Translating Chinese Romanized Name into Chinese Idiographic Characters via Corpus and Web Validation

Yiping Li — Gregory Grefenstette

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M)
Commissariat à l'Energie Atomique
Bat. 38-1; 18, rue du Panorama; BP 6; 92265 Fontenay aux Roses Cedex; France
li@zoe.cea.fr
gregory.grefenstette@cea.fr

ABSTRACT. *Cross-language information retrieval performance depends on the quality of the translation resources used to pass from a user's source language query to target language documents. Translation lists of proper names are rare but vital resources for cross-language retrieval between languages using different character sets. Named entities translation dictionaries can be extracted from bilingual corpus with some degree of success, but the problem of the coverage of these scarce bilingual corpora remains. In this article, we present a technique for finding Chinese transliterations for any Chinese name written in English script. Our system performs transliteration of Pinyin (the standard Romanization for Chinese) to Chinese characters via corpus and web validation. Though Chinese family names form a small set, the number and variety of multisyllabic first names is great, and treatment is complicated by the fact that one Pinyin transliteration can correspond to hundred of different Chinese characters. Our method finds the best translations of a Chinese name written in Pinyin by filtering out unlikely translations using a bigram model derived from a very large monolingual Chinese corpus, and then vetting remaining candidate transliterations using Web statistics. We experimentally validate our method using an independent gold standard.*

RESUME. *La performance en recherche d'information translingue dépend de la qualité des ressources de traduction utilisées pour passer de la langue source (requête d'utilisateur) vers la langue cible des documents. Les listes de traduction de noms de personnes sont rares, et constituent en même temps des ressources essentielles pour la recherche d'information translingue entre des langues utilisant des jeux de caractères différents. Les dictionnaires de traduction d'entités nommées peuvent être extraits des corpus bilingues avec un certain succès, mais le problème du recouvrement de ces corpus bilingues, rares, reste présent. Dans cet article, nous présentons une technique pour retrouver la translittération en chinois de tous les noms chinois écrits en anglais. Notre système effectue la translittération du Pinyin (la romanisation standard du chinois) en caractères chinois via des validations effectuée sur corpus et sur le Web. Bien que les noms de famille en chinois constituent un ensemble peu important, les variétés des prénoms multi-syllabiques sont très importantes. Le traitement s'avère d'autant plus compliqué qu'à une translittération du Pinyin peut correspondre jusqu'à plus de cent caractères chinois différents. Notre méthode sélectionne la meilleure traduction*

des noms chinois écrits en Pinyin en filtrant les traductions impossibles et en utilisant un modèle de bigrammes extrait d'un très grand corpus chinois monolingue, puis en éliminant les traductions candidates restantes à l'aide de statistiques recueillies sur le Web. Nous avons évalué notre méthode en utilisant une référence indépendante.

KEY WORDS: *transliteration English-Chinese, proper names, corpus validation, web validation, translation*

MOT-CLES: *translittération anglais-chinois, noms propres, validation du corpus, validation du Web, traduction*

1. Introduction

The quality of the translation resources used to pass from a user's source language query to target language documents has a great effect on the performance of multilingual applications such as the cross-language information retrieval (Hull *et al*., 1996; Grefenstette, 1998; Levow *et al*., 1999) or machine translation systems. Among these translation resources, resources providing the correct translation, or the correct transliteration, of proper names are central to practical applications involving texts referring to people, places or organizations. Translation lists of proper names are obligatory for cross-language retrieval between languages using different character sets, but such lists are rare. If enough bilingual texts, covering the same domains and same periods as the user is interested in, are available it is possible to extract named entity translation dictionaries with some degree of success (Huang *et al*., 2003). However, the problem of the coverage of these scarce bilingual corpora remains. Therefore the need for out-of-vocabulary word translation (very often named entities) from one language to another is increasing, especially between language using a Roman alphabet and languages using other alphabets such as Chinese, Japanese and Korean (Li *et al*., 2004).

