WorkShop on Artificial Intelligence

A HYBRID METHOD FOR AUTOMATICALLY TAGGING CHINESE SENTENCES

Abstract

Automatically tagging Part-of-Speech is an important task in natural language processing. A better tagging result can increase the accuracy and performance of sentences parsing and semantics analysis. The major problem of the tagging is that the ambiguity in multiple-tag words caused the accuracy of the tagging decrease. Many approaches were proposed to solve the problem. However, some drawbacks still existed in these methods. In this paper, according to Chinese linguistic properties, we presented a hybrid method from two-way adjoining relation method and temporally clustering modification method for improving tagging. In preliminary experiments, it can really increase the accurate rate of the tagging for Chinese**.**

Authors: Tao-Hsing Chang and Chia-Hoang Lee

Department of Computer and Information Science, National Chiao Tung University

Address: 1001 Ta Hseuh Road, Hsinchu, Taiwan 300, R.O.C.

Email: thchang@cis.nctu.edu.tw, chl@cis.nctu.edu.tw

Phone: 0921061743

Fax: (02)23948937

Contact author: Tao-Hsing Chang

Keywords: natural language processing, POS tagging.

A HYBRID METHOD FOR AUTOMATICALLY TAGGING CHINESE SENTENCES

Tao-Hsing Chang and Chia-Hoang Lee

Department of Computer and Information Science
National Chiao Tung University, Hsinchu, Taiwan
thchang@cis.nctu.edu.tw   chl@cis.nctu.edu.tw

ABSTRACT

Automatically tagging Part-of-Speech is an important task in natural language processing. A better tagging result can increase the accuracy and performance of sentences parsing and semantics analysis. The major problem of the tagging is that the ambiguity in multiple-tag words caused the accuracy of the tagging decrease. Many approaches were proposed to solve the problem. However, some drawbacks still existed in these methods. In this paper, according to Chinese linguistic properties, we presented a hybrid method from two-way adjoining relation method and temporally clustering modification method for improving tagging. In preliminary experiments, it can really increase the accurate rate of the tagging for Chinese.

1. INTRODUCTION

Automatically tagging part-of-speech is an important task in natural language processing. A better tagging result can increase the accuracy and performance of sentences parsing and semantics analysis. Tagging for single-tag words is easy because it only consults the dictionary in which the words have been tagged, but choosing correct tags for multiple-tag words is quite difficult. Many studies[1][2][3][4][5][6] present various techniques to recognize the correct one from the possible tags of a word.

However, some limits exist in these methods respectively. For example, rule-based methods determine the tag of a word by using the rules generated from linguistic knowledge. The major disadvantages of the methods are that they require human effort to build and maintain the rules. Oppositely, statistical-based methods estimate the likelihood of each possible interpretation of a sentence/word by statistical values that are automatically generated from corpora. Based on the estimates, the most likely tag is then chosen. Generally, the tagging accuracy of statistical-based methods depends on the size of the corpus.

Some studies try to integrate these two types of methods in order to increase the tagging accuracy. Liu et al.[3] proposed a hybrid method for tagging Chinese words. It uses 27 rules to identify the tag of a word in the first stage. If the word cannot be tagged in the first stage, it uses relaxation labeling method to decide the tag of the word in the second stage. The hybrid method achieves on average 84 percent correct rate for multiple-tag words in preliminary experiences. However, the accurate rate of tagging in the second stage of relaxation labeling method is only average 74 percent, and the part occupies most of executing time of the hybrid method. Apparently, improving the performance of the tagging in statistical stage can increase the performance of tagging overall.

In this paper, we will present a hybrid method based on Chinese linguistic properties, and show that it can increase the accurate rate of tagging.

2. REVIEW SOME METHODS

Statistical methods use statistical data to select the best tag for a word from many possible tags. The methods usually use different evaluation formula involving such factors as the frequency of a word appeared with different tags to make decision. On the other hand, some methods evaluate the probability of each tag sequence of the sentence, and the tag sequence with the highest probability is the solution. Some methods[3] find the best suitable tag of the words individually. These methods have their advantages and disadvantages respectively. Below, we will briefly discuss these methods in detail.

