International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC 2014)

Research on Modern Chinese Multi-category Words Part of Speech Tagging Based on Hidden Markov Model

Zhendong Song
Information Science and Technology Institute
Heilongjiang University
Harbin,China
E-mail：songzd2000@163.com

Abstract—In recent years, computer systems are widely used in the modern Chinese part of speech tagging. Modern Chinese part of speech tagging is a basic subject in the natural language processing. It is widely used in machine translation, natural language understanding, establishing of the Chinese corpus, information retrieval, text classification, text proofreading and speech recognition, among others. In the part of speech tagging, multi-category words part of speech (POS) tagging is always a difficulty. Although the total number of multi-category words in the modern Chinese is not high, the usage is fairly widespread. This paper, proposes an algorithm of multi-category words part of speech tagging. First, it is word segmentation according to the traditional method. And then, on this basis, we introduce a method based on the rules of multi-category words part of speech tagging. Finally, a detailed description of the Hidden Markov Model (HMM) used in the words part of speech tagging, and a statistical algorithm based on Hidden Markov Model.

Keywords- Computer systems;Chinese information processing; Multi-category words; Part of speech tagging; Hidden Markov Model

I INTRODUCTION

The main task of modern Chinese part of speech tagging is to mark ambiguous word part of speech through the use of computer, in a manner that is automatic and accurate. Presently, the language model of part of speech tagging can be divided into two parts: method based on rules and method based on statistics [2]. The method based on rules has poor adaptability, and needs a special large corpus support. Non statistical model is often used as an independent tagger. Consequently, it is difficult for it to be used as a component part of a larger probability model. However, the method based on the statistics is using different algorithms, according to the probability of word occurrences, to mark with probability statistics. This method can make up for the shortcomings of the method based on the rules. Among the methods based on statistical approaches, Hidden Markov Model is widely applied. It is one of the best language models in the statistical model.

In respect of Hidden Markov Model (HMM) used for part of speech tagging, the foreign scholars have done a lot of research work. For instance, in 1988, Church and Derose proposed the first hidden markov model English tagger based on probability and statistics of the English words. In 1994, Schvtze and Singer proposed a Variable Memory Markov Model. In 1999, Scott and Mary proposed fully Second Order Hidden Markov Model. SangZoo and Jun-ichi proposed a Hidden Markov Model based on Lexicalized Hidden Markov Model [3]. In recent years, domestic scholars also have done a lot of research. A lot of literature has written on the concrete analysis and improvement for the traditional Hidden Markov Model used in part of speech tagging.

This paper combines the rule-based method with the statistic based method to implement annotation on Modern Chinese ambiguity among words, focuses on the application of statistical method.

II WORD SEGMENT TECHNOLOGY

The word segment is the continuous word sequence of language according to certain standard reconfigured into a sequence of words.

After many years of research, word segment technology has been more widely entrenched presently. Maximum matching (MM) method is one of the most effective methods. Maximum matching method proposed by the Soviet Union experts in twentieth Century at the end of the 50's, is a kind of automatic word segment algorithm that appeared earliest. The basic idea of the algorithm is: first build lexicon, which contains all the possible words, the word to segment given by a string of Chinese characters S, according to a certain principle (positive or negative) get the substring of S. If the substring matched with an entry in the lexicon, the substring is the word segment. Continue to segment the rest, until the rest is empty. Otherwise, the substring is not a word, is to take S substring matching. The MM method is a widely used for mechanical word segmentation method. The "machine" means the algorithm only depend on the existing dictionary for matching [4].

According to each matching priority of long term or short term, mechanical word segment method is divided into the maximum matching method and the minimum matching method. According to the direction of the scan list and the interception of words by increasing words or decreasing words. Mechanical word segment method is divided into forward MM method, reverse MM method, adding word MM method and minus word MM method. Maximum matching method commonly used, assumes that the longest term in automatic word segment dictionary contains a number of I Chinese characters, and then take the material to be processed in the current string sequence of I characters as the matching field, matching by looking up word segment dictionary .