(Meng *et al*., 2001) used cross-lingual phonetic mapping to realize the transliteration of foreign names from English to Chinese characters in the context of cross-language spoken document retrieval. Similar techniques to address the out-of-vocabulary problem have been used in transliteration of foreign names (Gao *et al*., 2004), in foreign place name transliteration problem (Wan *et al*., 1998), in English-Chinese named entity alignment (Feng *et al*., 2004), and in the acquisition of English-Chinese transliteration word pairs from parallel texts (Lee *et al*., 2003). These referenced works especially address the transliteration of foreign names into a different, but the problem of back-transliteration of Chinese names has not been discussed until now.

Here we show that it is possible to generate the possible spellings of a Chinese name, and then to filter out unlikely spellings using a large corpus and/or the WWW. In this paper, we present our technique, similar to the technique proposed by (Qu *et al*., 2004) for transliterating Japanese names, for transliterating Chinese names written in Western script back into Chinese characters.

1.1 Background on Chinese name formation

Compared to Occidental names, Chinese name composition is quite flexible. Any combination of Chinese ideographs can be used to form a given name, although characters (we call Chinese ideograph as character in the rest of the paper) are chosen a priori to express a blessing or an expectation for the newborn. There are thus many varieties of personal names which share the same characters as in other common words.

Pinyin is the modern method used to translate Chinese names in English. Pinyin literally means "join together sounds" (a less literal translation being "phoneticize", "spell" or "transcription") in Chinese. It is a system of Romanization (phonetic notation and transliteration to Roman script) for Standard Mandarin (also called simplified Chinese) used in the People's Republic of China. In 1979 the International Organization for Standardization (ISO) adopted Pinyin as the standard Romanization for Modern Chinese. Pinyin is a Romanization and not an Anglicization; that is, it is equally applicable for transliteration into any language that uses a Roman alphabet.

Besides Pinyin which is the official and most widely used system in the People's Republic of China, Wade-Giles is another Romanization system for the Chinese language based on Mandarin used in Taiwan. It was the main system of transliteration in the English-speaking world for most of the 20th century. In this work, we choose Pinyin, which is now the dominant system.

1.2 Overview on our transliteration method

In this paper, we present a technique for finding Chinese translations for any Chinese name written in English script, more generally in any language that uses a Roman alphabet, that is, Chinese name written in Pinyin. Our system is based on a mapping table (Unihan Database[1]) between pronunciations (Pinyin) and simplified Chinese characters. From Chinese characters to Pinyin, most characters have only one pronunciation, though some characters have up to four, five pronunciations. However, from Pinyin to Chinese characters, the degree of ambiguity is much greater. Homonym is a very common phenomenon in Chinese. One Pinyin sound usually corresponds to many different Chinese ideographs (about 20 characters per Pinyin if the pinyin includes the vowel tones written as numbers, or 64 characters in average if the vowel tone is not represented as is in Chinese names written in a Roman alphabet).

Our system performs translation of Pinyin to Chinese characters in three steps which appear in the following sections. In section 2, we show how multisyllabic Pinyin is segmented and how the possible Chinese representations are produced,

1. http://www.unicode.org/charts/Unihan.html

using Unihan database. In section 3, we show how a unigram and a bigram model can filter possible Chinese translations and significantly reduce the number of possible representations to be tested. Section 4 shows how the Web can be used to validate the remaining transliteration candidates. An evaluation of our method against a golden standard is presented in Section 5. Error analysis is provided in Section 6. Conclusions are given in the last section.

2. Segmentation of multisyllabic Pinyin and Pinyin to character transliteration

The purpose of our method is to find Chinese script versions of Chinese names found in Western texts. Our method thus starts with names written in Pinyin, the common modern method for writing Chinese names in Western scripts. For each name, a syllabic Pinyin segmentation is applied. Using a mapping table between Pinyin and Chinese ideographs, we then transliterate Pinyin back into characters. These steps are described in this section.

