基于受限最大熵模型的汉语词性标注的研究

袁彩霞，王小捷

（北京邮电大学信息工程学院，100876，北京，中国）

摘　要： 本文研究了一种基于受限最大熵模型的汉语词性标注系统。论文提出了递增引入和错误驱动的特征选取方法，而后描述了标注任务中受限的模型学习特性。最后通过北京大学简体人民日报语料，我们验证了该方法的可行性。

关键词： 最大熵；汉语词性标注；特征选择

Chinese POS Tagging Using Restricted Maximum Entropy Model

Caixia Yuan, Xiaojie Wang, Junjie Zhai

(Center for Intelligence Science and Technology Research,
Beijing University of Posts and Telecommunications, 100876, Beijing, China)

Abstract: This paper presents issues related to POS tagging for Chinese language using maximum entropy technique, in which we introduce a novel feature selection strategy based on incremental trail and error-driven, then we describe the restricted method for model learning in tagging task. We test our method on the simplified Chinese corpus of Peking University China and show that a significant improvement is obtained.

Keywords: Maxent Entropy Model, Chinese POS Tagging, Feature Selection

1 Introduction

Part-Of-Speech (POS) tagging is to assign each word in the text a POS tag such as Noun and Verb, which is the base of some Natural Language Processing (NLP) tasks and helpful in improving performance of NLP applications, such as parsing, machine translation and information extraction.

There has been a great deal of work exploring methods for automatically training POS taggers. Brill introduced laboriously handcrafting rules for tagging [9]; many papers has reported Markov-model based tagger which has been used for both English POS tagging [7], [8] and Chinese POS tagging [6]. But as a probabilistic generative model, Markov tagger has shortcomings

The research work of this paper is supported by MOE funded project of "Tools for Chinese and Minority Language Processing" (No. MZ115-022).

作者简介：袁彩霞(1982，11-)，河南省项城市，北京邮电大学博士，E-mail:yuancaixia@gmail.com

for mining useful information latent in the long-distance dependent context due to its strict independence assumption.

In the past few years the maximum entropy (MaxEnt) approach has been recommended as effective probabilistic model for its flexible feature selection and its novel method to smoothing. The typical and successful applications of MaxEnt for NLP are referred in [1] and [5], used in machine translation and English language POS tag assignment and parsing.

In this paper, we apply MaxEnt model to Chinese POS tagging, and describe a novel feature selection method and restricted model learning for MaxEnt. When tested on the corpus of Peking University, our tagger achieves tag accuracy of 97.7946% for 44 tags and 98.5944% for 20 tags, which outperforms HMM method that reported 94.86% for 26 tags [15] and other MaxEnt model that reported 95% for 20 tags [14].

In the remainder of this paper, we first briefly describe the maximum entropy principle in section 2 and then discuss our method of feature selection in section 3. Section 4 introduces the restricted training for our Chinese MaxEnt POS tagger. Section 5 presents our experiments and discussion of the results. We conclude the paper in section 6.

2 Maximum Entropy Modeling

Maximum entropy model offers a measurable way to estimate the probability of a certain linguistic class y occurring with a certain linguistic context x. It is an approach to statistical modeling using log linear distributions. According to maximum entropy principle, we seek for a conditional distribution $p(y \mid x)$ that has the maximum entropy:

$$p^*(y|x) = \arg\max_{p \in P} H(p) = \arg\max_{p \in P} [-\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x)] \quad (12)$$

In this framework, events of a given context are represented as multiplicities of weighted features latent in the contextual predicates. In POS tagging task, to express the event that a word in some context is labeled as a certain tag, we define an indicator function [1] with the form as follows to numerically present a feature, where *cp* denotes the contextual predicates of a specific word:

$$f(x,y) = \begin{cases} 1 \ if \ tag = y, and \ cp = x \\ 0 \ \text{otherwise} \end{cases}$$

Then $p(y \mid x)$ is given by:

$$p(y \mid x) = \frac{1}{Z_\lambda(x)} \exp(\sum_{i=1}^{n} \lambda_i f_i(x,y)) \quad (2)$$

where, $Z_\lambda(x)$ is the normalizing factor:

$$Z_\lambda(x) = \sum_y \exp(\sum_{i=1}^{n} \lambda_i f_i(x,y)) \quad (3)$$

Some iterative algorithms [4] have been used for parameter estimation of the model. Complete review of such methods is beyond the scope of this paper. Rather, we concentrate on the aspects of feature selection and training procedures that most directly influence the model performance.

