A Novel Method for Chinese Named Entity Recognition Based on Character Vector

Jing Lu[1,2,3](✉), Mao Ye[2], Zhi Tang[1,2], Xiao-Jun Huang[2],
and Jia-Le Ma[2]

[1] Institute of Computer Science and Technology,
Peking University, Beijing, China
10548887@pku.edu.cn

[2] State Key Laboratory of Digital Publishing Technology,
Peking University Founder Group Co., Ltd., Beijing, China
{jing.lu,yemao.apb}@founder.com.cn

[3] Postdoctoral Workstation of the Zhongguancun Haidian Science Park,
Beijing, China

Abstract. In this paper, a novel method using for Chinese named entity recognition is proposed. For each class, A posteriori probability model is acquired by combing probabilistic model and character vector, which are acquired from each class by using training data. After segment Chinese sentence into words, the posteriori probability of every words in each class can be calculated by using model we proposed, and thus the type of word could be determined according to maximum posteriori probability.

Keywords: Named entity recognition · Word vector · Character vector

1 Introduction

The research of Named entity recognition (NER) is a basic work in nature language process (NLP). First, NER is an important technology not only in parsing but also in information retrieval and automatic answering system. Second, the result of recognition Named entity could affect subsequent work.

Usually, the dictionary can be acquired and map the word to appropriate type. But most named entities were not logged in dictionary and new words turn up every day in the network age. So it is necessary to recognize them correctly without excessive reliance on the dictionary and Make sure the subsequent work process would not be effect.

Generally speaking, the purpose of Chinese NER is to categorize the words in the text into appropriate classes (type), which including seven broad classes (1. person name, 2. institution name, 3. toponymy, 4. time, 5. date, 6. current and 7. percentage) [1, 3, 5, 6].

For obviously pattern makes it easier to recognize the classes of time, date, current and percentage, the NER often refer to recognize person name, toponymy, and institution name, which is also the main part of this article.

S. Guo et al. (Eds.): CollaborateCom 2015, LNICST 163, pp. 141–150, 2016.
DOI: 10.1007/978-3-319-28910-6_13

The research of NER has a history of 10 more years, the popular method being used nowadays is to recognize extracting features of specified entities, for instance, the family names or the words ahead of “县” (county), “市” (city) are the special features of names or area names, respectively. Then statistics model such as CRF [8, 11] or SVM [1] was trained by using training data and use for recognizing entities. Another method being used is analyze the latent semantics in Named Entity and use semantics to recognition Named Entity [4].

Although the above method could achieve good effect, there were obviously drawbacks. Firstly, it is difficult to acquire adequate training data. Secondly, the adapt ability of this method is low, which rely on training data and domain knowledge. Thirdly, it is unavoidable to extract feature by labor who usually is the expert in respective field and this makes inconvenience for subsequent research. Finally, the poor performance in test data leads to the difficult in generalization.

The method which called distributed representation map the atomic semantic unit to higher vector space. Atomic semantic units are mapped into points in vector space and the distance between points represent the distance in semantic means. In English, the word is the atomic semantic unit in a sentence. In Named Entity, the word is consisted atomic semantic, and there are many Named Entity Recognition research based on atomic semantic [2]. However, the definition of the atomic semantic unit is not unchangeable. For instance, the root could be considered as the atomic semantic unit in a English word. Unlike English, the word in Chinese is formed by characters and each character indicates a semantic means (type). So the Chinese character can be considered as atomic semantic unit. For instance, the means of “省长” (governor of a province) is consisted of “省” (province) and “长” (governor). The character “省” (province) refer to a scope and the character “长” (governor) means the social status in this scope. So when we see a word “市长” (governor of a city), we can classify “市长” (governor of a city) and “省长” (governor of a province) to a same class because the only one different character is the range of scope, and the semantic meaning of “省” (province) and “市” (city) is close.

Take a person’s name “张小明” (Zhang Xiao Ming) for example, “张” (Zhang) is close to “赵” (Zhao) and “王” (Wang), so “王小明” (Wang Xiao Ming), “赵小明” (Zhao Xiao Ming) can be attribute to a person’s name according the class “张小明” (Zhang Xiao Ming) belongs to.

Summing up, in Chinese named entity recognition, it is important to grasp the semantic meaning in the character and distances of meanings.

Based on the meaning and the performance of vector of character, this paper proposes the Chinese named entity recognize model based on character vector.