2.1 Unihan database

Since Pinyin describes the pronunciation of a Chinese character using Roman characters, to perform back transliteration we need a mapping table between each Pinyin sound and all Chinese characters corresponding to this pronunciation. The Unihan database, prepared by Unicode consortium, is a Web resource that establishes this mapping. This database contains rich information, divided into fields, about each CJK (Chinese, Japanese and Korean) character: its unicode encoding, its pronunciation in Chinese, Japanese and Korean, its historical meaning, etc. Two fields contain Pinyin pronunciations, the fields "kMandarin" and "kHanyuPinlu." The kMandarin field exists for 25394 characters. Besides simplified Chinese characters, traditional Chinese characters and some character radicals are also covered. Since we are concerned with modern Chinese in this work, these supplemental traditional characters can introduce unnecessary ambiguities for this translation task. Instead of this richer field, we chose only characters containing the kHanyuPinlu field which is based on Modern Standard Beijing Chinese Frequency Dictionary. It contains 3800 records. For example, for the character 和, the following information is presented: its Unicode hexadecimal 548C and its four possible pronunciations in Chinese with their respective frequencies in the Beijing Chinese Frequency Dictionary corpus. We did not use these frequencies in our experiments.

U+548C kHanyuPinlu he2 (9513), huo5 (38), he5 (24), he4 (9)

In our initial experiments, we found errors were caused by missing characters in our mapping table produced from the kHanyuPinlu field which does not cover the

entire set of simplified Chinese characters. For example the common family name 崔 is missing from this list. In order to complete our mapping table, the Structural Groups Table by Mary Ansell[2], which includes Chinese characters, pinyin and GB code, was added. The entry for family name 崔 in our completed mapping table is as follows:

U+B4DE cui1 崔

The final combined list has been enlarged to 7305 entries. In this list, the number of Chinese characters corresponding to one Pinyin varies considerably. Each Pinyin corresponds to 18.5 different ideographs on average, with a maximum of 113 and minimum of 1.

2.2 Segmentation of multisyllabic Pinyin into mono-syllables

The Unihan database, augmented by the Structural Groups Table, provides a mapping from Pinyin to simplified Chinese characters. But, as Chinese names are multisyllabic, before using the mapping table constructed with the Unihan database and Structural Groups Table, we need to perform a Pinyin segmentation of the Chinese name written in Roman script.

Each Chinese character corresponds to a one-syllable Pinyin. Chinese Han personal names (family name and given name) contain usually two or three characters, rarely four characters. Non-Han minority names can be composed of more characters. In English articles, Chinese names are often presented in two major ways: family name and then given name separated by a space, or family name and given name presented as one unit. For example the name of the current Chinese Prime Minister 温家宝 is written in two ways in English journals: “Wen Jiabao” or “Wenjiabao”. Given a Pinyin name, there may be more than one way to segment it into valid Pinyin characters found in our mapping table. For example, the name “Lianhong” can have both a two-syllable segmentation “Lian hong” and a three-syllable segmentation “Li an hong”. Of the 7870 Pinyin names in our gold standard (see section 5.1 below), 382 had more than one possible segmentation, and the remaining 7488 had only one possible segmentation.

2.3 Transliteration

After the Pinyin segmentation procedure, we get one or more possible segmented Pinyin for each name. Using our mapping table between Pinyin and Chinese characters, we can obtain all Chinese character combinations possibly corresponding to the segmented Pinyin. For a multi-syllabic name, the number of translating combinations of each syllable is exponential in the number of syllables. For example,

2 http://www.dbis.ns.ca/~stirling/phonor.html

for the name of Chinese Prime Minister segmented as “Wen Jia bao”, we have 15 characters pronounced as “wen”, 22 characters pronounced as “Jia”, and 16 pronounced as “bao”. This yields 15×22×16=5280 translation candidates, a considerable number to test. The next steps allow us to eliminate some candidates using lists of family names.

