

## Personal Name Recognition Based on Categorized Linguistic Knowledge

Weiguang Qu  
*School of Mathematics and  
Computer Science,  
Nanjing Normal University,  
Nanjing, Jiangsu, P.R. China*  
wgqu@njnu.edu.cn

Xuri Tang  
*School of Chinese Language  
and Literature,  
Nanjing Normal University,  
Nanjing, Jiangsu, P.R. China*  
xrtang@yahoo.cn

Bin Li  
*School of Chinese Language  
and Literature,  
Nanjing Normal University,  
Nanjing, Jiangsu, P.R. China*  
gothere@126.com

### Abstract

*This paper proposes an integrated approach for personal name recognition (PNR) in Chinese by utilizing both statistical language models and categorized linguistic knowledge. Various formulas are proposed for calculating personal name credibility and context credibility for different types of personal names. Experiment is conducted on large-scale corpus to evaluate the approach and the F-1 scores has reached 98.85% and 92.73% respectively in close and open test.*

### 1. Introduction

Recognition of OOV (Out of Vocabulary) words is a major challenge in Chinese morphology. In Roman languages, there are clear boundaries. However, in Chinese, characters in OOV words are often capable of combining with its adjacent characters to form legal words, which makes this problem even harder to deal with.

Among all OOV words, personal names occupy a large proportion. According to statistics, in January 1998's "People's Daily", the Chinese personal names take the share of 48.6% in OOV words[1], while the number of foreign personal names is 10.5%, about one-fifth of that of Chinese personal names[2]. Therefore, it is necessary to address this problem more vigorously.

There are currently three major approaches to the problem of personal name recognition: rules-based approaches[3], statistic approaches[4], and approaches integrating both rules and statistics[5]. In rule-based approach, the recognition process is firstly triggered by characters used for family names, and the detection of boundary is determined by elements adjacent to the family name character. This method enjoys a high rate of precision, but the rate of recall is not satisfactory. The statistical method makes use of corpus to obtain

the probability or credibility that a character string functions as personal name and the probability that a word function as a adjacent context to personal name. When the probability goes beyond a threshold, the string is recognized as a personal name. The probability of context is crucial for the recognition of those character strings which do not enjoy a high probability for names. The integrated approach is used in the hope to combine the advantages of both the rule-based approach and the statistics-based approach to enhance the performance of personal name recognition (PNR). Statistical methods build language model for PNR, which can reduce the complexity in manual rule establishment; rule-based approach can give full play to the existing knowledge of language, and reduce dependence on the scale of the corpus. The mainstream of current research is the integrated approach, which focused on how to establish a highly efficient statistical language model and on how to make full use of linguistic rules and linguistic knowledge to further improve system performance.

After several decades of development, great progress has been made in Chinese PNR. However, deficiencies still exist, as follows:

1. Lack of distinction when applying statistical models to different kinds of personal name structure. Japanese names and names from Europe and US are not identical in structure; names from Russia and Eastern Europe have a structure which is different from other countries; even names from Russia and other eastern European countries are different from each other. In Chinese, Han names, the names from Xinjiang and Tibetan names have their own characteristics. A uniform model based on statistics will not be able to make use of these characteristics.

2. Relevant linguistic knowledge database is not fully used. Statistical models and the role of language rules are often limited; the corpus size is also limited. Statistical models based on corpus are seriously

constrained by the size of corpus. Those linguistic phenomena which do not occur in corpus can not be captured in a statistical model. The use of existing knowledge can compensate for the lack of corpus and solve the problems that statistical models can not solve.

To address the above problems, we propose a series of solutions: statistic analysis is carried out on the structures of different types of personal names; different personal name knowledge database are built to compensate for the deficiency of corpus scale and to reduce its impact on language model. We carry out experiments on large-scale corpus, the result of which verifies the effectiveness of the methods.

## 2. Corpus-based inspection and analysis

PNR is often based on word segmentation of text. The fact that errors exist in word segmentation makes personal name recognition more difficult. Several factors contribute to this difficulty: (1) diversity of personal names; (2) embedded words inside personal names; (3) possible combination of characters of personal names and their adjacent characters.

