Automatic Recognition of Chinese Unknown Words[1] Based on Roles Tagging[2]

Kevin Zhang (Hua-Ping ZHANG) Qun LIU Hao ZHANG Xue-Qi CHENG
Email: {zhanghp,liuqun, zhaohao,cxq}@software.ict.ac.cn
Software Division, Institute of Computing Technology, Chinese Academy of Sciences
NO. 6, South Road, Kexueyuan, Zhongguancun, Haidian Dist. P.O. BOX 2704, Beijing, P.R. China, 100080

Abstract

This paper presents a unified solution, which is based on the idea of “roles tagging”, to the complicated problems of Chinese unknown words recognition. In our approach, an unknown word is identified according to its component tokens and context tokens. In order to capture the functions of tokens, we use the concept of roles. Roles are tagged through applying the Viterbi algorithm in the fashion of a POS tagger. In the resulted most probable roles sequence, all the eligible unknown words are recognized through a maximum patterns matching. We have got excellent precision and recalling rates, especially for person names and transliterations. The result and experiments in our system ICTCLAS shows that our approach based on roles tagging is simple yet effective.

Keywords: Chinese unknown words recognition, roles tagging, word segmentation, Viterbi algorithm.

Introduction

It is well known that word segmentation is a prerequisite to Chinese information processing. Previous research and work in word segmentation have made great progresses. However, cases with unknown words are not satisfactory. In general, any lexicon is limited and unable to cover all the words in real texts or speeches. According to our statistics on a 2,305,896-character news corpus from *the People's Daily*, there are about 1.19% unknown words. But they are difficult to be recalled and often greatly reduce the recognition rate of known words close to them. For example, the sentence “部长孙家正在工作。” (Pronunciation: “Bu Zhang Sun Jia Zheng Zai Gong Zuo.”) has two valid segmentations: “部长/孙家正/在/工作” (The minister Sun Jiazheng is at work) and “部长/孙家/正在/工作” (The minister Sun Jia now is at work). “孙家正” is a person name in the first, while “孙家” is another name in the latter. Meanwhile, the string “孙家正在” will lead to overlapping ambiguity and bring a collision between the unknown word “孙家正”(Sun Jiazheng) and “正在”(zheng zai; now). What’s more, the recognizing precision rates of person names, place names, and transliterations are 91.26%, 69.12%, and 82.83%, respectively, while the recalling rates of them are just 68.77%, 60.47%, and 78.29%, respectively. (Data from official testing in 1999) [Liu (1999)] In a word, unknown words recognition has become one of the biggest stumbling blocks on the way of Chinese lexical analysis. A proper solution is important and urgent.

Various approaches are taken in Chinese unknown words recognition. They can be broadly categorized into “one-for-one”, “one-for-several” and “one-for-all” based on the number of categories of unknown words, which can be recognized. One-for-one solutions solve a particular problem, such as person name recognition [Song (1993); Ji (2001)], place name recognition [Tan (1999)] and transliteration recognition [Sun (1993)]. Similarly,

[1] We define unknown words to be those not included in the core lexicon and unable to be generated by FSA, such as person name, place name. But numeric or common time word is not unknown because they can generate by a simple FSA.

[2] Related research in this paper is supported by Foundation of National Key Basic Research Project (ID: G1998030507-4 and G1998030510).

one-for-several approaches provide one solution for several specific categories of unknown words [Lv (2001); Luo (2001)]. One-for-all solutions, as far as we know, have not been applicable yet [Chen (1999); He (2001)].

Although currently practicable methods could achieve great precision or recalling rates in some special cases, they have their inherent deficiencies. First of all, rules applied are mostly summarized by linguists through painful study of all kinds of huge “special name libraries” [Luo (2001)]. It’s time-consuming, expensive and inflexible. The categories of unknown words are diverse and the amount of such words is huge. With the rapid development of the Internet, this situation is becoming more and more serious. Therefore, it’s very difficult to summarize simple yet thorough rules about their compositions and contexts. Secondly, the recognition process cannot be activated until some “indicator” tokens are scanned in. For instance, possible surnames or titles often trigger person name recognition on the following 2 or more characters. In the case of place name recognition, the postfixes such as “县”(county), “市”(city) will activate the recognition on the previous characters. What’s more, these methods tend to work only on the monosyllabic tokens, which are obvious fragments after tokenization [Luo (2001); Lv (2001)]. It takes the risk of losing lots of unknown words without any explicit features. Furthermore, this trigger mechanism cannot resolve the ambiguity. For example, unknown word “方林山” (Fang Lin Shan) maybe a person name “方/林山”(Fang Linshan) or a place name “方林/山”(Fanglin Mountain).

