Research on Method of Automatic Recognition of Chinese Place Name Based on Transformation*

TAN Hong-ye, ZHENG Jia-heng, LIU Kai-ying

(*Department of Computer Science, Shanxi University, Taiyuan* 030006, *China*)

E-mail: zxdthy@163.com

http://www.sxu.edu.cn

Received April 19, 2001; accepted July 12, 2001

Abstract: The automatic recognition of Chinese place names, a special case of the recognition of Chinese special nouns, is an important task in Chinese information processing. The method based on statistical technique can only ensure the recall to some degree, but the precision is relatively low. In this paper, an approach based on transformation is proposed, which can effectively overcome the deficiency caused by statistics. The performance of the approach is evaluated on a real data set, and the precision finally reaches 90.9%, improved by 7%.

Key words: Chinese place name; automatic recognition; transformation; special noun; Chinese information processing

The rapid development of information technology and World-Wide-Web has led to an increased interest in automatic information processing, for example automatic input and output of information, automatic proofreading and classification of texts, information retrieval and indexing, and machine translation, which all involve automatic recognition of special nouns. In general, special nouns include Chinese person names, Chinese place names, translation names, organization names, and brand names etc. We focus on the automatic recognition of Chinese place names in this paper.

We have already used the statistical technique to recognize Chinese place name, the recall and the precision are 97.5% and 83.6% respectively[1]. In this paper, we put forward a method of automatic recognition of Chinese place name based on transformation and the consequent precision reaches 90.9%, improved by 7%.

1 Limitation of the Method Based on Pure Statistical Measures

In the statistical measures, we first build a Chinese Place Name Base (CPB) based upon the largest, most standard book Chinese Place Name Set (CPN)[2], which includes 100 000 Chinese place names. Then we analyze the features of the place name characters and the place name words in CPB, and obtain the Chinese Place Name Character Base (CPCB) and the Chinese Place Word Base. After that, we calculate the weight of likelihood

* Supported by the National "Ninth Five-year plan" Social Science Foundation of China under Grant No. 97@yy001-2 (国家"九五"计划社科基金)

TAN Hong-ye was born in 1971. She is a lecturer at the Department of Computer Science, Shanxi University. She received her Master degree in computer application from Shanxi University in 2000. Her research interests are artificial intelligence, Chinese information processing and automatic recognition of special nouns. **ZHENG Jia-heng** was born in 1948. She is a professor and master supervisor of the Department of Computer Science, Shanxi University. Her current research areas are artificial intelligence and Chinese information processing. **LIU Kai-ying** was born in 1931. He is a professor of the Department of Computer Science, Shanxi University. His research areas are artificial intelligence and Chinese information processing.

(WOL) for each place name character in CPCB according to the real text training set containing 2 800 000 characters, and get a probability estimation value for each character. The WOL of a place name character represents the likelihood of a character being a part of a Chinese place name in a real corpus. Finally we design the probability estimation formula with the help of the statistical information in CPB and WOL.

After testing, we found out that this approach can ensure the recall to some extent, which dues to that the book Chinese Place Name Set includes most of Chinese place name characters and words. However, this approach cannot achieve a plausible result because of the following reasons: (1) Not all the place names are included in CPN, even though CPN already covers nearly 100 000 place names. In the real texts containing 2 800 000 characters, there are 1 793 place names occurring 11 590 times in total. Among those place names, 540 place names, about 30% of the total, are not included in CPN, such as "蟒河风景区", "小浪底", and "中关村". The reason is these new place names appear with the development of economy and society. Moreover, 5.5% of these place names not included in CPN cannot be recognized because the characters making up the place names aren't included in CPN; (2) According to our research, there are 3 685 Chinese place name characters in total and the characters maybe occur as a part of Chinese place name or a part of other words in real texts. As a result, a lot of incorrect place names must be caused based on the statistical measure. Table 1 gives some examples of place names being recognized incorrectly.

Table 1 Examples of place names being recognized incorrectly

Sentences	Incorrect place names
为贫困地区铺富路	富路
各大中城市都要自己负起综合平衡的责任	大中城市
从民政部到各省市区民政厅局,大兴调查研究之风	大兴
进入全国邮电百强县行列	百强县
为海峡两岸企业架金桥	金桥

We have tested the system based on the statistical method. The testing corpus consists of the articles from the People's Daily, containing 50 000 characters and 514 Chinese place names. After the corpus is tagged, 599 place names are recognized, among which correct place names are 501, boundary-error place names are 9, and entirely incorrect place names are 89. Meanwhile, there are 22 place names cannot be recognized. From the above, we conclude that the elimination of entirely incorrect place names is most important among all the errors caused by the statistical measure. Therefore, our further work focuses on the confirmation of correct place names and the elimination of incorrect place names.

In order to solve this problem, we have managed to make the data, used in the statistical model, more real by increasing the size of training corpus. Meanwhile, some rules have been summarized after we analyzed and studied a great number of place name examples being recognized correctly and incorrectly. So far, this kind of rules is more than 20, which can be classified into four types: initial recognition type choosing rules confirmation rules, negation rules and boundary modification rules. But the rules obtained by human summarization cannot cover the entire existing relationships between place names and their contexts. So in this paper we use transformation-based machine learning to overcome the deficiency. The experimental results show that it is effective and efficient in the confirmation of correct place names and the elimination of incorrect place names.

