Corpus-Based Pinyin Name Resolution

Kui-Lam KWOK
Computer Science Dept., Queens College,
City University of New York
65-30 Kissena Boulevard,
Flushing, NY 11367, USA
kwok@ir.cs.qc.edu

Peter DENG
Computer Science Dept., Queens College,
City University of New York
65-30 Kissena Boulevard,
Flushing, NY 11367, USA
deng@ntkk.cs.qc.edu

Abstract

For readers of English text who know some Chinese, Pinyin codes that spell out Chinese names are often ambiguous as to their original Chinese character representations if the names are new or not well known. For English-Chinese cross language retrieval, failure to accurately translate Pinyin names in a query to Chinese characters can lead to dismal retrieval effectiveness. This paper presents an approach of extracting Pinyin names from English text, suggesting translations to these Pinyin using a database of names and their characters with usage probabilities, followed with IR techniques with a corpus as a disambiguation tool to resolve the translation candidates.

Introduction

It is important for many applications to be able to identify and extract person names in text. For English, capital letter beginning of a word is an important clue to spot names, in addition to other contextual ones. When an English story refers to a foreign person, it is relatively easy to represent the person's name if the alphabets have approximate correspondences between the languages. When it refers to a Chinese person, this is not possible because Chinese language does not use alphabets. The most popular method for this purpose is Pinyin coding (see, for example, the conversion project at the Library of Congress website (2002)), China's official method of using English to spell out Chinese character pronounciations according to the Beijing Putonghua convention. Chinese characters are monosyllabic, and the large majority of them has one sound (ignoring tones) and hence one code. However, given a Pinyin it usually maps to multiple characters. Such an English Pinyin name raises ambiguity about the original Chinese characters that it refers to and hence the person. If the name is well known, such as Mao ZeDong, this is not an issue; if the name is less frequently seen, one would like to see or confirm the actual Chinese characters.

The situation is similar to many Chinese word processing systems that use Pinyin as one of their input methods. When a Pinyin is typed (sometimes with tonal denotation), many candidate characters will be displayed for the user to select. The character list can be ordered based on a language model, Chen & Lee (2000), or on the user's past habit. When one comes across names as input however, a language model is not as helpful because practically any character combination is possible for names.

Pinyin names also present difficulties in a cross language information retrieval (CLIR) scenario. Here, an English query is given to retrieve Chinese documents, and Pinyin names could be present as part of the query. In general, one can have three approaches to CLIR as discussed in Grefenstette (1998): translate the Chinese documents to English and do retrieval matching in English; translate the English query to Chinese and do matching in Chinese; or translate both to an intermediate representation. With the first approach, one could use standard table lookup to map the characters of a Chinese name to Pinyin after identifying a name for extraction. Chen and Bai (1998), Sun et.al. (1994) have shown that this extraction process is not trivial since Chinese writing has no white space to delimit names or words. A more general difficulty is that the document collection may not be under a user's control, but available for retrieval purposes only. This makes document translation to the query language (or to an intermediate language) not suitable. A more flexible approach is to translate a query to Chinese and do retrieval in Chinese. This has

been the more popular method to use for CLIR in TREC experiments: Voorhees and Harman (2001). Whichever translation direction one chooses, a bilingual dictionary is essential. This dictionary however can be expected to be incomplete, especially with person names. Missing their translations can adversely impact on CLIR effectiveness. This raises the question of how to render Pinyin names into Chinese characters for translingual retrieval purposes.

In the recent NTCIR-2 English-Chinese cross language experiments, Eguchi et.al. (2001), quite a few queries have names. Kwok (2001) found that these lead to good monolingual retrieval because the names are quite specific and have good retrieval properties. On the other hand, for CLIR that starts with English queries, not being able to translate Pinyin names correctly leads to substantial deficit in effectiveness. This causes comparisons with monolingual results particularly dismal.

In this paper, we propose an approach to resolve the characters from a Pinyin code. It is based on: 1) a rule-based procedure to extract Pinyin codes for Chinese person names in English text; 2) a database for proposing candidate Chinese character sequences for a Pinyin code based on usage probabilities; and 3) a target collection and IR techniques as a confirmation tool for resolving or narrowing down the proposed candidates. These are described in Sections 1, 2, and 3 respectively. Section 4 presents some CLIR results and a measure of the effectiveness of our procedures.

