Identification of Chinese Personal Names in Unrestricted Texts*

Lawrence CHEUNG, Benjamin K. TSOU
Language Information Sciences Research Ctr
Tat Chee Avenue, Kowloon Tong
Hong Kong
{rlylc, rlbtsou}@cityu.edu.hk

Maosong SUN
National AI Lab.,
Tsinghua University, Beijing,
China
lkc-dcs@tsinghua.edu.cn

Abstract

Automatic identification of Chinese personal names in unrestricted texts is a key task in Chinese word segmentation, and can affect other NLP tasks such as word segmentation and information retrieval, if it is not properly addressed. This paper (1) demonstrates the problems of Chinese personal name identification in some IT applications, (2) analyzes the structure of Chinese personal names, and (3) further presents the relevant processing strategies. The geographical differences of Chinese personal names between Beijing and Hong Kong are highlighted at the end. It shows that variation in names across different Chinese communities constitutes a critical factor in designing Chinese personal name identification algorithm.

Keywords: Chinese personal name identification, Chinese word segmentation, Chinese IT applications, Chinese linguistic differences

1 Introduction

Automatic identification of Chinese personal names in unrestricted texts plays a key role in Chinese NLP tasks such as word segmentation. Personal name identification in Chinese involves many different issues that are not found in similar tasks in English. Despite extensive research in the issue in recent years (Wang et al. 1992, Song et al. 1993, Sun et al. 1995), many NLP applications still suffer from the weakness of current available performance.

The paper has four sections: Section 1 illustrates some IT applications in which Chinese personal name identification plays important roles, to promote an appreciation of the significance of the research. Section 2 introduces the structure of Chinese personal names, and Section 3 discusses the relevant processing strategies. Lastly, we will highlight the significant differences of Chinese personal names between Beijing and Hong Kong in Section 4, showing the added difficulty caused by variations in different Chinese communities.

2 The Role of Automatic Identification of Personal Names in IT Applications

2.1 Information Retrieval (IR)

* This research is supported in part by the National Basic Research Scheme of China, under grant no. G199803050 and by Language Information Sciences Research Centre, City University of Hong Kong.

IR systems, including search engine on WWW, retrieval system of digital library, text mining etc., search and retrieve the requested information from large bodies of data stored in electronic form. With the advent of the Internet, search engine techniques have become most widely influential. However, the effectiveness of current systems is not satisfactory especially for retrieving Chinese web sites/pages. We have examined how personal names interfere with the correct retrieval of relevant results in *Sina*[1], a popular Chinese search engine, and *Google* [Big5 Chinese][2]. For instance, suppose we want to get some information about "中將" (lieutenant general) from WWW, and enter that word as a query to Sina, then some of the "relevant" web pages will be returned, as in (1)—(4):

(1) 陳中將與俄羅斯選手爭奪跆拳道女子67公斤以上級金牌 (*Sina*)
(2) 黃健中將與奧利弗-斯通合作中美合拍《賽金花》 (*Sina*)

(3) 副司令陳中將視導南測中心受訓部隊 (*Google* [Big5 Chinese])
(4) 屆時悉尼奧運會冠軍得主陳中將代表河南隊參賽。 (*Google* [Big5 Chinese])

These webpages are actually irrelevant to "lieutenant general" although the string "中將" is embedded. The string "中將" in the two sentences cannot be analyzed as the word meaning "lieutenant general". In (1) and (2), the character "中" is the given name, and "將" an adverb by itself meaning "will" for referring to the future. In (3) and (4), "中將" forms the two-character given name. The misinterpretation would have been avoided if the search engine could detect the presence of proper names and segment the words accordingly.

2.2 Text-to-Speech Conversion (TTS)

The text-to-speech conversion system accepts textual input and generates the corresponding speech signal. It finds applications in mail reading over the phone, generating spoken prompts in voice response systems, human-machine interface and so on. Lucent Technologies hosts a web site called *Bell Labs Mandarin Text-to-Speech Synthesis*[3] which allows the user to input text and outputs the corresponding speech. We tested sentences containing characters that are pronounced differently when it functions as a surname and as a character with other meaning. Any incorrect pronunciation will suggest that the system fails to differentiate personal names from other content words. Consider (5) first:

(5) 我的老闆 查金泰 不同意他弟弟 查建國 先生的看法。
zha *zha* (*cha*)
"My boss Zha Jin-Tai did not agree to the opinion of his younger brother, Mr Zha Jian-Guo."

In Mandarin, "查" can be pronounced as *cha* (a mono-syllabic word meaning "to check") or *zha* (surname). In (3), both occurrences of "查" serve as a surname. The synthesized speech correctly pronounces the first occurrence as *zha*, but mis-pronounces the second one as *cha*, indicating that the Chinese name "查金泰" is correctly identified whereas "查建國" is not (As can be seen in the Section 3, the identification of the former one is easier that that of the latter). In this case, the language processing module "under-recognizes" Chinese names.