This paper presents a one-for-all approach based on roles tagging to avoid such problems. The process is: tagging tokens after word segmentation with the most probable roles and making unknown words recognition based on roles sequence. The mechanism of roles tagging is just like that of a small and simple Part-Of-Speech tagger.

The paper is organized as follows: In section 2, we will describe the approach in general. Following that, we will present the solution in practice. In the final part, we provide recognition experiments using roles-tagging methods. The result and possible problems are discussed as well.

1 Unknown words recognition based on roles tagging

1.1 Lexical roles of unknown words

Unknown words are often made up of distinctive components, most of which are monosyllabic characters or short words; in addition, there are some regular relations between unknown words and their locality, especially with their left and right context. As we often write or speak, a Chinese person name is usually comprised of a one-or-two-character surname and a following given name of one or two characters, like “肖建群”(Xiao Jianqun) and “诸葛亮”(Zhu-Ge Liang). The previous words are mostly titles, occupations or some conjunctive words, such as “经理”(Manager), “司机”(Driver) and “对”(To). The following words tend to be verbs such as “说” (to say) , “表示”(to express). Similar components, contexts and relations can be discovered in place name, transliteration, organization name, or other types of unknown words.

We define unknown word roles with respect to varied internal components, previous and succeeding contexts and other tokens in a particular sentence. Various roles are extracted according to their functions in the forming of different unknown words. Person names roles and transliterations roles set are shown in table 1a and 1b respectively. Using the roles set of person name, the tokens sequence “馆/内/陈列/周/恩/来/和/邓/颖/超生/前/使用/过/的/物品/” (What Zhou Enlai and Deng Yunchao used before death are presented in the museum) will be tagged as “馆/A 内/A 陈列/K 周/B 恩/C 来/D 和/M 邓/B 颖/C 超生/V 前/A 使用/A 过/A 的/A 物品/A”.

Role	Significance	Examples
B	Surname or family name.	*张*/华/平/先生；*欧阳*/修
C	First Chinese char in the 2-char given name	张/*华*/平/先生
D	Last Chinese char in the 2-char given name.	张/华/*平*/先生

E	Given name with a single Chinese char.	张/*浩*
F	Prefix in the name.	*老*/刘、*小*/李
G	Postfix in the name.	王/*总*、刘/*老*、肖/*氏*
K	Previous context before person name.	又/*来到*/于/洪/洋/的/家
L	Succeeding context following person name.	新华社/记者黄文/*摄*
M	Parts between two person names.	编剧/邵/钧/林/*和*/稽/道/青/说
U	Known words generated by previous context and the first component of name.	这里/*有关*/天/培/的/壮烈/；现任/主席/*为何*/鲁/丽/。/
V	Known words generated by the last component and next context.	龚/学/*平等*/领导/, 邓/颖/*超生*/前
.....		
A	Others tokens not mentioned above.	*全国*/*人民*/*深切*/缅怀/邓/小/平/

Table 1a: Roles set of Chinese person names

Role	Significance	Examples
B	The first component of transliteration	*克*/林/顿
C	Middle component	史/*蒂*/*芬*/*·*/*斯*/*皮*/*尔*/*伯*/格
D	Last component	克/林/*顿*
.....		

Table 1b: Roles set of transliterations

1.2 Roles tagging and unknown words recognition

On the one hand, the sentence include words with different roles for a particular category of unknown words, on the other hand, such words can be recognized after identifying their roles sequence. That is: tagging tokens after word segmentation with the most probable roles sequence, then recognizing unknown words by maximum patterns matching on the final roles sequence.

