Personal Names in Unrestricted Chinese Texts: Nature and Identification

Benjamin K. TSOU, Lawrence Y. L. Cheung

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon Tong,
Hong Kong

{rlbtsou, rlylc}@cityu.edu.hk

Abstract

The detection of personal names as well as proper names, and the identification of unknown words in unrestricted texts are critical tasks in NLP for East Asian languages, especially for word segmentation, information retrieval and machine translation. This is even more critical for Chinese which uses almost exclusively only the Chinese script and has little overt morphological markings and no equivalent use of capital letters for proper nouns as in English. This paper: (1) discusses the extent of the problems in some relevant IT applications, (2) analyzes the structure of Chinese personal names, and (3) presents some relevant processing strategies and the supporting language resources in general. Differences among Chinese personal names in Beijing and in Hong Kong are highlighted. It is argued that the awareness of variation in names across different Chinese communities constitutes a critical factor in enhancing the effectiveness of Chinese personal name identification algorithms.

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

1. Introduction

Personal names constitute an important linguistic symbol in conveying meaning. They are anchors of ideas, events, cultural artifacts, etc., e.g. Nobel Prize, Newtonian physics, Clinton-like behaviour and Thatcherism. Personal names provide a rich source for terminology in many domains. Efficiency in personal name identification is important for improving the detection and extraction of terms in the field of computational terminology. The last few years have seen the growth of research in this area (Miller et al., 1999; Cucchiarelli and Velardi, 2001). Named entity recognition was highlighted as an evaluation task in the Sixth and Seventh Message Understanding Conferences (MUC-6 and MUC-7) and First and Second Multilingual Entity Task (MET-1 and MET-2).

Because of the diverse and important linguistic differences between Chinese and English, personal name identification in Chinese involves many more complex issues than in English, e.g., word segmentation, absence of capital/small letter distinction, morphological paucity, syntactic ambiguity, and significant social and cultural differences among Chinese communities (Tsou and Kwong, 2001). Recent statistics from Chinese corpora provides an indicative range of personal names appearing in different domains (Tsou, 2000; Tsou, 2001). Table 1 shows that personal names account for as much as 16.8% of all word types in the 3-year LIVAC[1] newspaper corpus. They represent up to 2.4% of the word tokens in the 29 million character corpus.

	Newspaper Headlines[2]		Newspaper Texts[2]	Court Proceedings
%	Hong Kong (1 yr)	Taiwan (1 yr)	6 Chinese Communities (3 yrs)	Hong Kong (1 case)
Type	4.5	4.2	12.8 to 16.8	4.6
Token	4.4	3.7	1.6 to 2.4	0.6

Table 1 Amount of personal names in different domains

Because of the inherent linguistic problems above, the processing of Chinese personal names (as well as other named entities) in NLP requires much more than itemized listing, and poses a serious challenge.

The rest of the paper will be divided into three main sections: (1) assesses some relevant basic problems encountered in IT applications, (2) introduces the structure of Chinese personal names and the relevant processing strategies, and (3) highlights the importance of building language resources for personal name extraction.

[1] LIVAC synchronous corpus collects newspaper texts every four days since 1995 from Chinese newspapers in 6 Chinese communities including Beijing, Hong Kong, Macau, Shanghai, Singapore and Taiwan.
(LIVAC website: http://www.rcl.cityu.edu.hk/livac)

[2] The estimation is based on the 3 year data (1995—98) from the LIVAC corpus. It contains 29 million characters.

2. Significance of Personal Names in IT Applications

Efficient identification of personal names is crucial in many IT applications. Poor management of personal names in these systems can compound the errors in other NLP modules, resulting in serious deterioration of system performance. Cheung et al. (2002) conducted tests showing that poor personal name processing results in serious webpage retrieval errors.

(1) 陳中將與俄羅斯選手爭奪跆拳道金牌 (Sina) [3]
Chen Zhongjiang will compete with a Russian athlete for the gold medal for Taekwondo.