The other case is Chinese name being "over-recognized". Input (6) to *Bell Labs TTS*. The character "華" acquires different tones in different contexts. It is pronounced as *hua1* (a morpheme meaning China) or *hua4* (surname). The character "曾" is another potential trap. It is pronounced as *ceng2* (an adverb meaning "once") or *zeng4* (surname).

[1] *Sina* Search Engine website: http://www.sina.com.cn

[2] *Google* [Big5 Chinese] website : http://www.google.com/intl/zh-TW/

[3] *Bell Lab TTS* website: http://www.bell-labs.com/project/tts/mandarin-gb.html

(6) 華國鋒 曾任 中華人民共和國國務院總理。
hua4 *zeng1* *hua2*
"Hua Guo-Feng is the former premier of the People's Republic of China."

The system correctly differentiates the two occurrences of "華" and pronounces the characters correctly. However, it misinterprets "曾任" as a surname and pronounces "曾" as *zeng1*. Apparently, the conversion system has not been able to identify personal name reliably and made errors in generating the correct speech output.

2.3 Machine Translation (MT) System

MT application is another important NLP application. The rise of the Internet gives rise to a great demand for translating web pages on the fly from a foreign language into the native language. Although MT systems are far from perfect, many free web-based MT systems have already been capable of doing reasonably good translation. We select the popular *Transtar* website[4] in Mainland China for testing. It offers Chinese-to-English Machine Translation. The MT system translates (7a) and (8a) as (7b) and (8b) accordingly.

(7) a. 我 看見鄧小平同江澤民打招呼。
b. "I see that Deng Xiao-Ping greets with Jiang Zemin."
(8) a. 我看見周星馳同張學友打招呼。
b. "I see week star Chi open together study friend greet."

At first glance, (7a) and (8a) are identical in syntactic and semantic structures, except the personal names. However, while (7b) is a good translation, (8b) is completely incomprehensible. An examination of (7b) and (8b) shows that the major difference lies in the system's failure to identify the two personal names in (8). It is likely that the two paramount political figures' names in (7a) are registered in the system's lexicon and the information can be called up in processing. It makes it possible to parse the Chinese string and to generate the English translation. In contrast, the names[5] in (6b) are relatively less well-known in Mainland China and are not in the system’ s lexicon, resulting in the brute-force translation of "周" to "week", "星" to "star", "馳" to "Chi"("馳" is a bound morpheme and the system simply output its romanization of the character), "張" to "open", "學" to "study", "友" to "friend" respectively.

3 Processing Chinese Personal Names: Challenges and Strategy

3.1 Structure of Chinese Personal Names

The basic structure of modern Chinese personal names is largely similar across different communities. The frequent length is 2 or 3 characters. The maximum can be as long as 6 characters. Table 1 shows the possible structures of Chinese personal name. Chinese personal name begins with the surname which can be one-character (as in (a) and (b)) or two-character (as in (c) or (d)). It is followed by the given name which can be one-character (as in (a) and (c)) or two-character (as in (b) or (d)). A unique structure exists in some Chinese communities such as Hong Kong: the name of a married female may be preceded by her husband's surname, as in (e) and (f). Such usage occurs in formal occasions or writing of formal register.

[4] *Transtar* Website: http://www.transtar.com.cn/

[5] They are famous movie star and singer in Hong Kong.

	Full Name	Husband's Surname			Surname			Given Name		Length
		H1	(H2)	+	S1	(S2)	+	G1	(G2)	
a.	李鵬				李			鵬		2
b.	鄧小平				鄧			小	平	3
c.	諸葛亮				諸	葛		亮		3
d.	東方聞櫻				東	方		聞	櫻	4
e.	陳方安生	陳			方			安	生	4
f.	諸葛東方聞櫻	諸	葛		東	方		聞	櫻	6

Table 1. Possible variations of Chinese personal names

3.2 Challenges

Several characteristics make it difficult for computers to recognize Chinese personal names. Unlike capitalization of personal names in English, there is no explicit marking for personal names in Chinese. Other problems include: (1) Chinese texts do not have explicit word boundaries; (2) the character set for surname and given name are strictly a subset of Chinese character set that forms common Chinese words; (3) some characters in personal names may be mono-syllabic words; (4) some multi-syllabic words can be involved in Chinese personal names starting at every possible position. All of these factors introduce a lot of potential ambiguities in Chinese personal name identification.