We like to stress that even if one obtains the correct Chinese characters for a Pinyin, they can still refer to different persons with the same name. We do not address this issue here.

1 Pinyin Name Extraction

Chinese person names in Pinyin have fairly predictable formats such as: first alphabet of the family name (surname) is capitalized, as is the first word (or second word) of a given name. Two-syllable given names may appear as one word or two. The latter may be hyphenated, a practice popular in places such as Taiwan or Hong Kong. Thus, one may find Chairman Mao's name in any of the following formats:

Mao Ze Dong
Mao ZeDong Mao Zedong
Mao Ze-Dong Mao Ze-dong

Some publications also place the given name in front of the surname to agree with Western name convention. This style is supported but not used in this paper.

While the surname character is pretty much closed, the given name is not. It is well known that the most popular Chinese surnames number to about 100. Including less frequent ones bring the number to about 400 which we use: see Hundred Surname website (2002). Sun, et.al. (1994) reported over 700 surnames in their studies when additional infrequent ones are included. Other than for a few exceptions, this set all have unique Pinyin codes. These surname codes constitute an important clue for spotting a name sequence. The capitalized word(s), and the monosyllabic nature of words immediately after (or before) the surname give further support of its existence. We also loosen name definition to detect entries that have a hyphen but without a surname. Some rare surnames can be two syllables long, and often pair with one syllable given names. A woman may include her own family name in addition to her husband's. For our current study, we limit testing to a sequence of two to three Pinyin syllables only. This seems sufficient for the large majority of names encountered. Fig.1 shows our procedure to identify possible Pinyin names without the need of a training corpus or a name dictionary.

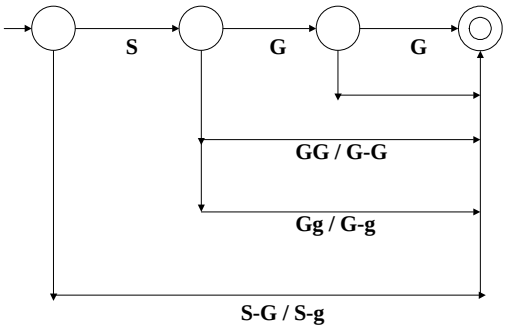

Figure 1: Pinyin Name Detection Algorithm (S,G = surname & given name syllable with upper-case first character; Gg, G-g = concatenated or hyphenated syllables, second one with lower-case)

2 Mapping Pinyin Name to Chinese

To suggest Chinese characters for the detected Pinyin, we downloaded about 200K Chinese names. This is augmented with another ½ million Chinese name usage isolated from the TREC-6 Chinese collection using BBN's

IdentiFinder (see Section 4). Last name and given name/characters are stored separately to form a database of name usage with frequencies. Two-character given names are stored both ways: as a single entry (observed) and as two separate characters. Observed usage items have their frequencies muliplied by a large factor to separate it from the unobserved type. A potential Pinyin surname is mapped to a set of possible characters. Existence of such characters in this surname database is the first step to decide that one may have a possible name sequence. Otherwise, we assume the Pinyin is not a name.

Knight and Graehl (1997) have proposed to compose a set of weighted finite state transducers to solve the much more complicated problem of back-transliteration from Japanese Katakana to English. Their concern includes all types of source Katakana terms (not just names), corruptions due to OCR, approximations due to modeling of English, Japanese pronunciations, and a language model for English. Proposing Chinese characters for Pinyin is like back-transliteration and can also be viewed probabilistically. Some unique considerations however lead to a much simpler problem.

Given an English Pinyin name E=EsEg (surname Es, given name Eg), our concern is to find the best Chinese name character sequence C=CsCg that maximizes P(C|E), or equivalently P(E|C)*P(C). Since surnames (Es,Eg) and given-names (Cs,Cg) can be considered independent, this probability can be re-written as: P(Es|C)*P(Cs)*P(Eg|C)*P(Cg).The conditioning on C can be replaced by Cs and Cg respectively since Chinese given names Cg should not influence English surname Es, and Cs should not influence Eg. As discussed before, other than a few exceptions Chinese characters have unique Pinyin, and hence P(Es|Cs) and P(Eg|Cg) is deterministic. Maximizing P(C|E) is equivalent to maximizing P(Cs')*P(Cg'), where Cs' and Cg' are sets of characters mapping from Es and Eg respectively. These probabilities are obtainable from frequencies in our database. Given names are limited to one or two syllables. In the latter case, the two characters are also assumed independent, and estimates of P(Cg') are smoothed between character pairs and their corresponding singles.

