Semantic Transliteration of Personal Names

Haizhou Li*, Khe Chai Sim*, Jin-Shea Kuo†, Minghui Dong*

*Institute for Infocomm Research
Singapore 119613
{hli,kcsim,mhdong}@i2r.a-star.edu.sg

†Chung-Hwa Telecom Laboratories
Taiwan
jskuo@cht.com.tw

Abstract

Words of foreign origin are referred to as borrowed words or loanwords. A loanword is usually imported to Chinese by phonetic transliteration if a translation is not easily available. Semantic transliteration is seen as a good tradition in introducing foreign words to Chinese. Not only does it preserve how a word sounds in the source language, it also carries forward the word's original semantic attributes. This paper attempts to automate the semantic transliteration process for the first time. We conduct an inquiry into the feasibility of semantic transliteration and propose a probabilistic model for transliterating personal names in Latin script into Chinese. The results show that semantic transliteration substantially and consistently improves accuracy over phonetic transliteration in all the experiments.

1 Introduction

The study of Chinese transliteration dates back to the seventh century when Buddhist scriptures were translated into Chinese. The earliest bit of Chinese translation theory related to transliteration may be the principle of "Names should follow their bearers, while things should follow Chinese." In other words, names should be transliterated, while things should be translated according to their meanings. The same theory still holds today.

Transliteration has been practiced in several ways, including phonetic transliteration and phonetic-semantic transliteration. By phonetic transliteration, we mean rewriting a foreign word in native grapheme such that its original pronunciation is preserved. For example, *London* becomes 伦敦 /Lun-Dun/[1] which does not carry any clear connotations. Phonetic transliteration represents the common practice in transliteration. Phonetic-semantic transliteration, hereafter referred to as semantic transliteration for short, is an advanced translation technique that is considered as a recommended translation practice for centuries. It translates a foreign word by preserving both its original pronunciation and meaning. For example, Xu Guangqi[2] translated *geo-* in *geometry* into Chinese as 几何 /Ji-He/, which carries the pronunciation of *geo-* and expresses the meaning of "a science concerned with measuring the earth".

Many of the loanwords exist in today's Chinese through semantic transliteration, which has been well received (Hu and Xu, 2003; Hu, 2004) by the people because of many advantages. Here we just name a few. (1) It brings in not only the sound, but also the meaning that fills in the semantic blank left by phonetic transliteration. This also reminds people that it is a loanword and avoids misleading; (2) It provides etymological clues that make it easy to trace back to the root of the words. For example, a transliterated Japanese name will maintain its Japanese identity in its Chinese appearance; (3) It evokes desirable associations, for example, an English girl's name is transliterated with Chinese characters that have clear feminine association, thus maintaining the gender identity.

[1] Hereafter, Chinese characters are also denoted in *Pinyin* romanization system, for ease of reference.

[2] Xu Quangqi (1562–1633) translated The Original Manuscript of Geometry to Chinese jointly with Matteo Ricci.

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 120–127,
Prague, Czech Republic, June 2007.

Unfortunately, most of the reported work in the area of machine transliteration has not ventured into semantic transliteration yet. The Latin-scripted personal names are always assumed to homogeneously follow the English phonic rules in automatic transliteration (Li et al., 2004). Therefore, the same transliteration model is applied to all the names indiscriminatively. This assumption degrades the performance of transliteration because each language has its own phonic rule and the Chinese characters to be adopted depend on the following semantic attributes of a foreign name.

(1) Language of origin: An English word is not necessarily of pure English origin. In English news reports about Asian happenings, an English personal name may have been originated from Chinese, Japanese or Korean. The language origin affects the phonic rules and the characters to be used in transliteration[3]. For example, a Japanese name *Matsumoto* should be transliterated as 松本 /Song-Ben/, instead of 马茨莫托 /Ma-Ci-Mo-Tuo/ as if it were an English name.