(2) 司令陳中將視導南測中心受訓部隊 (Google Chi.)[3]
Admiral Chen Zhongjiang inspected the trainee army force in Nance centre.

For example, the examined search engines mistook 中將 *zhongjiang* in (1) and (2) as the common noun for the military rank of "lieutenant general", whereas, in fact, they represent given names in the above contexts. The problem is similar to identifying *Dean Martin*, the well known American entertainer, as the head of a faculty in a university.

Tsou and Kwong (2001) also reported that the Chinese-to-English machine translation systems [4] have serious but unrecognized problems handling personal names. Table 2 shows that all four machine translation systems perform rather poorly in personal name identification. The probable cause for the errors is the use of static name list to identify personal names. The above demonstrates that IT applications need far more sophisticated algorithms than simple character matching and name database to adequately detect personal names in Chinese texts.

Data Source	Translation Accuracy			
	EWGate	*TongYi*	*Transtar*	*WorldLingo*
Hong Kong	24%	5%	9%	15%
Beijing	30%	6%	20%	56%
Taiwan	19%	0%	5%	16%

Table 2 Translation accuracy of personal names

3. Processing Chinese Personal Names: Challenges and Strategy

3.1. Challenges in Processing Chinese Personal Names

The basic structure of modern Chinese personal names is largely similar across different Chinese communities. Although the frequent length is 2 to 3 characters, the maximum can be as long as 6 characters. Table 3 shows the possible structures of Chinese personal name. Chinese personal names begin with a one- or two-character surname, followed by a one- or two-character given name. The name of a married female may be preceded by her husband's surname, as in (e) and (f). The unique structure

	Full Name	Husband's Surname			Surname			Given Name		Length
		H1	(H2)	+	S1	(S2)	+	G1	(G2)	
a.	李鵬 *Li Peng*				李 *Li*			鵬 *Peng*		2
b.	鄧小平 *Deng Xiaoping*				鄧 *Deng*			小 *Xiao*	平 *Ping*	3
c.	諸葛亮 *Zhuge Liang*				諸 *Zhu*	葛 *Ge*		亮 *Liang*		3
d.	東方聞櫻 *Dongfang Wenying*				東 *Dong*	方 *Fang*		聞 *Wen*	櫻 *Ying*	4
e.	陳方安生 *Chen Fang Ansheng*	陳 *Chen*			方 *Fang*			安 *An*	生 *Sheng*	4
f.	諸葛東方聞櫻 *Zhuge Dungfang Wenying*	諸 *Zhu*	葛 *Ge*		東 *Dong*	方 *Fang*		聞 *Wen*	櫻 *Ying*	6

Table 3 Structure of Chinese personal names

[3] *Google* [Big5 Chinese] URL: http://www.google.com/intl/zh-TW and *Sina* URL: http://www.sina.com.cn

[4] (1) Transtar V3.0, (2) TongYi '98, (3) *WorldLingo* (http://www.worldlingo.com), (4) EWGate: (http://www.EWGate.com/ewtranslite.html)

is found in speech or writing of formal register in some Chinese communities such as Hong Kong.

Apart from variable length, several characteristics make Chinese personal name processing difficult:

(a) There is no explicit morphological marking or capitalization for names in Chinese.

(b) Chinese texts do not have explicit word boundary.

(c) The character set for surnames and given names is a subset of Chinese characters for common Chinese words, and hence readily gives rise to structural ambiguity.

(d) Some personal names may be simple mono-syllabic words.

(e) Some polysyllabic words can be embedded in Chinese personal names, e.g. 王朝聞 *Wang Chaowen* (王朝 *wangchao* = dynasty), 馬勝利 *Ma Shengli* (勝利 *shengli* = victory) and 嚴肅 *Yan Su* (嚴肅 *yansu* = serious(ly)).

