Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):1474-1478

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Chinese named entity recognition algorithm based on the improved hidden Markov model

Jie Liu

School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, Shaanxi, China

ABSTRACT

This paper proposed an algorithm based on the improved Hidden Markov Model (HMM) to achieve the Chinese Named Entity Recognition (NER) in Nature Language Processing (NLP) .It put forward three kinds of familiar recognition methods on the basic of summarizing the trait and difficulty of Chinese NER. The application of the Hidden Markov Model (HMM) based on statistical method in NER was researched. It pointed out and analyzed the limitation of the conventional HMM. At the same time, some improvements have been made in addition.

Key words: Chinese Named Entity Recognition; statistical; Hidden Markov Model; limitation; improvement

INTRODUCTION

With the rapid development of information technology, the amount of information increases exponentially everyday. So it is more difficult to obtain useful information from those large numbers of information. Therefore, a lot of information is needed to deal with. For example Information Retrieval (IR) and Information Extraction (IE) .The NER is the basic work in NLP, and it plays a great role in many domains such as IR、IE、Text Classification、Question Answering System and so on.

The NE which appears frequently in text is the primary causation that restricts the enhancement of participial precision. The quality of recognition will influence the participial precision at first hand. Moreover, it influences the precision of Part Of Speech (POS) label and the syntax analyze. Relative to the English, it's a more arduous task to identify the NE because of the characteristic of Chinese Nature Language. So the research of recognition algorithm of named entity is more important on theoretical significance and practical value.

2. Chinese Named Entity Recognition

2.1 Background and Status quo

The Named Entity is the primary carrier of information. Therefore the NER becomes the chief task in information processing. The Message Understanding Conference(MUC) which is held by America Defense Advanced Research Projects Committee had performed 7 times from 1987 to 1998.In the 6th MUC held In September 1995, it introduced the evaluating task of the NER[1],It included English 、Chinese and Japanese mostly. In the 7th MUC held in 1998 ,the NE was divided into 7 kinds ,Person、Location、Organization、Data、Time、Percentage、Monetary value and so on[2].

The NER is a kind of especial pattern recognition. In recent years, many people are researching it actively. It's early to research the English NER abroad. Without participial problem, it just needs to consider the characteristic of words. The difficulty of English NER is lower relatively, so that the research had achieved a high level at present. In the MUC, the experimental results show that the precision and recall rates are all about 97%. Because of the intrinsic characteristic of Chinese, researchers must analyze the morphology at first when they process the text which increases the difficulty of NER. At present, the research of Chinese NER just enters an underway phase. The reports about the precision and recall rates are about 90% generally in inland and foreign country.

2.2 The Characteristic and Difficulty of the Chinese NER

The Chinese NER mainly relates to person、location and organization. The inherent character of grammar and morphology determines that it's more difficult to identify them. As follows the difficulties [3]:

① In different scene and field, there is much diversity in the extension of NE.
② There are so many entities that they can't be enumerated one by one. Without doubt, a lot of entities can't be embodied in dictionary. because person, location and organization are infinite aggregate,
③ The names of some kinds of entities change frequently, and there is no strict rule to follow.
④ There are many diversiform expression forms.
⑤ It's become abbreviation when it was used after the first time.

2.3 The Methods of Chinese Named Entity Recognition

The methods of Chinese NER are divided into three types as a whole. The first type is the method based on rules. The second type is the method which is based on statistic and the third type is the method based on rules and statistic together. The second and the third type are used generally.

The method based on rules:
This method need to construct rules template artificially with analyzing the intrinsic and extrinsic characteristic. It identifies these NE through matching those rules. The experiments have show that this method has high precision rate and efficiency in small rules. But by the reason of limited overlay rate of rules, it isn't transplantable. On the other hand, if the language experts need to compile and construct accurate rules, they must investigate and understand the context, which will cost much more manpower and resource. For instance the Proteus System [4] in New York University, the NetOwl System[5]of IsoQuest Inc, and the FACILE System[6] in Manchester University of Science and Technology.

The method based on statistic:
The method means that it not only need to label the tagged data but also count the probability of the word which is belong to the NE with training the sample. If the result is bigger than the certain numerical value, then this word is identified as the NE. Compared to the method based on rules, this method based on statistic is stronger and more flexible and more transplantable. For the moment, increasing models based on statistic are used in the NER. For example HMM, Maximum Entropy，Support Vector Machines , Decision Tree and so on.

The method based on rules and statistic together:
Making use of rules and statistic together, on the one hand, it can reduce the complexity and blindness of the method based on rules through counting the probability. On the other hand, it can play down the requirement of size of the tagged data set through reusing those rules. So this method is used generally in practice.