To illustrate, we use the Pinyin: Jiang ZeMin (correct Chinese name is 江泽民) as an example. This spelling is confirmed as a name because "Jiang" maps to five possible surnames, and "ZeMin" obeys given-name format, and have corresponding characters. Each surname character and all possible combinations of the given name characters are considered and probabilities evaluated based on the database of name usage frequencies. The top 16 candidates and estimated probabilities produced from our procedure are shown below:

.763 江泽民 .119 蒋泽民 .110 姜泽民 .005 江泽敏
.001 强泽民 .001 蒋泽敏 .001 姜泽敏 .000 江泽闽
.000 江则闵 .000 江泽棉 .000 江泽绵 .000 江泽冕
.000 江泽 .000 江泽悯 .000 江泽岷 .000 蒋泽闽

The probabilities are skewed because the first (correct) name has large usage frequency in the training data. However, every candidate is a possible name irrespective of probabilities because of the idiosyncracies of name forming.

Quite often, some places or organizations also sound like names. These will also be translated (see example in Section 4). A couple of notable failures are strings like 'So China', which our procedure decodes as a name 'So Chi-na', 'So' being a legitimate surname in Wade-Giles convention. 'Hong Kong' also passes our test with candidates: 洪孔, 红崆, 鸿空, etc. A 'stoplist' of such string patterns is employed to partially alleviate these errors.

3 Pinyin Name Resolution

Once candidate names for a Pinyin are available, one may output the top n ranked items as answers. However, selecting names based on probability may not be the best strategy. Quite often, people deliberately choose rare characters for naming purpose because they want to be differentiated from the usual run-of-the-mill names. Our strategy is to use IR techniques with a text collection to help in name selection. For cross language retrieval, it is especially helpful to use the target retrieval collection for resolution. This ensures that a translated name exists in the collection for retrieval. For general application, one could employ domain-relevant collections. Moreover, one can also use the occurrence frequency of the names in the collection to help narrow down the candidates: i.e. the higher the frequency, the more probable that the name is the intended one. This has the

advantage that selection is tailored more to the application, and less dependent on the name character database of Section 2. When the collection is well chosen, this process can whittle down the candidates to just a few with good accuracy.

4 Experimental Studies

We performed two studies to demonstrate our Pinyin resolution strategy. The first is to repeat retrieval on some queries in NTCIR-2 cross language experiments to see how Pinyin name resolution can affect effectiveness. A second experiment is to use BBN's IdentiFinder as a reference, and to compare how our procedures succeed in extracting Pinyin names and translating them with respect to a reference set.

4.1 CLIR with Pinyin Names

One of the NTCIR-2 cross language retrieval experiments (Eguchi, et.al. 2001) consists of 50 English topics and a Chinese target collection of about 200 MB. The purpose is to retrieve relevant Chinese documents using English text (topics) as queries. The Chinese counterparts to the English topics were also given so that CLIR results can be compared to monolingual. The original topics are lengthy; we limit our queries to a few words from the 'title' section of the topics. Three queries have Pinyin names and two contain non-person Pinyin entities that satisfy our Pinyin name detection format.

On running these 'title' queries through our procedure, the Pinyin codes were identified, candidates suggested, and resolved using the target collection. Listed in Table 1 are the queries. The Pinyin name in each 'Original English' and 'Original Chinese' query is bolded. Under the column 'Selected Names with Occurrence Frequency' are the resolved Pinyin names in Chinese, together with their occurrence frequencies in the retrieval collection. As discussed in Section 3, these selections are narrowed down from a large number of candidates in the intermediate step.

