting Chinese names and their contexts of the large-scale language corpus [4], so this method has a good portability and is the most common approach. Two other methods based on HMM (hidden Markov model) and based on MEM (maximum entropy model) have been proposed to identify Chinese names, in which problem of person name recognition is converted into a sequence annotation problem, but the strict independence hypothesis must set in HMM, while tag offset must be solved in MEM [5, 6].

CRF short for conditional random fields is a new probability graph model, which does not require strict independence assumption, can easily contain a variety of characteristics in the model, and can successfully solve the problem of the label position, so it has been widely applied to natural language processing, such as Chinese word segmentation, speech tagging, and Named Entity [7]. CRF model integrating

many features can significantly improve the performance of Named Entity Recognition [8]; therefore we present a method of Chinese personal name recognition based on CRF and knowledge base according to the characteristics of Chinese personal name. The method adopts Chinese character as processing unit to build multiple features CRF model, uses selection algorithm of the person name knowledge base and incremental feature template to select features, and finally implements the automatic person name recognition from Chinese document.

The remaining of this paper is organized as follows. Section 2 analyzes the characteristics of Chinese personal name. In Section 3, we describe CRF and its application on feature annotation. Section 4 builds the knowledge base according to the features of Chinese personal names. Section 5 describes the recognition process and illustrates the details of the key content. Section 6 describes our experiment based on our training corpus and open test corpus. We make a conclusion of our work and present direction we are forwarding in Section 7.

2. The Characteristics of Chinese Personal Names

The names of Chinese people have their own tradition and characteristics. Unlike westerners, the family name in China is put first, followed by the given name. The family name and the given name usually contain one or two Chinese characters in order to avoid confusion [9]. After statistical analysis on a large number of corpuses, we found that the main characteristics of Chinese personal names are manifested in the following aspects.

(1) The Usage of Characters in Family Names. Altogether, some 22,000 family names have been used in China in the past, some of them have been reserved, and only 3,500 are commonly used nowadays, so usage of words in family names is limited. Among all the family names, 100 common ones cover almost 87% of the total population. The most popular three are Li, Wang, and Zhang, respectively, occupying about 7.9%, 7.4%, and 7.1% of the whole Chinese population, so a few words are the most popular, such as Zhang, Wang, Li, Zhao, and Liu [10].

(2) The Usage of Characters in Given Names. The characters used as name are many, but Chinese names are meant to convey special meaning, and the given names often express the best of wishes for the newborn. Therefore, they are relatively concentrated. For example, the names formed by the characters with high frequency occupied 30.1% of the whole statistics' 3345 names.

(3) The Characteristics of Personal Name's Boundary Terms. The boundary terms were formed by adjacent character with personal name in the text, which include appellation words (such as Mr, Ms, and Dr), action verbs (e.g., said and notes), and punctuation (such as " , "). We can use these boundary terms to determine the boundary of names, so these are indexes for identifying person names.

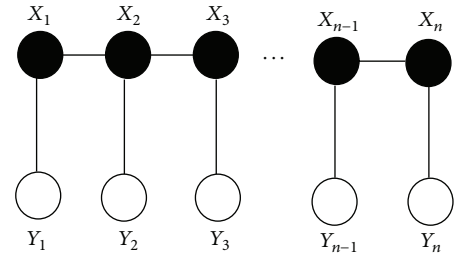

Figure 1: The structure of CRF.

3. CRF

CRF refers to predicting the most likely marked sequence by defining label sequences and observing conditional probability ($P(Y \mid X)$) of these sequences. The basic problem of recognizing Chinese personal names from Chinese text is to gain words tagging to serialized data. Therefore, it is very suitable to select CRF for recognizing Chinese personal names. CRF used to simulate the sequence data is a simple chain graph or chart, as shown in Figure 1.

The recognition algorithm of Chinese personal name is as follows.

