Studies on Automatic Recognition of Common Chinese Adverb's Usages Based on Statistical Methods

Hongying Zan
College of Information Engineering, Zhengzhou University
iehyzan@zzu.edu.cn

Junhui Zhang
College of Information Engineering, Zhengzhou University
zhangj.zzu@gmail.com

Xuefeng Zhu
Key Laboratory of Computational Linguistics(Peking University) of China Ministry Education
yusw@pku.edu.cn

Shiwen Yu
Key Laboratory of Computational Linguistics(Peking University) of China Ministry Education
yusw@pku.edu.cn

Abstract

The study on Automatic Recognizing usages of Modern Chinese Adverbs is one of the important parts of the NLP-oriented research of Chinese Functional words Knowledge Base. To solve the problems of the existing rule-based method of adverbs' usages recognition based on the previous work, this paper has studied automatically recognizing common Chinese adverbs' usages using statistical methods. Three statistical models, viz. CRF, ME, and SVM, are used to label several common Chinese adverbs' usages on the segmentation and part-of-speech tagged corpus of People's Daily(Jan 1998). The experiment results show that statistical-based method is effective in automatically recognizing of several common adverbs' usages and has good application prospects.

1 Introduction

Chinese vocabulary can be divided into functional words and notional words. In the field of Natural Language Processing(NLP), many studies on text computing or word meaning understanding are focused on the notional words, rarely involving functional words. Especially in some common NLP application fields, such as text summarization, text classification, information retrieval, and so on, the researchers mainly take notional words as features, and list some functional word as stop words without considering their influence on text meaning. This will impact the deep analysis of text semantic, especailly for chinese, and become the bottleneck of machine understanding on text content, and impede further improving the performance of application systems. Due to Chinese lacking morphological changes(Li X., 2005), Chinese functional words undertake the grammatical functions and grammatical meanings, and in other language these functions are mainly undertaken by morphological changes. So, functional words play an more important role in Chinese semantic understanding and grammatical analysis. The study on functional words of modern Chinese semantic in Chinese text processing and understanding has great significance.

Yu(Yu S., 2004), Liu(Liu, Y., 2004), et al, have defined the generalized functional words as adverbs, conjunctions, prepositions, modal particles, auxiliary, and localizer words. From the statistic, the number of modern Chinese adverbs is about 1000 with the broad definition standard. Compared with other fuctional words, the adverbs number is much larger. The function and usages of modern Chinese adverbs vary widely from each other, especially some common adverbs. Therefore for modern Chinese text understanding, adverbs are the important text features which can not be neglected. For the modern Chinese adverbs, only using the segmentation and part-of-speech tagging information for Chinese text automatic processing and understanding is not enough. So, particular study on the usage of adverbs in texts comprehensive is indispensable, and the automatic identification of adverbs' usage in some extend is of great significance.

2 Related Researches

The work of automatically recognizing usages of adverbs of modern Chinese is part of the NLP-oriented research of Modern Chinese Functional Words Knowledge Base. Yu et al. proposed the idea of building the "Trinity" knowledge-base of generalized functional words(Yu, S., 2004), and defined the generalized functional words as adverbs, conjunctions, prepositions, modal particles, auxiliary, and localizer words(Yu, S., 2004)(Liu, Y., 2004). Zan et al. described adverb's usages using formal rules(Zan, H., 2007a), and initially built the machine-oriented modern Chinese adverb dictionary and the usage rule base(Zan, H., 2007b),. Hao et al. imported the dictionary and rule base(Hao, L., 2007). Based on the previous work, Liu et al. realized an automatically rule-based recognizing system and got precision at 74.89%(Liu, R., 2008).

The rule-based method has the advantage of simple, intuitive, strong pertinence, etc, but it also has the shortcomings of lower coverage, and it is difficult to be further optimized or generalized. For example, there are some adverbs which different usages are difficult to describe using formal rules, such as:

（1）想睡觉**尽管**睡，反正是星期天。

[(1)It is Sunday, you can sleep in **at will**.]

（2）她们俩听报告时**尽管**说话，报告的内容根本没听见。

[(2)They were **always** talk**ing** while listensing report, so they catched nothing of the report content.]

