Research of POS Tagging Rules Mining Algorithm

Yin Shaohong
School of Computer Science & Software Engineering,
Tianjin polytechnic university
Tianjin, China
yinsha@tjpu.edu.cn

Fan Guidan
School of Computer Science & Software Engineering,
Tianjin polytechnic university
Tianjin, China
Fanguidan.1207@163.com

***Abstract*—Part of speech contains important grammatical information, so it has great significance for the natural language understanding while the words in the sentence are marked on the parts of speech. POS tagging rules based on statistical methods and rule-based method can mining effectively, but its marked accuracy need to be improved. This paper presents a statistical method and rules of the combination of speech tagging rule mining algorithm in order to improve the correct rate of marked.**

Keywords- Automatic annotation; Rule method; POS Tagging Rule

I. INTRODUCTION

As a basic research project in the field of Chinese information, Chinese Automatic Speech Tagging is a prerequisite for further Chinese syntax analysis, semantic understanding, and it can be used for information retrieval, machine translation, text-to-speech conversion, spell checking and other fields. In addition, the POS tagging can improve the effect of Chinese segmentation. So the research of POS Tagging is important. The purpose of POS tagging is to choose a most probable part of speech sequence for the words in a sentence.

The main research methods include based on statistic, rule-based [1]. The typical method of statistic-based approach is CLAWS algorithm and hidden Markov model HMM.Merialdo used HMM to POS tagging and it has achieved great success.

Rule-based approach has the following problems. First, from the application range of the rules, the method generated by artificial method can only be some common rules [2], rather than a large number of individual rules. The application scope of individual rules is very small, but it has a great influence on the correct rate. Second, method generated by artificial method is not entirely correct, its accuracy needs further verification. Therefore, the key problem of Chinese POS tagging is that we can obtain rules automatically and efficiently in the case of accuracy of statistical method improved hardly.

Based on the above analysis, the author proposes a hybrid algorithm of statistical method and rule method.

II. PROBLEM DEFINITION

Definition 1[3]: Word sets $DICT=\{WORD_i \mid i=1,2,...,n\}$,POS tag set $TEST=\{TAG_i \mid i=1,2,...,n\}$,itemsets $I = TEST \cup TAG$, where $WORD_i$ and TAG_i are a specific Chinese word or part of speech.

Definition 2 [3]: The labeled text $T=\{WORD_i,TAG_i \mid WORD_i \in DICT, TAG_i \in TEST\}$, where TAG_i is part of speech of $WORD_i$ in the marked text.

Definition 3[3]: Training set $E=\{(e_1,e_2,...,e_k) \mid e_i \in T$, the sentence of the text is marked as $SENTENCE=\{e_1,e_2,...,e_k, i=1,2,...,K, K \in N\}$. Where N is the set of natural numbers. Element $e_i(i=1,2,...,K,K \in N)$ is the sentence that has been marked in the set E.

Definition 4[3]: Pattern set $D=\{d \mid d \in I^+\}$ express the string that is composed by words and part of speech.

Definition 5[3]: if $X \in D$,length is $LEN(X)=K$, Mode X is K- mode.

Definition 6 [3]: if $X \in D, F=\{Y \mid Y \in D$ and $LEN(X)=LEN(Y)\}$, $X.SUP=\frac{freq(X)}{total(F)}$ is the support of mode X, $X.SUP$ is reflected in the proportion of the mode in the same length mode. Where $freq(X)$ means frequency of mode X, $total(F)$ means total frequency of $LEN(X)$ length of mode.

Definition 7[3]: Let $MINSUP$ be minimum support, $C=\{X \mid X \in D, X.SUP \geq MINSUP\}$ is common mode set. All the elements of C are called large mode.

Definition 8[3]: if X and Y are large mode, the association between X and Y is referred to as a rule $X \Rightarrow Y$, the credibility of the rules $(X \Rightarrow Y).CON$ defines as $PROB(Y/X)=\frac{freq(XY)}{freq(X)}$, and support defines as $(X \cup Y).SUP$.

It indicates the probability that contains X mode and Y mode in the pattern set D, where is occur probability.