Based on analysis of corpus, we obtain a table of attributes for Han personal names (Table 1) and a table of attributes for non-Han personal names (Table 2).

**Table 1. Han Name Feature List**

Single-Char or Dbl-char FN <sup>1</sup>	姚, 王, 于, 皇甫, 东方
Single-char GN	颖, 伟, 昊, 静, 之, 也
Beginning Char in Dbl-char GN	晓, 淑, 伟, 志, 亦
Ending Character in Dbl-char GN	华, 东, 花, 伟, 之
Word or word sequence with Single-char FN and GN	高峰, 汪洋, 白雪, 万/里
Word or word sequence with Single-char FN and Dbl-char GN	黄河/燕, 白天/明, 周长/喜, 金元/东, 陈列/雄, 盛世/才
Word or word sequence with dbl-char GN	陈/凯歌, 潘/长江, 魏/传统, 王/向/东, 徐/从/善, 张/长/路
Word with single-char GN or ending char in dbl-char GN and right adjacent character	龚学/平等 张/勇为, 小/平和, 褚时/健在, 沈国/放在
Word with single-char GN or ending char in dbl-char GN and right adjacent characters	张太/雷同/志 张学/良将/军, 宝/丹青/年, 沈国/放大/使
Word with Single-char and ending char in left-side dbl-char word	局/长沙/万里, 市/长孙/长林

**Table 2. Foreign Name Feature List**

Beginning char in foreign name	克, 施
Middle char in foreign name	德, 利
Ending char in foreign name	特, 斯, 夫
Single-char foreign name	琼, 金, 简, 乔

<sup>1</sup> Some acronyms are used in this paper. Here is the list for reference. FN: family name; GN: given name; Single-char: single character; Dbl-char: double character;

Foreign name with embedded word	马达/维基亚, 巴/班吉/达
Foreign name with ending char boundary ambiguity	奥/尔/布/赖/特等, 奥/尔布赖/特出/使

In the process of identifying personal names, the credibility of the string and the credibility of its context features are equally important. For example, under all circumstances “姚淑华” can only be a personal name, because “姚” is a character used only for family name and “淑华” is never used for purposes other than given name. However, some personal names are commonly used Chinese characters. For example, “于” can be used as a family name and can also be used as a preposition, meaning “in”. Further more, the frequency of the character used as preposition is far greater than the frequency as a family name. Some of the characters used as personal names are also words commonly used in the language. For this reason, the significance of context gets more salient for PNR. For example, “于广州” has different meanings in different context:

A 刘叔叔 1927 年生于广州。

B 商务部副部长于广州同志到黑龙江考察。

In Example A, “于广州” means “in Guangzhou province”. In Example B, it is a personal name, referring to a person who has that name. In this case, the character string “于广州” does not have a high credibility. However, in Example B, the context of the “副部长 (Deputy Minister)” and “同志 (comrade)” play a strong role to ensure that “于广州” in this context is a personal name. This shows that PNR requires not only consideration of the credibility of the character string itself, but also attention to the credibility of the context. Performance can only be improved by a comprehensive consideration of both the credibility of the target string and the credibility of context.

## 3. Approach to PNR

Based on analysis of corpus, we conducted research focused on those existing problems mentioned above.

### 3.1. Credibility Calculation of Personal Names

The calculation of the credibility of personal names is essential for recognition. A reliable method for credibility calculation can provide more accurate information for PNR, thereby improving system performance. We have conducted calculation of personal name credibility on two types separately: Han personal names and foreign personal names.

#### 3.1.1. Calculation of Han Names Credibility.

The credibility of Han names consists of two parts: family name credibility and given name credibility.

### 1. Family Name credibility ( $C_{hx}$ )

The formula used for calculating a character N which can be used as family name is given below:

$$C_{hx}(N) = \frac{\text{Count\_FN}(N)}{\text{Count}(N)}$$

Count\_FN(N): Frequency of N as family name in corpus; Count(N): Frequency of N in corpus.