3. The Markov Model And The Hidden Markov Model

The Markov Model (MM) and the Hidden Markov Model (HMM) based on statistic are used to describe stochastic process .They provide the technique of recognition system that is constructed automatically through training the probability of the data. These models are applied in various fields of NLP. They have become the main methods of NLP based on statistic and they are one of the great harvests certainly.

3.1 The Markov Model

The Russian chemist named Markov put forward the Markov Model in 1870, which is used to depict the sequence information of stochastic variable [7].It is regarded as stochastic and finite state automaton. There is a probability between the transitions of every state which is used to express the possibility of the state transition. If there are N numbers states S_1、S_2、…、S_n， with time go ahead, the system transfers from one state to the other state. When the time is *T*, the state is described as q_t. The current state and all expected states are described with this method in the system. Namely, the probability of the state *S* when the time is *T* lies on those states at the time of *1*、*2…*、*T-1*, and the probability is:

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ...) \tag{1}$$

In the especial condition, if the state of the system at the time of *t* correlates to the state at the time of *t-1* merely, then the system can form a discrete Markov chain:

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ...) = P(q_t = S_j \mid q_{t-1} = S_i) \tag{2}$$

More, with considering the stochastic course which is independent at the time of *t* :

$$P(q_t = S_j \mid q_{t-1} = S_i) = a_{ij}, 1 \le i, j \le N \tag{3}$$

The stochastic course is a Markov Model, and the probability of the state transition (a_{ij}) must ensure these conditions:

$$a_{ij} \ge 0, \quad \sum_{j=1}^{N} a_{ij} = 1 \tag{4}$$

For instance, at the period of time, the weather is described as three states Markov Model [8]:
State 1: rainy or snowy
State 2: cloudy
State 3: sunny

The possibility of the state transition process is described as this matrix:

$$A = [a_{ij}] = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

If the first day is sunny, according as this model, the possibility of the weather is “sunny ,sunny, rainy, rainy, sunny, cloudy, sunny” in succedent seven days:

$$\begin{aligned} P(O \mid M) &= P(S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3 \mid M) \\ &= P(S_3) \times P(S_3 \mid S_3) \times P(S_3 \mid S_3) \times P(S_1 \mid S_3) \times P(S_1 \mid S_1) \times P(S_3 \mid S_1) \times P(S_2 \mid S_3) \times P(S_3 \mid S_2) \\ &= 1 \times a_{33} \times a_{33} \times a_{31} \times a_{11} \times a_{13} \times a_{32} \times a_{23} \\ &= 0.8 \times 0.8 \times 0.1 \times 0.4 \times 0.3 \times 0.1 \times 0.2 \\ &= 1.536 \times 10^{-4} \end{aligned}$$

3.2 The Hidden Markov Model

In the MM, every state expresses a observable event, which limits the applicability of the model. In the HMM, the observable events are stochastic functions of states. Therefore, the model is a double stochastic process, and the state transition process can’t be observed. So nothing but the output values of the each moment [8] can be obtained.

The HMM can be expressed by the compages with five elements：（S,K,∏,A,B）. *S*={S_1,…,S_M}, The *S* denotes the finite muster of states；*K*= {K_1,…,K_N},The *K* denotes the observational sequence；∏={ π$_i$}、*i* ∈*S*, The ∏ denotes the primal states; *A*={a_{ij}}、*i*、*j*∈*S* ,The a_{ij} denotes the probability of the state transition from S_i to S_j; *B*={b_{ik}}、 *i*∈*S*、*k*∈*K* ,The b_{ik} denotes the output probability from S_i to the observational K_j.

On the assumption that there is a observational sequence in length *L*: *O*=(O_1，⋯O_L)，in the model μ={S，K，∏，A，B}, the state *X*=(X_1，⋯X_L) is a hidden stochastic process made up of *L* stochastic variables. In the HMM, the probability of state sequence is:

$$P = (X \mid O, \mu) = P(x_1 \mid o, \mu) \prod_{i=1}^{L} P(x_{i,} x_{i+1} \mid o, \mu) \tag{5}$$

3.3 The Viterbi Algorithm

The process of finding solution is to working-out the optimum state sequence making use of the Viterbi algorithm in the known HMM. If the position is at the *j* node at present, then the probability of the hidden state *i* is:

$$\delta_i(j) = \max\ P(x_1 \ldots x_{j-1}, o_1 \ldots o_{j-1}, x_i = i|\mu) \tag{6}$$

Begin:

1、 $\delta_i(1) = \pi_i$ (7)

2、 $\delta_i(t+1) = \max\limits_{1\le k\le N} \delta_j(t) a_{ji} b_{ik}, 1 \le i \le l+1$ (8)

3、Save the best process $\omega_i(t+1) = \max\limits_{1\le k\le N} \delta_j(t) a_{ji} b_{ik}, 1 \le i \le l+1$ (9)

4、If it isn't end, then switch to the second step, otherwise switch to the fifth step;

5、 $\omega = \max \delta_i(t+1), 1 \le i \le N$ (10)

3.4 The Improved Hidden Markov Model

The traditional HMM is based on the assumption that the next state depends on the previous and the current states of events .It only takes into account the current word influence and the current mark, and finds the relation between words with neglecting the direct influence of the context. Therefore, the model will be in face of participle and label difficulty in the NER.

