CHINESE POS TAGGING BASED ON MAXIMUM ENTROPY MODEL

JIAN ZHAO, XIAO-LONG WANG

School of Computer Science and Technology, Harbin, Harbin Institute of Technology, 150001, China
E-MAIL: zhaojian@insun.hit.edu.cn, xlwang@insun.hit.edu.cn

Abstract:

POS (part of speech) tagging is the basic work in Natural Language Processing (NLP). The tagging precision will have an important effect on the result of later process, such as syntax analysis. In this paper, a Chinese POS tagger based on maximum entropy model is presented, which trains from a large corpus annotated with Chinese POS tags and assigns a best tag sequence to a Chinese sentence to be annotated. In this model, all the features that are useful to predicate the POS tags are mined to make the model close to the real case. In addition, for the problem of overfitting, a smoothing method and a POS dictionary are maintained to reduce the model's dependence to training data and improve the efficiency of searching process. Open testing results shows that Chinese POS tagging with this method can achieve the state-of-art accuracy (96.8%).

Keywords:

Part of speech tagging; Maximum entropy model; Features selection; Smoothing model

1 Introduction

Many Chinese words are ambiguous in their part-of-speech. For example, in Chinese sentence "新华社记者前方报道", the word "报道" is a verb, nevertheless in another sentence "在这篇报道中...", the word "报道" is a noun. A Chinese POS tagger is a system that assigns a proper tag to each word in the sentence using word's context information.

POS tagging is the basic work in Natural Language Processing (NLP), which has an important role in many NLP domains, for example, in information extraction system, one of the main intentions is to find out the significant nouns (including compound noun). In syntax analysis system, the accuracy of POS tagging has a direct influence on the ultimate result. Recently several approaches have been used for building a POS tagging system. Generally they can be categorized into three classes. The first class is based on rules [1][2], in which the methods of data mining or knowledge discovery are adopted to acquire rules automatically from annotated text. The second class is based on statistics [3][4][5], in which Hidden Markov Model is widely used which assumes that a word depends probabilistically on just its part-of-speech category. This category depends on the proceeding two POS tags. Doung Cutting [6] has developed a POS-tagger based on a hidden Markov model whose accuracy exceeds 96%. The third class is based on neural network. Nakamura et al. [7] trained a 4-layer feed-forward network with up to three proceeding part-of-speech tags as an input to predicate the tag of next word. The accuracy of this tagger was similar to that of a trigram-based tagger.

Recently the maximum entropy has become a researching hotspot, which has been adopted in NLP for English text, including machine translation [8], language modeling [9] and text classification [10]. Adwait used maximum entropy model for English POS tagging firstly. He compared this model to other POS tagging techniques, which resulted in a conclusion that it is an extremely flexible technique for linguistic modeling, and it can perform at 96.6% accuracy for English POS tagging [11].

In this paper, the maximum entropy model is introduced for Chinese part-of-speech tagging. The rest of the paper is organized as follows: section 2 introduces the maximum entropy model; section 3 describes the features selection for Chinese POS tagging; model smoothing is discussed in section 4;section 5 states the results and analysis of the experiment; finally the conclusion is presented in section 6.

2 . Maximum Entropy

Many problems in natural language processing can be viewed as tasks of classification: given a context O, to make a decision that which class is fittest for the incoming event a . It can be implemented with a conditional probability distributions p , so the dual problem is to find $c \in C$ which maximizes the $p(c \mid O)$. Maximum entropy is a method for estimating probability distributions from training data. The principle is simple: given a collection of constraints, select a model, which satisfies all the facts, but is as uniform as possible synchronously. For example, in Chinese POS tagging scenario, suppose that there are 10 linguistic tags in our model. Without any concurrence information of the word "报道" with these tags, the chance of classifying "报道" into each tag is equal (10%). From the training data set, we are told that the occurrence of the case that "报道" is tagged with "NOUN" accounts for 40%. Intuitively, when the incoming word is "报道", it would be said that the word has a 40% probability of being a noun, and a 6.67% probability of being each of the other 9 classes. If a word does not appear

0-7803-7508-4/02/$17.00 ©2002 IEEE

in a trained corpus, it has 10% chance of being tagged with each of these tags, that is to say the class distribution is uniform.