Matching word segment technique was actually restricted by thesaurus capacity. However, after decades of research, the basic thesaurus capacity has reached the demand of practical application.

III METHOD BASED ON THE RULE

Rule-based method is a traditional method. Its advantage is to make full use of existing achievements in linguistics. For some special ambiguity combination, this method can obtain a high effect of eliminating different meanings by detailed description of the feature information in the context of words and part of speech.Obviously, the rule-based method requires a large corpus. The corpus is divided into raw corpus and mature corpus.Raw corpus refers to those original text without being processed, while mature corpus is the part of speech tagged corpus. Because calculation of the data needs to be words and part of speech information, it must be annotated and data statistics provided using the corpus information.

To this end, we accumulate years of study to form the existing corpus. Corpus is such as People's Daily, the famous novels, novels in China Mainland, Hong kong and Taiwan novels, the scientific and technological paper, and others. In the process of data collection, only "准备" as an example, we collected 62765 sentences. By giving each sentence manual annotation, we summarized the tagging rules.

For example:

①全胡同中，大家都高兴，都准备[v]着迎接胜利，只有冠晓荷心中不大痛快。【NEW->The famous novel ->Laoshe->Four Generations ->Chapter One->10】

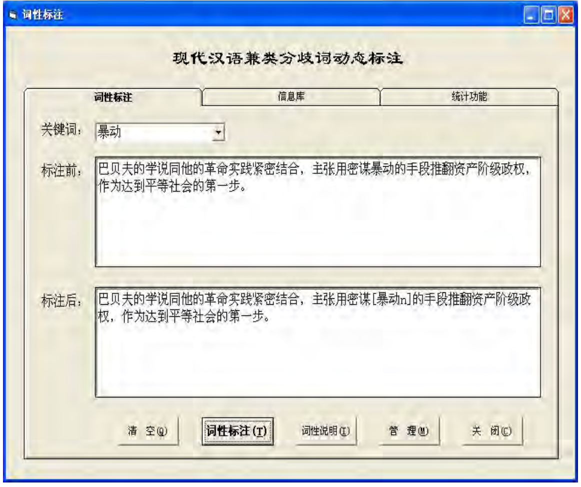
②拥有这样的条件，加上一年多的准备[n]，克拉姆尼克显然觉得胜券在握。【NEW->People's Daily ->2002 】

From example 1 , it is not difficult to conclude that if the word "准备" is followed by the word "着", then the word "准备" is a verb (v). And from example 2, it was not difficult to conclude that if the word "的" comes before the word "准备", then the word "准备" must be a noun(n).

In example 1 and example 2，"准备" is a noun / verb multi-category word.

The part of speech tagging program interface of the method based on rules refers to Figure 1.

The method based on rules is supported by corpus. If the corpus is large enough and the words to be tagged can be covered by corpus, then the tagging accuracy is extremely high, and can even reach 100%. However, with the new words appearing more and more commonly, even new collocation structure appears, the tagging accuracy may have change. Although the method based on rules is an important part of speech tagging method, it can not be the only method. It must be combined with the method based on statistics. There are many kinds of methods based on statistics, they have different characteristics. Hidden Markov Model is one of the methods that are based on statistical approach. And this model are widely used with better effect in statistical models.



关键词：暴动

标注前：巴贝夫的学说同他的革命实践紧密结合，主张用密谋暴动的手段推翻资产阶级政权，作为达到平等社会的第一步。

标注后：巴贝夫的学说同他的革命实践紧密结合，主张用密谋[暴动n]的手段推翻资产阶级政权，作为达到平等社会的第一步。

Figure 1.Program interface of part of speech tagging

Ⅳ HIDDEN MARKOV MODEL OVERVIEW

Markov model describes a class of important stochastic process. Assume the existence of a sequence of random variables (usually related to time), such that it satisfies the following conditions: each random variable is not mutually independent of each other, and each random variable only depends on the random variable before it in the sequence. In many similar systems, we can make the assumption that, to predict the future state based on the now state without consideing the past state. That is to say, the future random variables in the sequence have nothing to do with the past random variables. It conditional depends on the current random variables. Random variable sequences like this, often referred to a Markov chain. We can say the sequence has a Markov property.