Roles tagging is similar to Part-Of-Speech tagging. Our tagging process is based on Viterbi Algorithm [Rabiner and Juang (1989)], which is to select the optimum with maximum probability from all possible tag sequences. The methodology and its deduction is given as below:

Suppose that T is the tokens sequence after word segmentation and R is the roles sequence for T. We take the role sequence $R^{\#}$ with the maximum probability as the best choice. That is:

$T=(t_1, t_2, \dots, t_m)$,

$R=(r_1, r_2, \dots, r_m)$, m>0,

$R^{\#}=\arg\max_{R} P(R|T)$..................….........E1

According to the Bayes equation, we can get:

$P(R|T)= P(R)P(T|R)/P(T)$ E2

For a particular token sequence, P(T) is a constant. So, We can get E3 based on E1 and E2:

$R^{\#}=\arg\max_{R} P(R)P(T|R)$ E3

We may consider T as the observation value sequence while R as the state sequence hidden behind the observation. Now we introduce Hidden Markov Model [Rabiner and Juang (1986)] to resolve such a typical problem:

$$P(R)\,P(T|R) \approx \prod_{i=0}^{m} p(t_i \mid r_i)\, p(r_i \mid r_{i-1})$$

$$\therefore R^{\#} \approx \arg\max_{R} \prod_{i=0}^{m} p(t_i \mid r_i)\, p(r_i \mid r_{i-1}) \text{ ....E4}$$

$$\Leftrightarrow R^{\#} \approx$$

$$-\arg\min_{R} \sum_{i=0}^{m} \{\ln p(t_i \mid r_i) + \ln p(r_i \mid r_{i-1})\} \text{ ......E5}$$

E5 is simpler for computation than E4.

Now, we can find the most possible token sequence with equation E5. It’s a simple application of Viterbi Algorithm.

The final recognition through maximum pattern matching is not performed on the original texts but performed on roles sequence. The person patterns are {BBCD, BBE, BBZ, BCD, BE, BG, BXD, BZ, CD, FB, Y, XD}. Before matching, we should split the tokens whose roles are like “U” or “V”(which indicate that the related token is generated by internal components and the outside contexts of unknown words) into two proper parts. Such a processing can recall more unknown words and reduce the overlapping collision. As for the above sample sentence, the final roles sequence after splitting is “AAK*BCD*M*BCD*LAAAAAA”. Therefore, we can identify the possible person names “周恩来” and “邓颖超” according to the recognition pattern “BCD”.

1.3 Automatic acquisition of roles knowledge

As described in E5, the tag sequence $R^{\#}$ is decided by two kinds of factors: $p(t_i \mid r_i)$ and $p(r_i \mid r_{i-1})$. $p(t_i \mid r_i)$ is the probability of a token t_i given the condition of being tagged with role r_i, while $p(r_i \mid r_{i-1})$ is the transitive probability from role r_{i-1} to role r_i. Both factors are useful lexical knowledge for tagging and final recognition. According to laws of large numbers, if the training corpus is large enough, we can acquire the roles knowledge as following:

$p(t_i \mid r_i) \approx C(t_i,r_i)/C(r_i)$ ....……......... E6

Where C(t_i, r_i) is the count of token t_i being role r_i; and C(r_i) is the count of role r_i.

$p(r_i \mid r_{i-1}) \approx C(r_{i-1},r_i)/C(r_{i-1})$ ....…….......…E7

Where C(r_{i-1},r_i) is the count of role r_{i-1} followed by role r_i.

C(t_i,r_i), C(r_i) and C(r_{i-1},r_i) are extracted from corpus through a training process. The training corpus came from one-month news from the *People’s Daily* with 2,305,896 Chinese characters, which are manually checked after word segmentation and POS tagging (It can be downloaded at icl.pku.edu.cn, the homepage of the Institute of Computational Linguistics, Peking University).

However, the corpus is tagged with the Part-Of-Speech set. Before training, the original POS tags should be converted to the proper roles by analysing every token in the sentence.

2 Algorithm and implementation

The unknown words recognition based on roles tagging has three main steps: automatic acquisition of roles knowledge from the corpus; roles tagging with Viterbi algorithm and unknown words recognition through maximum pattern matching.

Viterbi algorithm is a classic approach in statistics. It aims to select the optimum roles sequence with maximum possibility from all possible results. Our evaluation function for decision-making is E5 given in sub-section 1.2. Considering the length limitation of this paper, we skip the details.

Therefore, we only provide algorithms for roles knowledge learning. In the last part, the entire process of unknown words recognition will be listed.