Firstly, we select our using feature function, which is a key process of model building and affects directly the recognition performance of the model. On the basis of the analysis on the characteristics of the Chinese people names, we introduce itself features of word and features of context to select the feature function ($f_k(y_{t-1}, y_t, x, t)$) in our model.

Secondly, weight estimation of feature function would be calculated, which refers to calculating weight (λ_j) of each feature function ($f_k(y_{t-1}, y_t, x, t)$), so this progress is called training process of CRFs model. We choose maximum likelihood estimate method to complete trainings and get the weight of characteristic function in the form of iteration.

Finally, the recognition process of CRFs model would be accomplished. If the observed sequence of segmentation strings $X = (x_1, x_2, \ldots, x_T)$ is given, the condition sequence maximum probability $Y = (y_1, y_2, \ldots, y_T)$ would be gotten in the process.

4. Knowledge Base of Chinese Personal Names

Experimental results indicate that global features are very effective in improving the performance of Named Entity Recognition [11]. The local features, such as word and term, can be directly extracted from the training and test corpuses, while the global feature could be extracted based on related knowledge base. In our knowledge base, some information should be included, such as the table used to record usage of characters in family names, the table used to record usage of characters in given names, the table used to record left boundary words of person names, and the table used to record right boundary words of person names. In corpus pretreatment, the global feature would be extracted by querying our knowledge base. We construct knowledge base of names in this paper according to tagged corpus of People's Daily published between January 2000 and June 2000.

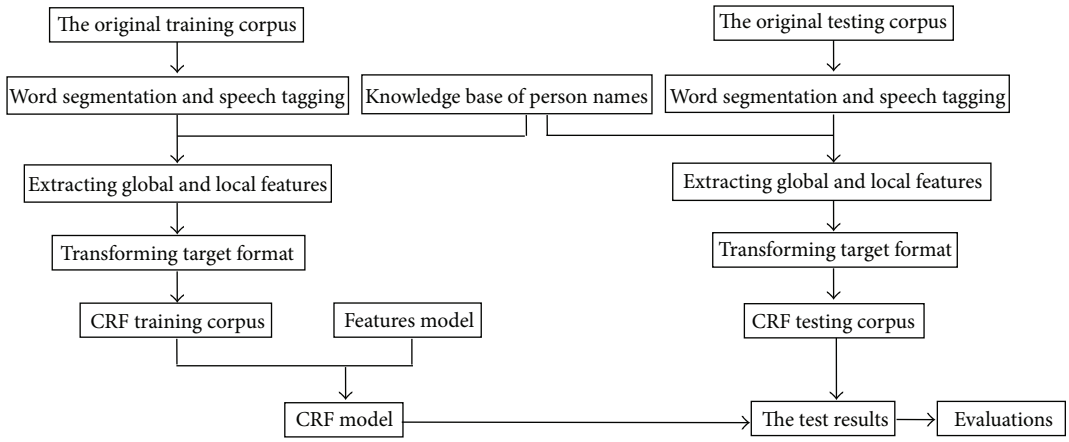

Figure 2: The process of Chinese names recognition.

(1) The table used to record usage of characters in family names: the 317 commonest family names have been included in this table, among which the number of single characters is 312 and the number of hyphenated characters is 4.

(2) The table used to record usage of characters in given names: the 1132 common characters of given names have been included in this table.

(3) The table used to record left boundary words of person names, which includes 1380 words: the words would fall into this table if the probability of left boundary words is greater than the specified threshold. The formulation for calculating probability is given. If we use Left_number to represent the number of times those words are seen as left boundary words and use Total_number to show the number of total times those words appeared, the probability of left boundary words would be gotten as follows:

$$\text{the probability of left boundary words} = \frac{\text{Left\_number}}{\text{Total\_number}} \times 100\%. \quad (1)$$

(4) The table used to record right boundary words of person names, which includes 1380 words. The words would fall into this table if the probability of right boundary words is greater than the specified threshold. The formulation for calculating probability is given. Of course, if we use Right_number to represent the number of times those words are seen as right boundary words and use Total_number to show the number of total times those words appeared, the probability of right boundary words would be gotten as follows:

$$\text{the probability of left boundary words} = \frac{\text{Right\_number}}{\text{Total\_number}} \times 100\%. \quad (2)$$

5. Chinese Personal Names Recognition by Combining CRF with Knowledge Base

5.1. Recognition Process. When using CRF model name recognition, the global features, which are extracted according to knowledge base, are introduced to improve the recognition performance of Chinese names. Because the experiment is accomplished using toolkit (CRF++), the format of corpus must be transformed in this paper. The process is shown in Figure 2.