2 Our Approach

Figure 1 shows the Flow chart of algorithm in this paper:

Firstly, it is need to prepare training data for each named entity classes, then the initial statistical model and character vector of each class are obtained respectively by using training data. Then the recognition models of each class are acquired by combing character vector and initial statistical model. When input a string, the recognition model

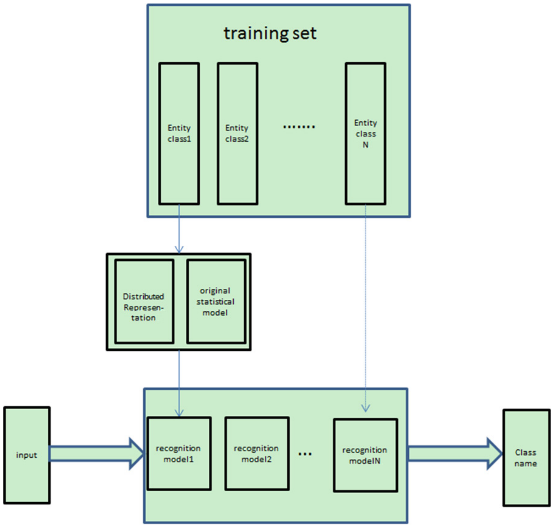

Fig. 1. The flow chart of paper

can output the posterior probability of corresponding class. Finally, the class of input string will be determined according maximum posterior probability.

In above steps, we could find out recognition models will be obtain independently. So, for easier presentation, the rest part of Sect. 2 would only discuss the class toponymy as an example of all other classes.

In Sect. 2.1, we will discuss the significance of character vector and the method to acquire them. In Sect. 2.2, the statistical model of named Entity will be proposed. In Sect. 2.3, the recognition models will be obtained by combining the character vector (Sect. 2.1) and statistical model (Sect. 2.2), and the recognition models will be applied in NER.

2.1 Training Character Vector

In Sect. 2, we discuss the importance of the semantic of character in Chinese NER. But Conventional methods treat more than 3000 common used Chinese characters as independent type, thus the semantic distance between characters is ignored. But in a same entity class, characters in close semantic distance can be replaced by each other in most conditions. Furthermore, through the replacement of characters, many Named Entities will be categorized, so it is urgent to calculate semantic distances of characters.

The Distributed representation is proposed in recent years. The basic idea of Distributed representation is mapping the word unit which is basic unit of sentence into high dimension vector space [8, 9]. The result of this method also be called word vector.

In vector space, the distances between word vectors refer their semantic distance. It is more likely they have same semantic meaning if they get close enough to each other.

The idea of word vector was got from English language which the word is the basic semantic unit in sentence. While in Chinese Named Entity, the character is the basic semantic unit. So it is necessary to modify the method of word vector and apply it in NER.

Take the idea of the word vector method, we treat every Chinese character as basic semantic unit and map them to high dimension vector space by the same way. The mapped result is called character vector. Then, we can calculate the distance between these character vectors and close distance usually means semantic similarity and replaceable.

Given that some Chinese character is polysemy, we get character vector by classes. The same character usually has different vector value in different class.

In this paper, character vector is obtained from Word2Vec offered by Google. After training, every character of entity string can be represented as vector with 8 dimensions. The number of dimensions could be set according to actual demand.

Take toponym for example, the vector of character “县” (county) and “市” (city) were:

$$V_{t\,oponymy,\text{"县"}} = [\,0.23, -0.15, 0.33, 1.23, -0.78, -0.28, -0.28, 0.2]$$

$$V_{t\,oponymy,\text{"市"}} = [\,0.27, 0.07, 0.0, 0.27, -0.42, 0.03, -0.04, 0]$$

The character vector of toponym are showed in 2-Dimension coordinate in Fig. 2. It is obviously that characters “县” (county), “市” (city), “省” (province) are closer, and apart from above three words, characters “东” (east), “南” (south), “西” (west), “北” (north) are close to each other. This fact indicates the substitutability of characters. The closer distance they have, the chance of they replace each other is higher. Furthermore, mapping the word to vector will strongly promote the generalization ability of the model.