(2) Gender association: A given name typically implies a clear gender association in both the source and target languages. For example, the Chinese transliterations of *Alice* and *Alexandra* are 爱丽丝 /Ai-Li-Si/ and 亚历山大 /Ya-Li-Shan-Da/ respectively, showing clear feminine and masculine characteristics. Transliterating *Alice* as 埃里斯 /Ai-Li-Si/ is phonetically correct, but semantically inadequate due to an improper gender association.

(3) Surname and given name: The Chinese name system is the original pattern of names in Eastern Asia such as China, Korea and Vietnam, in which a limited number of characters[4] are used for surnames while those for given names are less restrictive. Even for English names, the character set for given name transliterations are different from that for surnames.

Here are two examples of semantic transliteration for personal names. *George Bush* and *Yamamoto Akiko* are transliterated into 乔治⊔布什 and 山本 亚喜子 that arouse to the following associations: 乔治 /Qiao-Zhi/ - male given name, English origin; 布什 /Bu-Shi/ - surname, English origin; 山本 /Shan-Ben/ - surname, Japanese origin; 亚喜子 /Ya-Xi-Zi/ - female given name, Japanese origin.

In Section 2, we summarize the related work. In Section 3, we discuss the linguistic feasibility of semantic transliteration for personal names. Section 4 formulates a probabilistic model for semantic transliteration. Section 5 reports the experiments. Finally, we conclude in Section 6.

2 Related Work

In general, computational studies of transliteration fall into two categories: transliteration modeling and extraction of transliteration pairs. In transliteration modeling, transliteration rules are trained from a large, bilingual transliteration lexicon (Lin and Chen, 2002; Oh and Choi, 2005), with the objective of translating unknown words on the fly in an open, general domain. In the extraction of transliterations, data-driven methods are adopted to extract actual transliteration pairs from a corpus, in an effort to construct a large, up-to-date transliteration lexicon (Kuo et al., 2006; Sproat et al., 2006).

Phonetic transliteration can be considered as an extension to the traditional grapheme-to-phoneme (G2P) conversion (Galescu and Allen, 2001), which has been a much-researched topic in the field of speech processing. If we view the grapheme and phoneme as two symbolic representations of the same word in two different *languages*, then G2P is a transliteration task by itself. Although G2P and phonetic transliteration are common in many ways, transliteration has its unique challenges, especially as far as E-C transliteration is concerned. E-C transliteration is the conversion between English graphemes, phonetically associated English letters, and Chinese graphemes, characters which represent ideas or meanings. As a Chinese transliteration can arouse to certain connotations, the choice of Chinese characters becomes a topic of interest (Xu et al., 2006).

Semantic transliteration can be seen as a subtask of statistical machine translation (SMT) with

[3] In the literature (Knight and Graehl,1998; Qu et al., 2003), translating romanized Japanese or Chinese names to Chinese characters is also known as *back-transliteration*. For simplicity, we consider all conversions from Latin-scripted words to Chinese as transliteration in this paper.

[4] The 19 most common surnames cover 55.6% percent of the Chinese population (Ning and Ning 1995).

monotonic word ordering. By treating a letter/character as a word and a group of letters/characters as a phrase or token unit in SMT, one can easily apply the traditional SMT models, such as the IBM generative model (Brown et al., 1993) or the phrase-based translation model (Crego et al., 2005) to transliteration. In transliteration, we face similar issues as in SMT, such as lexical mapping and alignment. However, transliteration is also different from general SMT in many ways. Unlike SMT where we aim at optimizing the semantic transfer, semantic transliteration needs to maintain the phonetic equivalence as well.