3.2. Basic Strategies

The complexity of Chinese personal name identification task calls for a combination of different processing strategies. They can be broadly divided into statistical approach and linguistic approach.

3.2.1. Linguistic Approach

Linguistic context provides important cues to locate Chinese personal names. Syntactic structures and lexical collocation provide good indication on whether or not the character string immediately before or after it is a potential personal name, e.g. 張志偉先生 (*Mr.* Zhang Zhiwei) and 朱鎔基總理 (*Premier* Zhu Rongji). Sun et al. (1995) integrates features to detect frequently used patterns, lexical items and syntactic structures that are useful for identifying names. For example, personal names often precede verbs like 說 *shuo* (say), 指出 *zhichu* (point out), etc. Lü et al. (2001) detect personal names by evaluating the interaction between potential personal names and neighbouring words. The POS co-occurrence restriction is checked and the best segmentation for potential name string is computed so as to generate the most probable context. Luo and Song (2001) studied the structure of personal name and place name formation. The linguistic knowledge is represented as a set of generative rules in finite state automata. Additional exceptional handling is added to deal with easily confused ambiguous contexts.

3.2.2. Statistical Approach

Statistical approach has been the most popular approach for name identification task. Previous studies typically exploited the character distribution frequency in different parts of a name and designed algorithms to extract string patterns that match the distributional criteria. For example, Sun et al. (1995) and Song and Tsou (2001) reported that about 400 characters[5] could cover over 99% of all Chinese surnames in texts. Furthermore, some character combinations in given names are more frequent than others. Cheung et al. (2002) also pointed out that there are significant variations among Chinese communities. The character preference in given names varies depending on a range of factors like gender, geography, character position in a given name, social changes, etc. The character probability is approximated by frequency distribution from large text corpora or name databases.

Sun et al. (1995) and Zheng et al. (1999) evaluated every candidate string by computing mutually exclusive probability for the 3 characters in a name candidate string, as in (6).

(6) $p_{pn}(c_1\, c_2\, c_3) = p_{sur}(c_1) * p_{m1}(c_2) * p_{m2}(c_3)$

where
$p_{pn}(s)$ = probability of candidate string s being a personal name
$p_{sur}(x)$ = probability of character x being a surname
$p_{m1}(x)$ = probability of character x being the first character of a given name
$p_{m2}(x)$ = probability of character x being the second character of a given name

Lü et al. (2001) proposed to measure probability of a potential name string by considering the probability of the 3 characters in a name candidate string as mutually inclusive events, as in (7) adapted from Lü et al. (2001).

(7) $p_{pn}(c_1\, c_2\, c_3) = p_{1F}(c_1) + p_{1M}(c_2) + p_{nE}(c_3)$

where
$p_{pn}(s)$ = probability of candidate string s being a personal name
$p_{1F}(x)$ = probability of character x being a surname
$p_{1M}(x)$ = probability of character x being the first character of a given name
$p_{nE}(x)$ = probability of character x being the second character of a given name

Most Chinese personal name identification algorithms incorporate linguistic and statistical techniques. These hybrid systems have been reported to achieve 80—90% precision and recall rates (Sun et al., 1995; Lü et al., 2001; Luo and Song, 2001).

[5] There are 21,886 characters in the GBK Chinese character set.

4. Personal Name Language Resources for Terminology Extraction

Statistical frequency data, as discussed in Section 3, has to be based on empirical data from large text corpora. Thus relevant personal name databases become a critical resource to support name identification systems and to customize algorithms. At least four major dimensions should be adequately addressed in the construction of personal name language resources, including: (1) structural distribution, (2) character frequency of personal names, (3) character co-occurrence for given names, and (4) communal differences. The significance and relevance of appropriate personal name database cannot be overemphasized because of the rarely understood magnitude of variation of personal names among Chinese communities which is much greater than that existing in English speaking communities. We will illustrate the differences in personal name patterns by using name databases taken from Beijing and Hong Kong.[6]

4.1. Structural Distribution

Single-character surnames predominate both databases, accounting for over 99%, as in Table 4. This suggests that double-character surnames may be handled separately using item listing in view of its very limited number of types and tokens. The data shows a divergence in the preference for single- and double-character given names in Beijing and in Hong Kong. Single-character names account for 29% of the Beijing database.[7] In contrast, single-character given names only cover 2% of the data for Hong Kong. The findings are crucial to the prioritization of rules related to the length of personal names in identification algorithms.