Definition 9[3]: Let $MINCON$ be minimum confidence, if $(X \Rightarrow Y).CON \geq MINCON$, then rule $X \Rightarrow Y$ is reliable production rules.

Definition 10[3]: Take k-mode $I_i=a_1a_2...a_{k-1}a_k \in I$, and $a_k \in TEST$, a_k is POS tagging of word k, take the form of rules $\bigwedge_{j=1}^{k-1} a_j \Rightarrow (W_k,A_k)$. If it exists before the k-1 word, and its mark constituted mode is $a_1,a_2,...,a_{k-1}$, POS tagging of the k word （ W_k ） is a_k .

Definition 11[3]: There is a given positive number ε and $\varepsilon \rangle MINSUP$.

If support $MINSUP\langle(X \cup Y).SUP\langle\varepsilon$ of rule $X \Rightarrow Y$, then the rule is personalized rule.

If rule $X \Rightarrow Y$ is personalized rule and $(X \Rightarrow Y).CON \rangle MINCON$, then the rule is individuality reliable rules.

If support of rule $X \Rightarrow Y$ is $(X \cup Y).SUP \rangle \varepsilon$, then the rule is common rules.

If rule $X \Rightarrow Y$ is common rules and $(X \Rightarrow Y).CON \rangle MINCON$, the rule is common reliable rules.

III. ALGORITHM BASED ON RULE METHOD FOR POS TAGGING

We use association rule to excavate frequent itemsets of words and part of speech from annotated Chinese text corpus, and we use frequent itemsets to study the mode sequence of words and part of speech. Assume the judgment on part of speech based on the context by machine is consistent with people, we judge part of speech according to part of speech, word and combination of the two in the context. If statistical corpus is larger, we obtain general mode sets that are greater than the minimum support by mining algorithms for given MINSUP and MINCON, while we get association rules. If the credibility of this rule is higher than minimum confidence, we can get the part of speech rules. We define a high enough minimum confidence, the rules still can handle if there are multi-category. Rules obtained by this method can be used as a supplement to probabilistic methods, while the problem of Chinese POS tagging is resolved preferably.

Algorithm is expressed as follows: Given a training set E, according to two layers of different structural of WORD and TAG in E that influence on words and parts of speech，and we use frequent itemsets mining algorithm to calculate production rules which meet minimum support and minimum confidence, namely personality and common reliable rules in definition 11.

We decompose the problem into two smaller sub problems during the implement process. One sub problem calculates all common mode set in T which satisfy minimum support, another use common mode set to create association rules that meet the minimum confidence, the latter is relatively simple. One kind of constructive methods is based on definition 11 to each common mode. If $(a_1a_2...a_{k-1} \Rightarrow a_k).CON > MINCON$, then we add it to rule set. The key technology is to find common mode sets efficiently.

POS tagging mining is different from data mining in database. The combined tree of words and part of speech are altogether 2^i species in sentence length of i, the sum total of length patterns is exponential growth with the growth of the pattern length i, therefore, minimum support is diminishing vector. We must ensure that its credibility is higher than the short patterns because of smaller apply range of long pattern, therefore, credibility vector must be incremental vector.

Minimum support vector is $minsup[] = \{a_1a_2...a_n\}$,minimum confidence vector is $minconf[] = \{b_1b_2...b_n\}$, $total[i]$ represent the total number of i-mode(its initial value is 0), $MaxPatternSize = W$ represent size of the longest pattern specified by users, *pattern* is a pattern and *pattern.count* stores the number of occurrences of the pattern, *CandPatternSet* express candidate pattern set(its initial value is empty set {}), $largePatternSet = \{\}$ signify large pattern set finally formed(its initial value is empty set {}), $E=\{(e_1,e_2,...,e_k)|e_i \in T,i=1,2,...,k\}$ .Mining algorithm can be described as follows:

1) If $E = \varnothing$, then go to step *7)*.

2) Take $(e_1,e_2,...,e_k) \in E$, move pointer is $L=1, E \leftarrow E-\{(e_1,e_2,...,e_k)\} \in E)\}$.