In computational linguistic literature, much effort has been devoted to phonetic transliteration, such as English-Arabic, English-Chinese (Li et al., 2004), English-Japanese (Knight and Graehl, 1998) and English-Korean. In G2P studies, Font Llitjos and Black (2001) showed how knowledge of language of origin may improve conversion accuracy. Unfortunately semantic transliteration, which is considered as a good tradition in translation practice (Hu and Xu, 2003; Hu, 2004), has not been adequately addressed computationally in the literature. Some recent work (Li et al., 2006; Xu et al., 2006) has attempted to introduce preference into a probabilistic framework for selection of Chinese characters in phonetic transliteration. However, there is neither analytical result nor semantic-motivated transliteration solution being reported.

3 Feasibility of Semantic Transliteration

A Latin-scripted personal name is written in letters, which represent the pronunciations closely, whereas each Chinese character represents not only the syllables, but also the semantic associations. Thus, character rendering is a vital issue in transliteration. Good transliteration adequately projects semantic association while an inappropriate one may lead to undesirable interpretation.

Is semantic transliteration possible? Let's first conduct an inquiry into the feasibility of semantic transliteration on 3 bilingual name corpora, which are summarizied in Table 1 and will be used in experiments. E-C corpus is an augmented version of *Xinhua* English to Chinese dictionary –for English names (Xinhua, 1992). J-C corpus is a romanized Japanese to Chinese dictionary for Japanese names. The C-C corpus is a Chinese *Pinyin* to character dictionary for Chinese names. The entries are classified into surname, male and female given name categories. The E-C corpus also contains some entries without gender/surname labels, referred to as *unclassified*.

	E-C	J-C[5]	C-C[6]
Surname (S)	12,490	36,352	569,403
Given name (M)	3,201	35,767	345,044
Given name (F)	4,275	11,817	122,772
Unclassified	22,562	-	-
All	42,528	83,936	1,972,851

Table 1: Number of entries in 3 corpora

Phonetic transliteration has not been a problem as Chinese has over 400 unique syllables that are enough to approximately transcribe all syllables in other languages. Different Chinese characters may render into the same syllable and form a range of homonyms. Among the homonyms, those arousing positive meanings can be used for personal names. As discussed elsewhere (Sproat et al., 1996), out of several thousand common Chinese characters, a subset of a few hundred characters tends to be used overwhelmingly for transliterating English names to Chinese, e.g. only 731 Chinese characters are adopted in the E-C corpus. Although the character sets are shared across languages and genders, the statistics in Table 2 show that each semantic attribute is associated with some unique characters. In the C-C corpus, out of the total of 4,507 characters, only 776 of them are for surnames. It is interesting to find that female given names are represented by a smaller set of characters than that for male across 3 corpora.

	E-C	J-C	C-C	All
S	327	2,129	776	2,612 (19.2%)
M	504	1,399	4,340	4,995 (20.0%)
F	479	1,178	1,318	2,192 (26.3%)
All	731 (44.2%)	2,533 (46.2%)	4,507 (30.0%)	5,779 (53.6%)

Table 2: Chinese character usage in 3 corpora. The numbers in brackets indicate the percentage of characters that are shared by at least 2 corpora.

Note that the overlap of Chinese characters usage across genders is higher than that across languages. For instance, there is a 44.2% overlap

[5] http://www.cjk.org

[6] http://technology.chtsai.org/namelist

across gender for the transcribed English names; but only 19.2% overlap across languages for the surnames.

In summary, the semantic attributes of personal names are characterized by the choice of characters, and therefore their *n*-gram statistics as well. If the attributes are known in advance, then the semantic transliteration is absolutely feasible. We may obtain the semantic attributes from the context through trigger words. For instance, from "Mr *Tony Blair*", we realize "*Tony*" is a male given name while "*Blair*" is a surname; from "Japanese Prime Minister *Koizumi*", we resolve that "*Koizumi*" is a Japanese surname. In the case where contextual trigger words are not available, we study detecting the semantic attributes from the personal names themselves in the next section.