The Pinyin in Query 33 is for a kind of bean, while Query 44 has the name for a well known mountain, but they satisfy our definition of a name pattern. It can be seen that except for Query 46, the name with the largest occurrence agrees with the one intended in the monolingual query. In Query 46, the given name 'Yo-yo' is non-standard Pinyin, with suggested candidates like '唷唷' or '育育', and there are no such entries in the collection. If it were spelt 'You-you', the correct characters '友友' will be among the candidates and selected by the collection. When these Pinyin names with frequency>=5 were added to our MT software concatenated with dictionary translation procedure, Kwok (2001), the initial retrieval results in Table 2 are obtained. Here we follow the TREC convention to evaluate retrieval using the measures RR (relevant documents in top 1000 retrieved), Av.P (average precision), and P@20 (precision at the top 20 documents retrieved).

Substantial improvements were obtained for four of the queries when the names are correctly picked, and come closer to or even surpass the monolingual result. This demonstrates that our approach to Pinyin name resolution can work, but we need more queries of this type to confirm the effect. Query #15 has very high Av.P of .3287 because dictionary translation brought in useful content words not present in the monolingual query like: 绑票 (kidnapping), 杀害, 杀人案件 (murder criminal case). These

Table 1. Pinyin Name Resolution in 5 Queries (* denotes Correct)

Qry#	Original English	Original Chinese	Selected Names with Occurrence Frequency
15	**Bai Xiao-yan** kidnapping murder criminal case.	白**晓燕**绑架撕票案.	*白晓燕 66
33	**Bai-feng** Bean.	白**凤**豆.	*白凤 69, 白風 2
44	**Hua-shan** Art Zone.	华**山**艺术特区.	*华山 41, 華善 2, 花山 2
46	**Ma Yo-yo** cello recital.	马**友友**演奏会.	马 3792, 麻 234
47	**Jin Yong** kung-fu novels.	金**庸**武侠小说.	金永 5, 金勇 2, 金泳13, *金庸 186

Table 2: Effect of Pinyin Resolution on Retrieval Results of 5 Queries (Compared to Monolingual and Translation Only)

Qry#	Monolingual				Translation				Translation+ Pinyin		
	RR	Av.P	P@20		RR	Av.P	P@20		RR	Av.P	P@20
15	17	.1594	.15		12	.0611	.10		18	.3287	.30
33	13	.5277	.45		13	.2174	.10		13	.4579	.45
44	7	.3856	.30		5	.0082	.00		6	.1783	.15
46	7	.7543	.35		7	.0078	.00		7	.0077	.00
47	17	.5801	.45		17	.3179	.35		17	.6311	.50

expand the query and combine synergistically with the Pinyin name to provide precision surpassing the monolingual result. As a candidate name, 金庸 in Query #47 has very low probability compared to others because the character 庸 (meaning 'mediocre') is rarely used in names. It was pulled out by high occurrence frequency in the target collection. Thompson & Dozier (1997) have also shown that correctly indexing names in monolingual English retrieval leads to better retrieval.

4.2 Resolving Pinyin Names in Text

In another experiment we intended to test our Pinyin procedure with parallel collections that contain many paired names, but failed to locate one. We intend to evaluate how well our extraction procedure works, and whether candidate suggestion can recover correct Chinese names. A pair of collections was downloaded from the Peoples' Daily website (2001) Year 2001 English version (~17MB) and the Chinese version (~70MB) as our test collections. A sampling shows that they have very different content. Our aim is to isolate Pinyin names from the English collection, and create a list of their Chinese counterparts. We can then compare our Pinyin extraction against the English list. We also like to see how our database suggest Chinese candidates for this fairly recent name set. The evaluation is more approximate compared to doing an evaluation using parallel corpora with lots of names paired.

BBN's Identifinder, described in Weischedel et. al. (1996) was employed to process both collections independently. When given English or Chinese texts, IdentiFinder can bracket enttities of different types such as: PERSON, LOCATION, ORGANIZATION, etc. for later extraction. PERSON entities were isolated and two unique person name lists were produced: 4840 in English and 47621 in Chinese. They include Pinyin, non-Chinese and Chinese person names. The Chinese list contains many entries with one character (such as a surname 陈), translitered foreign names, and some with symbols. These we want to avoid. By capturing entries of length >=2 characters, without symbols, and having legitimate surnames, a filtered list of 23,863 Chinese entries were obtained. They were mapped into Pinyin and intersected with the English list. A total of 897 COMMON entries resulted, forming our reference set (Fig. 2). These are Chinese names obtainable by translating from the 4840-entry English list and which occur on the filtered list.