2.4 Family name list

We can use a list of Chinese family names because the list is nearly closed, a few hundred names cover almost all the possibilities. In general, Chinese Han family names have only one character, while few of them contain two characters. Family names more than two characters are non-Han minority family names. With this list of common family names, we can begin the reduction of possible transliterations starting from family name part of the name, which is given first in a Chinese name. Instead of generating all possible combinations by using the mapping table, we look at first the beginning characters and compare them with our family name list. Our list of family name was constructed using the classical Chinese family name list (百家姓) which was completed by other Web information[3]. In our final list we have 595 family names, including 510 one-character family names, 68 double character ones and 17 multi-character ones. Corresponding pronunciation in Pinyin for each family name has been added by hand in the list. For example, for the name of Chinese Prime Minister “wen jia bao”, only 闻, 文 and 温 in the family name list can be pronounced as “wen”. The number of translating combinations is cut down from 15×22×16=5280 to 3×22×16=1056.

Filtering by family names reduces the number of possible translation combinations, but many names still possess tens, even hundred of thousands transliterations. We decided to further eliminate translation possibilities using an additional unigram and bigram filter suggested by (Qu *et al*., 2004) for Japanese transliterations.

3. Filtering by unigram or bigram model derived from large monolingual corpus

Since any family name can be combined with any given name, we treat these two name parts separately. We further reduce the number of candidate given names to consider with unigram and bigram models in this section. In the case that family name is not in family name list, we also use the unigram model to reduce the number of candidates.

---

3 http://www.jpwz.com/gb2312/chinese/xingshi/xingshilist.asp

3.1 Establishment of unigram and bigram models

We derived a unigram model for Chinese characters from Web in the following way. Using each character in the mapping table as separate query, we sent off a number of queries to Google. The first 100 URLs of web pages containing the Chinese character and in GB2312 encoding were collected. We also stored the page count of each character to give a rough approximation of the characters frequency. We crawled the URLs corresponding to all the characters queried and in this way, we obtained a very large monolingual Chinese corpus containing all characters in our mapping table. From this corpus, we extracted overlapping bigrams of Chinese characters and calculated their frequencies. This corpus is very diverse; we have all kinds of different web pages, commercial, educative, journalistic, etc. The texts mix Chinese, Latin characters, symbols in one or two octets, only Chinese characters bigrams have been extracted.

Besides this large diverse monolingual corpus, we also obtained a large monolingual corpus from the texts of a Chinese journal BeijingWanBao[4] for a period of one year and a half (May 2003 - Oct 2004). This corpus is much more homogeneous. We also extracted all overlapping bigrams and calculated their frequencies, combining the frequencies from both sources.

3.2 Filtering effectuated by unigram and bigram frequency

As explained previously, family names and given names are processed separately. For family name not contained in our closed list, we used the unigram model to retain the most frequent characters. For the given name, we have either one-character or two-character types. We use bigram frequency to filter two-character given names, and unigram frequency to one-character names.

4. Validation by Web

The WWW is a big, mixed-lingual corpus, and it has been shown (Qu *et al.*, 2004) to deliver better validation than a limited-sized monolingual corpus. It is often possible to find both a word and its translation on the same Web page, and person names and specialized terminology are among the most frequent mixed-lingual items. We therefore used the Web as our ultimate validation stage.

For the items on our list of bigram and unigram filtered transliteration candidates, we searched for Web pages which contained both the family and given name in Pinyin, as well as the family and given name in Chinese characters. In our queries, we separated the family name from the given name in both Pinyin and ideographs,

4 http://www.ben.com.cn/BJWB/

because a whole name is too specific, and few transliterations could be validated except for famous persons or frequently used names. The aim of our work is not to find translation for special names, which might appear in special dictionaries, or be extractable from parallel sources. Instead we try to find a best translation in ideographs for any name. It is reasonable that persons with different family names share the same given name, although this phenomenon is less prevalent in China than in Western countries. We found that searching the Web with separated names increases the recall of our method without detriment to the precision, because the pages which contain the whole name are also in our search results.

