Automatic Chinese Aspectual Classification Using Linguistic Indicators[*]

Defang Cao[1],Wenjie Li[2], Chunfa Yuan[1],and Kam-Fai Wong[3]

[1,] State Key Laboratory of Intelligent Technology and System,
Tsinghua University, Beijing, China
cdf03@mails.tsinghua.edu.cn; cfyuan@tsinghua.edu.cn
[2] Department of Computing, Hong Kong Polytechnic University,
Hong Hum, Hong Kong, China
cswjli@comp.polyu.edu.hk
[3]Department of SE & EM, Chinese University of Hong Kong, China
kfwong@se.cuhk.edu.hk

Abstract

The lack of regular morphological tense markers makes Chinese aspectual classification complicated. An approach of Chinese aspectual classification based on linguistic indicators is proposed in this paper. Forty-nine indicators are used in the classification, and they are divided into three levels. Based on our approach, two three-layers classifiers are also constructed. In close test and open test, these two three-layers classifiers both achieve higher accuracy than single classifier. The indicators are then evaluated and selected by RELIEF-F evaluation algorithm for the further experiments. The experiments using pruned features selected by RELIEF-F can also achieve good results.

Keywords: aspectual classification, linguistic indicators, RELIEF-F

I. Introduction

Temporal information extraction has become a hot research area. The aspectual class of a verb plays an important role in temporal information analysis and semantic understanding. In general, the aspectual classification maps clauses to a small set of primitive categories in order to reason about time. A clause consists of a main verb and related components. So the verb is not lonely considered in the aspectual classification, other components near the verb must be also considered.

In English, many aspectual classification methods have been proposed, eg. corpus-based linguistic indicators method [1,2], utilizing argument structure [3] or grammatical features [4]. Eric [1,2] used fourteen indicators that measure the frequency of lexico-syntactic phenomena linguistically related to aspectual class to do aspectual classification. The aspectual class of Eric's work is defined by two aspectual distinctions, stativity and completedness (i.e., telicity). Paolo and Suzanne [3,4] concentrate more on structural and grammatical information, so the verb class is defined as transitive/intransitive, or accusative/unaccusative.

[*] This research is supported by NSFC (60573186), Research Grants Council of Hong Kong (RGC reference number PolyU5085/02E) , 863 project (2001AA114210) of China and partially supported by the CUHK strategic grant (# 4410001).

In Chinese, there are also some considerable works. Chen [5] and Ma [6] did some theory research in verb classification. Ma [6] stated that the situation of a sentence is fully determined by the situation of the main verb of the sentence. He used three aspectual properties: static, durative, and telic to classify verbs into four situational types. Chen [5] stated that the situation of a sentence not only depends on the main verb of the sentence but also on other parts of the sentence. Ma and Chen just proposed their classification criterion, no practical experiment or system. Xiaodan Zhu [7] proposed an algorithm based on rule sets and reported good result. However, some inherited problems, such as rule set construction, rule selection and contradiction elimination between rules, are still remain unsolved. Wei Li [8] proposed a classification algorithm using Fuzzy Sets and Genetic Algorithm, the verbs are classified into four classes: attribute, mentality, activity and instantaneous. Wei Li used probability of verb being in one situation, the part of speeches of surrounding words and special words that have great impact on determining the verb situation as features in verb classification.

The rest of this paper is organized as the follows. Section 2 introduces aspectual classes in this paper. Section 3 describes the linguistic indicators in Chinese. Section 4 describes classification approach and proposes two three-layers classifiers. Section 5 evaluates attributes by RELIEF-F. Section 6 gives experiments results. Section 7 concludes this paper.

II. Aspectual Classes

In this paper, three aspectual distinctions are used: static/dynamic, telic/nontelic and durative/ instantaneous. These three distinctions compose four aspectual classes.