Assume that $X=\{X_1, X_2, \ldots\ldots,X_T\}$ is a value for $S=\{s_1,s_2,\ldots\ldots,s_N\}$ random variable sequence, the properties of the Markov model are as follows:

(1) $P(X_{t+1}=S_k|X_1,X_2,\ldots,X_t)=P(X_{t+1}=S_k|X_t)$

(2) $P(X_{t+1}=S_k|X_t)=P(X_2=S_k|X_1)$

X random variable sequence is called a Markov chain, this model is called the Markov model. In the

Markov model, each state represents an observed event. This limits the applicability of the model. When the "state" in the Markov model is not visible to the outside world, it turned into a hidden Markov model. In the hidden Markov model, the observed events are the random function of the state. So the model is a dual random process, the state transition process of the model is not observable (is hidden), the random process of the observable events is a random function of the hidden state transition process. We can also understand that hidden Markov model refers to the internal state of the Markov model is not visible to the outside, outside can only see the output value of each time.

Hidden Markov Model is composed of the basic Markov model, a hidden Markov model can be made up of a five-tuple (S, K, Π, A, B), defined as follows:

(1) $S= \{s_1, s_2,..., s_N\}$: state set;

(2) $K= \{k_1,k_2,...,k_m\}$: output value set;

(3) Π is the initial state;

(4) $A= \{a_{ij}\}$, $i\in S$, $j\in S$, $a_{ij}=P(X_{t+1}=S_j|X_i=s_i)$: the state transition probability;

(5) $B= \{b_{ik}\}$, $j \in S$, $k \in K$, $b_{jk}=P(O_t=K_k|X_t=s_j)$: symbol emission probability.

Here $x= (x_1,x_2,..., x_{t+i})$ is the state sequence and $O= (O_1,O_2,...,O_t)$ is the output sequence [5].

Ⅴ THE APPLICATION OF THE HIDDEN MARKOV MODEL TO PART OF SPEECH TAGGING

Modern Chinese part of speech tagging problem can be described as a state sequence whose part of speech sequence $t_1, t_2,\ldots,t_n$ is hidden, under the condition of a known word sequence $w_1,w_2,\ldots,w_n$. The result of our study is to count the part of speech transfer matrix $[a_{ij}]$ and the output matrix $[b_{jk}]$ from part of speech to words. The problem solving process is to find the most likely state sequence.

According to the word segment method mentioned above, we assume that W is the word sequence after the word segment, T is a possible part of speech tagging sequence of W, T* is the final result of tagging, that is the maximal probability part of speech sequence. Thus we can get:

$W= (w_1,w_2,\ldots,w_n)$

$T= (t_1,t_2,\ldots,t_n)$

$T^*= \text{argmax}P (T|W)$

So, the question turns out to be the question of W search, the best hidden Markov state sequence T under the condition of a given observed sequence.

Applying Bayes algorithm, we obtain,

$$P (T|W)=\frac{P(T)P(W|T)}{P(W)}$$

Conclusion:

$$T^*=\text{argmax}\frac{P(T)P(W|T)}{P(W)}$$

For a given word sequence, the word sequence probability P (W) is same for any part of speech tagging sequence. So we can ignore it in the calculation of T*, then

$T^*=\text{argmax}\ P(T)P(W|T)$

According to the N element syntax hypothesis, there is

$P(T)P(W|T)\approx\prod_{i=1}^{m} p(w_i|\ w_1t_1\ldots w_{i-1}\ t_{i-1}t_i)p(t_i\ |\ w_1t_1\ldots w_{i-1}t_{i-1})$

