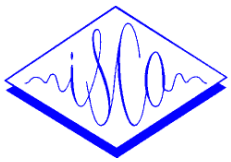

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

CHINESE PERSON NAME IDENTIFICATION BASED ON RULES AND STATISTICS

Wenjie CAO,

National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing
caowj@nlpr.ia.ac.cn

Chengqing ZONG

National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing
cqzong@nlpr.ia.ac.cn

*Juha Iso-Sipilä**

*Nokia China R&D Center, Beijing
juha.iso-sipila@nokia.com

Bo XU

National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing
xubo@nlpr.ia.ac.cn

ABSTRACT

This paper describes our strategies for automatic identification of Chinese person names in text. In our approach, we use bound words, bound rules and linguistic information, including parts of speech, dependency between words, etc., to represent the external context features of names. Bound rules are trained by real corpus. Based on one million Chinese person names, we have developed a probability model to represent the internal features of Chinese names. In the identification process, firstly, a potential Chinese person name is extracted by using the rules and characters that can be used as surnames. Secondly, the weight of the potential name is computed with the probability model. The potential names whose weights are below the threshold will be output as the real Chinese person names. Through open test, the precision rate of the system is 83.66%, and the recall rate is 93.50%.

1. INTRODUCTION

Person name identification is an important issue in the natural language processing. In Chinese text, Chinese person name takes only 1-2%, but sometimes the parsing errors of Chinese person names take up over 50% of the total parsing errors [4]. Besides, compared to those of most western languages, Chinese names are more flexible, and person name identification in text has less information (blank between words, capital of the first letter in names). A character in a person name, together with the neighboring character, can also constitute a word, which may introduce difficulties for the following text analysis. All these bring more challenges for researchers.

Contemporary researches on Chinese person name identification are mainly based on rules and statistics, and researchers have done many useful tries in these aspects. In reference [4], the researchers utilized the results of automatic parsing, and made the job of extracting the Chinese names a part of post-processing of automatic parsing. In the paper, the authors used the frequencies of tetrasyllabic strings, the distances between potential last names and the nearest titles, and also the bound information to find the Chinese person names in texts. Reference [5] classified the left and right bound vocabulary into 3 grades, or 4 grades, including the non-bound words. Potential names with different grade bound words will be given different threshold to be confirmed as a person names. Also, the paper gives more specific rules to check the bound words. In [1], the authors brought forward an INF (Inverse Name Frequency) model. The model is different from the traditional ones, in that in the model, probability distribution of non-names is somewhat like normal distribution, which is more popular in classification system.

To ensure the high precision of a person name identification system, Researchers should take efforts to find out and represent more information about the external and internal features of Chinese person names, then improve the accuracy rate, and ultimately improve the understanding accuracy rate. In this paper, we want to make some contributions to this target. Much information has been utilized in the identification process, including the bound information, formation information of Chinese person names, statistic information of arrangement of bound words and person names in a whole text, and statistic information of surnames and given names of Chinese person names based on 1-million person name corpus.

In the second part of this paper, firstly, we give some results on an elaborate study on one-million-person-name corpus; and then, we present our statistics, rules and strategies towards the identification of Chinese person names. In the third part, the paper describes the experiment results. The fourth part is the conclusion.

2. IDENTIFICATION OF CHINESE PERSON NAMES

2.1 Person Name Corpus Analysis

We have collected one million person names, in which about 99.8% are Chinese person names. Entities of the corpus were mainly extracted from many newspapers, and some were inputted manually. We have analyzed the composition of the corpus. Also we got the frequencies of surnames, monosyllabic given names, the first and the second characters of disyllabic given names, and the bi-grams of disyllabic given names. Probability information above has been used in our Chinese person name identification system. Totally, we found out 974 surnames from the name corpus. 4,064 characters are involved in the corpus. In the surname list, 23 are compound, and others are monosyllabic. Before the identification process, we removed some of the characters from the surname list that rarely used as surnames in texts.

2.2 Identification of Chinese Person Names

Through the analysis of real text corpus and the name corpus, we've found the following points

- Predominant number of Chinese person names, especially in the newspapers, could be marked by their left and right bound words;
- The characters and their distribution of Chinese names are different from other strings in a text, and for disyllabic given names, the two characters may not be independent of each other;
- Bound information and the name's internal features together may indicate whether a string is a Chinese person name;
- A person name often appears more than one time in an article.