	Surname		Given Name	
%	Beijing	HK	Beijing	HK
Single-Character	99.9	99.6	29.1	2.1
Double-Character	0.1	0.4	70.9	97.9

Table 4 Distribution of name structures

[6] The Beijing name database has 125,033 names, and is drawn from a county in Beijing. They are representative of names in Mainland China because the county population is composed of migrants coming from different provinces of China. The Hong Kong database contains 11,358 names. They are student and staff names taken from the Registrar's Office, City University of Hong Kong.

[7] Sun et al. (1995) reported that single-character given names account for about 37% of the name database for all students' names (10 years) at Tsinghua University in Beijing.

4.2. Character Frequency of Personal Names

Not all characters are equally probable in being different parts of a Chinese personal name. All studies mentioned in Section 3 have exploited such characteristics to different extent. Table 5, 6 and 7 show that the ten most frequently used surnames, first character and second character of given names.

	Beijing				Hong Kong		
Rank	Surname	%	Cum. %	Rank	Surname	%	Cum. %
1	王 *Wang*	9.1	9.1	1	陳 *Chen*	10.2	10.2
2	張 *Zhang*	8.3	17.4	2	黃 *Wang*	6.7	16.9
3	李 *Li*	7.9	25.3	3	李 *Li*	5.9	22.8
4	劉 *Liu*	6.5	31.8	4	梁 *Liang*	4.6	27.4
5	陳 *Chen*	3.2	35.0	5	林 *Lin*	4.2	31.6
6	趙 *Zhao*	3.2	38.2	6	張 *Zhang*	3.6	35.2
7	楊 *Yang*	3.0	41.2	7	劉 *Liu*	3.0	38.2
8	孫 *Sun*	2.0	43.2	8	吳 *Wu*	3.0	41.2
9	馬 *Ma*	1.7	44.9	9	何 *He*	2.8	44.0
10	吳 *Wu*	1.6	46.5	10	鄭 *Zheng*	2.1	46.1

Table 5 10 most frequent single-character surnames in Beijing and Hong Kong

(Shaded items appear in both columns.)

	Beijing				Hong Kong		
Rank	G1	%	Cum. %	Rank	G1	%	Cum. %
1	淑 *shu*	3.2	3.2	1	嘉 *jia*	3.8	3.8
2	玉 *yu*	3.1	6.3	2	偉 *wei*	3.7	7.5
3	秀 *xiu*	2.9	9.1	3	志 *zhi*	3.5	11.0
4	曉 *xiao*	2.6	11.7	4	家 *jia*	2.8	13.8
5	文 *wen*	2.3	14.0	5	詠 *yong*	2.2	16.0
6	建 *jian*	2.2	16.2	6	慧 *hui*	2.1	18.1
7	志 *zhi*	1.9	18.0	7	國 *guo*	2.0	20.1
8	小 *xiao*	1.8	19.8	8	文 *wen*	2.0	22.0
9	桂 *gui*	1.7	21.5	9	佩 *pei*	1.9	24.0
10	春 *chun*	1.4	22.8	10	麗 *li*	1.9	25.9

Table 6 10 most frequent first characters (G1) of double-character given names

(Shaded items appear in both columns.)