By using the bigram model, there are

$P(w_i|w_1t_1\ldots w_{i-1}t_{i-1}t_i\ )=P(w_i|t_i\ )$

$P(t_i|w_1t_1\ldots w_{i-1}t_{i-1})=P(t_i|t_{i-1})$

So

$T\ ^*=\text{argmax}\prod_{i=1}^{m}\ P(w_i|t_i)P(t_i|\ t_{i-1})$

Here, $P(w_i|t_i)$ is the probability of the word w_i whose part of speech is t_i, $P(t_i|\ t_{i-1})$ refers to the transition probability of the part of speech t_{i-1} to part of speech t_i [6].

Thus we get the final part of speech tagging results of the Multi-category words.

Through these calculations, we use maximum likelihood estimation to estimate the two probabilities from the relative frequency:

$P(w_i|t_i)=C(w_i,t_i)/C(t_i)$

$P(t_i|t_{i-1})=C(t_{i-1},t_i)/C(t_{i-1})$

Here $C(w_i, t_i)$ is the number of occurrences of the word w_i as the part of speech t_i in modern Chinese part of speech tagging corpus. $C(t_i)$ is the number of occurrences of the part of speech of t_i; $C(t_{i-1},\ t_i)$ is the number of two adjacent part of speech for T_{i-1} and t_i.

Using statistics of January 1998 People's Daily idiom material (8621K), we obtain the POS transfer matrix and part of speech frequency table[7]. Referring to the following data:

回顾 (v, non multi-category)发展(n/ v, multi-category)的(u, non multi-category), 历史 (n, non multi-category).

The calculation process according to the following presentation:

P(n| v)=C (v, n)/C(v)=47 009/ 184 765= 0. 25

P(v | v)=C (v, v)/C(v)=30 484/ 184 765= 0. 16

P(发展|n)=C(发展，n)/ C (n)=1 647/ 353 270= 0. 005

P(发展|v)=C(发展，v) / C(v)=1 568/ 184 765= 0. 008

p (n| v) p(发展| n)=0. 25* 0.005= 0. 00 125

p (v| v) p(发展| v)=0. 16* 0. 008= 0. 00 128

We can see, the probability of v is greater than the probability of n for 发展. Therefore, the part of speech tagging result is as follows:

回顾/v发展/v的/u历史/n

With the help of above hidden Markov model, we put the algorithm into the multi-category word part of speech tagging computer system and tag the existing corpus. The tagging result can basically meet the needs of multi-category words part of speech tagging in the modern Chinese.

Ⅵ COMPARED TO OTHER METHODS

In addition to the hidden Markov model, the cascaded hidden Markov model, maximum entropy method and decision tree methods are methods based on the statistical method. These methods are used in the different POS tagging methods.These methods are as follows.

A. *Cascaded Hidden Markov Model*

Cascaded Hidden Markov Model (referred to as CHMM) is a multi-layer hidden Markov model.It is actually a combination of simple HMM. Several ways between the layers of hidden Markov models associated with each other, forming a close coupling relationship between each layer. Each HMM layer share a word segmentation map as the public data structure; each layer of hidden Markov model using N-Best strategy, a plurality of the best results will produce to the word map for the higher level model; low layer HMM provides data to high layer HMM at the same time, and also provides the support for the parameter estimation of the data. The time complexity of the system and HMM are the same [8].

B. *Maximum entropy Method*

The maximum entropy principle was first proposed by E. T Jaynes. Many things show a certain randomness, the results are often not identified. We do not know the probability distribution of the random phenomenon, only know a few samples or sample characteristics. How to make a reasonable judgment on the distribution, it is necessary to solve the problem. Maximum entropy model is based on a sample of information on a method for an unknown distribution of inference. When we have known some of the information, the most rational probability distribution is consistent with the given information and the maximal entropy distribution [9].