This separation is suitable for given names with two or more characters. On the contrary, one-character given names are too general to be separated as query since the most frequently used character in the Web will certainly be our answer. This is very often not what we expect for a person's given name. In consequence, we used entire Pinyin and entire name in characters as query names consisting of a family name and a one character given name.

For the previously given example of the Chinese Prime Minister, we sent the following query combinations to the Google search engine:

wen jiabao 闻 家宝	wen jiabao 闻 家抱	wen jiabao 闻 加宝
wen jiabao 文 家宝	wen jiabao 文 家暴	wen jiabao 文 加保
wen jiabao 文 加宝	wen jiabao 文 佳保	wen jiabao 文 佳宝
wen jiabao 温 家宝	wen jiabao 温 家抱	wen jiabao 温 家暴
wen jiabao 温 加宝	wen jiabao 温 加堡	wen jiabao 温 假报

……

And we obtained the following web statistics:

wen jiabao 温 家宝	7,210 pages	wen jiabao 闻 家抱	1 page
wen jiabao 文 家宝	219 pages	wen jiabao 闻 加宝	1 page
wen jiabao 闻 家宝	32 pages	wen jiabao 文 家暴	1 page
wen jiabao 温 加宝	7 pages	wen jiabao 文 佳宝	1 page
wen jiabao 文 加宝	4 pages	wen jiabao 温 家抱	1 page
wen jiabao 文 佳保	3 pages	wen jiabao 温 加堡	1 page
wen jiabao 温 家暴	3 pages	wen jiabao 温 假报	1 page
wen jiabao 文 加保	2 pages		

Among these research statistics, we have the greatest number of answer pages containing the combination of "wen jiabao 温家宝", and 温家宝 is the correct translation for the name of the Chinese Prime Minister.

5. Evaluation

With an independent gold standard, we have evaluated our method with the classic measures: precision, recall and F-measure.

5.1 Creation of gold standard

We composed a gold standard of Chinese names and their Pinyin transcriptions in the following way. From the Web, we collected different name lists written in Chinese characters: student names, journalist names[5], lawyer names[6] and some leader names[7]. We collected 7870 names in all. Each Chinese name on this list was then automatically transliterated into Pinyin using the DimSum Chinese Tools[8]. The results of this automatic transliteration contained some errors, since it is difficult to choose the correct pronunciation out of context for multiple pronunciation characters (Bao, 1999). However, characters which can be both family name and common word have fixed pronunciations when they are used as family names. We first corrected these errors so that family names produced only Pinyin corresponding to these fixed pronunciations. Then we manually verified and corrected the cases in which the DimSum translation produced a Chinese character-Pinyin mapping missing from the enlarged Unihan mapping (see section 2.1). At this stage, we then had a verified list of 7870 pairs of Pinyin and Chinese character names. We used this list to test the recall of our method, since our system should be able to find every one of these pairs. It is not possible to use this gold standard list to test the precision of our method because a name written in Pinyin can have more than one transliteration in Chinese characters. Our system described in sections 2 and 3 takes one Pinyin name in input and produces a ranked list of Chinese character names in output.

5.2 Evaluation measures

All results were evaluated with the family name and given name as an entire word. We used the following classic measures.

Precision: number of correct Chinese translations / total number of obtained translations.

Recall: number of correct Chinese translations / total number of translations in gold standard.

5.3 Evaluation of translation

In our gold standard, each Pinyin entry corresponds to only one Chinese translation. In reality one name in Pinyin can be translated in many different ways. In order to measure precision, since we lacked all the possible correct translations in the gold standard, we simulated a complete gold standard in the following way: if a

5 http://www.xinhuanet.com/reporter/list1.htm
6 http://www.zhls.net/lawyer/lawyer2.asp
7 http://news.xinhuanet.com/misc/2002-11/15/content_630633.htm
8 DimSum Chinese Tools v0.7.2, http://www.mandarintools.com/dimsum.html

Chinese character corresponding to a family name followed by one or two Chinese characters was found on the Web as a contiguous unit, then that entire sequence name was considered as a valid name for our tests.