### 2. One-character Han Given Name Credibility ( $C_{hd}$ )

Given a word N, its credibility can be calculated using the following formula:

$$C_{hd}(N) = \begin{cases} \frac{\text{Count\_GN}(N)}{\text{Count}(N)} & \text{If } N \text{ in corpus} \\ 0.2 & \text{Otherwise} \end{cases}$$

Count\_GN(N): Frequency of N as given name in corpus; Count(N): Frequency of N in corpus.

It is noted that for single-character given name, a new word is often used, for example “玓”. This observation provides motivation that a word which does not appear in corpus has a credibility of 0.2.

### 3. Double-character Han Given Name Credibility

Given a double-character given name MN, credibility is calculated on the condition whether MN appears in the corpus or not. If MN appears in corpus, its credibility  $C_{hs1}$  is calculated using the following formula:

$$C_{hs1}(MN) = \frac{\text{Frequency of } MN \text{ as real name}}{\text{Frequency of } MN \text{ in corpus}}$$

If MN is never used in the corpus as a given name, calculation is conducted using all double-character given name strings. A set of double-character given name string is formed, and calculation is done on characters in the set. Firstly, a maximum left character frequency (max\_freq\_left) and a maximum right character frequency (max\_freq\_right) are obtained from the set. Thus the left character M has a credibility  $C_{hs_l}$ :

$$C_{hs_l}(M) = \frac{\text{Count\_left\_GN}(M)}{\text{Max\_freq\_left}}$$

where Count\_left\_GN(M) is the frequency of M as left character in given name. The right character N has a credibility  $C_{hs_r}$ :

$$C_{hs_r}(N) = \frac{\text{Count\_right\_GN}(N)}{\text{Max\_freq\_right}}$$

where Count\_right\_GN(N) is the frequency of N as right character in given name.

Therefore, the credibility  $C_{hs2}$  of MN is calculated using the following formula:

$$C_{hs2} = \frac{(C_{hs_l}(M) + C_{hs_r}(N))}{2}$$

### 3.1.2. Calculation of Foreign Name Credibility

Given a character string  $M_1M_2\dots M_N$ , and credibility  $C_i$  for  $M_i$  in the string, the credibility of the string can be calculated using the following formula:

$$C_{NM} = \frac{1}{N} \sum_{i=1}^N C_i$$

Foreign Name set can be formed from training corpus. Then we can get the beginning character list (BEG\_list), ending character list (END\_list), middle character list (MID\_list). With these list,  $C_i$  can be calculated.

For  $i=1$  or  $i=N$ ,

$$C_i = \begin{cases} 2 * \frac{\text{Count\_B}\{E\}(M_i)}{\text{Maximum\_Count\_B}\{E\}} & \text{if } M_i \in \text{BEG\_list or END\_list} \\ -3 & \text{Otherwise} \end{cases}$$

otherwise,

$$C_i = \begin{cases} \frac{\text{Count\_M}(M_i)}{\text{Maximum\_Count\_M}} & \text{if } M_i \in \text{MID\_list} \\ -1 & \text{Otherwise} \end{cases}$$

Where Count\_B{E,M}(M<sub>i</sub>) is the frequency of M<sub>i</sub> as beginning{ending,middle} character.

Maximum\_Count\_B{E,M} is the maximum frequency of beginning{ending, middle} characters. From the formula, it can be seen that during the calculation of foreign name credibility, the beginning word and the ending word play a very important role. It should be noted that, because of the introduction of coefficient of beginning and ending character, the value of credibility may be larger than 1.

### 3.2. Credibility calculation of context

RFR (Relative Frequency Ratio) is a parameter proposed for calculation of collocation, which has proved to be effective in collocation extraction and sense disambiguation[6]. In PNR, we have adapted it in the following ways. Given a personal name A and its context:  $W_{-1} A W_1$

Where,  $W_{-1}$  is the first word to the left of A,  $W_1$  is the first word to the right of A.