3) If $K - L + 1 = 0$, then $(e_1, e_2, ..., e_k)$ has been processed, go to step *1)*.

4) If $K-L+1>W$, take $(WORD_L, TAG_L), ..., (WORD_{L+W-1}, TAG_{L+W-1}), LEN \leftarrow W$, otherwise take $(WORD_L, TAG_L), ..., (WORD_K, TAG_K), LEN \leftarrow K - L + 1$.

5) Loop through $J = 2...TO...LEN$,and implement pattern string length of j $pattern, pattern.count \leftarrow pattern.count + 1$, $CandPatternSet \leftarrow CandPatternSet \cup \{pattern\}$ $total[j] \leftarrow total[j] + 1$.

6) If $L \leftarrow L + 1$,then go to step *3)*.

7) If $CandPatternSet = NULL$,then go to step *11)*.

8) Take $pattern \in CandPatternSet$, $CandPatternSet \leftarrow CandPatternSet - \{pattern\}$.

9) If $(pattern.count / total[i]) > \min sup[i]$ Then $L\arg ePatternSet \leftarrow L\arg ePatternSet \cup \{pattern\}$.

10) Go to step *7)*.

11) If $L\arg ePatternSet = NULL$,then algorithm ends.

12) Take $pattern = a_1a_2...a_k \in L\arg ePatternSet$, $L\arg ePatternSet \leftarrow L\arg ePatternSet - \{pattern\}$.

13) If $a_1a_2...a_k \in Tag \cap (a_1a_2...a_{k-1} \Rightarrow a_k).CON > MINCON[k]$,we add rule "if $a_1a_2...a_{k-1} \Rightarrow (W_k, a_k)$ " to rule base.

14) Go to step *11)*.

The first six step of the algorithm is to obtain the length pattern of all the examples in the training set E, and they are deposited in candidate pattern set *CandPatternSet* . We can get large pattern set $L\arg ePatternSet$ by calculation from step *7)* to *10)*. It can obtain rule from large pattern and adds rule to rule base from step *11)* to *14)*.

We sort from largest to smallest according the length of rules at the appearance of rule base, the purpose is easy to use context information in apply rules and help to improve of the part of speech disambiguation accuracy. If the rules have the same length, then we sort from largest to smallest according to support, and it is easy to more use common rules in matching, so it further improve the matching efficiency of rules.

IV. A HYBRID ALGORITHM OF STATISTICAL METHOD AND RULE METHOD

This paper presented a hybrid algorithm of statistical method and rule method to increase the tagging accuracy. First we use rules to processing when we mark multi-category. If we can use rules to processing, then we mark

part of speech and realize high efficiency and high accuracy by rule-based method. Otherwise we utilize probabilistic method to mark. The tagging algorithm is as follows:

1) First we use part of speech dictionary to mark all possible part of speech of non multi-category and multi-category for be marked text which has been segment.

2) Take fragment SECTION. The fragment includes a multi-category sequence, several non multi-categories before sequence and serialized non multi-category. As follows:

$$Word_1 \quad Word_2 \quad Word_3 \ldots Word_{i-1} \quad Word_i \quad Word_{i+1} \ldots Word_j \quad Word_{j+1}$$

$$Tag_1 \quad Tag_2 \quad Tag_3 \ldots Tag_{i-1} \quad Tag_{i,1} \quad Tag_{i+1,1} \ldots Tag_{j,1} \quad Tag_{j+1}$$

$$Tag_{i,2} \quad Tag_{i+1,2} \ldots Tag_{j,2}$$

$$Tag_{i,3} \quad Tag_{j,3}$$

3) If there is no SECTION, algorithm ends.

4) $L \leftarrow 1$.

5) We structure mode from $(WORD_L, TAG_L),\ldots,(WORD_{i-1}, TAG_{i-1})$, and use match rule to match it. We mark the word $Tag_{i,1}$ with its part of speech according rule $Word_i$ if the match is successful. We suggest $Word_i$ was non multi-category. Go to step *2),* reselect SECTION.