Table 1. Aspectual class properties in three aspectual distinctions

Aspectual Class	static/dynamic	telic/nontelic	durative/instantaneous
State	static	-[1]	-
Achievement	dynamic	telic	instantaneous
Accomplishment	dynamic	telic	durative
Process	dynamic	nontelic	durative

A state is defined as a stable situation which does not involve changes. It refers to a fact or a property. For example,

(1) 桌上放着一顶帽子(A cap is put on the desk).

A process refers to an activity not involving a culmination or an anticipated result. But it must have durations. For example:

(2)他在跑步(He is running).

An achievement describes an action which involves an inherent culmination but occur instantaneously. There is no duration as in an achievement. For example:

(3)炸弹爆炸了(The bomb exploded).

An accomplishment describes an action which involves an inherent culmination and lasts a duration time in time line. For example:

(4)他跑了三千米(He has run 3 kilometers).

[1] The ‘-‘represents no meaning. State is static and just over there, not change. It’s hard to say when it achieves a result or it can last how long in time line.

III. The Linguistic Indicators in Chinese

Compared with English, there is a special way to express aspectual information in Chinese. This paper extracts three level linguistic indicators. These indicators are applied to aspectual classification as features.

Word Level Indicators

The linguistic indicators in word level are single Chinese words that imply aspectual or situation information. Each word is a linguistic indicator. The Table 2 shows word level indicators. The words in bracket {} imply same aspectual function, and are clustered into one indicator.

Table 2. word level linguistic indicators

Word level Indicators	Indicator Tag	List
Temporal Auxiliary word	ut	着,了,过,起来,下去
Temporal Adverb	dt	只,{已,已经,早,早已},就,先,方才,立刻,一向,就近,才,{在,正,正在},刚,从来,新近,近来,一早,曾,再,{渐渐,逐渐}
Frequency Adverb	df	{常常,常},反复,天天,刻刻
Degree Adverb	dd	太,最,{极,十分,很},略,稍微,更
Particle word	y	了
Aspectual verb	vav	开始,继续,{续,陆续}
Privative word	dn	{不,没,没有}

Phrase Level Indicators

Phrase level indicators mainly consist of adverbial modifier and complement. In Chinese, adverbial modifier and complement have direct relation with the main verb in a sentence. The phrase level linguistic indicators are shown in Table 3.

Table 3. phrase level linguistic indicators

Phrase level Indicators	Indicator tag	Example
Time interval adverbial modifier	tia	以后[这一个多月里/tia]，他见了唐小姐七八次
Time point adverbial modifier	tpa	[八月九日下午/tpa]，船到上海
Location adverbial modifier	la	六点钟在[吃早点地馆子里/la] 聚会
Time complement	tc	胳膊酸[半天/tc]
Location complement	lc	还坐在[亭子里/lc]
Measure complement	mc	鸿渐支吾掩饰了[两句/mc]
Frequency complement	fc	可是盘算[一下/fc]
Trend verb complement	tvc	把辛楣桌上六七本中西文书全搬[下来/tvc]了

Sentence Level Indicators

Some structures of a sentence are also considered as linguistic indicators. There are three cases that list as Table 4.

Table 4. sentence level linguistic indicators

Sentence level Indicators	Indicator tag	Example
Ba sentence	pba	刘小姐[把/pba]她拉进去了
Bei sentence	pbei	[被/pbei]教师痛骂一顿
Has /has not object	vgn/vgi	胳膊[酸/vgi]半天

IV. Our Approach

Based on the property of aspectual distinction, the classification is organized in three steps.

Step 1. All instances are classified into two sub classes (c1 and c2) based on static/dynamic distinction. Then c1 is a class of state instances and c2 is a class of dynamic instances.

Step 2. The dynamic instances are classified into two sub classes (c3 and c4) again based on telic/nontelic distinction. In this case, c4 is a class of process instances, and c3 is a class of dynamic&telic instances.

Step 3. As similar, the dynamic&telic instances can be also classified into class c5 and c6 based on durative/instantaneous distinction. C5 is a class of accomplishment instances, and c6 is a class of achievement instances.