2 Transformation-Based Machine Learning

Eric Brill[3] proposed the method of the transformation based, error-driven learning in automatic Part of Speech (POS) tagging and improved the precision. This method consists of two steps: (1) Obtaining rules. At first, the most probable POS tagger of each word, gained in a tagged corpus, is used to tag the training corpus. Then, the most valuable transformation rules are selected from all the transformations after comparing the tagging

results with the standard texts. In the end, the rules are used to tag the training corpus again and again until no more new transformation rules are found; (2) Using transformation rules to tag texts. At first texts are tagged initially. Then, the tagged texts are corrected by utilizing the transformation rules. Thus, the final tagging texts are formed.

In this paper, we use the similar idea to confirm and eliminate place name candidates.

2.1 Types of transformation rules

At present, the transformation rules have 2 types: confirmation transformation rules and negation transformation rules.

Suppose:

CpString: a Chinese place name character sequence, which is initially recognized as a place name candidate,

$Word_{Pre}$: the preceding word of CpString,

$Word_{Next}$: the next word of CpString,

Validity: a Boolean mark used to indicate whether CpString is valid,

W_p and W_n: certain words,

Confirmation rules are as the following:

(1) if $CpString.Word_{Pre}=W_p \&\& CpString.Word_{Next}=W_n$, then $CpString.Validity:=true$;

(2) if $CpString.Word_{Pre}=W_p$, then $CpString.Validity:=true$;

(3) if $CpString.Word_{Next}=W_n$, then $CpString.Validity:=true$;

Negation rules are as the following:

(1) if $CpString.Word_{Pre}=W_p \&\& CpString.Word_{Next}=W_n$, then $CpString.Validity:=false$;

(2) if $CpString.Word_{Pre}=W_p$, then $CpString.Validity:=false$;

(3) if $CpString.Word_{Next}=W_n$, then $CpString.Validity:=false$;

Confirmation rules are used to confirm a place name candidate to be a Chinese place name, and negation rules are used to eliminate a place name candidate initially recognized.

2.2 Estimation function of a rule

We estimate the performance of a rule based upon the relative correction ratio of a rule. The estimation function of a rule is defined as follows:

$$F_i=(Num_{corr}-Num_{err})/Total,$$

where F_i is the score of the i^{th} rule, *Total* is the total number of the i^{th} rule having been used, and Num_{corr} and Num_{err} are the number of the i^{th} rule having been used correctly and incorrectly respectively. When setting the threshold $T=0.5$, about 2 000 transformation rules are learned from the training corpus.

2.3 Obtaining of transformation rules

Figure 1 illustrates how to obtain transformation rules in our approach. We use the place names being recognized correctly and incorrectly as driven source to get the rules, and thus the transformation rule set is created. The obtaining of the transformation rules consists of two steps:

(1) Obtaining of training corpus and standard corpus

We first choose 500 000-character real texts from the People's Daily as the training corpus. Then, we perform the initial recognition of place names on the training corpus and obtain the standard corpus by correcting the initial tagged corpus manually. We set the initial tagged corpus or the tagged corpus using the transformation rules as the current tagged corpus.

(2) Obtaining of transformation rules

. Compare the current tagged corpus with the standard corpus to learn initial rules, which are then used to tag

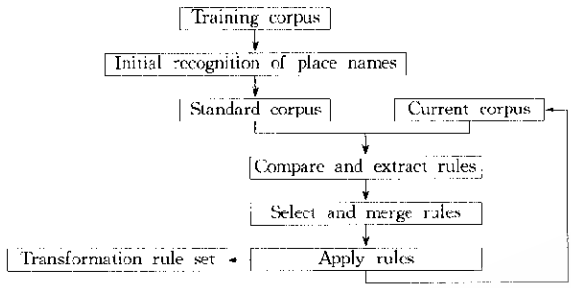

Fig. 1 Transformation Based machine learning

and correct the current corpus. Meanwhile, numbers of correction and errors for each rule are recorded respectively, which are used to calculate the score to each rule according to the estimation function. Repeat the above steps until the score for each rule is stable.

Only the rules with scores greater than a threshold value T are retained in rule set.

(3) Merge of rules

After obtaining the transformation rules, we merge some rules according to the following strategies:

• If the preceding or next words in two rules are punctuations which indicate the start or end of a subsentence, then the two rules are merged into one rule. The new rule has the information of the start or end of a sentence.

• If the preceding or next words in two rules are numerals, then the two rules are merged into one rule, which has the numeral information.