4 Formulation of Transliteration Model

Let *S* and *T* denote the name written in the source and target writing systems respectively. Within a probabilistic framework, a transliteration system produces the optimum target name, T^*, which yields the highest posterior probability given the source name, *S*, *i.e.*

$$T^* = \underset{T \in \mathcal{T}_S}{\arg\max}\, P(T \mid S) \qquad (1)$$

where $\mathcal{T}_S$ is the set of all possible transliterations for the source name, *S*. The alignment between *S* and *T* is assumed implicit in the above formulation. In a standard phonetic transliteration system, $P(T \mid S)$, the posterior probability of the hypothesized transliteration, *T*, given the source name, *S*, is directly modeled without considering any form of semantic information. On the other hand, semantic transliteration described in this paper incorporates language of origin and gender information to capture the semantic structure. To do so, $P(T \mid S)$ is rewritten as

$$\begin{aligned} P(T \mid S) &= \sum_{L \in \mathcal{L}, G \in \mathcal{G}} P(T, L, G \mid S) && (2) \\ &= \sum_{L \in \mathcal{L}, G \in \mathcal{G}} P(T \mid S, L, G) P(L, G \mid S) && (3) \end{aligned}$$

where $P(T \mid S, L, G)$ is the transliteration probability from source *S* to target *T*, given the language of origin (*L*) and gender (*G*) labels. $\mathcal{L}$ and $\mathcal{G}$ denote the sets of languages and genders respectively. $P(L, G \mid S)$ is the probability of the language and the gender given the source, *S*.

Given the alignment between *S* and *T*, the transliteration probability given *L* and *G* may be written as

$$\begin{aligned} P(T \mid S, L, G) &= \prod_{i=1}^{I} P(t_i \mid T_1^{i-1}, S_1^i) && (4) \\ &\approx \prod_{i=1}^{I} P(t_i \mid t_{i-1}, s_{i-1}, s_i) && (5) \end{aligned}$$

where s_i and t_i are the i^{th} token of *S* and *T* respectively and *I* is the total number of tokens in both *S* and *T*. S_j^k and T_j^k represent the sequence of tokens $\left(s_j, s_{j+1}, \ldots, s_k\right)$ and $\left(t_j, t_{j+1}, \ldots, t_k\right)$ respectively. Eq. (4) is in fact the *n*-gram likelihood of the token pair $\langle t_i, s_i \rangle$ sequence and Eq. (5) approximates this probability using a bigram language model. This model is conceptually similar to the joint source-channel model (Li et al., 2004) where the target token t_i depends on not only its source token s_i but also the history t_{i-1} and s_{i-1}. Each character in the target name forms a token. To obtain the source tokens, the source and target names in the training data are aligned using the EM algorithm. This yields a set of possible source tokens and a mapping between the source and target tokens. During testing, each source name is first segmented into all possible token sequences given the token set. These source token sequences are mapped to the target sequences to yield an *N*-best list of transliteration candidates. Each candidate is scored using an *n*-gram language model given by Eqs. (4) or (5).

As in Eq. (3), the transliteration also greatly depends on the prior knowledge, $P(L, G \mid S)$. When no prior knowledge is available, a uniform probability distribution is assumed. By expressing $P(L, G \mid S)$ in the following form,

$$P(L, G \mid S) = P(G \mid L, S) P(L \mid S) \qquad (6)$$

prior knowledge about language and gender may be incorporated. For example, if the language of *S* is known as L_S, we have

$$P(L \mid S) = \begin{cases} 1 & L = L_S \\ 0 & L \neq L_S \end{cases} \qquad (7)$$

Similarly, if the gender information for *S* is known as G_S, then,

$$P(G \mid L,S)=\begin{cases}1 & G=G_S\\ 0 & G\neq G_S\end{cases} \tag{8}$$

Note that personal names have clear semantic associations. In the case where the semantic attribute information is not available, we propose learning semantic information from the names themselves. Using Bayes' theorem, we have

$$P(L,G \mid S)=\frac{P(S \mid L,G)P(L,G)}{P(S)} \tag{9}$$

$P(S \mid L,G)$ can be modeled using an *n*-gram language model for the letter sequence of all the Latin-scripted names in the training set. The prior probability, $P(L,G)$, is typically uniform. $P(S)$ does not depend on *L* and *G*, thus can be omitted.