Our system of identifying Chinese person names is based on these clues. Approaches to the identification of Chinese person names can be described as following.

We use the characters that can be surnames as the entrance criteria of the name identification system. Then a bound checking module gets into operation. In the module, bound words and bound rules are used to select potential Chinese person names. For those strings that have clear left bound and right bound, the module directly returns the syllable length of the strings. But sometimes, the bound information is confusing: two bound pairs (left and right bounds) can be passed by the checking of bound rules. It indicate that the string may be a disyllabic or trisyllabic Chinese person name. In this case, the module returns another nonzero numeral. If the bound checking process indicates that a string is a person name, the module returns zero.

After a potential Chinese person name is obtained through bound checking, a probability model is used to check if the weight of the potential Chinese person name below the given threshold. If returned value by bound checking process is 2 or 3, we just compute W_2 or W_3. If the result is smaller than the given threshold, the potential name, together with its bound words, could be confirmed and be added to a name list. If the returned value is another nonzero numeral, we need to compute both w2 and w3, and select the lower one, to see if it is lower than the threshold. If so, then it is confirmed as a Chinese person name, and will be added to the name list together with the bound words.

2.3 Acquisition of Bound Words & Bound Rules

Till now, we have collected 1580 left bound words and 2090 right bound words. Some of the left and right bound words were collected manually from dictionaries, and some were acquired from the training data. The training data are articles from the People's Daily of 1996, Contemporary China, and some other newspapers (the size is 8.2 Mb). The Chinese person names of the training data were identified by the bound words that we've collected, statistic results of the one-million name corpus, and rules established by linguistic information. During the identification process, the left bound words and the right bound words of person names were labeled and extracted. Then, we classified the words according to the extracted bound words, the semantics and the parts of speech of the bound words, and manually labeled the bound words. Like reference [4], the classified words have different grade levels.

Rules for person name identification are mainly trained by the bound pairs (the left bound words and the right bound words) of the names that we extracted from the training data, including the correct ones and the wrong ones. They regulate that which kind of left bound words, together with which kind of right bound words, could indicate that the string between the two bound words may be or may not be a person name. To avoid data sparse problem, we represented the rules after training by the two-tuples of the classes of the bound words. Because of the limitation of the size of the training data, the extracted bound pairs cannot comprehensively cover the cases. So we add some linguistic information to the rule base to make up for lack of training data. The linguistic information includes the part of speech of a word, name-building information, and dependency information between words.

Rules described above can be seen as general rules. But it is not enough, because there are also some specific cases. For example, there are some settled arrangements. For example,

1. "任命赵尔陆为部长"
2. "以李四光为代表"
3. "以江泽民同志"

In these cases, it's unwise to use classes to represent these words. We should record the patterns (任命,为), (以,为), (以,[TITLE]) themselves. We can see that in these patterns, the bound pair may use the bound words directly, or be represented by the combination of bound word and class.

For another example: a string "[str1]、[str2]、…", maybe all the strings are Chinese person names, but they also could be other proper noun, or even phrases. So, we need to specially design rule for such a pattern.

2.4 Probability Model of Chinese Person Names

To check if a string is a Chinese person name, we give a probability measurement. Suppose the surnames and given names are statistically independent, the name model

$\mathrm{str}=\mathrm{s}_1\mathrm{s}_2$, $W_2(str)=-LOG(Ps(s_1)P_1(s_2))+C_2<\text{Threshold}$ (1)

$\mathrm{str}=\mathrm{s}_1\mathrm{s}_2\mathrm{s}_3$, $W_3(str)=-LOG(Ps(s_1)P_2(s_2,s_3))+C_3<\text{Threshold}$ (2)

Where, s1, s2 and s3 are syllables, $\mathrm{str}=\mathrm{s}_1\mathrm{s}_2(\mathrm{s}_3)$ is a syllable sequence in a text;

$\mathrm{W}_2(\mathrm{str})$ or $\mathrm{W}_3(\mathrm{str})$ is the weight of str when it is disyllabic or tribyllabic;

$Ps(s_1)$ is the probability of s_1 (compound surnames temporarily not concerned) as a surname, it equals to the ratio of its frequency as a surname to its frequency as a monosyllabic word in real text corpus;

P1(s2) (P2(s2,s3)) is the probability of s2 (s2s3) as monosyllabic (disyllabic) given name. Here, P2(s2,s3) used bi-gram. It equals to the probability of "s2s3" appearing as given name in the name corpus;

C2 and C3 are constants. C2 or C3 is the negative common logarithm of the probability that disyllabic or trisyllabic person names takes up in the name corpus. The ratio of the frequency of disyllabic names to that of trisyllabic names is 1:8.82, estimated from the name corpus. These constants are used to make the criterion of the weights of disyllabic and trisyllabic names be consistent.