In the adverb usage dictionary, the adverb "jinguan/尽管" has two meanings: <d_jin3guan3_1> and <d_jin3guan3_2>. The meaning of "jinguan/尽管" in sentence (1) is belong to <d_jin3guan3_1>, it means the action or behavior can be without any limitations; the meaning of "jinguan/尽管" in sentence (2) is belong to <d_jin3guan3_2>, it means the action or behavior is continuously. This two meanings are very easy to distinguish manually, but they are hard to identify automatically. The two meanings' discrimination cannot accurately describe using formal rules.

Moreover, the rule-based method also exists some other problem, for example, some adverbs' usages require modifying verb phrase, or clauses, or used in imperative, and so on. These problems need deep syntactic even semantic knowledge to solve. But this is lack in the segmentation and part-of-speech tagging corpus. So, the rule-based method will be unable to identify the adverbs' usages in such situations.

To solve the problems of the existing rule-based method of adverbs' usages recognition, based on the foundation of the previous work, this article considers using statistical method to recognize adverbs' usages automatically. This method can be continuously optimized according to actual training data and language model, it will avoid the limitations of rule-based method.

3 Studies on Automatic Recognition of Adverbs' Usages Based on Statistical methods

In NLP, the research can be divided into three questions: point, sequence, and structure(Vapnik V., 1998). For the Chinese adverbs' usages recognition task, it can be taken as a point question which classify the context of adverbs, and also can be taken as a sequence question which recognize the adverb sequence in the sentence. So, we choose three statistical models: Conditional Random Fields(CRF), Maximum Entropy(ME), and Support Vector Machine(SVM), which have good performance and used widely in the field of machine learning. CRF and ME model can be used in sequence tagging, and SVM is a better statistical models in categories.

3.1 Statistical models

CRF is advanced by J. Lafferty(Lafferty J., 2001). It is one of the undirected graph models. Given input sequence corresponding conditional probability of label sequence, this model's training target is to find the maximum of conditional probability. It has been widely used in the field of NLP, such as Chinese Word Segmentation(Miao X., 2007), Named Entity Recognition(Chen W., 2006)(Shi S., 2006)(Guo J., 2007)(Zhang J., 2006), Syntactic Analysis(Fei Sha, 2003), and so on.

ME has been widely used for classification problem. The basic idea of ME is to dig the potential constraint conditions in the

known event sets, and then choose a model which must satisfy the known constraint conditions, while possibly let the unknown event uniform distribution. In the NLP applications, the language model based ME does not dependent on domain knowledge, and is independent of the specific task. It has been use in many key fields of NLP, and has achieved good results in Named Entity Recognition(Wang J., 2005), POS tagging(Zhang L., 2008), Chunking Analysis（Li S., 2003）, Text Emotional Tendencies Classification(Liu, K. 2008).

SVM is a statiscal machine learning method and has good performance in classification(Vapnik V., 1998). In NLP, SVM is widely used in Phrases recognition(Li, G., 2005), Word Sense Disambiguation(Yu, K., 2005)(Lu, Z., 2006), Text classification, and so on. SVM has good generalization ability, and can well classify the data in the training sample limited circumstances. To the usage recognition of adverbs, the available data is limited, so using SVM may be good.

CRF, ME and SVM are the outstanding statistical models in machine learning. CRF can well consider the mutual influence between usage marks, and overcomes the problem of marker offset. This is good for some rare usage recognition of adverb. The language model built by ME method is independent to specific tasks, and domain knowledge. ME can effectively use context information, and comprehensively evaluate the various characteristics. SVM has good generalization ability, and can well classify the data in the training sample limited circumstances. The advantages of these models are beneficial to recognize adverbs' usages correctly.

In this paper, we use CRF++[1], the ME toolkit maxent[2] of Zhang Le, and LibSVM[3] toolkit as the automatic tagging tool in our experiments.

3.2 Feature Selection of Models

Linguists Firth once said "You shall know a word by the company it keeps"(Firth, 1957). This refers to the mean of a word can only be judged and identified from the words associated with it. To the adverbs' usage recognition, it also needs to get the word's usage knowledge from the contexts. Through analyzing some examples, we found that words and part of speech in the contexts are useful to identify adverbs' usages. Therefore, in our experiment, to CRF and ME model, we select 3 template features as table 1. The value of n can take 2, 3, 4, 5, 6, and 7.