	Beijing				Hong Kong		
Rank	G2	%	Cum. %	Rank	G2	%	Cum. %
1	華 *hua*	3.6	3.6	1	儀 *yi*	3.1	3.1
2	英 *ying*	3.4	7.0	2	華 *hua*	2.3	5.4
3	蘭 *lan*	2.1	9.1	3	明 *ming*	2.2	7.6
4	平 *ping*	1.9	11.0	4	敏 *min*	2.2	9.8
5	珍 *zhen*	1.8	12.8	5	文 *wen*	2.1	11.9
6	明 *ming*	1.7	14.5	6	玲 *ling*	1.9	13.7
7	榮 *rong*	1.6	16.1	7	珊 *shan*	1.7	15.4
8	生 *sheng*	1.5	17.6	8	欣 *xin*	1.6	17.0
9	芳 *fang*	1.3	18.9	9	輝 *hui*	1.6	18.6
10	琴 *qin*	1.3	20.1	10	雯 *wen*	1.6	20.1

Table 7 10 most frequent second characters (G2) of double-character given names

(Shaded items appear in both columns.)

The character type for Chinese surnames is fairly limited in actual data. In Table 5, the ten most frequent surnames cover over 46% of the name tokens though the ranking of surnames is quite different in both databases. For example, the most frequent surname 王 *Wang* in Beijing is ranked as 14th in the Hong Kong. In contrast, the character types for given names are far more diverse. In the Hong Kong database, there are over 820 character types for given names as opposed to 257 character types for surnames. As shown in Table 6 and 7, the ten most frequently used G1 and G2 character cover no more than 26% of all name tokens in both databases respectively.

4.3. Character Co-occurrence for Given Names

Apart from localized character preference in given names, our data also reveals that the character combinations of double-character given names are far from being random. Previous research tended to consider the probabilities of each character position in isolation, and ignored interesting patterns of character co-occurrence in given names. The information is useful for resolving ambiguity given rise by the diverse character types in given names. Here are two examples for the two most common G1 characters from Hong Kong database: 嘉 *jia* and 偉 *wei*. Given G1 = 嘉 *jia* / 偉 *wei*, there is about 30% of chance that the given name is one of the combinations in a—e and f—j respectively. (Table 8)

	Combina-tion	%	Cum. %		Combina-tion	%	Cum. %
a	嘉 + 敏 *jia + min*	11.2	11.2	f	偉 + 強 *wei + qiang*	6.4	6.4
b	嘉 + 儀 *jia + yi*	6.5	17.7	g	偉 + 文 *wei + wen*	5.7	12.1
c	嘉 + 雯 *jia + min*	4.6	22.3	h	偉 + 雄 *wei + xiong*	5.7	17.7
d	嘉 + 琦 *jia + qi*	4.3	26.6	i	偉 + 傑 *wei + jie*	*5.4*	23.2
e	嘉 + 慧 *jia + wei*	3.6	30.1	j	偉 + 明 *wei + ming*	5.2	28.3

Table 8 *5* most frequent combinations provided G1 = 嘉 *jia* / 偉 *wei*

4.4. Communal Differences

Previous studies do not seem to pay much attention to the sociolinguistic aspects of name variation among Chinese communities. It has been mentioned in Section 4.1 that there are far more single-character given names in the Beijing database. If we further compare the columns for Beijing and Hong Kong in Table 6 and 7, only two characters overlap. The divergence in character preference is obvious in the two databases. The implication is that name identification algorithms using statistical approach should maintain character probability derived from various Chinese communities in order to maximize the performance. Other sociolinguistic differences such as married female's names, nicknames, etc, have yet to be studied. The assumption that personal name identification can be simplistically tackled on the basis of personal name language resources from a single community will certainly be problematical for NLP applications that have to process unrestricted texts from different geographical locations.