Given a sentence with personal names, the frequencies of the left adjacent word and the right adjacent word is calculated as LocFrq<sub>m</sub>(word), ( $m = -1, 1$ ), the sum of left adjacent and right adjacent words LocTotal<sub>m</sub>, ( $m = -1, 1$ ). Therefore the ratio of each word in the wordlist LocRatio<sub>m</sub>(word) can be obtained as below:

$$\text{LocRatio}_m(\text{word}) = \text{LocFrq}_m(\text{word}) / \text{LocTotal}_m$$

Calculation is conducted on all existing corpus (Named as the corpus bank), and the frequency of each word, known as GlobFrq (word), is obtained. The overall frequency of words in the corpus bank known as GlobTotal is also obtained. Thus the proportion of each word in the wordlist, which is named GlobRatio(word) is calculated using the formula as below:

$$\text{GlobRatio}(\text{word}) = \text{GlobFrq}(\text{word}) / \text{GlobTotal}$$

The relative frequency of a word in the position  $m$ ,  $f_m(\text{word})$  is calculated using the following formula:

$$f_m(\text{word}) = \text{LocRatio}_m(\text{word}) / \text{GlobRatio}(\text{word})$$

Relative Frequency  $f_m$  is a good indication of the mutual attraction between word at position  $m$  and personal name and can be used as Credibility of personal name context.

In order to calculate more accurately the contextual features of different types of personal names, the following situations are considered:

- (1). Han Family Name + Han Given name (nrf + nrg): Full Han name;
- (2). Han Family Name (nrf): Short form consisting of only Han family name;
- (3). Han Given name (nrg): Short form consisting only of Han given name;
- (4). Other types (nr): foreign names, names of Xinjiang, Tibetan name, and so on.

Table 3 gives the credibility of the right context  $f_m$ . It can be seen that these words have different credibility in different types of personal names. And the advantage of such type of differentiation makes it possible to match selectively with the type of personal names appear before them and improve recognition results.

**Table 3. Right Context Credibility of Some Words**

	nrf+nrg	nrf	Nrg	nr
总理	11.70	232.65		14.26
小姐	2.17	287.75		50.13
外相				317.49
同志	47.33	11.23	71.33	1.74

For example, in the two strings "韩/外相 (Korea / Foreign Secretary)" and "韩/小姐(Han/ Miss)", "韩" is a family name, which can also be used as a short name for Korea. Using the method mentioned above, context credibility can be used to distinguish these two different situations. The word "外相 (Foreign Minister)" requires a family name of a country in its left side, but not a personal name. The word "小姐 (Miss)" requires its left side to be a foreign name, or a full Han name, or Han family name. Credibility tells us that it expects more a word which is a single-character

family name. Thus the word "韩" can be easily recognized as a single-character family name.

Besides, we collect and construct a list of the knowledge databases (Table 4).

**Table 4. Knowledge Databases and their scale**

Types of Knowledge database	Item scale
Japanese surname list	8000
Xinjiang personal name list	2000
Tibetan personal name list	500
Foreign personal name list	17000

Statistical analysis and application of knowledge obtained though this process can improve coverage of the test corpus.

In an article, the same person is often repeatedly mentioned. The full name is often used to introduce a person, and then short forms like given name or family name are used. In such circumstances, if we can identify correctly the name when it firstly appears, the remaining short forms will be easy to identify. The approach adopted here is to treat a discourse as a unit of personal name recognition, and save all the full names recognized in the discourse into a name list. And then re-scan the text and tag all the strings which are identical to strings in the name list as personal names.

## 4. Experiment and result

In order to test the effectiveness of the methods proposed above, an experiment is conducted on large scale corpus. The corpora which is built from text of People's Daily published during January, 1998 is selected for experiment. Texts spanning from the 1st to 25th are used as training corpora and texts spanning from 26th to 31st as test corpora<sup>2</sup>.

Experiment consists of two parts: training and test. Details can be described as below:

1. Using training corpora, we obtain context credibility list for nrf+nrg, nrf, nrg, and nr. At the same time we construct wordlist (word\_list) that does not contain personal names, and bi-gram wordlist (bi\_word\_list) that neither contains personal names;
2. Training to obtain Han name credibility;
3. Using training corpora and knowledge database, credibility of personal names of Xinjiang, Tibetan names, Japanese names, and names of European and US are obtained;
4. Using training corpora to obtain relationship between personal name credibility and context credibility. For example, if the personal name

<sup>2</sup> The corpora can be downloaded from website <http://icl.pku.edu.cn>.

credibility is high, its reliance on context is lower. At the same time, other factors such as whether the string is contained in word\_list, whether the string and its context is contained in bi\_word\_list etc. 26 rules are obtained in this step. The following are two of them:

rule1:

$C_{hx}(X) > 0.1 \ \&\& \ C_{hsi}(MN) > 0.1$ , XMN is personal name

rule2:

$C_{hx}(X) > 0.2 \ \&\& \ C_{hd}(M) > 0.3$  XM is personal name

Using the above two rules, 73.4% of all the personal names are recognized and the error rate is only 0.26%.