2.1 Roles knowledge learning

Input: Corpus which is segmented and POS tagged

T: the type of unknown words;

R: Roles set of T

Output: C(t_i,r_i), C(r_i) and C(r_{i-1},r_i)

Algorithm:

(1) Get one sentence S from corpus C;

(2) Extract all tokens and POS tags from S;

(3) Convert all POS tags to roles in T after role analysis.

(4) Store the tokens whose role is not ‘A’ into the recognition lexicons of unknown words T, where ‘A’ is not internal components nor context role.

(5) Calculate the total number C(t_i,r_i) of token t_i being role r_i. At the same time, count C(r_i), which is the number of role r_i appearances.

(6) Sum C(r_{i-1},r_i) which is the times of role r_{i-1} followed by role r_i.

(7) If no more sentences in the corpus C, exit; else go to (1)

First of all, we must explain step (3). Our corpus is tagged with POS and person, place or organization name are tagged with ‘nr’, ‘ns’ or ‘nt’ respectively; Such POS are unique and different from noun. Transliterations can be extracted from words tagged with ‘nr’ or ‘ns’ and through analysing its component chars. So we can easily locate such kinds of words. Meanwhile, we can judge whether a word is unknown by looking it up in the core lexicon. Then we can identify roles of words according to their locality, which are before or following a particular unknown word.

Here we provide a sample sentence from our corpus like “本报/r 蚌埠/ns １月/t １日/t 电/n 记者/n 黄/nr 振中/nr 、/w 白/nr 剑峰/nr 报道/v”. In step (2), we can extract tokens and tags like “本报”/ ‘r’; “蚌埠”/ ‘ns’ and so on. When we train person recognition roles, firstly, we locate person name “黄/nr 振中/nr” and “白/nr 剑峰/nr” just by searching POS ‘nr’; Secondly, judge whether they are unknown after looking them up in the core lexicon; At last we can tag unknown words component and their context near their locality. So the final roles after conversion are “本报/A　蚌埠/A　１月/A 1 日/A 电/A 记者/K 黄/B　振/C 中/D 、/M 白/B 剑/C 峰/D 报道/L”. Then we can train the parameters based on new segmentation and person recognition roles sequence.

In addition, we train every different kind of unknown word on the same corpus individually. That is: person roles, place roles and other roles are acquired respectively. Therefore, the unknown place recognition roles sequence of the above sentence may like “本报/K　蚌/B 埠/D １月/L　１日/A　电/A　记者/K　黄/A 振中/A　、/A　白/A 剑峰/A　报道/A”. Such a mechanism can greatly reduce the problem of sparse data.

2.2 The entire process of Unknown words recognition

Input: Original sentence S;
R: the roles set of unknown words;
P: pattern sets for recognition.
Output: Possible unknown words of type T.
Algorithm:

(1) Word segmentation (we segment words on sentence S with N-shortest paths method [Hua-Ping ZHANG, Qun LIU (2002)]);
(2) Tag tokens sequence with roles in set R using Viterbi algorithm. Get the roles sequence $R^{\#}$ with maximum possibility.
(3) Split tokens whose role is like ‘U’ or ‘V’ in the person roles. These roles indicate that the internal components glue together with their context.
(4) Maximum match final roles sequence to the recognition patterns P and record their position.
(5) Generate the candidate unknown words according to the result of pattern matching.
(6) Exclude those candidates which are fit for the exclusive rules.(For example, Chinese person name can not include non-Chinese chars.)
(7) Output the possible unknown words.

Now, we take person recognition on the sentence “本报蚌埠１月１日电记者黄振中、白剑峰报道” as exemplification. In the first place, we can get the sequence “本报/蚌埠/１月/１日/电/记者/黄/振/中/、/白/剑/峰/报道/” after rough word segmentation; Then we tag it with Viterbi algorithm using person recognition roles lexicon and transitive array. So, the most probable roles sequence is “AAAAAK<u>*BCD*</u>M<u>*BCD*</u>L”.Therefore, candidate perosn names “黄振中” and “白剑峰” can be recognized after maximum string matching.