5.2. Tagging Sets. In our method, we transformed the Chinese names recognition problem into an equivalent sequence tagging problem, so the tagging set is used to record the result of sequence annotation, which is defined as $\{B, I, E, S, O\}$, where B refers to the first word of personal name, I refers to the middle word of person name, E refers to the final word of personal name, S represents person name with a word character, and O represents word which is not used for person name.

5.3. Feature Selection. In theory, CRF may contain any features. Feature selection can affect greatly the training effect of CRF, so an effective feature set should be selected to obtain ideal accuracy and recall rate [12]. Based on the features of the Chinese names above, 6 types of features are selected in this paper, including the word, the speech of word, and four Boolean values which refer to whether word falls into the table used to record usage of characters in family names, whether word falls into the table used to record usage of characters in given names, whether word falls into the table used to record left boundary words of person names, and whether word falls into the table used to record right boundary words of person names. For 6 types of features, the first two categories belong to local features, and the four Boolean values belong to global features extracted according to the names of knowledge base. The meaning and the corresponding values of various features are shown in Table 1.

Segments of training corpus marked by the above features are shown in Table 2, where the first column represents Ch, the second column represents POS, the third column represents LN, the fourth column represents FN, the fifth column

Table 1: The feature and the corresponding values of Chinese names.

Feature	Value
The word (Ch)	The word itself
The speech of word (POS)	n_B, n_I, n_E, v_S, ...
Belongs to family names (LN)	LN or 0
Belongs to given names (FN)	FN or 0
Belongs to left boundary words (LB)	LB or 0
Belongs to right boundary words (RB)	RB or 0

Table 2: The segments of training corpus.

0	1	2	3	4	5	6
新	nt_B	0	0	0	0	0
华	nt_I	0	0	0	0	0
社	nt_E	0	0	0	0	0
记	n_B	0	0	LB	0	0
者	n_E	0	0	LB	0	0
胡	nr_B	LN	0	0	0	B
晓	nr_I	0	FN	0	0	I
梦	nr_I	0	FN	0	0	E
报	v_B	0	0	0	RB	0
道	v_E	0	0	0	RB	0

represents LB, the sixth column represents RB, and the final column represents the results of the manual annotation.

5.4. Feature Templates Selection. For training learning of conditional random fields, choosing suitable feature templates is the key to identify Chinese personal name. Because there are many selected features of person name, choosing suitable feature templates is a complex process that requires repeated experimental comparison. In this brief, we use incremental selection algorithm to obtain the best feature templates, and the specific implementation is described below.

(1) Set the feature of a word as the base feature of atomic feature template. We would do, respectively, some experiments combining the feature of a word with other features of names and then put these features into atomic feature template if the identification effect is improved. Based on the many experimental results, the suitable atomic feature template was determined.

(2) Build items of compound feature template according to the characters of atomic feature template. We would do, respectively, some experiments combining atomic feature template with an item of compound feature template and then put this item into the items set of candidate compound feature templates if the identification effect is improved. Based on the many experimental results, items of compound feature template were determined.

Table 3: The optimal feature template.