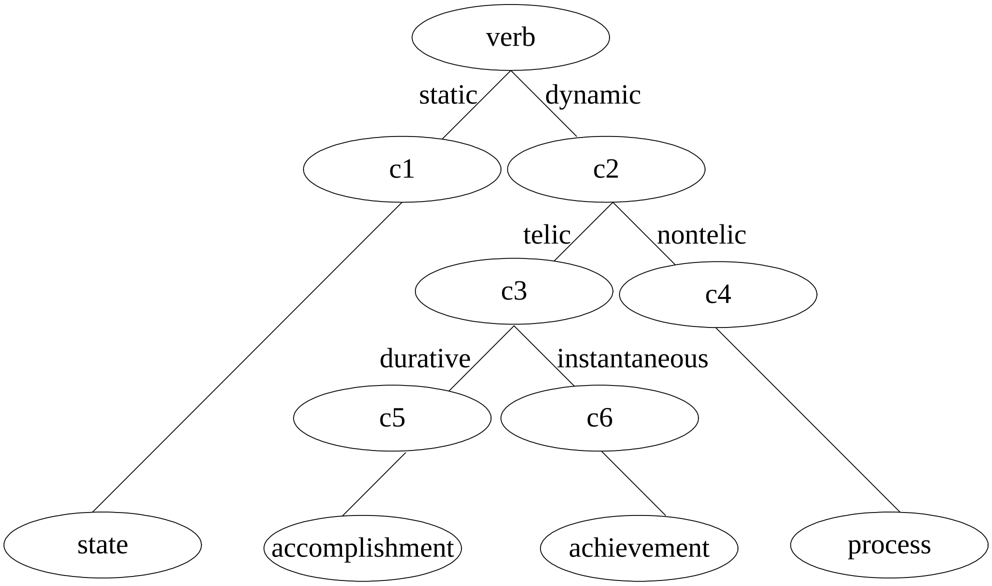

Fig. 1. Telic/nontelic prior three-layers classifier

From above three steps, all instances are classified into four aspectual classes. The three-layers classifier is shown in Fig.1, it is called telic/nontelic prior classifier. If the classification in Step 2 is based on durative/instantaneous distinction and the classification in Step 3 is based on telic/nontelic distinction, another three-layers classifier is also constructed. It is called durative/instantaneous prior classifier.

In each layer of classifier, SVM was used to do binary classification; this is because SVM is the most suitable for binary classification problems.

V. Attributes Evaluation

5.1 The RELIEF-F Evaluation Algorithm

Kira and Rendell [11] developed an algorithm called RELIEF, which is very efficient in estimating attributes. The key idea of RELIEF is to estimate attributes according to how well their values distinguish among instances that are near each other. Original RELIEF can deal with discrete and continuous attributes and is limited to only two-class problems. Kononenko, I. [12] analyzed and extended RELIEF, then developed RELIEF-F, which is able to deal with noisy and incomplete data sets and can efficiently deal with multi class problems. Each layer of the multi-layers classifier in Section 4 is a binary classification, and our data is noisy in some degrees, so we choose RELIEF-F as attributes evaluation algorithm. The attributes evaluation is done in three binary classifications separately.

5.2 Attributes Evaluation in Static/Dynamic Classification

The corpus contains total 1003 instances. In static/dynamic classification, there are 146 static instances and 857 dynamic instances.