In HMM, the observed event's stochastic process is stochastic function with the hidden state transition. It corresponds to the output probability matrix [b_{ik}] from states to events in the HMM. The observable event is influenced by fore-and-aft *N* states, and it is a stochastic function of these *N* states. Similarly, a hidden state can influence fore-and-aft *N* events. The influence space of the hidden state is *N=2S+1*, and there are *2S+1* events influenced in the fore-and-aft window. Respectively, anterior *S* states 、 current event and latter *S* states.

The observable sequence is $O=(O_1,\cdots,O_L)$.In the model μ,the hidden state sequence is $X=(x_l,\cdots,x_L)$. The probability of this state sequence can be expressed:

$$P = (X|O,\mu) = P(x_1|o,\mu)\prod_{i=1}^{L} P(x_{i+1}|x_i,o,\mu) \tag{11}$$

Lookup the most possible count route: $\arg\max\limits_x P(X|O,\mu)$, When stochastic event sequence is ascertained , the maximum is found: $\arg\max\limits_x P(X,O|\mu)$ .In this text, the current state is influenced by fore-and-aft *N=2S+1* events: $O_{i-s}\cdots O_{i+s}$. Well then, the optimum state sequence is found by making use of the Viterbi algorithm, and the second step is transformed to the form thereinafter:

$$\delta_i(t+1) = \max_{1\le k\le N} \delta_j(t) a_{ji} b'_{ik} \tag{12}$$

$$b'_{ik} = f_1(b_{i(k-s)}) * \ldots * f_{s+1}(b_{ik}) * \ldots * f_{2s+1}(b_{i(k+s)}) \tag{13}$$

$$f_{s+1}(b_{ik}) = b_{ik} \tag{14}$$

$$f_1(b_{i-(k-s)}) = \frac{1}{\log(b_{i-(k-s)})} \tag{15}$$

$$f_{2s+1}(b_{i(k+s)}) = \frac{1}{\log(b_{i(k+s)})} \tag{16}$$

DISCUSSION

The diversity is the most significant features of the output model comparing with traditional HMM. The improved HMM makes many outputs of events aiming at the current state, and the optimum state sequence can be found through counting the statistical dependent connection between the NE and the context .Except for the characteristic of word, the influence that the context acts on the word must be taken into consideration. It is helpful in doing correct labeling and the lexical analysis. Without question, the accuracy of the NER has been enhanced in this algorithm.

CONCLUSION

The improved HMM can give dual attention to the characteristic information of the context in the participle and the labeling process. Therefore, it displays better performance and higher accuracy comparing with traditional HMM. Meanwhile, it has higher precision and recall rates in the NER. The experimental results show that it has more important practice significance in the NLP technology.

Acknowledgements
On the course of writing this paper till the moment of its completion, there are those who contributed to this work that I would like to express my gratefulness here. One is Mr. LI Pengfei, my husband, having been undertaking much more homework so that I could put all my heart into composing this paper. The other one I should make a grateful acknowledgement is Mr. DANG Xianmin, one of my best friends, who helped me a lot in translation work.

REFERENCES

[1]YU Hong-kui; ZHANG Hua-ping. *Journal on Communications*.**2006**,2(2),87-93
[2]YU Hong-kui. The National Network and Information Security Technology Symposium.**2005**,22(4),541-547
[3]HE Zhong-yang; YANG Bai-wei; LI Ou .*Journal of Information Engineering University*.**2011**,10(5),596-600
[4]FENG Yuan-yong; SUN Le. *Journal of Chinese information processing*.**2008**,22(1),104-110
[5]QIAO Wei; SUN Mao-song. *Journal of Tsinghua University*.**2010**,50(5),758-762
[6]LIN Guo-yu; BAI Yun; ZHANG Wei-song. *Journal of southeast university(Natural Science Edition),* **2013**, 43(6), 1217-1221
[7]EirlysE; Davies; Abdelali Bentahila. *Journal of Multilingual and Multicultural Development* . **2012** (3),213-235
[8]LI Guang-yi; WANG Hou-feng. *Journal of chinese information processing*.**2013**,27(5),29-34