The objective of MEM (maximum entropy model) is to find a best probability distribution to maximize the conditional entropy:

$$H(p) = -\sum \tilde{p}(O)p(c \mid O)\log p(c \mid O) \qquad (1)$$

The features in MEM are a set of binary-value functions, i.e.

$$f_i(c,O) = \begin{cases} 1 & if : c = c' \ and \ \Pr ed(O) \\ & = TRUE \\ 0 & else \end{cases} \qquad (2)$$

where $\Pr ed(O)$ is a logistic function which denotes whether the context of current word is consistent with the specific predictive information. The constraints put on this model are that the empirical feature expectations equal the model expectations [11], i.e.,

$$P = \{p \mid E_p f_i = E_{\tilde{p}} f_i, 1 \le i \le k\} \qquad (3)$$

where k is the total number of features, and

$$E_p f_i = \sum_{c,O} \tilde{p}(O)p(c \mid O)f_i(c,O)$$

$$E_{\tilde{p}} f_i = \sum_{c,O} \tilde{p}(c,O)f_i(c,O) = \frac{1}{N}\sum f_i(c,O) \qquad (4)$$

where N is the total number of training samples, e.g., $\{(c_1,O_1),\cdots,(c_N,O_N)\}$. The distribution of $p(c \mid O)$ is always of parametric exponential form [8]:

$$p(c \mid O) = \frac{1}{Z(O)}\prod_{i=1}^{k}\alpha_i^{f_i(c,O)} \qquad (5)$$

$$Z(O) = \sum_c \prod_{i=1}^{k} \alpha_i^{f_i(c,O)} \qquad (6)$$

3 Features Selection

The features used in MEM should embody any information that might be helpful for predicating tag *t*. If the feature exists in the feature set of the model, its corresponding coefficient will contribute towards the conditional probability $p(c \mid O)$. In order to mine these features, a context window with five words length is maintained when the training data is scanned, which comprises n-gram and n-pos information, e.g.

Table 1. the context window

W_{i-2} W_{i-1} W_i W_{i+1} W_{i+2}
t_{i-2} t_{i-1} t_i

In addition, the information of word formation is also been explored as the feature to predicate the tags. For example, there is a Chinese sentence in corpus:

身高/n 165/m 厘米/q 的/u 他/r 是/v 世界/n 上/f 最好/a 的/u 左/f 后卫/n。

When the word “的” is to be tagged, the current context window is made up of “165、厘米、的、他、是” five words. We can get such following features:

Table 2. features from a sentence

w_i=的	t_i=u
w_{i-1}=厘米	t_i=u
w_{i-2}=165	t_i=u
w_{i-1} w_i =厘米 的	t_i=u
$w_{i-2}w_{i-1}w_i$ =165 厘米 的	t_i=u
t_{i-1}=q	t_i=u
t_{i-2}=m	t_i=u
t_{i-2} t_{i-1}=m q	t_i=u

If the scanned word contains number, such as the word “165” in above sentence, another kind of feature can be added:

IsWordContainNum(w_i)=true t_i=m;

Above features used for Chinese POS tagging are the same as those used for English. Other than these, some peculiar features of Chinese are mined. Chinese words are constructed by many morphemes in three manner: reduplicating 、 deriving and compounding. Although Chinese words are not of rich morphology, many rules are still discovered which are useful for POS tagging. We discuss these respectively:

1) Chinese reduplicate word: The reduplicate phenomenon of Chinese word is more complicated, which has several kinds of forms: AA、ABB、AAB、A 里 AB、AABB、ABAB、A 一 A、A 了 A、A 了一 A(where A or B Represents a Chinese character). We have got some features by researching Chinese corpus, such as,

Table 3. features from reduplicate word

w_i has a form of ABB (or A里AB、AABB) and AB is an adjective e.g. 孤单单、糊里糊涂	t_i=z
w_i has a form of AA and A is a monosyllabic adjective. e.g. 好好、慢慢	t_i=ad
w_i has a form of AA的 and A is a monosyllabic adjective. e.g. 酸酸的	t_i=z
w_i has a form of AA and A is a monosyllabic noun morpheme. e.g. 狗狗、猫猫	t_i=n
w_i has a form of AA and A is a monosyllabic number. e.g. 万万	t_i=ad
w_i has a form of AA and A is a monosyllabic Adverb. e.g. 刚刚	t_i=ad

2) Derivative: some Chinese words are constructed by adding a prefix or suffix to root. There are some features based on prefix or suffix to predicate the POS tags of some derivative.