3 Experiments and Discussions

Both close and open recognition test were conducted. In the close test, we tested our system within the training corpus, which is the knowledge base for recognition. Open test, however, is more realistic, because it is performed on arbitrary real texts outside the training corpus. The corpus in our experiments is from 2-months news in 1998 from *the People’s Daily*.

In this paper, we only provide the recognition results of Chinese person and transliterations. The recognition of place names and other kind of unknown words can get similar performance.

3.1 Recognition experiment of Chinese person name

Test Type	Close	Open
Corpus (news date)	1.1-2.20	2.20-2.28
Corpus Size	14,446K	2,605K
Num of Chinese person names	21,256	3,149
Num of recognized person names	27,813	4,130
Num of correctly recognized names	20,865	2,886
Precision rate	75.02%	69.88%
Recalling rate	98.17%	91.65%
F-measurement	85.05%	79.30%

Table 2 Experiment results of Chinese person names recognition

In Tables 2, precision rate and recalling rate are defined as equations E6 and E7 respectively. In addition, F-measurement is a uniformly weighted harmonic mean of precision rate and recalling rate as shown in E8.

Precision rate=

$$\frac{\text{num of correctly recognized words}}{\text{num of recognized words}} \quad \text{……..E6}$$

Recalling rate=

$$\frac{\text{num of correctly recognized words}}{\text{num of total unknown words}} \text{……..E7}$$

F-measurement =

$$\frac{\text{Recalling rate} \times \text{Precision rate} \times 2}{\text{Recalling rate} + \text{Precision rate}} \text{...….E8}$$

3.2 Recognition Experiments of transliterations

Test Type	Close	Open
Corpus (news date)	1.1-2.20	2.20-2.28
Corpus Size	14,446K	2,605K
Num of transliterations	9,059	1,592
Num of recognized transliterations	10,013	1,930
Num of correctly recognized transliterations	8,946	1,496
Precision rate	89.35%	77.52%
Recalling rate	98.75%	93.97%
F-measurement	93.85%	84.96%

Table 3 Results of transliterations recognition

3.3 Discussions

The traditional ways to test unknown words recognition is to collect sentences including unknown words and to make recognition experiments. Those sentences that haven't the type of unknown words will be excluded from experiments in the pre-processing. In our experiments, we just take the realistic corpus and make no filtering. Therefore, the precision rates may be lower but closer to the realistic linguistic environment than previous tests. We have made experiments in the traditional way and the precision rate can be improved by less than 15%. In a word, there is no comparable with precision rates of previous unknown words recognition experiment.

In addition, our experiments show that the unknown words recognition based on role tagging can achieve very high recalling rates. For such a problem, recalling is more essential than precision. Low recalling rate means that we have no chance to recognize many unknown words through any efforts in the following steps, although words recognized are mostly valid; However, precision rate can be greatly improved in other processes, such as POS tagging or sentence simple parsing. In our system ICTCLAS (Institute of Computing Technology, Chinese Lexical System), we can exclude most invalid unknown words during POS tagging. The precision rate of Chinese person names recognition can achieve over 95% after POS tagging while the recalling rate is not reduced.

Our approach is purely corpus-based. We all know that, in any usual corpus, unknown words are sparsely distributed. If we depend totally on the corpus, the problem of sparse data is inevitable. But in the fine-tuning of our system, we found some countermeasures and successfully solved the problem.

Lexical knowledge from linguists can be incorporated into the system. This does not mean that we fall back to the old ways. We just demand for those general rules about name formation to avoid apparent mistakes. As to person name recognition, there are several strict restrictions, such as the length of name, the order between surname and given name.

Except for enlarging the training corpus, we provide two more counteractions:

Firstly, a "best n" approach [Hua-Ping ZHANG, Qun LIU (2002)], which provides n (n>1) possible tag sequences with leading probabilities, is feasible. Usually the desired tag sequence could be re-targeted or constructed from the best n sequences. In this way, we improved the recalling rate at the cost of precision rate. But given a better recalling, we could remedy in latter stages of language processing. When 3 most probable sequences are employed, the recalling and precision of unknown words in ICTCLAS can be enhanced obviously.

The second resolution is training on a name library in addition to training on a corpus. As we all know, it's easier and cheaper to get a person name library or other special name libraries than to segment and tag a corpus. We could extract the inner components relations from the unknown words libraries, and then merge these data into the roles information from the original corpus. When the special name libraries were introduced, both precision and recalling rates can be improved.