2.1 The Hidden Markov Models

Hidden Markov models (HMM), a powerful statistical modeling, are frequently used in modern speech recognition systems. [2] applied Markov models to solve tagging problem and achieved fairly good result. Viterbi algorithm[7] is an approach which can implement Markov models efficiently. The algorithm provides an evaluation function to estimate the probability of each possible tag sequence, and select tag sequence with the highest probability as the tag sequence of the sentence. The method uses two factors to estimate the tag for a word: the relation between the tag of the word and the tag of previous word, and the probability of the word appeared with different tags. The relation is called preceding adjoining relation, and the probability is called lexical information.

Before using the algorithm, the method would need three different kinds of data from corpus: the number of times of each tag appeared in corpus, the number of times of each word occurred with different tags in corpus, and the number of times of each tag following the other tags. With these data, the method can compute the probability of preceding adjoining relation and lexical information by Equations (1) and (2).

$$P(t^j \mid t^k) = \frac{C(t^k, t^j)}{C(t^k)} \tag{1}$$

$$P(w^l \mid t^j) = \frac{C(w^l, t^j)}{C(t^j)} \tag{2}$$

where $C(t^k)$ is the number of times of tag t^k appeared in corpus, $C(t^k, t^j)$ the number of times of tag t^k following the tag t^j, and $C(w^i, t^j)$ the number of times of word w^i occurred with tag t^j in corpus

Using Equations (1) and (2), the method can compute the probability of each word appeared with different tags by Equations (3) and (4). The $\delta_i(t^k)$ in Equation (3) is the probability of the word w_i appeared with t^k in the sentence. The $\varphi_{i+1}(t^j)$ in Equation (4) is the tag of word w_i which provides the maximum probability to word w_{i+1} appeared with tag t^j. In addition, the method supposes the PERIOD is the start word of each sentence, and defines $\delta_1(\text{PERIOD}) = 1.0$, $\delta_1(t) = 0.0$ for $t \neq$ PERIOD .

$$\delta_{i+1}(t^j) = \max_{1 \le k \le T}[\delta_i(t^k) \times P(w_{i+1} \mid t^j) \times P(t^j \mid t^k)], 1 \le j \le T \tag{3}$$

$$\varphi_{i+1}(t^j) = \arg\max_{1 \le k \le T}[\delta_i(t^k) \times P(w_{i+1} \mid t^j) \times P(t^j \mid t^k)], 1 \le j \le T \tag{4}$$

Finally, the method can obtain the tag sequence of the sentence by Equations (5), (6) and (7).

$$X_n = \arg\max_{1 \le j \le T} \delta_n(t^j) \tag{5}$$

$$X_i = \varphi_{i+1}(X_{i+1}), 1 \le i \le n-1 \tag{6}$$

$$P(X_1 \ldots X_n) = \max_{1 \le j \le T} \delta_{n+1}(t^j) \tag{7}$$

where $X_1, \ldots, X_n$ are respectively the tags of the words $w_1, w_2, \ldots, w_n$.

Hidden Markov models are the first method which uses the lexical information in addition to preceding adjoining relation. The experiments[2] show that the accurate rate of HMM for tagging is over 90% for all words.

2.2 The Relaxation Labeling Method

Relaxation labeling method is a popular approach applied to image processing and background analysis. [3] presents a method which integrate relaxation labeling method and rule-based method to increase the accuracy of the tagging. It uses 27 rules to identify the tag of a word in first stage. If the word cannot be tagged in first stage, it uses relaxation labeling method to decide the tag of the word in second stage. The experiment in [3] shows that there are only average 74 percent accurate rates for tagging multiple-tag words using the relaxation labeling method in second stage. Below we briefly discuss relaxation labeling method.