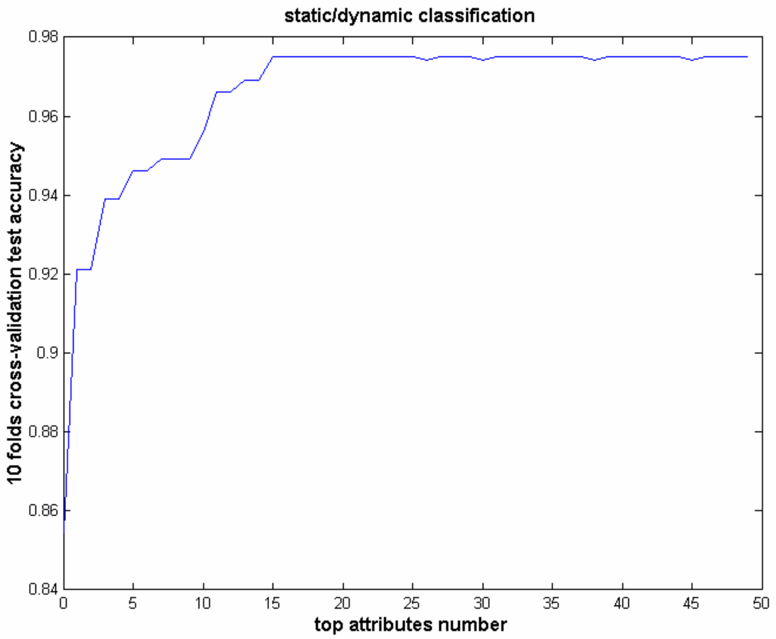

Fig. 2. 10 folds cross validation classification test result in static/dynamic distinction with top x attributes. The x-axis is the number of top attributes in RELIEF-F rank. The y-axis is the 10 folds cross validation test accuracy with according numbers of top attributes in x-axis

The RELIEF-F evaluation algorithm gives all attributes an average merit score and an average rank according to the importance of attributes in the classification.

As the rank list is too long, so it is not presented in this paper.

Based on the RELIEF-F rank, fifty times 10 folds cross-validation classification test are processed. In test N, top (N-1) attributes are selected. The result is shown in Fig. 2. It is clear that only top 15 attributes can achieve the highest accuracy, 97.5075 %. From top 15 attributes to top 49 attributes, there is a little undulation that can be ignored.

5.3 Attributes Evaluation in Telic/Nontelic Classification

There are total 857 dynamic instances, 457 telic (375 achievement and 82 accomplishment) and 400 nontelic (400 process).

The RELIEF-F evaluation algorithm ranks all attributes. The final rank list is too long to present here.

Based on the RELIEF-F rank, fifty times 10 folds cross-validation classification test are processed. In test N, top (N-1) attributes are selected.

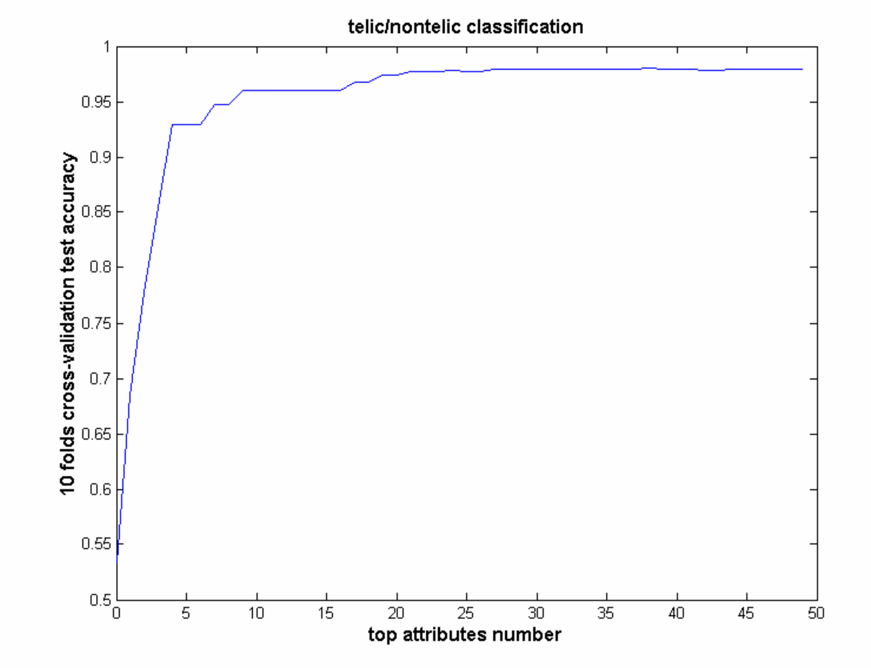

Fig. 3. 10 folds cross validation classification test result in telic/nontelic distinction with top x attributes. The x-axis is the number of top attributes in RELIEF-F rank. The y-axis is the 10 folds cross validation test accuracy with according numbers of top attributes in x-axis

The first highest accuracy is got at top 38 attributes, it is 98.0163 %, but only this point's accuracy is 98.0163 %. At top 27 attributes, the second highest accuracy is 97.8996%, and persists nearly from 27 to 49, except the highest point at 38 and a little descent at 42, 43.