We chose every 7th name from the gold standard to test our approach, 1124 Pinyin names in all. Each name was transcribed into a ranked list of Chinese names as described in sections 2 to 4, that is, (i) the pinyin was segmented, (ii) the possible Chinese sequences corresponding to the Pinyin were generated, (iii) these sequences were ranked using bigrams and unigram frequencies, (iv) the top 1000 frequent Chinese sequences were retained, (v) each sequence and its segmented Pinyin was queried on the Web (with family and given names separated), (vi) the Pinyin and Chinese sequences was scored according to the number of pages in which the terms were found, (vii) the Pinyin name and its candidate Chinese transliterations were ranked in descending order according to this score, as shown in section 4.

To measure precision, we then queried Google again for the whole Chinese translation (family name and given name with no spaces between the characters). If some Web pages containing this Chinese word were found, then we considered the Chinese word as a valid transcription of the original Pinyin name. Otherwise, if the entire name was not found, the transcription was considered as incorrect. For the list of 1124 names tested, the highest ranking Pinyin-Chinese combination (using separated family and given names) yielded a valid name in 79.8% of the cases. In this list of validated words we find 32.7% of the gold standard Pinyin-Chinese pairs. In few cases, only the highest ranking Pinyin-Chinese combination was the only combination found. For the words in which more than one combination was found, if we include the second most frequent combination, the precision decreases to 76.9%, and the recall over the gold standard rises to 45.7%.

We remark that the precision decreases and the recall increases by taking one more possible transliteration validated by the Web. And the Precision is much higher than recall. This means that we can propose reliable ideograph presentations for each Pinyin, but it is difficult to find comparatively rare transliterations.

6. Discussion

6.1 Bigram model parameters

After Pinyin to character transliteration, unlikely translating candidates are filtered out by bigram and unigram frequency. Each candidate is scored by the frequency of the bigram or unigram that composes it. We did not have any three character names (that is, either given name or family name) to consider. In order to determine how far down the ranked list of bigram candidates to go, we studied how many bigrams would have to be considered in order to find the correct translation in our gold standard. Considering all 7870 entries of our gold standard, by varying the number

of bigrams taken from the 30 most frequent to the 1600 most frequent, we can get an idea of how many alternatives we have to consider. The result is showed in Table 1.

Nb of bigrams	Coverage of good transl.	Coverage percentage
30	4293	54.54
50	4561	57.94
100	4943	62.79
200	5469	69.47
300	5855	74.37
400	6127	77.83
500	6364	80.84
600	6527	82.91
700	6651	84.48
800	6746	85.69
900	6834	86.81
1000	6897	87.61
1200	7003	88.96
1500	7100	90.19
1600	7128	90.54

Table 1. *Relationship between number of bigrams and the coverage of good transliterations.*

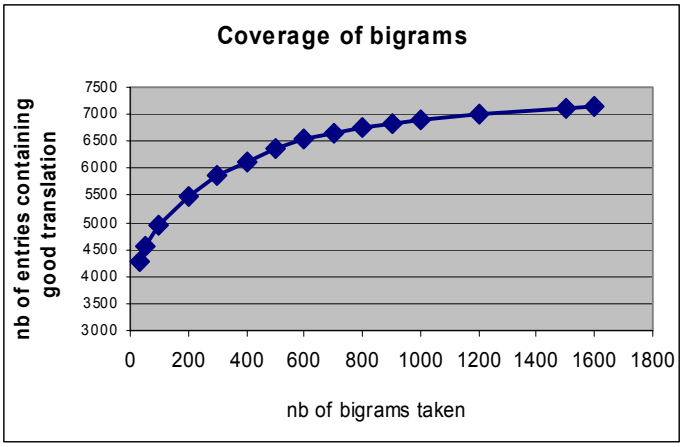

Figure1. *Bigram coverage*

The above curve based on this table in Figure 1 shows that at the beginning coverage of correct transliterations increases greatly with the augmentation of the number of bigrams taken. When bigram number is greater than 1000, the curve slope becomes smaller. The curve tends to be stable.