5. Further Works

Based on Section 4.2 and 4.3, further investigation into the statistical distribution of personal names can be done. First, it seems that previous studies tended to have overlooked character co-occurrence phenomenon. Co-occurrence probability can be used to improve existing algorithms for Chinese personal name tagger. For example, instead of merely utilizing the probability of a candidate character being part of a name, the tagger may give a higher rating to those candidate strings whose G1 and G2 combination is commonly found in Chinese names. Statistical studies like those in Section 4.3 will be conducted for all character combination in the two databases to identify high frequency patterns.

Second, as we mentioned earlier and noted by a reviewer, gender is a significant factor in character choice for given names. Such data may find applications in transcription system and speech recognition application such as caller name identification. The recognition engine may first determine the caller's gender based on the speaker's voice pitch and then select the appropriate probability database for name identification task accordingly.

Third, the communal differences revealed by our preliminary analysis suggest that name databases from other Chinese communities are important language resources. For example, more name databases will be collected (e.g. Shanghai, Taiwan and Singapore) for comparison.

6. Conclusion

Personal names provide an important source for new terms in text processing. Personal name identification is crucial to terminology extraction. This paper discusses the challenge and basic strategies in personal name identification in unrestricted Chinese texts. The review of IT applications shows that reliability in the processing of Chinese personal names is still far from acceptable. This situation contributes to serious errors in other NLP tasks such as incorrect data retrieval and parsing. Current Chinese personal name identification systems capitalize on linguistic and statistical techniques to deal with the processing. To adequately support such systems, personal name language resources are critical. Four dimensions have been highlighted in the construction of such resources, including (1) structural distribution, (2) character frequency of personal names, (3) character co-occurrence for given names, and (4) communal differences. Despite the potential contribution to the identification task, the latter two dimensions seem to have gone largely unnoticed in the literature. More empirical study of personal names will be beneficial to the performance improvement of personal name identification systems.

7. Acknowledgement

This research study is supported by the Language Information Sciences Research Centre, City University of Hong Kong and by a Competitive Earmarked Research Grant (CityU 1238/00H) from the Research Grant Council of Hong Kong and supported by NTT Service Integration Laboratory. We would also like to thank 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.

8. References

Cheung, L., B. K. Tsou and M. Sun. (2002) Identification of Chinese Personal Names in Unrestricted Texts. *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, Cheju, Korea, pp. 28—35.

Cucchiarelli, A. and P. Velardi. (2001) Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence. *Computational Linguistics* 27 (1): 123—131.

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

Lü, Y., T. Zhao, M. Yang, H. Yu and S. Li. (2001) Leveled Unknown Chinese Words Resolution by Dynamic Programming. [in Chinese] *Journal of Chinese Information Processing*, 15 (1), Beijing, China.

Miller, D., R. Schwartz, R. Weischedel, and R. Stone. (1999) Named entity extraction from broadcast news." *Proceedings of DARPA Broadcast News Workshop*. Herndon, VA.

Song, R. and B. K. Tsou. (2001) Preliminary Study on Chinese Proper Noun. [in Chinese] *Proceedings of the 20th Anniversary Conference of the Chinese Information Processing Society of China*. November 2001. pp. 14—19.

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

Tsou, B. K. (2000) Lexical Variation in Chinese: The Windows Approach. (Invited paper) Annual Research Forum, Linguistic Society of Hong Kong. December 2000.

Tsou, B. K. (2001) Corpus, Information Mining and the New Global Village. (Keynote speech) *Proceedings of 6th Natural Processing Pacific Rim Symposium*, Tokyo, November 2001. pp. 9—18.

Tsou, B.K. and Kwong, O.Y. (2001) Evaluating Chinese-English Translation Systems for Personal Name Coverage. *Proceedings of the MT Summit VIII Workshop on MT 2010 -- Towards a Road Map for MT*, Santiago de Compostela, Spain.

Zheng, J., X. Lin and H. Tan. (2000) The Research Chinese Names Recognition Method Based on Corpus. [in Chinese] *Journal of Chinese Information Processing*, 14 (1), Beijing, China.