Incorporating $P(L,G \mid S)$ into Eq. (3) can be viewed as performing a soft decision of the language and gender semantic attributes. By contrast, hard decision may also be performed based on maximum likelihood approach:

$$\overline{L}_S=\underset{L\in\mathcal{L}}{\arg\max}\,P(S \mid L) \tag{10}$$

$$\overline{G}_S=\underset{G\in\mathcal{G}}{\arg\max}\,P(S \mid L,G) \tag{11}$$

where $\overline{L}_S$ and $\overline{G}_S$ are the detected language and gender of *S* respectively. Therefore, for hard decision, $P(L,G \mid S)$ is obtained by replacing L_S and G_S in Eq. (7) and (8) with $\overline{L}_S$ and $\overline{G}_S$ respectively. Although hard decision eliminates the need to compute the likelihood scores for all possible pairs of *L* and *G*, the decision errors made in the early stage will propagate to the transliteration stage. This is potentially bad if a poor detector is used (see Table 9 in Section 5.3).

If we are unable to model the prior knowledge of semantic attributes $P(L,G \mid S)$, then a more general model will be used for $P(T \mid S,L,G)$ by dropping the dependency on the information that is not available. For example, Eq. (3) is reduced to $\sum_{L\in\mathcal{L}} P(T \mid S,L)P(L \mid S)$ if the gender information is missing. Note that when both language and gender are unknown, the system simplifies to the baseline phonetic transliteration system.

5 Experiments

This section presents experiments on database of 3 language origins (Japanese, Chinese and English) and gender information (surname[7], male and female). In the experiments of determining the language origin, we used the full data set for the 3 languages as in shown in Table 1. The training and test data for semantic transliteration are the subset of Table 1 comprising those with surnames, male and female given names labels. In this paper, J, C and E stand for Japanese, Chinese and English; S, M and F represent Surname, Male and Female given names, respectively.

L	Data set	# unique entries			
		S	M	F	All
J	Train	21.7k	5.6k	1.7k	27.1k
	Test	2.6k	518	276	2.9k
C	Train	283	29.6k	9.2k	31.5k
	Test	283	2.9k	1.2k	3.1k
E	Train	12.5k	2.8k	3.8k	18.5k
	Test	1.4k	367	429	2.1k

Table 3: Number of unique entries in training and test sets, categorized by semantic attributes

Table 3 summarizes the number of *unique*[8] name entries used in training and testing. The test sets were randomly chosen such that the amount of test data is approximately 10-20% of the whole corpus. There were no overlapping entries between the training and test data. Note that the Chinese surnames are typically single characters in a small set; we assume there is no unseen surname in the test set. All the Chinese surname entries are used for both training and testing.

5.1 Language of Origin

For each language of origin, a 4-gram language model was trained for the letter sequence of the source names, with a 1-letter shift.

Japanese	Chinese	English	All
96.46	96.44	89.90	94.81

Table 4: Language detection accuracies (%) using a 4-gram language model for the letter sequence of the source name in Latin script.

[7] In this paper, surnames are treated as a special class of gender. Unlike given names, they do not have any gender association. Therefore, they fall into a third category which is neither male nor female.

[8] By contrast, Table 1 shows the total number of name examples available. For each unique entry, there may be multiple examples.

Table 4 shows the language detection accuracies for all the 3 languages using Eq. (10). The overall detection accuracy is 94.81%. The corresponding Equal Error Rate (EER)[9] is 4.52%. The detection results may be used directly to infer the semantic information for transliteration. Alternatively, the language model likelihood scores may be incorporated into the Bayesian framework to improve the transliteration performance, as described in Section 4.