This model is used after potential Chinese person names have been selected by the rules. Then, we use this model to compute the weight of each potential Chinese person name, and put those whose weights are within a certain threshold into a name list as real Chinese person names.

Yet, there exists data sparse problem when we cannot find s1 or s2 or s3 or s2s3 in the corresponding statistic data. In these cases, if we can't find s1 in the surnames, or we can't find s2 in the monosyllabic given names of the name corpus, or we can't find s2 or s3 in the first or second characters of disyllabic given names of the name corpus, we need to take some measures. Since the size of the name corpus is quite large, and the characters that appear in the Chinese person names of real corpus, but don't appearing in the name corpus may be very few, we just omit the influence of these characters to the size of the total name corpus, and use the reciprocal of the total number of corresponding statistic data as the probability of these characters. If s1, s2 and s3 all appear in the corresponding statistic data, but s2s3 doesn't appears in the bi-grams of disyllabic given names, we just look on the two characters as independent. The name model of trisyllabic names then changed to following:

$\mathrm{str}=\mathrm{s}_1\mathrm{s}_2\mathrm{s}_3$,

$W_3(str)=-0.5LOG(Ps(s_1)P(s_2)P(s_3))+C_3<\text{Threshold}$ (2*)

Where, P(s2) (P(s3)) is the probability of s2 (s3) in the first (second) characters of disyllabic given names.

2.5 Some Consideration in the Identification System

2.5.1 Filter of special surnames

Some surnames can also be quantifiers, so before bound checking, we firstly check if there are numerals or "多", "来" , "余"… in the left of such words. Only if there are no such words in the left, could we go on to bound checking process. In addition, "多", "来", "余", "万", "千" can be surnames. However, they sometimes used as numerals or between the numerals and the quantifiers. So the principle above is also effective to these characters. Totally, 58 such surnames have been labeled in the surname list.

2.5.2 Loosing of bound rules

Sometimes, it is difficult to find the bound words of Chinese person names in the real corpus. To improve the recall rate, we loose the bound rules for one side of the potential names. That is, at least one side bound words of the potential names should be in the corresponding bound word list. This will For example,

"２０多年来，汪樟宝一辆自行车、…调解纠纷。"
(People' Daily, 1996)

In the above sentence, "汪樟宝" is a Chinese person name, and "，" is a left bound words of names. But "一辆" is not a right bound word of names. After loosing the bound rules, the name will be included in the name list. It is not difficult for us to conceive that there are so many such cases in real text. So the measure is quite effective to improve the recall rate of the system. Yet, the precision rate of the system will inevitably be decreased. In case the accuracy will be decreased greatly, we decrease the threshold for such cases.

2.5.3 Statistic methods in the identification

Considering each name may appear many times in an article, we compute the frequency of each name when it is added to the name list. For each potential Chinese person name, we check its frequency as an extracted name. If its frequency has reached to a certain value, we directly extracted it as a name, and noted its left and right bound words. Thus, on one hand, we increase the efficiency. On the other hand, we can also increase the recall rate for some real names that otherwise maybe cannot be passed by the bound checking rules.

The method can also increase the accuracy. After the identification process has been completed, we scan the extracted name list once more, and delete the repeated names that have the same left bound or right bound words in each extraction. To ensure the recall rate first, we request that repeated times is no less than 3. We consider the "names" may be parts of transliterate names, or parts of other proper nouns. Most such cases have been eliminated from the name list after the process. For example,

The string "尚志国有林场公司" (the People' Daily, 1996) appears no less then 10 times in a piece of report in People' Daily, 1996. Almost all of them can be extracted "尚志国" as a

Chinese person name. Each time, the left bound words of "尚志国" may be different. But the right bound words "有" are the same without exception. So, after the scan, the algorithm will consider that "尚志国" is a part of a certain proper noun, and will delete all of it in the name list.