5.4 Attributes Evaluation in Durative/Instantaneous Classification

There are total 857 dynamic instances, 375 instantaneous (375 achievement) and 482 durative (82 accomplishment and 400 process).

The RELIEF-F evaluation algorithm ranks all attributes in this distinction classification. The final rank list is too long to present here.

Based on the RELIEF-F rank, fifty times 10 folds cross-validation classification test are processed. In test N, top (N-1) attributes are selected.

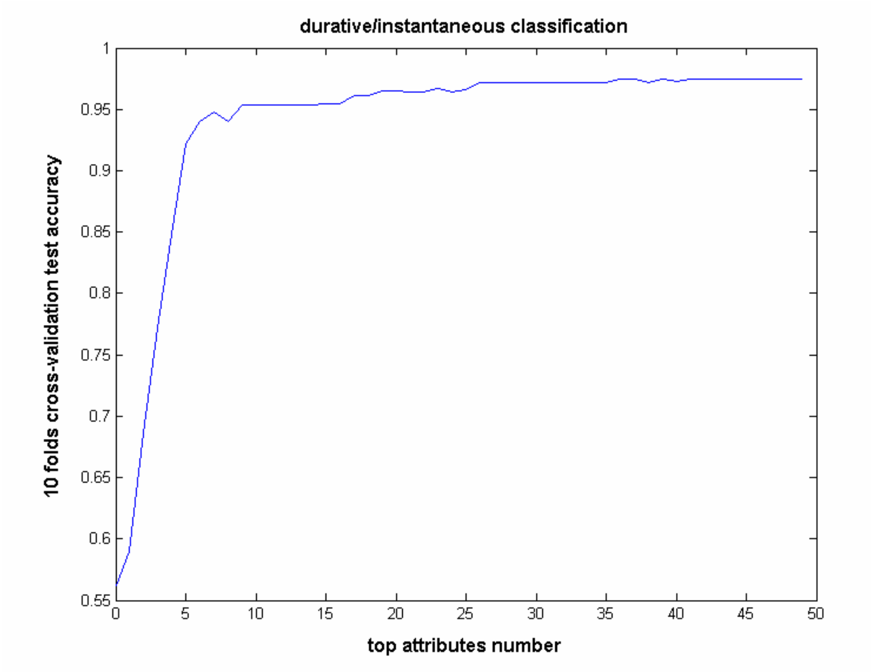

Fig. 4. 10 folds cross validation classification test result in durative/instantaneous distinction with top x attributes. The x-axis is the number of top attributes in RELIEF-F rank. The y-axis is the 10 folds cross validation test accuracy with according numbers of top attributes in x-axis

The first highest accuracy 97.4329 % is at top 36 attributes. Then it persist this highest accuracy except at 38 and 40.

For the static/dynamic classification, it achieves the first highest accuracy at top 15 attributes. But for telic/nontelic and durative/instantaneous classifications, it achieves the first highest accuracy at top attributes 38 and 36 separately. It seems that static/dynamic classification need fewer attributes than the other two classifications.

This phenomenon accords with people's common sense. The static/dynamic classification is easier than the other classifications.

Three figures all converge to a point in macro, but have little undulation in micro. RELIEF-F evaluates each attribute's merit and ranks them. The convergence of three figures proves the RELIEF-F's efficiency. The little undulation is caused by negative effect of some attributes. Some attributes have positive effect and others have negative effect, RELIEF-F rank give the most positive attributes a high score. The positive attributes can counteract the negative effect of negative attributes, so each figure converges though it exist little undulation.