• If the number of the preceding or next word occurring in a rule is far greater than that of concurrence of the preceding and next words in it, then the next or preceding word information of the rule is neglected. Here, we use the following formula to measure the far greater degree:

$$\frac{Occur(Word_{Pre})-Concurrent(Word_{Pre},Word_{Next})}{Concurrent(Word_{Pre},Word_{Next})}>10$$

$$\frac{Occur(Word_{Next})-Concurrent(Word_{Pre},Word_{Next})}{Concurrent(Word_{Pre},Word_{Next})}>10$$

where $Occur(Word_{Pre})$ and $Occur(word_{Next})$ are the total number of the preceding word and the next word occurring in a rule respectively, $Concurrent(Word_{Pre},Word_{Next})$ is the number of concurrence of the preceding word and the next word in a rule.

For example, the preceding and next words of a rule are like:

$Word_{Pre}$	$Word_{Next}$
在	当

In this rule, $Occur$("在")=384, $Concurrent$("在","当")=1. Accordingly, the rule changes into:

$Word_{Pre}$	$Word_{Next}$
在	(null)

After merged, more than 600 rules remain in the transformation rule set, among which there exist 70 negation rules. The examples of the rules are showed as follows:

• if $CpString.Word_{Pre}$=到, then $CpString.Validity$:=$true$;
e.g.:1992年7月到/蚌城/这家专科医院治疗,

• if $CpString.Word_{Pre}$=新华社 &&$CpString.Word_{Next}$=电, then $CpString.Validity$:=$true$;
e.g.:新华社/北京/电

• if $CpString.Word_{Pre}$=各 &&$CpString.Word_{Next}$=都, then $CpString.Validity$:=$false$;

e.g.:各/大中城市/都要自己负起综合平衡的责任.

· if $CpString.Word_{i-1}$=邮电 && $CpString.Word_{i+1}$=行列,then $CpString.Validity:=false$;

e.g.:进入全国邮电/百强县/行列.

3 Experimental Results and Discussions

As shown in Fig. 2, the system consists of three modules: the module of initial recognition of place names based on statistics, the module of recording the contexts of place names, and the module of confirming and eliminating place name candidates by applying rules. The system combines the method based on statistics and the idea of transformation, which not only keeps the high recall, but makes the system have the ability of automatic error correcting and improves the precision.

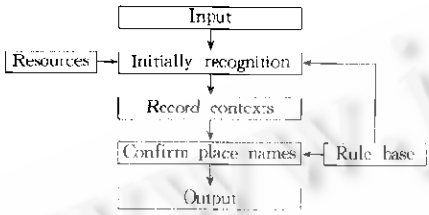

Fig. 2 The system of automatic recognition of Chinese place names

The system is tested on a real news corpus from the People Daily, which contains 50 000 characters and 514 Chinese place names. After the corpus is tagged by initial recognition module, 599 place names are recognized, among which correct place names are 501, boundary error place names are 9, and entirely incorrect place names are 89. Meanwhile, there are 22 place names cannot be recognized. At this time, the precision is 83.6%. Then, we apply the transformation rules to process these place name candidates, and 49 incorrectly recognized place names are removed. Finally, the recall and precision obtained are 97% and 90.92% respectively. The precision has been improved by 7%.

Some examples, in which the character sequences are recognized as place name candidates but are eliminated successfully in the further processing, are given in the following:

· 为贫困地区铺/富路/.

· 各/大中城市/都要自己负起综合平衡的责任.

· 进入全国邮电/百强县/行列.

· 为海峡两岸企业架/金桥/.

Some errors cannot be overcome by the system. For example, the character sequence "大兴" in the sentence "从民政部到各省市区民政厅局,/大兴/调查研究之风" is recognized as a place name incorrectly. This kind of errors cannot be removed because of the simplicity of the rule types, which are confined to the neighboring word form rules. We will develop rules with POS information, syntax information and semantic knowledge to solve the problem. In the future, we will increase the size of training corpus, make the system have the ability of self-learning and self-adjustability and develop rules with semantic knowledge to improve the performance of the system.

References:

[1] Tan, Hong-ye, Zheng, Jia-heng, Liu, Kai ying. Research on the method of automatic recognition of Chinese place names. In: Huang, Chang ning, Dong, Zhen-dong, eds. Proceedings of Computational Linguistics. Beijing: Tsinghua Publishing Company, 1999. 174~179.

[2] National Place Names Committee of China. Chinese Place Name Set. Beijing: Society Publishing Company of China, 1994 (in Chinese).

[3] Eric, Brill. Transformation-Based error-driven learning and natural language processing: a case study in part-of-speech tagging. Computational Linguistics, 1995,21(4):118~133.

附中文参考文献:

[2] 中国地名委员会. 中国地名录. 北京:中国社科出版社,1994.

基于变换的中国地名自动识别研究

谭红叶, 郑家恒, 刘开瑛

(山西大学 计算机科学系,山西 太原 030006)

摘要: 专有名词中的中国地名的自动识别是中文信息处理中要解决的一个重要问题.完全依靠统计方法只能保证一定的召回率,而准确率偏低.提出了一种基于变换的策略,可以有效地克服这一缺陷.经测试,系统最终的准确率提高了7%,达到了90.9%.

关键词: 中国地名;自动识别;变换;专有名词;中文信息处理

中图法分类号: TP391 **文献标识码**: A