3 Features for POS Tagging

The model depends on the distribution of the features and the information represented by them, thus it is significant that the features used be pertinent to the task. Much research has employed the features for POS tagging corresponding to the morphological and contextual evidence. Other than generating features based on the predefined template [5], we introduce our

feature selection approach that bases on incremental trail and is driven by error detection.

3.1 Feature Selection Based on Incremental Trails

The POS tag of a word is not only related to the word itself, but also closely related to words surrounding it and their POS tags, which is the context that word appears. Upon such knowledge, we choose the window length of context varies from 1 to 3 centered on the current word (i.e., the word under examination) to choose features that are most contributive for the POS tagging.

Firstly, the feature set with narrowest window (we name it by *seed features*) is formed as "current word, previous word, succeeding word", then based on the *seed features*, we add or cut down features with respect to increase or decrease of the tagging accuracy.

In each trail, we try on one predicate that falls into the context window. If the tagging accuracy increases due to adding this feature, we reserve it to the next pass of trail; otherwise, we delete it from the window. When predicate candidates in the window are all exhausted, or being added into the feature set, we expand length of the window and begin a new pass of trail until the length reaches 3. Besides words themselves that surround the center word, POS tags and POS tag sequence of the previous words are also induced as pragmatic features, and experiments show that the POS tags information increase tagging accuracy greatly. Table 1 shows some of feature sets the model tests in the incremental trails. In the table, *curword* denotes the current word, *label* is POS tag of the *curword*, $preword_i$ denotes the i-th word before current word, and $pretag_i$ is tag of $preword_i$, $sucword_i$ is the i-th word succeeding current word.

Table 1. Some feature set used in our increametal trail.

Trail pass	Feature set
1	label curword $preword_1$ $sucword_1$
2	label curword $preword_1$ $pretag_1$ $sucword_1$
...	...
n	label curword $preword_1$ $pretag_1$ $prewors_2$ $pretag_2$ $preword_3$ $pretag_3$ $sucword_1$ $sucword_2$ $sucword_3$

Based on the incremental trail, we obtain the optimal features as in table 2, in which, feature "$pretag_{12}$" represents the joint tag sequence of $pretag_1$ and $pretag_2$.

Table 2. The optimal features of events based on the increamental trail.

	Optimal features
Event Representing	label curword $preword_1$ $pretag_1$ $pretag_{12}$ $sucword_1$

3.2 Feature Induction Driven by Error

According to the number of different POS tags that a word owns in language applications, words are clustered into two types: words with more than one POS tags and those with single POS tag. In some sense, the former refers to trans-classed word in traditional word morphology.

By detecting the label errors when training and testing using the features in table 2, we observe that error rate for the multi-tag words is highest, and label being given to them is most closely related to POS tag of word previous to them. For instance, in corpus of Peking University, word "*报道(Report)*" has three different tags of "*n (noun), v(verb), vn(noun verb)*". But when taking into account tag of the previous word near to it (named by $pretag_1$ as in table 1), we observe that when $pretag_1$ is "*q*", "*Report*" is labeled as "n" with probability of 20.37%, of 0.15% for "v" and of 0.0% for "*vn*". So we can assume that when $pretag_1$ is "*q*", the word "*Report*" is labeled as "*n*" with highest probability, and as "*vn*" with lowest probability. Such probability is viewed as "*discriminating feature*" when choosing among the multi tags for multi-tag words.

For multi-tag words, we computer the probability as follows:

$$p(pretag_1 = t_i \mid tag = t_j) = \frac{count(pretag_1 = t_i, tag = t_j)}{count(tag = t_j)} \quad (4)$$

where $count(tag=t_j)$ is the number of a word being labeled as t_j, and $count(pretag_1 = t_i, tag = t_j)$ is the number of co-occurrence of $pretag_1 = t_i$ and $tag=t_j$.