Table 4. features from Derivative

The suffix of w_i is "头". e.g. 对头	t_i=n
The suffix of w_i is "子" e.g. 耙子	t_i=n
The suffix of w_i is "机" e.g. call机	t_i=n
The suffix of w_i is "家" e.g. 银行家	t_i=n
The suffix of w_i is "员" e.g. 海员	t_i=n
The suffix of w_i is "者" e.g. 生还者	t_i=n
The suffix of w_i is "生" e.g. 实习生	t_i=n
The suffix of w_i is "性" e.g. 不确定性	t_i=n
The suffix of w_i is "化" e.g. 电器化	t_i=v
The suffix of w_i is "们" e.g. 老人们	t_i=n
The prefix of w_i is "老" e.g. 老王	t_i=n

4 Smoothing the Model

In general, probability models estimated under the maximum entropy principle perform well in practice, but they have certain limitations that may lead to poor predication. The more common limitation is overfitting, that is to say that the results of tagging depend on the training data excessively. Especially when the volume of training data is not large enough, the problem of sparse data is very serious and the model is not reliable. For example, suppose that the data we want to sample is a pair of (w_i, t_i), and let the training data be all the Chinese words and their corresponding POS tags occur in some corpus of Chinese text. Consider the pair of (怪, v) which does not appear in the training data, we can get $E_{\ddot{p}} = \tilde{p}(c,O) = \tilde{p}(v,怪) = 0$,but in fact the Chinese word "怪" has some chance of concurring with the tag "v", e.g. in sentence "都/ad 怪/v 噪声/n 太/ad 大/a". So let $E_p = \tilde{p}(O)p(c \mid O) = \tilde{p}(怪)p(v \mid 怪) \neq 0$. In order to make the model more predicable, it must be capable of dealing with the circumstance that the expected value of some features over the model does not match exactly with that of the empirical distribution which we deem unimportant. In this paper, we solve this problem by adopting a smoothing method of Good-Turing discounting.

Good-Turing discounting can be viewed as the maximum entropy analog to Katz smoothing for conventional n-gram models [13]. Rosenfeld argued in [14] that the target value of the model $p(c \mid O)$ should be discounted. Instead of giving MEM's constraints by equation (3), the following constraints are proposed:

$$P = \{p \mid E_p f_i = E_{\tilde{p}GT} f_i, 1 \le i \le k\} \qquad (7)$$

$$E_{\tilde{p}GT} f_i = \frac{r^*}{N} \qquad (8)$$

We can find the differences between equation (8) and equation (4) that the frequency of the event is replaced by r^*, which is defined as follows [15]:

$$r^* = (r+1)\frac{E(N_{r+1})}{E(N_r)} \qquad (9)$$

where r is the frequency of event, and N_r is the frequency of frequency r or the number of members of the population with frequency r , $E(x)$ represents the expectation of the random variable x. when r is equal to 0, $E_{\tilde{p}GT} f_i$ should be calculated as follows:

$$E_{\tilde{p}GT} f_i = \frac{N_1}{N} \qquad (10)$$

The result of smoothed model has been compared with that of an ordinary model, which will be discussed in detail in next section.

5 Experimental Results

The tagger based on maximum entropy principles is trained on one-month articlés of People's daily with 1.2 million words. Open and close tests are carried out to validate the performance of the tagger.

In table 5 the accuracy rate of MEM-based tagger is compared to that of trigram-based tagger, which is trained on the same data. In order to observe the influence of the size of training corpus, the taggers are trained on the data of different volumes. The results are presented in figure 1.From these experimental results, we can find that the performance of MEM-based tagger is better than that of trigram-based tagger for Chinese text. They further indicate that accuracy of the MEM-based tagger is less affected by a smfall amount of training data than that of trigram-based tagger.