5.2 Gender Association

Similarly, gender detection[10] was performed by training a 4-gram language model for the letter sequence of the source names for each language and gender pair.

Language	Male	Female	All
Japanese	90.54	80.43	87.03
Chinese	64.34	71.66	66.52
English	75.20	72.26	73.62

Table 5: Gender detection accuracies (%) using a 4-gram language model for the letter sequence of the source name in Latin script.

Table 5 summarizes the gender detection accuracies using Eq. (11) assuming language of origin is known, $\overline{G_S} = \arg\max_{G \in \mathcal{G}} P(S \mid L = L_S, G)$. The overall detection accuracies are 87.03%, 66.52% and 73.62% for Japanese, Chinese and English respectively. The corresponding EER are 13.1%, 21.8% and 19.3% respectively. Note that gender detection is generally harder than language detection. This is because the tokens (syllables) are shared very much across gender categories, while they are quite different from one language to another.

5.3 Semantic Transliteration

The performance was measured using the Mean Reciprocal Rank (MRR) metric (Kantor and Voorhees, 2000), a measure that is commonly used in information retrieval, assuming there is precisely one correct answer. Each transliteration system generated at most 50-best hypotheses for each word when computing MRR. The word and character accuracies of the top best hypotheses are also reported.

We used the phonetic transliteration system as the baseline to study the effects of semantic transliteration. The phonetic transliteration system was trained by pooling all the available training data from all the languages and genders to estimate a language model for the source-target token pairs. Table 6 compares the MRR performance of the baseline system using unigram and bigram language models for the source-target token pairs.

	J	C	E	All
Unigram	0.5109	0.4869	0.2598	0.4443
Bigram	0.5412	0.5261	0.3395	0.4895

Table 6: MRR performance of phonetic transliteration for 3 corpora using unigram and bigram language models.

The MRR performance for Japanese and Chinese is in the range of 0.48-0.55. However, due to the small amount of training and test data, the MRR performance of the English name transliteration is slightly poor (approximately 0.26-0.34). In general, a bigram language model gave an overall relative improvement of 10.2% over a unigram model.

L	G	Set	J	C	E
×	×	S	0.5366	0.7426	0.4009
		M	0.5992	0.5184	0.2875
		F	0.4750	0.4945	0.1779
		All	0.5412	0.5261	0.3395
✓	×	S	0.6500	0.7971	0.7178
		M	0.6733	0.5245	0.4978
		F	0.5956	0.5191	0.4115
		All	0.6491	0.5404	0.6228
	✓	S	0.6822	0.9969	0.7382
		M	0.7267	0.6466	0.4319
		F	0.5856	0.7844	0.4340
		All	**0.6811**	**0.7075**	**0.6294**
○	○	S	0.6541	0.6733	0.7129
		M	0.6974	0.5362	0.4821
		F	0.5743	0.6574	0.4138
		All	0.6477	0.5764	0.6168

Table 7: The effect of language and gender information on the overall MRR performance of transliteration (L=Language, G=Gender, ×=unknown, ✓=known, ○=soft decision).

Next, the scenarios with perfect language and/or gender information were considered. This com-

[9] EER is defined as the error of false acceptance and false rejection when they are equal.

[10] In most writing systems, the ordering of surname and given name is known. Therefore, gender detection is only performed for male and female classes.

parison is summarized in Table 7. All the MRR results are based on transliteration systems using bigram language models. The table clearly shows that having perfect knowledge, denoted by "✓", of language and gender helps improve the MRR performance; detecting semantic attributes using soft decision, denoted by "○", has a clear win over the baseline, denoted by "×", where semantic information is not used. The results strongly recommend the use of semantic transliteration for personal names in practice.

Next let's look into the effects of automatic language and gender detection on the performance.