Table 1 Feature Template

ID	Meanings
T1	words, within the destined context window *n*
T2	the part of speech, within the destined context window *n*
T3	the words + part of speech + the combination of both, within the destined context window *n*

In the SVM experiment, the feature is numeric characteristics. To the adverb in the sentence, through selecting the window size of the context, and then calculating the mutual information(MI) of the features in the window and the adverb, the result of MI as feature vector. The MI between word *w* and word *u* can be calculated as follows,

$$I = \log \frac{p_1 * p_2}{p} \qquad (1)$$

Where:
p1: the frequency of *u* in the corpus
p2: the frequency of t in the corpus
p: the co-occurrence frequency of *w* and *u*

4 Experiments and Results Analysis

4.1 Experimental Corpus

The experimental data is the segmentation and part-of-speech tagged corpus of People's Daily(Jan 1998). First, we use the rule-based method(Liu, R., 2008) to tag the adverbs' usages in the experimental data. Then, we manually check the tagging results and get he standard corpus for experiment data. Observing the experiment data, the usage distribution of many adverbs' is very imbalance. Some adverbs have hardly appeared, and some usages of some adverbs have hardly appeared. If we choose this kind of adverbs for statistical experiment, it will bring great effect to the experiment results. Therefore, after analyzing the corpus, we consider to

[1] CRF++: Yet Another Toolkit[CP/OL]. http://www.chasen.org/~taku/software/CRF++

[2] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm

choose seven common Chinese adverbs which usage distribution is somewhat balanced in the corpus as the object of statistical learning.

4.2 Performance Evaluation

In the experiment, we use the precision(P) as the evaluation measure of the experimental results. To the word *W* and its usage *i*, we define P as followed:

$$P = \frac{the\ correct\ tag\ number\ of\ usage\ i}{the\ tag\ number\ of\ usage\ i} \quad (2)$$

4.3 Analysis of Experimental Results

In order to verify the performance of models, to every adverb, we use 4 fold cross-validation experiments. The results are the average results of cross-validation.

Experiment 1: Performance comparison experiment of Statistical methods and rule method

Aiming at the different statistical models, by selecting different feature, we did 3 groups experimental separately. For CRF and ME, we select T1 while n=2. To SVM we take MI as feature while the window size is 2. Results are shown in Table 2.

Table 2 The experiment result of rule-based method and the statistic-based method

Method / Adverb	Rule-based	CRF	ME	SVM
bian/便	0.409	0.459	0.453	0.876
fenbie/分别	0.506	0.673	0.679	0.905
Jiu/就	0.339	0.776	0.608	0.59
tebie/特别	0.697	0.783	0.652	0.932
yi/已	0.511	0.91	0.71	0.974
shifen/十分	0.712	0.95	0.865	0.993
xianhou/先后	0.963	0.575	0.59	0.846
average precision	**0.55**	**0.729**	**0.66**	**0.885**

From Table 2 we can see that the statistic-based results are better than the rule-based results on the whole. The average precision has been raised from 55% to 88.5%. It can clearly be seen that the statistical method has better adaptability and good application prospect in automatic identification of modern Chinese adverbs' usages.

At the same time, we can see that the statistical result of adverb "xianhou/先后" is obviously lower than the rule-based method. This is because the different usage of it can be easily distinguished from its rule, so the precision of rule-based method is higher than statistic-based method. To these words, we consider to use the method that combines the statistics-based and rules-based method.

Experiment 2: Statistical experiment under different feature template

By choosing different feature templates, this experiment to analyze the influence of different feature to the statistical method. Figure 1 is the average results of 6 adverbs(removing adverb "xian hou/先后") using three models. The abscissa 1-6 is the feature in the template T1 while n take 2, 3, 4, 5, 6, 7 separately. Figure 2 is the average results of these adverbs using CRF and ME with template T1, T2, and T3(see Table 1). The abscissa 1-3, 4-6, 7-9 ,10-12, 13-15, 16-18, is T1, T2, T3 while n take 2, 3 ,4 ,5, 6, 7.

From Figure 1 and Figure 2, we can see that the precision of statistical results have not great changes by choosing different context window. In general it can be achieved the best result within the window size (-4, +4) of the context. So, in the current scale of corpus, big window size may be not better when recognizing usages of adverbs, and it may bring more noise for recognizing with the increase of window size. But observing experimental results of specific words, we found that it's not all of the words exist this phenomenon. Figure 3 and Figure 4 is the result of adverb "jiu/就" and "bian/便" using three models with T1(n=2,…,7).