The result presented in section 5 was issued of a test effectuated with the threshold 1000.

6.2 Web validation and filtering by bigrams

We have discussed the relationship between number of bigrams taken and the coverage of good transliteration. To economize search time, it is very interesting to know the relationship between the performance of the Web validation and the bigram model. If Web validates with the same performance as the bigram model, with the Web ranking results in relation to the frequent bigrams, then we do not need to consider as many bigrams while still retaining a good performance, because the bigrams proposed with relatively little frequencies will be eliminated later by the Web. On the contrary, if Web validation is not related to our corpus bigram frequencies, it would be worth passing more time to search the translating candidates with less frequent bigrams, because they could be the good transliteration.

The distribution of bigrams validated by Web is presented in Table 2 as follows:

rank of bigram validated	nb of entries	accumulated percentage
=1	280	20.38%
<=10	462	54.0%
<=100	257	72.71%
<=1000	257	94.47%
>1000	76	100%

Table 2. D*istribution of bigrams*

We observe that for 20.38% of the cases, we validate the first most frequent corpus bigram by the Web statistics, for 54.0% we take our choice inside the first ten most frequent bigrams, for 72.71% in the first 100, 94.47% in the first 1000. When the precision of 72.71% can satisfy the request of an application, 100 most frequent bigrams are enough to obtain the desired result. In this case Web validation for other less frequent bigrams is not necessary.

6.3 Role of family name list

The list of family names was used for cutting down the number of translating combinations to consider in our method. At the same time, using this list improves the result. Unigram or bigram model filtering and Web validation favour more or less frequent characters. However, family names are not always very frequent. Using this list we have more chances to find the good transliteration of family names. For example, for Pinyin "Shi", we have 69 corresponding ideographs among which "是" (verb "be") is a very frequent character. By taking all 69 possible characters, we got it as transliteration of family name, but it cannot be a family name. However, for "Shi" we have five possible corresponding characters in our list of family names: 石, 史,时,师 and 施. Exploiting this information about family name transliteration helps us find good transliterations. If we apply the same method to transcriptions of names other than family names (for example, Chinese place names) we may be able to use other indicators (for example, mountain, lake, city, etc.) to perform the same function as family names in our described method.

6.4 Origin of Errors

On analysis, we can categorize the cause of errors into two types.

First, although we have enriched our mapping table with Structural Groups Table on basis of the Unihan database, there are still some missing characters. These characters are often rare, but some of them appear quite frequently in Chinese person's names. Besides these missing characters, our mapping table contains "wrong" Pinyin for certain characters. For example, for character 思 (si1), we have "sai1" and "sai5" as Pinyin. Some other resources can be use to complete and "correct" our mapping table for the use of this work.

Second, in our study, we use all bigrams from the corpus containing both proper names bigrams and common word bigrams. This improves our chances of finding the transliteration for a given name in Pinyin, as long as one given name can be pronounced in the same way as a very popular common word. Obviously, this fact favors the selection of frequent common noun as transliteration. Some of these common words can be given names, but some of them cannot. For instance, 光明 which means *brightness* is a frequent word. It can be a given name in proper name for example as 张光明. On the contrary, 练习 meaning *exercise* is also a frequent bigram (word). It has logically been selected as given name for Pinyin "Lian xi", while it is not an actual given name. The good transliteration should probably be 莲喜. In this case, a list of characters frequently used in given names could help improve our results. If the first ranking candidate contains characters out of this list, one more possible transliteration should be considered as a good transliteration, and a human user can be the final judge.