Template type	Items
Atomic template	Ch(n) ($n = -2, -1, 0, 1, 2$) POS(n) ($n = -2, -1, 0, 1, 2$) LN(n) ($n = -2, -1, 0, 1, 2$) FN(n) ($n = -2, -1, 0, 1, 2$) LB(n) ($n = -2, -1, 0, 1, 2$) RB(n) ($n = -2, -1, 0, 1, 2$)
Compound template	Ch(n)/Ch($n+1$) ($n = -1, 0$) Ch(n)/POS(n) ($n = -2, -1, 0, 1, 2$) Ch(n)/LN($n+1$) ($n = -2, -1, 0, 1$) Ch(n)/FN($n+1$) ($n = -2, -1, 0, 1$) POS(n)/LN($n+1$) ($n = -2, -1, 0, 1$) POS(n)/FN($n+1$) ($n = -2, -1, 0, 1$) POS(n)/LB($n+1$) ($n = -2, -1, 0, 1$) POS(n)/RB($n+1$) ($n = -2, -1, 0, 1$)

(3) Set atomic feature template as initial value of the optimal feature template. We would do, respectively, some experiments combining the optimal feature template with items of candidate compound feature template and then put this item into the optimal feature template if the identification effect is improved. Based on the many experimental results, the optimal feature template was determined.

The optimal feature template is shown in Table 3. In order to make full use of context information, we set the size of feature window to 5 considering the influence of large window on system performance.

5.5. The Training Methods of Model Parameters. For CRF model, the weight is used to measure the importance of feature function, which can be obtained by learning from training data. The GIS and IIS are the two earliest algorithms used to train model parameters of CRF, and they are iterative gradient methods. Iterative gradient method is simple and easy, but slow to convergence, so L-BFGS algorithm is widely used to train model parameters of CRF, which is more effective in convergence, and decreases space complexity and time complexity compared with iterative gradient algorithm. Therefore, we select the L-BFGS algorithm to train model parameters in this paper.

6. The Testing Research and Analysis

6.1. The Experimental Data. We split corpus into two parts in accordance with the ratio of 5 : 1, which is tagged on People's Daily published in January 1998 by Computational Linguistics Institute of Peking University, and then choose one as our training corpus and the other as our open test corpus. Training corpus contains 1,558,558 words including 1,086 words used in names and test corpus contains 270,270 words including 2,125 words used in names. In addition, the words, extracted randomly from training corpus, are set as of a closed test corpus. Finally, P (accuracy), R (recall rate), and

Table 4: The experimental results.

Methods	Test types	The number of			Accuracy	Recall rate	F value
		In text	Identified	Identified correctly			
CRF + knowledge base	Closed	1876	1867	1806	96.74%	96.29%	96.51%
	Open	2125	2074	1986	95.76%	93.46%	94.59%
CRF	Closed	1876	1847	1770	95.86%	94.37%	95.11%
	Open	2125	2102	1944	94.57%	92.61%	93.58%

F value are selected as evaluation standards of identification performance, which could be calculated as follows:

$$P = \frac{\text{the number of names identified correctly}}{\text{the number of names identified}} \times 100\%,$$
$$R = \frac{\text{the number of names identified correctly}}{\text{the number of names in the text}} \times 100\%, \quad (3)$$
$$F = \frac{2 \times P \times R}{P + R} \times 100\%.$$

6.2. The Experimental Results and Analysis. According to the relation between training set and test set, testing can be divided into closed testing and open testing. We made four groups of recognition experiments to evaluate objectively the recognition effect using the method combining CRF with knowledge base. Among them, the two former groups of experiments are closed testing and open testing based on CRF + knowledge base, and the two latter groups are closed testing and open testing based on CRF. The experimental results are shown in Table 4.

There are two local characteristics, words and speech of words, in the CRF model, but we introduce global features based on knowledge base (which is manually constructed) to the CRF model. From the experimental results, it can be seen that the accuracy, the recall rate, and F value increased using the method based on CRF + knowledge compared with that based on CRF. F value increased by 1.50% in the close test and by 1.01% in the open test. Above all, the recognition effect of personal names would be improved by introducing the global features.