Recognition is straight forward on the basis of above training:

1. Recognition is conducted on the unit of discourse;
2. For a word in a sentence, credibility of personal name and contextual credibility above mentioned is calculated;
3. Apply the rules for PNR;
4. Re-process text using the names obtained in the above steps to enhance the recall.

Three parameters: Precision (P), Recall (R) and F-1 measure, are used in the experimental evaluation, which are very common in NLP evaluation.

Using the training and test corpus mentioned above, the result of the experiment is in table 5:

**Table 5. Experiment Result**

	Corpus date	P	R	F-1
Close	98.1.1-1.25	99.60%	98.13%	98.85%
Open	98.1.26-1.31	94.08%	91.42%	92.73%

In [2] and [7], the same training and test corpus is used. Their F-1 scores for close test and open test are 97.30%/84.53% and 98.09%/84.47% respectively. Compared to the two results in [2] and [7], our result does not enjoy a good advantage in the close test, but the advantage is obvious in open test. One of the reasons that contribute to this improvement is that various types of knowledge database play a major role in personal name recognition. In the open test, personal names such as “宫川、谷内、井坂、武藤” does not appear in the training corpora, but are included in the dictionary we built for Japanese names. As a result, we are able to obtain a 8% higher F-1 in the open test.

## Conclusions

In this paper, we proposed formula for calculating personal name credibility and context credibility for various types of personal names and utilize knowledge database and discourse information to improve system performance. We use large-scale corpus for experiment to verify the validity of the method.

In the future, we plan to build more detailed and larger personal knowledge database to further enhance

the accuracy of statistics so as to further improve system performance. We also plan to build system that handles personal names, toponyms and organization names simultaneously, so that the ambiguity existing between them are considered, and system performance can be improved.

## Acknowledgements

This work is funded by the Chinese National Fund of Social Science under Grant 60773173, Chinese National 973 Program under Grant 2004CB318102, and Jiangsu Social Science Fund (06JSBYY001) and Chinese National Social Science Fund under grant 07BYY050.

## References

- [1] Zhang Huaping, Liu Qun, Automatic Recognition of Chinese Personal Name Based on Role Tagging. Chinese Journal of Computers. Vol. 27. No. 1, 2004
- [2] Gao Hong, Huang Degen, Yang Yuansheng. Foreign Person Names and Chinese Person Names Recognition in Chinese Texts. Mini\_Micro Systems, Vol. 27 No4, 2006
- [3] Lv Yajuan, Zhao Tiejun et al. Levelled Unknown Chinese Words Resolution by Dynamic Programming, Journal of Chinese Information Processing. Vol.15 No.1 2001
- [4] Zhang Yuejie, Xu Zhiting, Zhang Tao, Fusion of Multiple Feature for Chinese Named Entity Recognition Based on CRF Model. In: Li Hang et al. ed. AIRS 2008, LNCS 4993,2008
- [5] Wang Sheng, Huang Degen, Yang Yuansheng. Chinese Person Name Recognition Based on Mixture of Statistics and Rules. In: Huang Changning, Dong Zhendong ed. Corpora of Computational Linguistics. Beijing. Tsinghua University Press, 1999
- [6] Qu Weiguang, Sui Zhifang, Ji Genlin et al, A Collocation-Based WSD Model: RFR-SUM, in Proceedings of IEA/AIE 2007, LNAI, Vol. 4570, 2007
- [7] Gao Hong, Huang Degen, Yang Yuansheng. A Method of Chinese Personal Names Recognition Synchronized with Chinese Word Segmentation. Computer Engineering, Vol. 32 No19, 2006