7. Conclusion

In this study, we have examined a technique to transliterate Chinese names written in a Roman alphabet script (in Pinyin) back into Chinese ideographs. The technique uses a mapping table between sound in Pinyin and their corresponding ideographs to effectuate the back transliteration. Due to numerous homonyms, the number of proposed translation combinations is exponential in the number of characters in the given name. A list of family names has been employed to reduce the number of unnecessary translation combinations and also to favor the choice of correct transliteration for family names. For the given name, models of bigram and unigram serve to filter out unlike combinations. Finally, candidates selected by unigram and bigram frequencies are validated by Web statistics. We class our candidates according to the number of pages which contain both their Roman alphabet script and possible ideographs representations. In this work, we obtain a precision of 79.8%, and a recall of 32.7% using the most frequent combination found on the Web.

Future work will involve improving transliterations for given names which are pronounced in the same way as some frequent common words by limiting the character set available for Chinese proper names. As the tests we performed were mostly with Chinese Han names, we will study the performance of our method for Chinese minority non-Han names. Chinese has two writing systems, that is traditional Chinese and simplified Chinese. Besides Pinyin, several different Romanization systems exist for Chinese, such as Wade-Giles, Tongyong Pinyin for mandarin and Penkyamp for Cantonese. It will be interesting to test our system using different mapping tables between Chinese characters (both traditional Chinese and Simplified Chinese) and Latin scripts of other Romanization systems. In the future, we are planning to extend our work by testing its results in real cross lingual information retrieval applications.

8. Reference

Bao J., Arranging Polysyllabic Chinese Characters in Phonetic Alphabet Order with a Semi-automatic Computer Program, *Proceedings of the 1st ASIALEX Regional Symposium*, 1999.

Feng D., Lv Y., Zhou M., A New Approach for English-Chinese Named Entity Alignment, *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP*, Barcelona, Spain, July 25-26 2004.

Gao W., Wong K., Lam W., Phoneme-based Transliteration of Foreign Names for OOV Problem, *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, Hainan, China, 2004.

Grefenstette, G., Evaluating the Adequacy of a Multilingual Transfer Dictionary for the Cross Language Information Retrieval, *proceedings of* the *1st International Conference on Language Resources and Evaluation,* pages 755–758, May 1998.

Huang, F. and Vogel S., Improved Named Entity Translation and Bilingual Named Entity Extraction, *proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI'02)*, Pittsburgh, Pennsylvania, 2002.

Hull, David A., Grefenstette, G., Experiments In Multilingual Information Retrieval, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

Lee C., Chang J, Parallel- Aligned Texts using a Statistical Machine Transliteration Model, *Proceedings of HLT-NAACL*, Edmonton, Canada, May 27 - June 1 2003.

Levow G-A., Oard DW., Evaluating lexicon coverage for cross-language information retrieval, *Proceedings of Workshop on Multilingual Information Processing and Asian Language Processing*, 1999.

Li H, Zhang M., Su J., A Joint Source-Channel Model for Machine Transliteration, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, 25-26 July 2004.

Meng H., Lo W., Chen B., Tang K., Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval, *Proceedings of Automatic Speech Recognition and Understanding Workshop ASRU*, Madonna di Campiglio Trento Italy, December 9-13 2001.

Qu, Y., Grefenstette, G., Finding Ideographic Representations Of Japanese Names Written In Latin Script Via Language Identification And Corpus Validation, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona. July 21-26, 2004.

Virga P., Khudanpur S., Transliteration of Proper Names in Cross-lingual Information Retrieval, *Proceedings of the ACL Workshop on Multi-lingual Named Entity Recognition* Combining Statistical and Symbolic Models, Sapporo, Japan, July 12 2003.

Wan S., Verspoor C., Automatic English-Chinese Name Transliteration for Development of Multilingual Resources, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics(ACL)*, Montreal, Quebec, Canada, 1998.