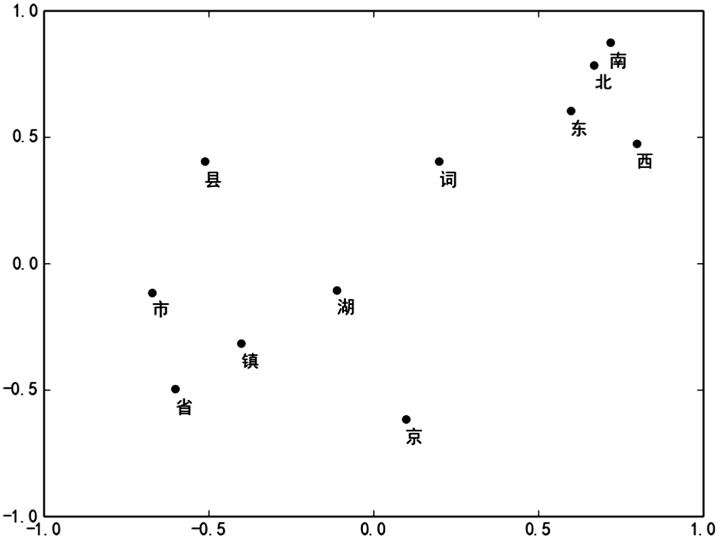

Fig. 2. The character in 2-Dimensions

2.2 Probabilistic Model

In above section, we discuss the importance of the semantic of character in Chinese NER. But Conventional methods treat more than 3000 common used Chinese characters as independent type, thus the semantic distance between characters is ignored. But in a same entity class, characters in close semantic distance can be replaced by each other in most conditions. Furthermore, through the replacement of characters, many Named Entities will be categorized, so it is urgent to calculate semantic distances of characters.

In this paper, the character is divided into three categories according to their location in the word, which is the First Character,the Inner Character and the End Character, respectively. These three types of models plus one-order Markov Process makes one kind of class named entity can be represented into four sub-models.

Still take toponymy for example. For string str = " *上海市* " (Shang Hai City), we define the probability of the model generate it shows as follow:

$$\begin{aligned} & P'(\text{" 上海市"} \mid class = Toponymy) \approx P'_B(\text{" 上 "} \mid class = Toponymy) \\ & \prod_{i=2}^{end-1} P'_M(str(i) \mid class = Toponymy) \\ & P'_E(\text{" 市 "} \mid class = Toponymy) \\ & \prod_{i=1}^{end-1} P'_{Markov}(str(i) \mid str(i-1)，class = Toponymy) \end{aligned} \tag{1}$$

The $str(i)$ represent the i-th character in str.

$$\prod_{i=2}^{end-1} P'_M(str(i) \mid class = Toponymy) = \prod_{i=2}^{end-1} P'_M(\text{海} \mid class = Toponymy)$$

$$\prod_{i=1}^{end-1} P'_{Markov}(str(i) \mid str(i-1), class = Toponymy) =$$

$$P'_{Markov}(\text{" 海 "} \mid \text{" 上 "}, class = Toponymy) P'_{Markov}(\text{" 市 "} \mid \text{" 海 "}, class = Toponymy)$$

Every sub-model will be calculated by using training data, the formulas shows as follow:

$$\begin{gathered} P'_B(cc|class = AreaName) = \frac{1}{N}\sum_{n=1}^{N} f_B(cc, data_n) \\ f_B(cc, str) = \begin{cases} 1 & str(1) = cc \\ 0 & Otherwise \end{cases} \end{gathered} \tag{2}$$

$$\begin{aligned} P'_M(cc|class = AreaName) &= \frac{1}{N}\sum_{n=1}^{N} f_B(cc,\ data_n) \\ f_M(cc, str) &= \begin{cases} 1 & str(i) = cc \\ 0 & Otherwise \end{cases} \end{aligned} \tag{3}$$

$$\begin{aligned} P'_E(cc|class = AreaName) &= \frac{1}{N}\sum_{n=1}^{N} f_B(cc, data_n) \\ f_E(cc, str) &= \begin{cases} 1 & str(end) = cc \\ 0 & Otherwise \end{cases} \end{aligned} \tag{4}$$

$$\begin{aligned} P'_{Markov}(cc|class = AreaName, cc_{pre}) &= \frac{1}{N}\sum_{n=1}^{N} f_{Markov}(cc|cc_{pre}, data_n) \\ f_{Markov}(cc|cc_{pre}, str) &= \begin{cases} 1 & str(i+1) = cc\,\&\&\,str(i) = cc_{pre} \\ 0 & Otherwise \end{cases} \end{aligned} \tag{5}$$

N is the size of training data in toponymy. cc_{pre} means the one character ahead of character cc.