VI. Experiments

A corpus containing 1003 sentences was constructed. These sentences are extracted from a Chinese novel named *Fortress Besieged* by Zhongshu Qian.

We construct three classifiers: a single SVM classifier and two three-layers classifiers. Close test with all attributes, open test with all attributes and open test with pruned attributes by RELIEF-F are applied to three classifiers separately.

Close test results in three classifiers using all 49 attributes are shown in Table 5.

Table 5. Close test in three classifiers with all 49 attributes

Aspectual classes	**Test Data number**	**Single SVM (correct number)**	**T/N prior classifier (correct number)**	**D/I prior classifier (correct number)**
State	146	123	123	123
Achievement	375	373	373	372
Accomplishment	82	67	66	69
Process	400	387	390	388
Accuracy	-	94.72%	94.92%	94.92%

In Table 5, Test Data column shows the number of instances in each aspectual class of Test Set. The other three columns show the number of instances that are correctly classified to each aspectual class by the associated classifier. T/N prior classifier is an abbreviation for Telic/nontelic prior classifier; D/I prior classifier is an abbreviation for Durative/instantaneous classifier.

In order to do an open test, the 1003 instances in corpus are divided into approximately equal two parts. The training set has 502 instances; the test set has 501 instances. Not only the instance sizes in two parts are approximately equal, but also instances sizes in each aspectual class are also approximately equal. The open test results are shown in Table 6, the column's meanings of Table 6 are same with the Table 5's.

Table 6. Open test in three classifier with all 49 attributes

Aspectual classes	**Test Data Number**	**Single SVM (correct number)**	**T/N prior classifier (correct number)**	**D/I prior classifier (correct number)**
State	73	57	57	57
Achievement	187	187	187	187
Accomplishment	41	31	33	33
Process	200	189	191	191
Accuracy	-	92.61%	93.41%	93.41%

The pruned attributes are selected based on the fig.2, fig.3 and fig.4. The stable points with the highest accuracy are selected. In fig.2, 15 is the first stable and highest point. In the fig.3, the highest point is 38, but this point is lone, not stable, so 27 is chosen. In the fig.4, two points are chosen, 36 and 26. 26 is another stable point that has a little lower accuracy than the point 36. Another reason choosing point 26 is to test the efficiency of fewer attributes.

There are two open tests with attributes numbers (15,27,36) and (15,27,26) in each layer.

Table 7. Open test in two three-layers classifier with pruned attributes

Aspectual classes	Test Data Number	T/N prior classifier (correct number)	D/I prior classifier (correct number)
State	73	57	57
Achievement	187	187	187
Accomplishment	41	33	33
Process	200	191	191
Accuracy	-	93.41%	93.41%

It may be a coincidence that these two open tests with pruned attributes have the same results.

It can achieve same accuracy with fewer attributes in classification. Fewer attributes need fewer computing resource, and have more quick computing speed. It is an important considering factor in huge data classification problem.

The three-layers classifiers have higher accuracy than the single classifier. In the close test, single SVM achieve 94.72% accuracy, telic/nontelic prior classifier and durative/instantaneous prior classifier both achieve 94.92% accuracy. The accuracy improves 0.2 %. In the open test, single SVM achieve 92.61% accuracy, telic/nontelic prior classifier and durative/instantaneous prior classifier achieve 93.41%. The accuracy improves 0.8 %. It proves that three-layers classifier have better performance in open test than that in close test.

The three-layers classifiers divided the complex classification problem into three simple classifications. For each simple binary classification, it can achieve high accuray with little confusion; the combination of simple classifiers gets good result. The three-layers classifiers proposed in this paper reorganize the structure of the SVM classifier used in Multi-classes problem. The single SVM classifier actually does a vote in classifying multiple class problems. It divides the classification into several binary classifications, and does a vote over all binary classifications. The three-layers classifier optimizes the whole classification process in structural level and gets better result as we assumed.