The situation becomes even more complicated regarding the fact that some polysyllabic common words are also possible in names, e.g. "王朝聞" ("王朝"= dynasty), "馬勝利" ("勝利" = victory) and "嚴肅" ("嚴肅" = serious(ly)). In these cases, the algorithm may easily mistake names as common words and produce irrecoverable errors. This is illustrated in (9) regarding "嚴肅" :

(9) … 同時嚴肅 指出了張毅的錯誤，希望他加以改正。
(Somebody)… meanwhile, seriously pointed out the mistake of Zhang Yi, hoping he could correct it.

Although "嚴肅" is interpreted as a common word meaning "serious", it is possibly a personal name. So another interpretation seems to be reasonable if larger context (perhaps the structure of the whole sentence) is referred to:

(10) … Meanwhile, Yan Su pointed out the mistake of Zhang Yi, hoping he could correct it.

3.3 Basic Strategies

Chinese personal names are so elusive that locating them in running texts usually relies on a combination of techniques and knowledge to achieve satisfactory recall and precision rate. In the following, several identification strategies will be outlined. They are broadly divided into two types, namely, inherent characteristics of personal names and contextual information of texts.

3.3.1 Inherent Characteristics of Personal Names

Probability
Certain combinations of characters are more likely to form a name. Supported by statistics derived from a large scale Chinese personal name database, the identification procedure can guess name candidates by figuring out the probability of any character string being a name.

log(prob(呂欽))= -5.02 >> log(prob(和廣))= -6.20

From the calculation, it is possible to predict that the combination "呂欽" is more likely to be a name than the combination "和廣".

Construction
For example, reduplication of characters can be found in Chinese names such as "媛媛", "強強", "毛毛" and "瀟瀟". The program may alert the identification procedure when reduplication patterns are encountered.

3.3.2 Contextual Information of Texts

Context can often provide important cues to locating Chinese personal names. Title is a good indicator that the characters immediately before or after it is a potential personal name, e.g. 張志偉先生 (Mr.), 王大文博士(Dr.) and 總理朱鎔基(Premier). Frequently used patterns or syntactic structures may also be useful for identifying names, for example,

(11) 以 <name> <title1> 的{title2}
e.g. 以<潘杜泉> { 團長的香港工會代表團 } 昨日訪問北京 …

(12) { {organization} {title} } <name>
e.g. { {國家} {主席} } <江澤民>

National AI Lab at Tsinghua University (Sun et al. 1997) developed a Chinese word segmentation prototype system that integrates all the above strategies for name identification. It achieves 95.0% recall and 87.6% precision in open test for Chinese personal name identification.

4 Geographical Differences of Chinese Personal Names

Generally speaking, different Chinese communities may share similar convention of naming, but we do find divergence in specific personal name structure and choice of characters. In this section, we will compare personal name data drawn from Beijing and Hong Kong. The statistical findings not only bear sociolinguistic interest about divergence of difference Chinese communities but also present essential cues to accurate personal name identification in texts from different Chinese communities. The knowledge derived can be applied to devise better techniques and customize identification algorithms for different Chinese texts.

4.1 Data

Two databases of Chinese personal names are collected from Beijing and Hong Kong. The Beijing database is drawn from a name list of a county in Beijing. The database that contains 125,033 names is representative of names in Mainland China because the county population is composed of migrants from different provinces of China. The Hong Kong database contains 11,358 names. They are student and staff names taken from Registrar's Office, City University of Hong Kong.

4.2 Findings

4.2.1 Distribution of Surnames

The Chinese surname consists of one or two characters. Single-character surnames are predominant, accounting for over 99% in both databases.

	Beijing (%)	HK (%)
Single-Char. Surname	99.92	99.56
Double-Char. Surname	0.08	0.44

Table 2. Distribution of single- and double-character surnames

The frequency of the surnames is ranked. Table 3 shows that the ranking of the top ten surnames in the two databases are quite different. The most striking difference is that the most frequent surname 王 (Wang) in Beijing is ranked as 14th in the Hong Kong.

	Beijing				Hong Kong		
Rank	Surname	%	Cum. %	Rank	Surname	%	Cum. %
1	王	9.12	9.12	1	陳	10.16	10.16
2	張	8.33	17.45	2	黃	6.71	16.87
3	李	7.92	25.37	3	李	5.87	22.74
4	劉	6.47	31.84	4	梁	4.57	27.32
5	陳	3.17	35.01	5	林	4.19	31.51
6	趙	3.15	38.16	6	張	3.64	35.15
7	楊	2.99	41.15	7	劉	3.03	38.19
8	孫	2.03	43.18	8	吳	2.96	41.15
9	馬	1.71	44.89	9	何	2.81	43.96
10	吳	1.62	46.51	10	鄭	2.09	46.05

Table 3. Ten most frequent single-character surnames in Beijing and Hong Kong
(Surnames found on both columns are shaded.)