	J	C	E	All
×	0.5412	0.5261	0.3395	0.4895
◇	0.6292	0.5290	0.5780	0.5734
○	**0.6162**	**0.5301**	**0.6088**	**0.5765**
✓	0.6491	0.5404	0.6228	0.5952

Table 8: The effect of language detection schemes on MRR using bigram language models and unknown gender information (hereafter, ×=unknown, ✓=known, ◇=hard decision, ○=soft decision).

Table 8 compares the MRR performance of the semantic transliteration systems with different prior information, using bigram language models. Soft decision refers to the incorporation of the language model scores into the transliteration process to improve the prior knowledge in Bayesian inference. Overall, both hard and soft decision methods gave similar MRR performance of approximately 0.5750, which was about 17.5% relatively improvement compared to the phonetic transliteration system with 0.4895 MRR. The hard decision scheme owes its surprisingly good performance to the high detection accuracies (see Table 4).

	S	M	F	All
×	0.6825	0.5422	0.5062	0.5952
◇	0.7216	0.4674	0.5162	0.5855
○	**0.7216**	**0.5473**	**0.5878**	**0.6267**
✓	0.7216	0.6368	0.6786	0.6812

Table 9: The effect of gender detection schemes on MRR using bigram language models with perfect language information.

Similarly, the effect of various gender detection methods used to obtain the prior information is shown in Table 9. The language information was assumed known *a-priori*. Due to the poorer detection accuracy for the Chinese male given names (see Table 5), hard decision of gender had led to deterioration in MRR performance of the male names compared to the case where no prior information was assumed. Soft decision of gender yielded further gains of 17.1% and 13.9% relative improvements for male and female given names respectively, over the hard decision method.

L	G	MRR	Overall Accuracy (%)	
			Word	Character
×	×	0.4895	36.87	58.39
✓	×	0.5952	46.92	65.18
	✓	0.6812	58.16	70.76
◇	◇	0.5824	47.09	66.84
○	○	**0.6122**	**49.38**	**69.21**

Table 10: Overall transliteration performance using bigram language model with various language and gender information.

Finally, Table 10 compares the performance of various semantic transliteration systems using bigram language models. The baseline phonetic transliteration system yielded 36.87% and 58.39% accuracies at word and character levels respectively; and 0.4895 MRR. It can be conjectured from the results that semantic transliteration is substantially superior to phonetic transliteration. In particular, knowing the language information improved the overall MRR performance to 0.5952; and with additional gender information, the best performance of 0.6812 was obtained. Furthermore, both hard and soft decision of semantic information improved the performance, with the latter being substantially better. Both the word and character accuracies improvements were consistent and have similar trend to that observed for MRR.

The performance of the semantic transliteration using soft decisions (last row of Table 10) achieved 25.1%, 33.9%, 18.5% relative improvement in MRR, word and character accuracies respectively over that of the phonetic transliteration (first row of Table 10). In addition, soft decision also presented 5.1%, 4.9% and 3.5% relative improvement over hard decision in MRR, word and character accuracies respectively.

5.4 Discussions

It was found that the performance of the baseline phonetic transliteration may be greatly improved by incorporating semantic information such as the language of origin and gender. Furthermore, it was found that the soft decision of language and gender outperforms the hard decision approach. The soft decision method incorporates the semantic scores $P(L,G\,|\,S)$ with transliteration scores $P(T\,|\,S,L,G)$, involving all possible semantic specific models in the decoding process.

In this paper, there are 9 such models (3 languages × 3 genders). The hard decision relies on Eqs. (10) and (11) to decide language and gender, which only involves one semantic specific model in the decoding. Neither soft nor hard decision requires any prior information about the names. It provides substantial performance improvement over phonetic transliteration at a reasonable computational cost. If the prior semantic information is known, e.g. via trigger words, then semantic transliteration attains its best performance.