VII. Conclusion

In this paper, we propose an aspectual classification method using Chinese linguistic indicators. Forty-nine linguistic indicators were used in our work. A corpus containing 1003 instances with linguistic indicators was created. Based on the aspectual distinctions, two three-layers classifiers were put forward. The experiments prove that our three-layers classifiers have a high accuracy than single classifier. The attributes selected by RELIEF-F can achieve the same accuracy compared to all attributes in open test. More work in attributes selection will be done in the future.

References

[1] Eric V. Siegel: Corpus-Based Linguistic Indicators for Aspectual Classification. In Proceedings of the 37th Conference on Association for Computational Linguistics. June 20-26, 1999, College Park, Maryland

[2] Eric V. Siegel and Kathleen R. McKeown: Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights. Computational Linguistics, Volume 26, Issue 4(December 2000), pp.595–628

[3] Paola Merlo and Suzanne Stevenson: Automatic Verb Classification Based on Statistical Distributions of Argument Structure. Computational Linguistics, Volume 27, Issue 3 (September 2001), pp.373–408

[4] Suzanne Stevenson and Paola Merlo: Automatic Verb Classification Using Distributions of Grammatical Features. European Chapter Meeting of the ACL, in Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Bergen, Norway, pp.45–52

[5] Chen ping: Discussion On Temporal System of Contemporary Chinese. Chinese Languages and Writings, Vol.6, 1998

[6] Ma Qingzhu: Time Quantity Phrase and Categories of Verbs. Chinese Languages and Writings, Vol.2, 1981

[7] Xiaodan Zhu, Chunfa Yuan, K.F.Wong and Wenjie.Li: An Algorithm for Situation Classification of Chinese Verbs. In Proceedings of the Second Chinese Language Processing Workshop. Oct. 2000, Hong Kong, pp.140-145

[8] Wei Li, Chunfa Yuan, Kam-Fai Wong, Chun-Hung Cheng and Wenjie Li: Using Fuzzy Sets and Genetic Algorithm for Chinese Verb Classification. In proceeding of 2004 World Congress on Intelligent Control and Automation (WCICA 2004). June 14-18, 2004, Hangzhou, China, pp2112-2118

[9] Kam-Fai Wong, Wenjie Li, Chunfa Yuan and Xiaodan Zhu: Temporal Representation and Classification in Chinese. Int. J. Comput. Process. Oriental Lang. 15, 2 (2000), pp.221-230

[10] Kam-Fai Wong, Wenjie Li and Chunfa Yuan: Classifying Temporal Concepts in Chinese for Information Extraction. In Proceedings of 5th Natural Language Processing Pacific Rim Symposium (NLPRS'99). November 5-7, 1999, Beijing, pp.172-177

[11] Kira K. and Rendell L.: A practical Approach to Feature Selection. In Proc. Intern. Conf. on Machine Learning. (Aberdeen, July 1992) D.Sleeman and P.Edwards (eds.), Morgan Kaufmann 1992, pp.249-256

[12] Igor Kononenko: Estimating Attributes: Analysis and Extensions of RELIEF. In Proceedings of the European conference on machine learning on Machine Learning. Catania, Italy, (1994), pp.171-182

Defang Cao
09/1999 - 07/2003: Dept. of Computer Science & Technology, Tsinghua University. ***Bachelor***.
09/2003 - present: Dept. of Computer Science & Technology, Tsinghua Univ. ***Master Candidate***

Wenjie Li
BSc(Tianjin University); MSc(Tianjin University); Ph.D.(CUHK). Research interests include natural language processing, information extraction, question answering and text summarization.

Chunfa Yuan
BSc(Tsinghua University); MSc(Tsinghua University).
Research interests include natural language processing, information extraction

Kam-Fai Wong
Kam-Fai Wong recieved his PhD from Edinburgh University in 1987. He is now the Associate Dean of Engineering and Professor, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. He is the editor in chief of the ACM Transaction on Asian Language Processing and International Journal on Computer Processing of Oriental Languages.