From Figure 3 and Figure 4, we can see that to different adverbs, the results of three models are not same, and even have big difference. To adverb "jiu/就", CRF is the best, SVM is the worst. To adverb "bian/便", SVM is the best, and the difference between CRF and ME is not very large. (Ma Z., 2004) also pointed out that every adverb needs to be synthetically analyzed and researched.

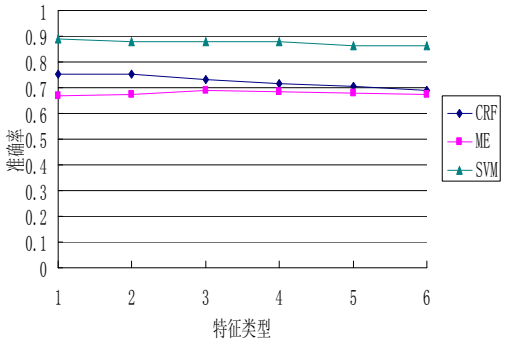

Figure 1 Average result of three models with T1(n=2,…,7)

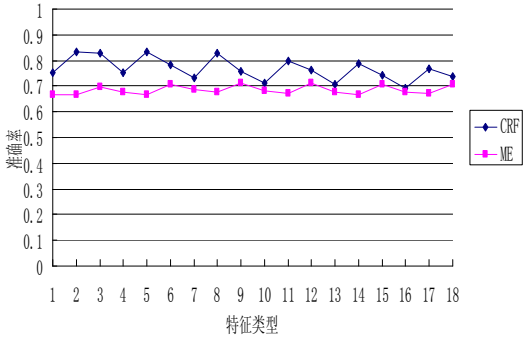

Figure 2 Average result of CRF and ME with T1, T2, T3(n=2,..,7)

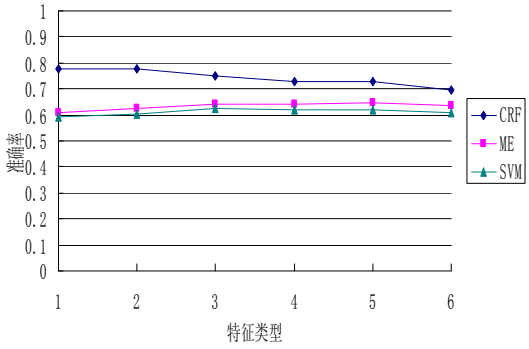

Figure 3 Adverb Result of adverb “jiu/就” using three models with T1(n=2,…7)

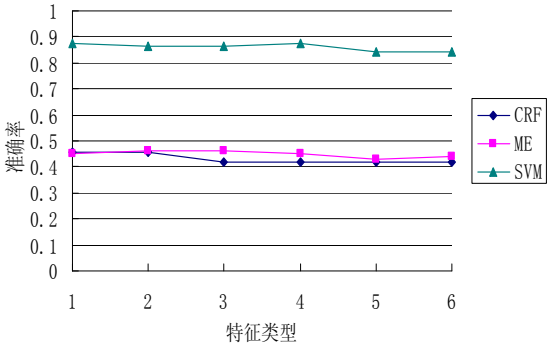

Figure 4 Adverb Result of adverb “bian/便” using three models with T1(n=2,…7)

So, to different adverb, we may be select different statistical model based on its own characteristics. For some common Chinese adverb, it’s very important to study and contrast case-by-case.

5 Conclusions

The article makes a preliminary study on automatically recognizing common adverbs’ usages. From the experimental results wen can see, compared with the rule-based method, statistic-based method has obvious advantages.

This article is a continuation of the work of Functional Word Knowledge Base. Furthermore, we will study the method that combines the rule-based method and the statistic-based method to automatically recognizing adverbs’ usages, and further enhance the recognition precision. We hope our study can help the Chinese lexical semantic analysis, and make a good base to the Chinese text machine understanding and the application of natural language processing.

Acknowledgement

The work in this paper was supported by the China National Natural Science Foundation (No. 60970083), the Open Project Program of the Key Laboratory of Computational Linguistics(Peking University)(KLCL 1004) of China Ministry Education, and the Outstanding Young Talents Technology Innovation Foundation of Henan Province(No. 104100510026).