2.3 Probabilistic Model Based on Character Vector

Using training data, we acquire the four sub-models of toponymy. To further improve the generalization ability of the probability model, the character vector is applied. The application of character in probabilistic model shows as follows:

$$P_B(cc|class = toponymy) = \frac{1}{Z_B}\sum_{cc' \in Set} P'_B(cc'|class = toponymy)dis(cc, cc') \tag{6}$$

$$P_M(cc|class = toponymy) = \frac{1}{Z_M}\sum_{cc' \in Set} P'_M(cc'|class = toponymy)dis(cc, cc') \tag{7}$$

$$P_E(cc|class = toponymy) = \frac{1}{Z_E}\sum_{cc' \in Set} P'_E(cc'|class = toponymy)dis(cc, cc') \tag{8}$$

$$P_{Markov}(cc|class = toponymy, cc_{pre}) = \frac{1}{Z_{Markov}}\sum_{cc' \in Set} P'_{Markov}(cc'|class = toponymy, c_{pre})dis(cc, cc') \tag{9}$$

Where Z_B, Z_M, Z_E, Z_{Markov} is the normalized coefficient, function $dis(cc, cc')$ is using to calculate distance between character cc and cc'

$$dis(cc, cc^{'}) = \frac{V_{toponymy,cc} V^{T}_{toponymy,cc'}}{\left|V_{toponymy,cc}\right| \left|V_{toponymy,cc'}\right|} \tag{10}$$

Where $V_{toponymy,cc}$ is the vector of character CC in toponymy.

So we can acquire the posterior probabilistic model of the toponymy type based on character vector, the formula shows as follow:

$$\begin{aligned} L(class = toponymy|str) &= Log(P(str|class = Toponymy)) + Log(P(class = Toponymy)) - Log(P(str)) \\ &\propto Log(P(str|class = Toponymy)) \propto \\ &= coefficient_{B,toponymy} Log(P_B(str(1)|class = toponymy)) \\ &+ coefficient_{M,toponymy} \sum_{i=2}^{end-1} Log(P_M(str(i)|class = toponymy)) \\ &+ coefficient_{E,toponymy} Log(P_E(str(end)|class = toponymy)) \\ &+ coefficient_{Markov,toponymy} \sum_{i=1}^{end-1} Log(P_{Markov}(str(i)|class = toponymy)) \end{aligned} \tag{11}$$

In the above formulas, each sub-models is weighted with coefficients and finally get the posterior probabilistic. The coefficients are calculate as follows:

$$coefficient_{B,toponymy} = \frac{1}{sum} \frac{1}{|entropy(B, toponymy)|} \tag{12}$$

$$coefficient_{M,toponymy} = \frac{1}{sum} \frac{1}{|entropy(M, toponymy)|} \tag{13}$$

$$coefficient_{E,toponymy} = \frac{1}{sum} \frac{1}{|entropy(E, toponymy)|} \tag{14}$$

$$coefficient_{Markov,toponymy} = \frac{1}{sum} \frac{1}{|entropy(Markov, toponymy)|} \tag{15}$$

where sum is the normalized coefficient, $entropy(B, toponymy)$ is the entropy of sub-model:

$$entropy(B, toponymy) = \sum_{cc \in Set} P_B(cc|class = toponymy) \log P_B(cc|class = toponymy) \tag{16}$$

Each posterior probabilistic $L(classname|str)$ can be acquired according to above method. We can judge the class string “str”s by using the follow formula:

$$class^{*} = \mathop{Arg\max}_{class \in EntityNames} (L(class|str)) \tag{17}$$

3 Experimental Results

In practical application, some words, such as “如果” (if), “其它” (otherwise), “但是” (but), which do not belong to any classes of named entity, need to be rejected. These words constitute of reject class, and the model of reject class will be acquired by using the same method.

In this paper, words randomly selected consist of reject class. Although they could be the named entity, the dispersion and rare characteristic makes they rarely affect the result. By using ICTCLAS offer by Chinese Academy of Sciences, we segmented sentences into words and randomly selected 20000 words as reject words for model training.

We compare our method with the method which called Chinese Named Entity Recognition via Joint Identification and Categorization (JIC) [13].

We train the Model by choosing People’s Daily In February to April 1998 data as pattern,and still use data in People’s Daily January 1998 as test data, the result was shown in Table 1.