There are two local characteristics, words and speech of words, in the CRF model, but we introduce global features based on knowledge base (which is manually constructed) to the CRF model. From the experimental results, it can be seen that the accuracy, the recall rate, and F value increased using the method based on CRF + knowledge compared with that based on CRF, and F value increased by 1.50% in the close test and by 1.01% in the open test. Above all, the recognition effect of person names would be improved by introducing the global features.

The method based on statistical language model is used to identify Chinese names from the large-scale corpus [5, 6]. The training corpus and test corpus used in literature [5, 6] are the same as this paper, so we make a series of experiments using different methods. The experimental results are shown in Table 5, which indicates that recognition effect is the best using method proposed in this paper.

Table 5: The experimental results.

Literature	Method	Accuracy	Recall rate	F value
This paper	CRF + knowledge base	95.76%	93.46%	94.59%
Literature [5]	HMM	84.00%	94.00%	88.72%
Literature [6]	MEM + rules	69.88%	91.65%	79.30%

7. Conclusion

According to characters of Chinese personal name, we presented the recognition method combining CRF and knowledge base to identify Chinese personal names from Chinese document and carried out some experiments using this method on open and closed testing. The results of our open and closed experiments show that its performance is competitive. In the future, we will try to do some works to improve the system performance, including introducing semantic analysis, choosing word segment tool with precision, and perfecting dictionary related names.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61064010 and 61364022).

References

[1] W.-T. Hu, Y. Yang, H.-F. Yin, Z. Jia, and L. Liu, "Organization name recognition based on word frequency statistics," *Application Research of Computers*, vol. 30, no. 7, pp. 2014–2016, 2013.

[2] L.-S. Li, Y.-Z. Dang, W.-P. Liao, D.-G. Huang, and Y. Zhang, "Recognition of Chinese location names based on CRF and rules," *Journal of Dalian University of Technology*, vol. 52, no. 2, pp. 285–289, 2012.

[3] Y.-X. He, C.-W. Luo, and B.-Y. Hu, "Geographic entity recognition method based on CRF model and roles combination," *Application Research of Computers*, vol. 32, no. 1, pp. 179–185, 2015.

[4] C.-Z. Jiang, H. Wang, and H. Yao, "Chinese name recognition based on HowNet and Bayesian classifier," *Journal of Nanjing University*, vol. 48, no. 2, pp. 147–153, 2012.

[5] H.-P. Zhang and Q. Liu, "Automatic recognition of Chinese personal name based on role tagging," *Chinese Journal of Computers*, vol. 27, no. 1, pp. 85–91, 2004.

[6] N. Jia and Q. Zhang, "Identification of Chinese names based on maximum entropy model and rules," *Computer Engineering and Applications*, vol. 43, no. 35, pp. 1–4, 2007.

[7] Y.-L. Li, Z. Zhou, and W. Wu, "Scene parsing based on a two-level conditional random field," *Chinese Journal of Computers*, vol. 36, no. 9, pp. 1898–1907, 2013.

[8] L.-N. He, Z.-H. Yang, H.-F. Lin et al., "Drug name entity recognition based on feature coupling generalization," *Journal of Chinese Information Processing*, vol. 28, no. 2, pp. 72–77, 2014.

[9] J.-D. Yu, D. Sui, and X.-Z. Fan, "Word-position-based tagging for Chinese word segmentation," *Journal of Shandong University of Technology*, vol. 40, no. 5, pp. 117–122, 2010.

[10] X.-F. Zhao, D. Zhao, and Y.-G. Liu, "The automatic gender recognition of Chinese name using conditional random fields," *Microelectronics & Computer*, vol. 28, no. 10, pp. 122–124, 2011.

[11] M.-S. Sun, C.-N. Huang, H. Gao, and J. Fang, "Identifying Chinese names in unrestricted texts," *Journal of Chinese Information Processing*, vol. 19, no. 2, pp. 16–27, 1995.

[12] A. Ritter, C. Sam, and O. Etzioni, "Named entity recognition in tweets: an experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pp. 1524–1534, July 2011.