But for sole-tag words, no matter what the $pretag_1$ is, the label for them is always be the same one. So we define the "*discriminating feature*" for such kinds of words as:

$$p(pretag_1 = t_i, i = 1...n \mid tag = t) = 1 \quad (5)$$

Such "*discriminating feature*" is supplemental to features in table 2 as element "*preprob*", that is, the new feature set of events has the format as in table 3. Experiment shows that with such particular feature, the predicate accuracy is significantly improved, which will be demonstrated in section 5.

Table 3. The integrated features of the optimal features and probability feature.

	Integrated features
Event Representing	label curword $preword_1$ $pretag_1$ $pretag_{12}$ $sucword_1$ preprob

4 Restricted Maximum Entropy Model

Performance of a statistical model is largely influenced by corpus and modeling algorithm. In the framework of supervised learning, MaxEnt model is trained by the tagged corpus, which is used for tag predicating in test procedures. But tagged training corpus is never enough to reliably specify $p(y \mid x)$ for all possible (x, y) pairs, since the words in x are typically sparse. The challenge is then to find assistance for external evidence about x and y to reliably estimate the probability distribution.

Let Y be a finite set of output values ever appearing in the labeled training corpus, the MaxEnt model predicates a conditional probability $p(y \mid x)$ for every $y \in Y$ given a certain linguistic context x. But in POS tagging task, the label options for a word are usually a small subset of the whole tag candidates. That is, the probabilities should be distributed among tags that the word ever owns, but not on the whole tag set.

In order to resolve two problems above, we introduce the external knowledge for the limited tagged corpus and construct a sun-classifier for word to reduce the searching dimension when labeling it.

The external knowledge is from Grammatical Knowledge-base of Contemporary Chinese [3] (GKCC). Besides automatically constructed events of the words that occur in labeled training data (*internal word*), we also create events for words that occur in GKCC dictionary but not appear in the training data (*external word*). But unlike self-contained contextual predicates of the *internal words*, such kind of events is created only with one feature, that is the word itself, and other features is deemed as null. Table 4 shows different event representations of the two different types of words.

Table 4. The Event Representing Format for Internal Words and External Words

Internal words	External words

Event Representing	label curword preword pretag sucword	label curword

MaxEnt model learns from events both of labeled corpus and of GKCC during training period. Meanwhile, the model constructs a tag bag for every word w occurring in either labeled corpus or GKCC, or in both of them. Let $Tag_c = \{tag_s..tag_t\}$ be w's tag set learning from labeled corpus, and $Tag_v = \{tag_u..tag_v\}$ be its tag set deducted from the GKCC vocabulary, then w's tag bag Tag_w is the conjunction of Tag_c and Tag_v ($Tag_w = Tag_c \cup Tag_v$). When the search procedure needs to tag word w and w exists in training data, only tags from w's tag bag are considered as tag candidates for it; if w is not in training data (w is *new word*), the search procedure explores all tags in overall tag set. By this way, we construct a virtual sub-classifier for each word and restrict the model in both training and testing.

5 Experiments

To test the effectiveness of our features selection and restricted modeling method, we conduct experiments and observe the accuracy of different models with different features and modeling methods.

We choose People's Daily corpus of January in 1998 from Peking University (PKU'9801) as our experimental data. PKU'9801 consists of 44 POS tag labels in total. But for validation and comparison, we combine some of the semantic related tags, for example, *vd*(adverbial verb), *vn*(noun verb), *Vg* (verb morpheme) to *v* (verb), and then obtain a new corpus with 20 tags.

The PKU'9801 has about 1,120 thousand words (include punctuations), and is split into two disjoint sections: 80% of them serve as training data, and the other 20% serve as test data. The experiments for 44 tags and 20 tags are conducted respectively. We conduct *baseline* experiment on events purely from the training data with features in table 2, and other enhancing strategies are compared with the baseline setting.

To observe the influence of each modeling factor presented in section 3 and 4, we design incremental experiments analogous to that of feature selection. First, we conduct the baseline experiment (*Baseline Model*), and then we apply the restricted train and test method to baseline model to confirm efficiency of external events and sub-classifying technique (*Restricted Model*).
To validate the impact of probability features on model performance, we add such "discriminating features" to events of both Baseline Model and Restricted Model. Figure 1 shows the flow of our experiments.