C. *Decision tree Method*

Decision tree is induction algorithm widely used in data mining. At the same time it is also a powerful tool for solving classification problems. The decision tree by constructing a two fork tree or multi fork tree, make the examples from the root node to a leaf node arrangement, and make the examples classification. A leaf node is the classification of examples. Each node on the tree for example specifies an attribute test, and each branch of the node Correspond with a possible value. The method of example classification is started from the root node. The node test attribute is specified, then according to the attribute of the given example value to grow down. This process repeats in the subtree that is the root again the new node [10].

To make a long story short, there are many methods of modern Chinese multi-category words part of speech tagging based on statistics.But after our experiment, we think that the hidden Markov model is the best method.

Ⅶ CONCLUSIONS

Part of speech tagging is very significant for the understanding of natural language especially in the syntactic analysis and the semantic analysis of the input text. This paper mainly introduces the part of speech tagging methods of modern Chinese multi-category word based on HMM model. Of course, the multi-category words part of speech tagging based on HMM model is just one of the different multi-category words part of speech tagging methods. There are also a lot of methods can solve these problems, and each methods has its own merits.

The exploration of the construction of special multi-category word information database is also the focus of our study in the future. The difficulty of this work is dynamic parts of speech tagging by computer, eliminate ambiguity, and make full use of the existing corpus tagging technology and the parts of speech research to solve the unsolved problems. We shall explore the corresponding rules of parts of speech and functions to build the functional system along with the existing system of parts of speech, to build a fully functional multi-category words information datebase.

ACKNOWLEDGEMENTS

Project sponsored by the natural science foundation of Heilongjiang province "Research of Modern Chinese Multi-category Words Part of Speech Dynamic Tagging System Based on Corpus" (No. F201215) ;

Project sponsored by the National Social Science Foundation "Research of Modern Chinese Multi-category Words Part of Speech Dynamic Tagging and Establishing of Corpus " (No. 10BYY066)

REFERENCES

[1].A.Liu Kaiying. “Automatic segmentation and part of speech tagging of Chinese text”. Beijing: The Commercial Press.2000, 5,pp.162-166.

[2].S.Huang Degen, Zhang Lijing, Zhang Yanli, Yang Yuansheng. “Disambiguation echanism Using Rule Techniques and Statistics Techniques”. Mini micro computer systems, Issue 7,2003, pp. 1252-1255.

[3].S.Wang Min, Zheng Jiaheng. “Chinese POS Tagging Based on Improved Hidden Markov Model”. Computer Applications, Issue 12,2006, pp. 197-199.

[4].S.Liu Qian, Jia Huibo. “A View of Chinese Word Automatic Segmentation Research in the Chinese Information Disposal”. Computer Engineering and Applications, Issue 3,2006,pp. 175-177.

[5].S.Wang Donghai, Zhao Wei, Chen Jie, Liang He. “Analysis and Improvement for the Automatic Part of Speech Tag of Chinese Based on Hidden Markov Model”. Journal of Changchun University of Technology(Natural Science Edition), Issue 3,2007, pp. 48-52.

[6].S.Yuan Lichi, Zhong Yixin. “A Novel POS Tagging Model”. Microelectronics and Computer, Issue 9,2005,pp. 1-2.

[7].S.Wen Rui, Zhu Qiaoming, Li Peifeng. “The Application of HMM and Negative Feedback Model in Part of Speech Tagging”. Journal of Suzhou University (Natural Science Edition), Issue 3,2005, pp. 39-42.

[8].S.Liu Qun, Zhang Huaping, Yu Hongkui, Cheng Xueqi. “Chinese Lexical Analysis Using Cascaded Hidden Markov Model”. Journal of Computer Research and Development. Issue 8,2004, pp.1421-1429.

[9].S. Lin Hong, Yuan Chunfa, Guo Shujun. “Chinese part of

speech tagging method based on maximum entropy principle". Computer application, Issue 1,2004, pp. 14-16.

[10].S.Wang Pengcheng. "Improved part of speech tagging of hidden Markov model". Journal of Henan College of Finance & Taxation. Issue 8,2009, pp. 88-89.