Assuming the sentence $S = w_1 w_2 ... w_n$ will be processed. Furthermore, the method assumes the word w_i appeared with *j* possible tags $t_i^1, t_i^2, ..., t_i^r, ..., t_i^j$, the word w_{i-1} appeared with *k* possible tags $t_{i-1}^1, t_{i-1}^2, ..., t_{i-1}^p, ..., t_{i-1}^k$, and the word w_{i+1} appeared with *m* possible tags $t_{i+1}^1, t_{i+1}^2, ..., t_{i+1}^q, ..., t_{i+1}^l$. Initially, we give each possible tag of the words a probabilistic value. For instance, the probability of word w_i appeared with any possible tag is $1/j$, the probability of the word w_{i-1} appeared with any possible tag is $1/k$. Then, the method computes the probability of each tag following other tags in corpus using Equation (8). Finally, the method can get the new probabilities of the word appeared with different tags by Equation (9).

$$P(t_{i+1}^q \mid t_i^r) = \frac{\sum_{\textit{for all } w_j \textit{ in the corpus}} P(t_j = t_i^r) P(t_{j+1} = t_{i+1}^q)}{\sum_{\textit{for all } w_j \textit{ in the corpus}} P(t_j = t_i^r)} \tag{8}$$

$$P^{new}(t_i^r) = \frac{Q(t_i^r)}{\sum_{y=1}^{j} Q(t_i^y)}, \tag{9}$$

where $Q(t_i^r) = \sum_{s=1}^{k} \sum_{t=1}^{m} P(t_{i-1}^p) P(t_i^r \mid t_{i-1}^p) P(t_{i+1}^q) P(t_{i+1}^q \mid t_i^r)$.

After the new probabilities of all words appeared with different tags in corpus have been generated, the method can compute the probability of the tags following other tags in corpus by

Equation (8) again, and the new probabilities of the words appeared with different tags in corpus are also generated by Equation (9). The cycle that generates the new probabilities from Equations (8) and (9) is called a training iteration. [3] indicates that the variation of the probability of the tags will be converge after several training iterations, and the tag of the word can be decided with highest probability..

The tagging accuracy of the relaxation labeling method is not satisfactory because it does not use lexical information. But, the relaxation labeling method states that the probability of the tag of the word can be estimated based on not only preceding adjoining relation, but also the relation between the tag of the word and the tag of succeeding word. We called the latter as relation succeeding adjoining relation.

3. THE HYBRID METHOD

In the Section 2, we discuss two statistical models for evaluating the probability of each possible tag of the multiple-tag words in sentences. Hidden Markov models have more accuracy for tagging because it uses to lexical information and structural information simultaneously. On the other hand, although relaxation labeling method also indicates the importance of the adjoining relation of the words should contain the anteceding word of the words and the preceding words of the words simultaneously.

Thus, we present a hybrid method for tagging Chinese sentence. It uses lexical information and structural information. More importantly, the structural information contains both preceding adjoining relation and succeeding adjoining relation. We called the statistical-based tagging method as two-way adjoining relation method. On the other hand, after tagging by two-way adjoining relation method, each word should be given a tag and the tags of some words are wrong. To solve the problem, we designed a temporally clustering modification method to modify these wrong tags of the words.

3.1 The Two-way Adjoining Relation Method

As hidden Markov models, two-way adjoining relation method would need three different kinds of data from corpus: the number of times of each tag appeared in corpus, the number of times of each word occurred with different tags in corpus, and the number of times of each tag following the other tags. Using these statistical data, the method can compute the probability of the word appeared with different tags in sentences. To compute the probability, assuming the sentence $S = w_1 w_2 ... w_n$ will be processed. Furthermore, assuming the word w_i appeared with *j* possible tags $t_i^1, t_i^2, ..., t_i^r, ..., t_i^j$, and the word w_{i-1} appeared with *k* possible tags $t_{i-1}^1, t_{i-1}^2, ..., t_{i-1}^p, ..., t_{i-1}^k$. For the possible tag t_i^r of word w_i, the new method computes the probability of tag t_i^r following t_{i-1}^p by Equation (10). Similarly, the method can obtain the probability of word w_i appeared with tag t_i^r by Equation (11).