The original English collection was next processed through our Pinyin identification procedure, and 1769 unique entries were detected to satisfy our criteria. Comparison with IdentiFinder's English list shows that 1467 (83%) are the same, and 302 (17%) different. The non-overlap can be due to: i) non-person entities that sound like names on our list; ii) non-Chinese names on the IdentiFinder list; iii) legitimate Chinese names detected by one and not the other; or iv) errors on either procedures. Candidate Chinese names were suggested for our 1769-entry Pinyin list, and afterwards resolved with the Chinese COMMON list. This tests how well our database suggests names for Pinyin. The result is shown in Table 3. We show suggestions of 1, 5, up to 50 candidates, and recall of the reference set improves steadily from 35.3% to 93.9% (missing 55 of those 897 in COMMON) at 50 suggested. This shows the

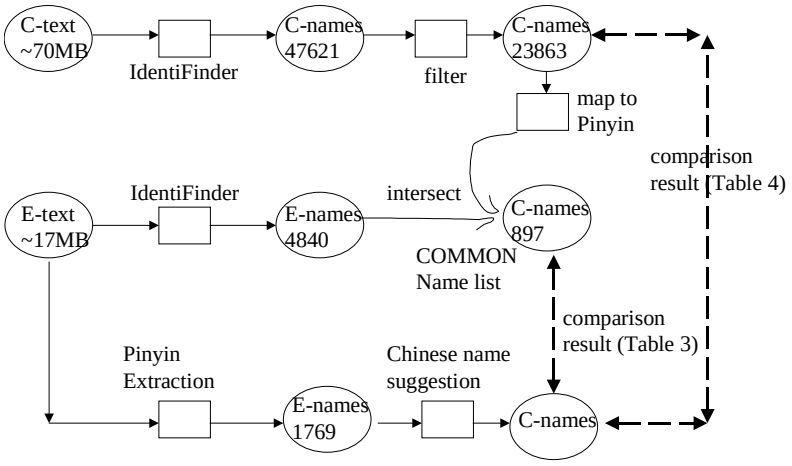

Fig.2 Pinyin Name Extraction & Suggestion: Comparison with IdentiFinder (BBN)

Table3 : 1769 Pinyin Names Resolved Against COMMON Name List (Size=897)

# of Candidates	Breakdown of 1769 Pinyin Names	Recall
1	1452 + **317 in COMMON**	**35.3% of 897**
5	1155 + **614** "	**68.5%** "
10	1041 + **728** "	**81.2%** "
30	949 + **820** "	**91.4%** "
50	927 + **842** "	**93.9%** "

Table 4: 1769 Pinyin Name Resolved Against Filtered Chinese List (Size=23863)

# of Candidates	Breakdown of 1769 Pinyin Names	NewNames Recovered
1	1422 + **347 in Filtered Chinese List**	**347-317=30**
5	1052 + **717** "	**717-614=103**
10	912 + **857** "	**857-718=129**
30	783 + **986** "	**986-820=166**
50	753 + **1016** "	**1016-842=174**

difficulty of suggesting a correct name: only ~35% recall at top 1, ~68% at top 5. In general, small 'top n' is not sufficient to recover a correct name translation, while using too many lead to noise. Hence there is a need to resolve candidates on a relevant collection.

We further compare the suggested Chinese names for the 1769 Pinyin against the filtered Chinese list (23863 entries) to see whether our Pinyin extraction can recover additional Chinese names not obtained by IdentiFinder (from the same English text). We found that at each suggestion level (Tables 4 & 3), more names were found by our Pinyin procedure that were missing in IdentiFinder: 30 at suggestion level 1, up to 174 (~19%) more names at the level of 50. These 174 are names in the filtered portion of the Chinese list but not included in COMMON because the English list from IdentiFinder does not have their corresponding Pinyin. The rest (1769-1016=) 753 on our list could be non-person entities that sounded like names, wrongly identified entries, or person names that do not exist in IdentiFinder's Chinese list. IdentiFinder may fail to extract some Chinese names as well. For example, some Pinyin names with 'An' as surname were missed. This study demonstrates

the ability of our approach to locate Pinyin names in English text and translate them.