4.2.2 Distribution of Given Names

There is a divergence between Beijing and Hong Kong in the preference for single- and double-character given names. Single-character names account for 29% of the Beijing database. In contrast, single-character given names are not common in Hong Kong. They cover only 2% of the data.

	Beijing (%)	HK (%)
Single-char.	29.07	2.13
Double-char.	70.93	97.87

Table 4. Distribution of single- and double-character given names

The preference for characters in given names is rather different in the two databases as well. Table 5 and 6 list the most common characters found in given names. When the columns for Beijing and Hong Kong are compared, only one or two characters overlap. Apparently, the two Chinese communities display considerable difference in the choice of characters in naming.

	Beijing				HK		
Rank	G1	%	Cum. %	Rank	G1	%	Cum. %
1	淑	3.18	3.18	1	嘉	3.76	3.76
2	玉	3.08	6.26	2	偉	3.66	7.42
3	秀	2.88	9.14	3	志	3.54	10.97
4	曉	2.57	11.71	4	家	2.78	13.75
5	文	2.28	13.99	5	詠	2.22	15.97
6	建	2.19	16.17	6	慧	2.13	18.10
7	志	1.86	18.03	7	國	1.99	20.09
8	小	1.78	19.81	8	文	1.95	22.04
9	桂	1.65	21.46	9	佩	1.93	23.97
10	春	1.35	22.80	10	麗	1.88	25.85

Table 5. Ten most frequent first characters (G1) of double-character given names (Surnames found on both columns are shaded.)

	Beijing				HK		
Rank	G2	%	Cum. %	Rank	G2	%	Cum. %
1	華	3.63	3.63	1	儀	3.07	3.07
2	英	3.37	7.00	2	華	2.28	5.34
3	蘭	2.10	9.09	3	明	2.23	7.57
4	平	1.91	11.01	4	敏	2.15	9.72
5	珍	1.83	12.84	5	文	2.08	11.80
6	明	1.66	14.50	6	玲	1.88	13.68
7	榮	1.56	16.06	7	珊	1.74	15.42
8	生	1.50	17.55	8	欣	1.58	17.00
9	芳	1.32	18.87	9	輝	1.57	18.57
10	琴	1.25	20.13	10	雯	1.55	20.12

Table 6. Ten most frequent second characters (G2) of double-character given names (Surnames found on both columns are shaded.)

4.2.3 Co-occurrence of Characters in Double-character Names

Our data reveals that some characters tend to co-occur to form a double-character given name. The information is useful for automatic identification. Given G1 = 嘉 and 偉 (the two most common G1 characters from Hong Kong), there is about 30% of chance that the given name is one of the combinations in 1a—1e and 2a—2e (Table 7).

	Name combination	%	Cum. %		Name combination	%	Cum. %
1a	嘉 + 敏	11.2	11.2	2a	偉 + 強	6.4	6.4
1b	嘉 + 儀	6.5	17.7	2b	偉 + 文	5.7	12.1
1c	嘉 + 雯	4.6	22.3	2c	偉 + 雄	5.7	17.7
1d	嘉 + 琦	4.3	26.6	2d	偉 + 傑	5.4	23.2
1e	嘉 + 慧	3.6	30.1	2e	偉 + 明	5.2	28.3

Table 7. Five most frequent combinations given G1 = 嘉 or 偉.

5 Conclusion

This paper discusses the challenges and basic strategies of personal name identification in unrestricted Chinese texts. The reviewed IT applications are still not reliable in processing personal names. The failure proliferates into further errors in other NLP tasks such as speech generation and parsing. The distributional and structural differences between names in Beijing and Hong Kong are also presented. The findings imply that personal name identification systems, especially those that are statistically-based, have to take into account the differences of personal names among various Chinese communities in order to refine the precision and recall.

Acknowledgement

We would like to thank Prof Rou SONG for his kindness in providing us with the Beijing personal name database, and Registrar's Office, City University of Hong Kong, for their contribution to our Hong Kong personal name database.

References

Song, R, H. Zhu, W. Pan and Z. Yin. (1993). Automatic Recognition of Person Names Based on Corpus and Rule Base. *Research and Application of Computational Linguistics*, Beijing Language Institute Press, Beijing, China.

Sun M., C. Huang, H. Gao and J. Fang. (1995). Identifying Chinese Names in Unrestricted Texts. *Journal of Chinese Information Processing*, 9 (2), Beijing, China.

Sun M., D. Shen, and C. Huang. (1997). "CSeg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts," *Proceedings of the 5th Int'l Conference on Applied Natural Language Processing*, Washington DC, USA.

Wang, L.J., W. C. Li and C. H. Chang. (1992). Recognizing Unregistered Names for Mandarin Word Identification, *Proceedings of COLING-92*, Nantes, France.