Table 5. comparison of accuracy rate

method	Open test	Close test
MEM	96.8%	97.1%
trigram	94.06%	95.8%

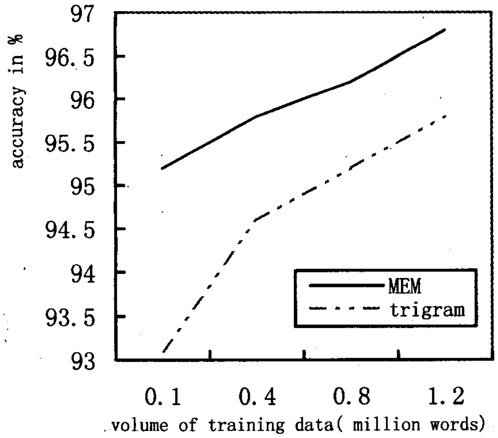

Fig.1. accuracy on varying volumes of training data

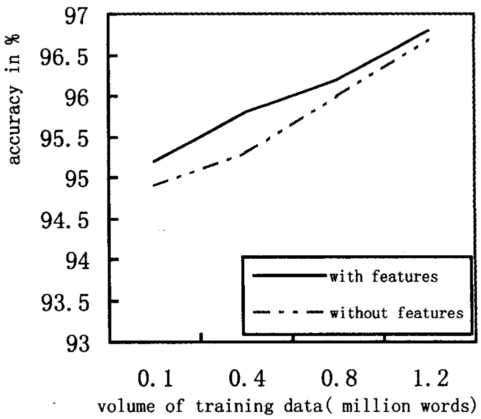

Fig.2. the influence of peculiar features on accuracy

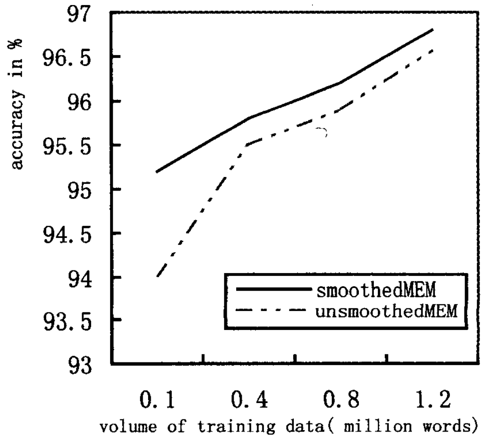

Fig.3. the result of comparison smoothed model to unsmoothed model

It is further tested whether the peculiar features for Chinese text have an effect on the tagging accuracy. The curves in figure 2 show that when the size of training data is small (0.1 million words), the peculiar features can shift the system accuracy from 94.9% to 95.2%. When the volume of training data is large, they have little influence.

Figure 3 indicates that smoothing algorithm is necessary for Chinese MEM-based tagger, which can improve the system performance with an increase of 0.5 percent.

6 Conclusions

In this paper, a Chinese POS tagger is presented, which is based on maximum entropy model. The experimental results show that MEM-based Chinese POS tagger can achieve the state-of-art accuracy (96.8%). The smoothing method can improve the performance of MEM to a certain extent. It is proved that the maximum entropy principle has a superior performance for Chinese linguistic modeling, because it mines and combines all the features that are useful for predicating the POS tags.

References

[1] Eric Brill. A simple rule-based part of speech tagger. In Proceedings of the Third Conference on Apllied Computational Linguistics, 1992.

[2] xiaoli li, zhongzhi shi. A data mining method to acquire part of speech rules in Chinese text. Journal of Computer Research & Development, Vol. 37, No. 12, 2000.

[3] S.DeRose, Grammatical Category Disambiguation by Statistical Optimization. Computational Linguistics, 1988.

[4] André Kempe. A Probabilistic Tagger and an Analysis of Tagging Errors. Technical report, 1993.

[5] Guohong fu. Research on Statistical Methods to Chinese Syntactic Ambiguity Resolution. A dissertation for the doctoral degree. 2001.

[6] Doug Cutting, Julian Kupiec, etc. A Practical Part of Speech Tagger. Xerox Palo Alto Research Center, 1992.

[7] M.Nakamura, K.Maruyama, et al. Neural Network Approach to Word Category Predication for English Text. In proceedings of the international conference on computational linguistics, 1990.

[8] Adam Berger, et al. A Maximum Entropy Approach to Natural Language Processing. Computational linguistics, 1996.

[9] Ray Lau, Ronald Rosenfeld, et al. Adaptive Language Modeling Using the Maximum Entropy principle. In proceedings of the human language technology workshop, 1993.

[10] Kamal Nigam, John Lafferty, et al. Using Maximum Entropy for Text Classification. In proceedings of the IJCAI-99 workshop on information filtering, Stockholm, SE, 1999.

[11] Adwait Ratnaparkhi. A Maximum Entropy Model for PArt-of-speech Tagging. In proceedings of conference on empirical method in natural language processing, university of Pennsylvania, 1996.

[12] S.M.Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Trans. on Acoustics, speech and signal processing, 1987.

[13] Stanley F.Chen, Ronald Rosenfeld. A Gaussian Prior for Smoothing Maximum Entropy Models. Technical report, CMU-CS-99-108, 1999.

[14] R.Rosenfel. Adaptive Statistical Language Modeling: Amaximum Entropy Approach. PH.D thesis, computer science department of Carnegie Mellon University, 1994.

[15] W Gale. Good-Turing Smoothing Without Tears. Technical report. AT&T Bell laboratories, 1996.