Table 1. Our method vs. JIC (People’s Daily offer training data)

	Our method			The result based on JIC
	Right rate	Recall rate	F-Measure	F-Measure
Person name	93.1 %	90.2 %	91.6 %	91.3 %
Toponym	90.2 %	91.1 %	90.64 %	89.7 %
Organization name	91.1 %	86.2 %	88.58 %	87.4 %

The result shows that the method proposed in this paper could recognize the Chinese Named Entity with higher right rate than JIC.

Second, in order to test the effectiveness of this method applying in other data set, the data set is acquired from Apabi Company and words are classified to 1. person name, 2. toponymy, 3. organization name or 4. official position. The number of each class in the training data was shows in Table 2:

Table 2. The number of pattern from Apabi Company

Entity class	Train pattern	Test pattern
Person name	20000	200
Toponym	20000	200
Offical position	20000	200
Organization name	20000	200
Reject	20000	200

The result in test data was showed in Table 3.

Table 3. The Right and Recall rate on data of Apabi Company

Entity class	Right rate	Recall rate
Person name	97.2 %	67.2 %
Toponym	98.1 %	70.1 %
Official position	99.5 %	72.1 %
Organization name	95.3 %	69.3 %

It could be finding that the right rate could also reach a higher level.

Compared with the conventional method, the method we proposed could not only recognize 3 big classes, but also other type entity, such as official Position, literary works, without extracting feature based on Specialized processing.

4 Conclusion and Prospect

In this paper, we introduce a new method character vector into the Named Entity Recognition. The result of experiment from this paper shows that this new method could achieve good effectiveness in Chinese Entity Character recognition and the most advantage of this method is to recognize Named Entity without type limit.

The method mentioned in this paper has obtained ideal recognition rate. The reasons are as follows. First, we described Chinese named entity in the form of probability by using markov process, and parameters of the probability model were obtained from training data. Second, the semantic distance of Chinese characters is proposed to be represented by the space distance of character vectors in this paper. This idea liberated the human labor from manually identifying the similarity distance and also removed the man-made interference during the process. Third, in order to improve the generalization ability of the model, we incorporate the character vectors with markov model. This incorporated model could achieve higher recognition rate. In summary, the method proposed in this paper theoretically could be used widely in most type of named entities for it removing human interfere during the process.

There were still some unsatisfied aspects need to be improve. First, a number of each type of entity should be prepared for training the model before used. Second, the effectiveness would be weakening if the training data and test data has big differences. Third, it lack of proper rejection mechanism to reject class which do not included in any class and that could lead to the misjudgment to the reject string. So the follow work could process with this three points and further improve this method.

References

1. Qi, Z., Zhao, J., Yang, F.: A new method for open named entity recognition of Chinese (2009)
2. Iwakur, T.: A named entity recognition method based on decomposition and concatenation of word chunks. ACM Trans. Asian Lang. Inf. Process. (TALIP) **12** (2013)

3. Pan, S.J.: Transfer joint embedding for cross-domain named entity recognition. ACM Trans. Inf. Syst. **31**(2), 7:1–7:27 (2013)
4. Konkol, M., Brychcín, T., Konopík, M.: Latent semantics in Named Entity Recognition. Expert Syst. Appl. **42**(7), 3470–3479 (2015)
5. Zhang, H., Liu, Q.: Automatic recognition of Chinese personal name based on role tagging. Chin. J. Comput. **27**(1), 85–91 (2004)
6. Yu, H.: Recognition of Chinese organization name based on role tagging. In: Advances in Computation of Oriental Languages, pp. 79–87 (2003)
7. Wang, N., Ge, R.: Company name identification in Chinese financial domain. J. Chin. Inf. Process. **16**(2), 1–6 (2002)
8. Zeng, G.: CRFs-based Chinese named entity recognition with improved tag set. Master degree theses of master of Being University of Posts and Telecommunications (2009)
9. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. (JMLR) **12**, 2493–2537 (2011)
10. Tomáš, M.: Statistical language models based on neural networks. PhD thesis, Brno University of Technology (2012)
11. Yao, J.: Study on CRF-based Chinese named entity recognition. Master degree theses of master of Suzhou University (2010)
12. Yu, H.: Chinese named entity identification using cascaded hidden Markov model. J. Commun. **27**(2), 87–94 (2006)
13. Zhou, J.: Chinese named entity recognition via joint identification and categorization. Chin. J. Electron. **22**(2), 225–230 (2013)