Conclusion and Discussion

A procedure to translate any Pinyin name into possible Chinese characters with probabilities based on usage frequencies is proposed. Candidates can further be resolved against a text collection to narrow down the possibilities. This leads to better CLIR results. For a recent English news collection, 83% of Pinyin names identified agrees with names found by BBN's IdentiFinder. Chinese name candidates for these Pinyin cover between 35.3 to 93.9% of a COMMON name set for the IdentiFinder names when suggestions varies between 1 to 50. But additional Chinese names not extracted by IdentiFinder can be located using our procedure. Pinyin is an official coding used in China and getting popular elsewhere. Names from other places such as Taiwan use different Pinyin conventions like Wade-Giles. We had some provision for them, but plan to expand our coverage for these names more completely in the future.

Some web search engines offer advanced techniques that allow users to input English key terms and display results from Chinese documents, selecting items that have the English term and Chinese counterpart. These engines serve like giant bilingual dictionaries providing for entity translation. However, web pages usually contain current data and popular names only (like Ma Yo-yo). Lesser known names (like Bai Xiao-yan) are not available. Our approach can suggest Chinese names for Pinyin even if web search fails, or the relevant collection employed does not further resolve the suggested translations. For CLIR, our procedure ties translated names to the retrieval collection. We envisage each of these approaches has its own advantages, and that employing both together may help provide more accuracy for the issue of how to translate Pinyin names.

Acknowledgements

This work was partially sponsored by the Space and Naval Warfare Systems Center San Diego, under Grant No. N66001-00-1-8912. We thank BBN for the use of their IdentiFinder software.

References

Chen, K-J. & Bai, M-H. (1998) *Unknown word detection for Chinese by a corpus-based learning method.* Intl. J. of Computatinoal Linguistics & Chinese Language Processing. 3:27-44.

Chen, Z. & Lee, K-F. (2000) A new statistical approach to Chinese Pinyin input. (available at http://www.microsoft.com/china/research/dload_files/g-nlps/NLPSP/n8.pdf)

Eguchi, K., Kando, N. and Adachi, J. (eds.) (2001) *Proc. of Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval, and Text Summarization.* NII: Tokyo. (available at http://research.nii.ac.jp/ntcir/)

Grefenstette, G. (1998) *Cross language Information Retrieval*. Kluwer Academic Publishers, Boston.

Hundred Surname website. (2002) (available at http://www.geocities.com/Tokyo/3919/hundred.html)

Knight, K. and Graehl, J. (1997) *Machine transliteration.* Proc.of 35th Annual Meeting of ACL, pp. 128-135.

Kwok, K.L. (2001) *NTCIR-2 Chinese, cross language retrieval experiments using PIRCS.* In: Proc. of Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval, and Text Summarization. pp. 111-118. NII: Tokyo. (available at http://research.nii.ac.jp/ntcir/).

Library of Congress Website. (2002) (available at http:// www.loc.gov/catdir/pinyin/outline.html)

People's Daily Website. (2001) (available at http://www. peopledaily.com.cn)

Sun, M.S., Huang, C.N., Gao, H.Y. and Fang, J. (1994) *Identifying Chinese names in unrestricted texts.* Comm. COLIPS, 4, pp. 113-122.

Thompson, P. and Dozier, C.C. (1997) *Name searching and information retrieval.* Proc. 2nd Conf. on Empirical Methods in NLP, pp. 134-140.

Voorhees, E. and Harman, D.K. (eds). (1998) *The Sixth Text Retrieval Conference (TREC-6).* NIST Special Publication 500-249. (available at http://trec.nist.gov/)

Voorhees, E. and Harman, D.K. (eds) (2001) *The Ninth Text Retrieval Conference (TREC-9).* NIST Special Publication 500-249. (available at http://trec.nist.gov/)

Weischedel, R. Boisen, S., Bikel, D., Bobrow, R., Crystal, M., Ferguson, W., Wechsler, A. & the PLUM Research Group. (1996) *Progress in Information Extraction.* Proceedings of Tipster Text Program (Phase II). pp. 127-142.