6 Conclusion

Transliteration is a difficult, artistic human endeavor, as rich as any other creative pursuit. Research on automatic transliteration has reported promising results for regular transliteration, where transliterations follow certain rules. The generative model works well as it is designed to capture regularities in terms of rules or patterns. This paper extends the research by showing that semantic transliteration of personal names is feasible and provides substantial performance gains over phonetic transliteration. This paper has presented a successful attempt towards semantic transliteration using personal name transliteration as a case study. It formulates a mathematical framework that incorporates explicit semantic information (prior knowledge), or implicit one (through soft or hard decision) into the transliteration model. Extending the framework to machine transliteration of named entities in general is a topic for further research.

References

Peter F. Brown and Stephen Della Pietra and Vincent J. Della Pietra and Robert L. Mercer. 1993, The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, 19(2), pp. 263-311.

J. M. Crego, M. R. Costa-jussa and J. B. Mario and J. A. R. Fonollosa. 2005, N-gram-based versus Phrase-based Statistical Machine Translation, In *Proc. of IWSLT*, pp. 177-184.

Ariadna Font Llitjos, Alan W. Black. 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In *Proc. of Eurospeech*, Denmark, pp 1919-1922.

Lucian Galescu and James F. Allen. 2001, Bi-directional Conversion between Graphemes and Phonemes using a Joint N-gram Model, In *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Scotland, pp. 103-108.

Peter Hu, 2004, Adapting English to Chinese, *English Today*, 20(2), pp. 34-39.

Qingping Hu and Jun Xu, 2003, Semantic Transliteration: A Good Tradition in Translating Foreign Words into Chinese Babel: *International Journal of Translation, Babel*, 49(4), pp. 310-326.

Paul B. Kantor and Ellen M. Voorhees, 2000, The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Informational Retrieval*, 2, pp. 165-176.

K. Knight and J. Graehl. 1998. Machine Transliteration, *Computational Linguistics* 24(4), pp. 599-612.

J.-S. Kuo, H. Li and Y.-K. Yang. 2006. Learning Transliteration Lexicons from the Web, In *Proc. of 44th ACL*, pp. 1129-1136.

Haizhou Li, Min Zhang and Jian Su. 2004. A Joint Source Channel Model for Machine Transliteration, In *Proc. of 42nd ACL*, pp. 159-166.

Haizhou Li, Shuanhu Bai, and Jin-Shea Kuo, 2006, *Transliteration*, In *Advances in Chinese Spoken Language Processing, C.-H. Lee, et al. (eds)*, World Scientific, pp. 341-364.

Wei-Hao Lin and Hsin-Hsi Chen, 2002, Backward machine transliteration by learning phonetic similarity, In *Proc. of CoNLL* , pp.139-145.

Yegao Ning and Yun Ning, 1995, *Chinese Personal Names*, Federal Publications, Singapore.

Jong-Hoon Oh and Key-Sun Choi. 2005, An Ensemble of Grapheme and Phoneme for Machine Transliteration, In *Proc. of IJCNLP*, pp.450-461.

Y. Qu, G. Grefenstette and D. A. Evans, 2003, Automatic Transliteration for Japanese-to-English Text Retrieval. In *Proc. of 26th ACM SIGIR*, pp. 353-360.

Richard Sproat, C. Chih, W. Gale, and N. Chang. 1996. A stochastic Finite-state Word-segmentation Algorithm for Chinese, Computational Linguistics, 22(3), pp. 377-404.

Richard Sproat, Tao Tao and ChengXiang Zhai. 2006. *Named Entity Transliteration with Comparable Corpora*, In *Proc. of 44th ACL*, pp. 73-80.

Xinhua News Agency, 1992, *Chinese Transliteration of Foreign Personal Names*, The Commercial Press.

L. Xu, A. Fujii, T. Ishikawa, 2006 Modeling Impression in Probabilistic Transliteration into Chinese, In *Proc. of EMNLP 2006*, Sydney, pp. 242–249.