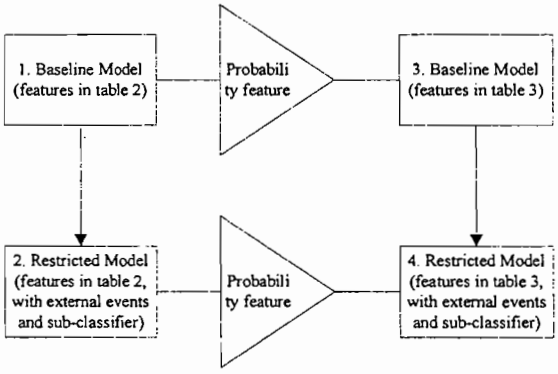

Fig. 4. Flow of our experiments.

The tagging accuracy for 4 experiments (shown in Fig. 1) is demonstrated in Table 5 and table 6. From table 5, we can see that, restricted model outperforms baseline model by about 1.15% with the same features. Inducing of external words from the GKCC dictionary into the training data reduces the number of new words and decreases the error rate of them remarkably, but increases the burden of training task due to the enlargement of training events. Even though the use of sub-classifier yields an insignificant (0.78% of overall 1.15 by further analysis) improvement in accuracy, it is used in further experiments since it reduces the average number of tags that are explored for each word, and thus significantly speeds up the tagger

Table 5. Performance of the Baseline model and Restricted Model both with features in table 2 。

Features	label curword $preword_1$ $pretag_1$ $pretag_{12}$ $sucword_1$			
Models	Baseline Model		Restricted Model	
Accuracy	44 tags	20 tags	44 tags	20 tags
(%)	94.8139	96.3334	95.9601	97.2588

Table 6. Performance of the Baseline model and Restricted Model both with feature in table 3.

Features	label curword $preword_1$ $pretag_1$ $pretag_{12}$ $sucword_1$ preprob			
Models	Baseline Model		Restricted Model	
Accuracy	44 tags	20 tags	44 tags	20 tags
(%)	96.1122	97.4359	97.7946	98.5944

References:

[1] Berger, A., Della Pietra, S. and Della Pietra, V.: A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics. 22(1) (1996) 39-71

[2] Pietra, S. D., Pietra V. D. and Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (1997) 19:380-393

[3] Yu, S. W., et al.: The Grammatical Knowledge-base of Contemporary Chinese-A Complete Specification, Edition 2. Tinghua University Press (2003)

[4] Malouf, R., et al.: A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In Proceedings of the Sixth Conference on Computational Language Learning (CoNLL-2002), Taipei (2002)

[5] Ratnaparkhi, R.: Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. thesis, University of Pennsylvania

[6] Chang, C. H. and Chen, C. D.: HMM-based Part-of-Speech Tagging for Chinese Corpora, workshop on Acquisition of Lexical Knowledge from Text (1993)

[7] Kupiec, J. M. Robust Part-Of-Speech Tagging Using a Hidden Markov Model. Computer Speech and Language (1992) 6:225-242

[8] Scott M. Thede and Mary P. Harper A Second-Order Hidden Markov Model for Part-of-Speech Tagging, In Processing of the 37th Annual Meeting of ACL (1999)

[9] Brill, E.: Some advances in rule-based part of speech tagging. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa.(1994)

[10] Mullen, T. and Osborne, M.: Overfitting avoidance for stochastic modeling of attribute valued grammars. In Proceedings of the Fourth Conference on Computational Natural Language Learning, Lisbon (2000) 49-54

[11] Chen, S. and Rosenfeld, R.: A Gaussian Prior for Smoothing Maximum Entropy Models. Technical Report, CMU-CS-99-108, Carnegie Mellon University, Pittsburg, February (1999)

[12] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. 18th International Conf. on Machine Learning (2001)

[13] Miller, S., Guinness, J., and Zamanian, A.: Name Tagging with Word Clusters and Discriminative Training. In Proceedings of HLT-NAACL (2004)

[14] Lin, H., Yuan, Ch. F., Guo, S. J.: A Chinese Part of Speech Tagging Method Based on Maximum Entropy Principle. Computer Applications, China, Vol. 24, No. 1(2004)

[15] Zhang, M., Li, S., et al.: Part of Speech Tagging Chinese Corpus Based on Statistics and Rules. Journal of Software China, Vol l. 9, No. 2 (1998)