$$P(t_{i-1}^p \mid t_i^r) = \frac{C(t_{i-1}^p, t_i^r)}{C(t_i^r)}, 1 \le r \le j, 1 \le p \le k \quad (10)$$

$$P(w_i \mid t_i^r) = C(w_i, t_i^r), 1 \le r \le j \quad (11)$$

That is, we estimate the probability of preceding adjoining relation for a word by Equation (10), and estimate the probability of lexical information for the word by Equation (11). The probability of lexical information is the same as HMM for tagging, but the probability of preceding adjoining relation is different from HMM. On the other hand, a word’s tag should also depend on the tag of following word. Actually, for Chinese, sometimes the succeeding adjoining relation for a word is closer than the preceding adjoining relation for the word.[10] Therefore, the new method includes the reverse reading direction of the sentences as well as the original left-to-right direction. Based on the observation, the following word of the word in sentences becomes the previous word of the word. Hence, we can use Markov model again to evaluate succeeding adjoining relation for the word.

We further assume the word w_{i+1} appeared with *m* possible tags $t_{i+1}^1, t_{i+1}^2, ..., t_{i+1}^q, ..., t_{i+1}^l$. For the possible tag t_{i+1}^q of word w_{i+1}, we can get the probability of tag t_t^r followed by t_{t+1}^q using Equation (12).

$$P(t_{i+1}^q \mid t_i^r) = \frac{C(t_i^r, t_{i+1}^q)}{C(t_i^r)}, 1 \le r \le j, 1 \le q \le l \quad (12)$$

Using Equations (10), (11), and (12), we can now consider a new evaluation function to compute the probability of the word appeared with the particular tag in a sentence. The function is the product of the probability of the preceding adjoining relation, the probability of lexical information, and the probability of succeeding adjoining relation for a word. Finally, the tag with highest probability will be chosen to be the tag of the word. Equations (13) and (14) show the evaluation function.

$$\delta(t_i^r) = \max_{1 \le p \le k, 1 \le q \le l} [P(t_{i-1}^p \mid t_i^r) \times P(w_i \mid t_i^r) \times P(t_{i+1}^q \mid t_i^r)], 1 \le r \le j \quad (13)$$

$$\varphi_i = \arg\max_{1\le r\le j}[\delta(t_i^r)] \qquad (14)$$

To use above equations, we must define such boundary values as $P(t_1^r \mid t_0^p)$ and $P(t_n^r \mid t_{n+1}^q)$. Assuming the beginning and ending of each sentence are periods, we can obtain the number of occurrences of each tag following period and that of each tag followed by period from training corpus. Given that the t_0 and t_{n+1} of all sentences are periods, both $P(t_1^r \mid t_0^p)$ and. $P(t_n^r \mid t_{n+1}^q)$ then be computed using Equation (12).

The third term in Equation (13) describes the conditional probability of the word given every possible tag of the succeeding word. Since the best possible tag of the succeeding word can be easily computed, our method would only consider the conditional probability of the word given the best possible tag of the succeeding word as Equation (15).

$$\delta(t_i^r) = \max_{1\le p\le k}[P(t_i^r \mid t_{i-1}^p) \times P(w_i \mid t_i^r) \times P(t_i^r \mid \varphi_{i+1})], 1 \le r \le j \qquad (15)$$

where φ_{i+1} is the best possible tag of word w_{i+1}.

3.2 The Temporally Clustering Modification

After tagging by two-way adjoining relation method, each word should be given a tag, and the tags of some words are wrong. The mistakes indicate that the tagging method needs more features to tag the words. The words which tend to co-occur sequentially can provide extra lexical information, and the grouping words is called as temporally clustering[8]. For instance, a sentence tagged by statistical-based phase is as follow:

回到	家	打開	門	一	看
VC	Nc	VC	Na	Db	VE

In the sentence, the tag of the word “看” should be not the tag “VE” but the tag “VC“. In the modification phase, the method counts the number of the word “看” in training corpus which occurred with different tags and adjoined by the word “一” which occurred with the tag “Db”. Because the word “看”, adjoined the word “一“ occurred with the tag “Db”, always occurred with tag “VC”, the method modify the tag of the word to be tag “VC”. We call the procedure as temporally clustering modification.