References

Chen Wenliang, Zhang Yujie, Hitoshi Isahara. *Chinese named entity recognition with conditional random fields.* In 5th SIGHAN Workshop on Chinese Language Processing, Australia, 2006.

Fei Sha , Fernando Pereira. *Shallow parsing with conditional random fields.* In: the proceedings of Human Language Technology/North American chapter of the Association for Computational Linguistics annual meeting, 2003: 213-220.

Firth J R., *A Synopsis of L inguistic Theory 1930 - 1955* In Studies on L inguistic Analysis. L ondon: B lackwell 1957：101-126

Guo Jiaqing, *Studies on the Chinese Named Entity Recognition based on conditional random fields.* Doctoral dissertation of the Shenyang Aviation Industry Colledge, China. 2007.

Hao, Liping, Zan, Hongying, Zhang, Kunli, *Research on Chinese Adverb Usage for Machine Recognition.* In：Proceedings of the 7th International Conference on Chinese Computing (ICCC2007): 122-125

Lafferty, J., McCallum, A., Pereira F,. *Conditional random fields: probabilistic models for segmenting and labeling sequence data.* In the Proceedings of International Conference on Machine Learning, 2001: 282-289.

Li, Xiaoqi, et al. *The teaching materials on the modern Chinese funciotnal word*. Peking University press, Beijing, China, 2005. (in Chinese)

Li, Guozheng, Wang, Meng, *Introduction on the Support Vector Machine*. The electronic Industry Press. Beijing, China, 2005.

LI, Sujian, Liu, Qun, Yang Zhifeng, *Chunk Parsing with Maximum Entropy Principle*, Chinese Journal of Computers, 2003(12), 1722-1727.

Liu, Kang; Zhao, Jun, *Sentence Sentiment Analysis Based on Cascaded CRFs Model*, Journal of Chinese Information Processing, 2008(1), 123-128.

Liu, Rui,. et al. *The Automatic Recognition Research on Contemporary Chinese Language*, Computer Science, 2008(8A): 172-174. (in Chinese)

Liu, Yun, *The construcion of Chinese funtional words konwledge base*. Peking University. Postdoctoral reports of Peking University. 2004.

Lu, Zhimao, Liu, ting, *Survey of the statitical word sense disambiguation study*. Jounal of Electroniics, 2006.2

Ma,.Zhen, *Study Methodology of the Modern Chinese Function Words*. Commercial Press.2004.(in Chinese)

Miao Xuelei. *A Random Conditional Fields Based Method to Chinese Word Sense Disambiguation Research.* Shenyang Institute of Aeronautical Engineering. 2007.

Shi Shumin, Wang Zhiqiang, Zhou Lang, *Chinese Named Entity Recognition based on conditional random fields.* In the Proceedings of the 3rd students computational linguistics conference . 2006.(In Chinese)

Vapnik V., *Statistical Learning Theory*. Wiley-Interscience ublication. John Wiley&Sons, Inc,1998

Wang, Jiangwei, *Chinese named entity recognition Based on Maximum Entropy*, Doctoral dissertation of Nanjing University of Science and Technology, 2005.

Yu, Kun, Guan, Gang, Zhou, Ming. *Resume information extraction with cascaded hybrid model*. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, Michigan. 2005： 499-506

Yu, Shiwen, et al. Knowledge-base of Generalized Functional Words of Contemporary Chinese[J]. Journal of Chinese Language and Computing, 13(1): 89-98. 2004.

Zan, Hongying, Zhang Kunli, Chai,Yumei Yu, Shiwen. *The Formal Description of Modern Chinese adverbs' usages.* In Proceedings of the 9th Chinese Lexical Semantics Workshop(CLSW-2007), 52-56. 2007. (in Chinese)

Zan, Hongying, Zhang, Kunli, Chai,Yumei, Yu, Shiwen. *Studies on the Functional Word Knowledge Base of Contemporary Chinese*. Journal of Chinese Information Processing,2007(5): 107-111. (in Chinese)

Zhang Jian, *Studies on the English Named Entity Recognition based on conditional random fields.* Doctoral dissertation of the Harbin Industry University, China. 2006.

Zhang, Lei, *Study of Chinese POS Tagging Based on Maximum Entropy*, Doctoral dissertation of Dalian University of Technology, 2008.