6) If $L \le i-1$,then $L \leftarrow L+1$,go to step *5).*

7) Take multi-category fragment between two non multi-categories, as follows:

$$Word_{i-1} \quad Word_i \quad Word_{i+1} \ldots Word_j \quad Word_{j+1}$$

$$Tag_{i-1} \quad Tag_{i,1} \quad Tag_{i+1,1} \ldots Tag_{j,1} \quad Tag_{j+1}$$

$$Tag_{i,2} \quad Tag_{i+1,2} \ldots Tag_{j,2}$$

$$Tag_{i,3} \quad Tag_{j,3}$$

We calculate the probability of all paths of POS tagging and sort it to obtain the maximum path. Then we mark the $Word_i ... Word_j$ according to marker of the maximum path.

8) Go to step *2).*

Data mining systems has certain openness and excavate adding new corpus at any time, and it generates new patterns and rule. But adding new corpus may lead to failure of the original rules because its dependence credibility and support reduce the repaired mode before cumulative excavation may become final rules due to low support or low confidence. Therefore, we must use management module of rule base to manage rules effectively (such as increase and delete rule, or modify the credibility and support of rules).

V. EXAMPLE AND RESULTS ANALYSIS

We can gain POS tagging set trough some adjustment based on the content of the 2008 People's Daily. And we use statistical methods and statistics-based and rules-based method to mark. Its accuracy, as shown in Table Ⅰ:

TABLE I. TABLE TYPE STYLES

	Statistical methods	**Statistics-based and rules-based method**
accuracy	91.8%	96.2%

The accuracy rate is 91.8% when we use statistical methods to mark test corpus, but the statistics-based and rules-based method can achieve an accuracy rate of 96.2%, it is much better than the former. If the success rate of a certain method is much higher than 90%, then we think that the method is effective. Therefore, statistics-based and rules-based method can improve the accuracy of mark obviously, and it is a very effective method.

REFERENCES

[1] Terry Winograd. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language[J]. MIT AI Technical Report 235,1971,2.

[2] Harabagiu.S, M. Pasca, and S. Maiorano. Experiments with Open Domain Textual Question Answering. Proceedings of the 18th COLING Conference. Saarbrucken, Germany. 2000:292-298.

[3] James Allen, Donna Bayon, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda stent. Towards Conversational Human-Computer Interaction[J]. AI Magazine, 2001.

[4] Bos, Johan, and Malte Gabsdil. First-order intference and the interpretation of questions and answers. In Proceedings of Gotalog 2000, ed. Massimo Poesio and David Traum, 43-50. Gothenburg Papers in Computational Linguistics 1-5.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] U. Hermjakob. Parsing and Question Classification for Question Answering[C]. In Proceedings of the ACL Workshop on Open-Domain Question Answering, Toulouse, France, 2001.

[6] Androutsopoulos.Natural Language Interfaces to Data bases an Introduction[J]. Natural Language Engineering,1995:21-89.

[7] Weizenbaum. ELIZA-A Computer Program for the Study of Natural Language Communication Between Man and Machine.Communications of the ACM. 1966,9:36-45.

[8] E. Brill, J. Lin, M. Banko, S. Dumais and A. Ng.Data-Intensive Question Answering[C]. Proceedings of the Tenth Text Retrieval Conference (TREC 2001).

[9] Ravichandran, D. and E. H. Hovy. Learning Surface Text Patterns for a Question Answering System[C]. In 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002) Conference. Philadelphia, PA, July 2002.

[10] Paul Cohen, Robert Schrag, Eric Jones, Adam Pease,Albert Lin, Barbara Starr,David Gunning,and Murray Burke.The DARPA high-performance knowledge bases project[C].AIM agazine,1998,12:25-49.

[11] Marius Pasca and Sanda Harabagiu. High performance question answering[C]. In Proceedings of the 24th SIGIR Conference on Research and Development in Information Retrieval.2001:366-37.

[12] F. Rinaldi, J. Dowdall, K. Kaljurand, M. Hess, D. Mo11 G.Exploiting Paraphrases in a Question Answering System[C]. Proc.Workshop in Paraphrasing at ACL2003, Sapporo, Japan. 2003, 7:25 -32..