Temporally clustering modification revises the tags of some words to be correct by the adjoining words and the tags of adjoining words. The major advantage of temporally clustering modification method is that it do not need very huge tagged corpus to build training sets, but the accuracy of the method using the larger corpus for tagging will be higher than hidden Markov models.

4. EXPERIMENTS

In this section, we compare the tagging performance of different methods. Before tagging a Chinese sentence, we need a method for segmenting the sentence into several words because the words in the sentence is composed of one or several characters and without blanks on both ends to indicate its boundaries. Many studies[10][11][12] presented different methods to solve the problem. In our experiments, the Chinese sentences are segmented into words by ideally segmentation and maximum matching algorithm with Sinica Electronic Dictionary, respectively. The experiments compare the difference between ideal segmentation method and maximum matching segmentation method. Because maximum matching algorithm cannot achieve the exactly same result as the manual segmentation, we only extract the segmented words which also exist in manually segmented sentences.

In addition, we generate the training and testing sets from Sinica Treebank. Sinica Treebank, composed of 9 files retrieved from Sinica Copus, includes 38,725 Chinese sentences and 239,532 Chinese words. The words in Sinica Treebank have been manually tagged. We choose one of 9 files to be a testing set, and the rest of the files to be a training set. Using the procedure, the experiments can produce 9 test sets and 9 training sets. Table 1 lists the 9 files and shows the number of the words and multiple words of each file which segment into words by ideal segmentation and maximum length segmentation respectively.

From training sets, the experiments use Equations (10), (11), and (12) to compute the frequency of the occurrences of the word appeared with different tags, the frequency of the occurrences of each tag, and the frequency of each adjoining relation. Based on these statistical data, the experiments tag each word in test sets using hidden Markov models, the two-way adjoining relation method, HMM included temporally clustering modification, and the hybrid method includes two-way adjoining relation method and temporally clustering modification. Table 2 shows the accuracy rates of four methods with ideal segmentation for tagging multiple-tag words.

Table 1: The number of words and multiple-tag words generated by different segmentation methods in 9 files

	Ideal Segmentation		Maximum Length Segmentation Method	
	The number of words	The number of multiple-tag words	The number of words	The number of multiple-tag words
Cheng	1302	529	1094	468
F79109	13830	6846	12659	6389
F79119	12301	5712	10440	5327
F79119a	21545	9227	17513	8673
F79119b	27884	11736	23682	11066
F79119c	16850	7087	14315	6661
Gtrvl1	7230	2972	5925	2760
Gtrvl2	7890	3337	6396	3136
Gtrvl3	26744	11176	22431	10467

In Table 2, the average accurate rate of the tagging for multiple-tag words in test sets using the hybrid method is 91.2 percent. It is 2.2 percent higher than that using HMM, and 1.5 percent higher than using adjoining relation method. Furthermore, the accurate rate of the tagging for multiple-tag words in testing sets using HMM included temporally clustering modification is 90.6 percent. It is also 1.6 percent higher than using HMM. Clearly, in the preliminary experiments, the hybrid method has higher performance than other methods.

Table 2: The accuracy rates of four methods with ideal segmentation for tagging multiple-tag words

The methods The test set	Hidden Markov Models	The two-way adjoining relation method	HMM included temporally clustering modification	the hybrid method
Cheng	89.4%	88.5%	90.9%	90.7%
F79109	86.9%	88.9%	88.0%	89.5%
F79119	89.6%	89.8%	90.4%	90.6%
F79119a	88.3%	88.8%	90.8%	91.4%
F79119b	89.5%	90.3%	91.3%	91.9%
F79119c	89.0%	89.4%	91.0%	91.6%
Gtrvl1	89.6%	91.2%	91.3%	92.6%
Gtrvl2	89.7%	89.8%	91.2%	91.5%
Gtrvl3	89.1%	90.4%	90.6%	91.5%
Average	89.0%	89.7%	90.6%	91.2%