Conclusion

The paper presents a one-for-all approach for Chinese unknown words recognition based on roles tagging. At first, we define roles set for every category of unknown words according to the function of tokens, such as internal component or contexts. Unknown words are recognized on roles sequence, tagged with the roles set using Viterbi algorithm. The knowledge about roles is extracted from the learning on corpus. Experiments on large size corpus verify that the approach based on role tagging is simple and applicable.

Acknowledgements

First of all, our thanks go to the Institute of Computational Linguistics, Peking University for providing the corpus. And we owe lots of thanks to our colleagues: Li Sujian, Li Jifeng, Li Shengtao, Zhu Hailong, Zhao Hongchao, Wang Shuxi and Dr. Zhou Lixin. We would especially express gratitude to the chief scientist Bai Shuo.

References

K.Y. Liu (1999) *The Assessment to Automatic Word Segmentation and POS Tagging Software.* Proceedings of the 4th Conference on Chinese Computer Intelligent Interface and Application, Beijing.

Z. Luo and R. Song (2001) *Integrated and Fast Recognition of Proper Noun in Modern Chinese Word Segmentation.* Proceedings of International Conference on Chinese Computing 2001, Singapore, pp. 323-328.

H. Luo and Z. Ji (2001) *Inverse Name Frequency Model and Rules Based on Chinese Name Identifying.* In "Natural Language Understanding and Machine Translation", C. N. Huang & P. Zhang, ed., Tsinghua Univ. Press, Beijing, China, pp. 123-128.

R. Song (1993) *Person Name Recognition Method Based on Corpus and Rule.* In “Computational Language Research and Development", L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press.

H. Y. Tan *(1999) Chinese Place Automatic Recognition Research.* In "Proceedings of Computational Language ", C. N. Huang & Z.D. Dong, ed., Tsinghua Univ. Press, Beijing, China.

M.S. Sun *(1993) English Transliteration Automatic Recognition.* In "Computational Language Research and Development", L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press.

Y.J. Lv, T. J. Zhao (2001) *Levelled Unknown Chinese Words Resolution by Dynamic Programming.* Journal of Chinese Information Processing. 15, 1, pp. 28-33.

X. H. Chen (1999) *One-for-all Solution for Unknown Word in Chinese Segmentation.* Application of Language and Character, 3.

Y. He (2001) *Identification of Unlisted Words on Transitive Probability of Monosyllabic Words.* In "Natural Language Understanding and Machine Translation", C. N. Huang & P. Zhang, ed., Tsinghua Univ. Press, Beijing, China, pp. 123-128.

Hua-Ping ZHANG, Qun LIU (2002) *Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method.* Journal of Chinese Information Processing. 16, 5, pp. 77-83.

L. R.Rabiner (1989) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of IEEE 77(2): pp.257-286.

L.R. Rabiner and B.H. Juang, (Jun. 1986) *An Introduction to Hidden Markov Models.* IEEE ASSP Mag., Pp.4-166.

基于角色标注的中文未登录词自动识别研究[3]

张华平　刘 群　张 浩　程学旗
Email : zhanghp@software.ict.ac.cn
中国科学院计算技术研究所软件室
北京2704信箱海淀区中关村科学院南路6号，中国

摘要：本文提出了一种基于角色标注的中文未登录词识别通用方法。角色指的是未登录词的内部组成成分、上下文以及句子中的其他部分。我们识别未登录词依据的是句子角色序列。首先，采取 Viterbi 算法对切分后的 Token 序列进行角色标注，这和词性标注类似。然后得到一个概率最大的角色序列，在此基础上，采取模式最大匹配识别未登录词。和角色相关的这些语言学知识可以从对语料库的训练学习当中抽取得到。作者做了几个识别实验，取得了较好的准确率和召回率，尤其是中国人名和音译名的识别。我们将基于角色标注识别未登录词的方法集成到了计算所汉语词法 ICTCLAS 系统，取得了较好的效果。实验结果表明：基于角色标注的未登录词识别方法是简单可行、令人满意的。

关键词:中文未登录词识别，角色标注，分词，Viterbi 算法.

[3] 本文有关研究得到国家重点基础研究项目（编号：G1998030507-4；G1998030510）资助。