2.6 The Flow Chart of the Name Identification System

According to the depiction above, we can describe the flow of the system. When we got the entire potential name list, we filter the list with the special surnames as described in 2.5.1. And then, if the potential names have already enough iterations in the real name list, for example: 4 times, we put the potential names directly into the real name list; If not, we use bound rules and linguistic information to check each string in the potential name list, and use the probability model to compute the selected ones. If the weight is lower than the given threshold, the string and its bound words will be put into the real name list. Re-scan the real name list, and eliminate those may be a part of other proper nouns, we will get the ultimate name list.

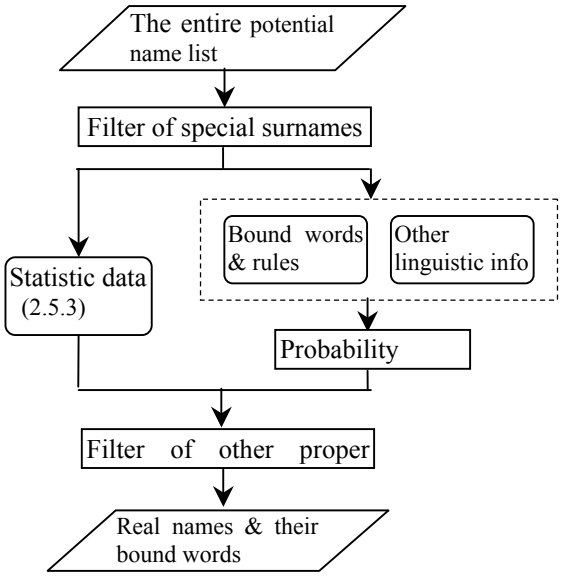

Figure 1. Flow chart of the name identification system

3. EXPERIMENTAL RESULTS & CONCLUSION

Results

We extract some articles from People' Daily 1996. This part of corpus contains 1,322 Chinese person names. The system identified 1,460 "Chinese person names", in which 1,266 are correct names. The precision rate of the system is: 1266/1460=84.66%, and the recall rate of the system is: 1266/1322=93.75%.

Conclusion

In this paper, we present approaches for identification of Chinese person names. It is used before segmentation and POS tagging. It will be more convenient to be used as the post-processing of segmentation by slightly changing the expression of the classes of bound words. We use the combination of statistics, bi-part bound rules, and other linguistic knowledge hoping to improve the performance of the system. We think the not high enough Precision/Recall rate is mainly because of lacking of tagged text corpus as the training data.

For the future, we plan to improve the performance system in three aspects: firstly, collect more labeled text corpus and retrain the bound rules. If it is necessary, we will change the definition of classes of bound words according to the statistic results and the weights of the Chinese person names. Secondly, we may find some new models to represent Chinese person names combining some bound information. At last, we hope to introduce more linguistic knowledge.

4. ACKNOWLEDGEMENTS

The research work described in this paper is supported by the National Nature Science Foundation of China under grant number 9835003 and 60175012. It is also supported by the National Key Basic Research Program of China under grant number G1998030504, and Europe LC-Star project under grant number IST-2001-32216.

5. REFERENCES

[1] Heng JI, Zhensheng LUO. "Inverse Name Frequency Model and Rules Based Chinese Name Identifying", The 6th Countrywide Computational Linguistics Unite Academic Conference (Written in Chinese)

[2] Maosong SUN, Changning HUANG, Haiyan Gao, Jie FANG "Automatic Identification of Chinese person name", Journal of Chinese Information Processing Vol.9, No.2

[3] Hsin-His Chen, Yung-Wei Ding, Shih-Chung Tsai, Guo-Wei Bian. "Description of the NTU System Used for MET2",Proceedings of First International Conference on Language Resources and Evaluation, Granada, Spain, 1998

[4] Xing WANG, Degen HUANG, Yuansheng YANG, "Identifying Chinese Names based on Combination of Statistics and Rules", JSCL-99.

[5] Bingwei LIU, Xuanjing HUANG, Yikun GUO, Lide WU, "Statistical Chinese Person Names Identification", Journal of Chinese Information Processing Vol.14 No.3.

[6] Chinese Academy of Social Sciences. "Analysis and Statistics of characters in surnames' and given names' usage", Yuwen Press (1990, Written in Chinese).

[7] Xiaohe CHEN, Automatic Analysis of Contemporary Chinese Using Visual C++, Beijing Language and Culture University Press, 2000.3.

[8] Jiahen ZHENG, Xin LI, Hongye TAN, "The Research of Chinese Names Recognition Method Based on Corpus", Journal of Chinese Information Processing Vol.14 No.1.