Table 3: The accuracy rates of four methods with maximum length segmentation method for tagging multiple-tag words

The methods / The test set	Hidden Markov Models	The two-way adjoining relation method	HMM included temporally clustering modification	the hybrid method
Cheng	88.7	88.2	90.4	90.0
F79109	86.6	88.7	87.7	89.1
F79119	89.2	89.4	90.1	90.3
F79119a	88.4	88.9	90.3	91.0
F79119b	89.2	89.7	90.8	91.3
F79119c	88.4	89.1	90.2	90.9
Gtrvl1	88.8	89.6	90.6	91.6
Gtrvl2	88.2	88.4	89.6	90.1
Gtrvl3	88.6	89.6	90.0	90.7
Average	88.4	89.0	90.0	90.5

Table 3 shows the accuracy rates of four methods with maximum length segmentation for tagging multiple-tag words. In Table 3, the average accurate rate of the tagging for multiple-tag words in test sets using the hybrid method is 90.5 percent. It is 2.1 percent higher than that using HMM, and 1.5 percent higher than using adjoining relation method. Furthermore, the accurate rate of the tagging for multiple-tag words in testing sets using HMM included temporally clustering modification is 90.6 percent. It is also 1.6 percent higher than using HMM. Clearly, in the experiments, the hybrid method still has higher performance than other methods.

5. CONCLUSIONS

In this paper, we presented a new hybrid method to automatically tag part-of-speech for Chinese sentence. The two-way adjoining relation method simultaneously refers to preceding adjoining relation, succeeding adjoining relation, and the lexical information to choose the most likely tag. In addition, the hybrid method applies temporally clustering modification to revise the wrong tag of the words which tag by two-way adjoining relation method. According to the result of the preliminary experiments, the method can increase the accurate rate of tagging for multiple-tag word.

REFERENCES

[1] Charniak, E. et al., “Tagger for Parsers,” *Artificial Intelligence*, 85: 45-57, 1996.

[2] Charniak, E. et al., “Equations for Part-of-Speech Tagging,” *Proc. of the 11th Nat’l Conf. on Artificial Intelligence*, 784-789, 1993.

[3] Liu, S. H. et al., “Automatic Part-of-Speech Tagging for Chinese Corpora,” *Computer Processing of Chinese and Oriental Languages*, 9(1): 31-47, 1995.

[4] Merialdo, B., “Tagging English Text with A Probabilistic Model,” *Computational Linguistics*, 20: 155-171, 1994.

[5] Perez-Ortiz, J. A. and Forcada, M. L., “Part-of-Speech Tagging with Recurrent Neural Networks,” *Proc. of 2001 IEEE Int’l Joint Conf. on Neural Networks*, 3: 1588-1592, 2001.

[6] Weischedel, R. et al., “Coping with Ambiguity and Unknown Words through Probabilistic Models,” *Computational Linguistics*, 19(2): 359-382, 1993.

[7] Manning, C. D. and Schutze, H., “Foundations of Statistical Natural Language Processing,” The MIT Press, MA, 1999.

[8] Meng, Helen M. and Siu, K. C., “Semiautomatic Acquisition of Semantic Structures for Understanding Domain-Specific Natural Language Querues,” *IEEE Trans. on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 172-181, 2002.

[9] Lee, L. S. et al., “An Efficient Natural Language Processing System Specially Designed for The Chinese Language,” *Computational linguistics*, 17(4): 347-374, 1991.

[10] Dai, Y. et al., “A New Statistical Formula for Chinese Text Segmentation Incorporating Contextual Information,” *Proc. of the 22nd annual int’l ACM SIGIR conf. on Research and development in information retrieval*, 82-89, 1999.

[11] Nie, J. Y. and Ren, Fuji, “Chinese Information Retrieval: Using Characters or Words?” *Information Processing and Management*, 35: 443-462, 1999.

[12] Yeh, C. L. et al., “Rule-based Word Identification for Mandarin Chinese sentences – A Unification Approach,” *Computer Processing of Chinese and Oriental Language,* 5(2): 97-118, 1991.