J Inf Process Syst, Vol.13, No.4, pp.716~730, August 2017
https://doi.org/10.3745/JIPS.02.0065

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Using Semantic Knowledge in the Uyghur-Chinese Person Name Transliteration

Alim Murat*,**,***, Turghun Osman*,**,***, Yating Yang*,**,
Xi Zhou*,**, Lei Wang*,**, and Xiao Li*,**

Abstract

In this paper, we propose a transliteration approach based on semantic information (i.e., language origin and gender) which are automatically learnt from the person name, aiming to transliterate the person name of Uyghur into Chinese. The proposed approach integrates semantic scores (i.e., performance on language origin and gender detection) with general transliteration model and generates the semantic knowledge-based model which can produce the best candidate transliteration results. In the experiment, we use the datasets which contain the person names of different language origins: Uyghur and Chinese. The results show that the proposed semantic transliteration model substantially outperforms the general transliteration model and greatly improves the mean reciprocal rank (MRR) performance on two datasets, as well as aids in developing more efficient transliteration for named entities.

Keywords

Gender, Language Origin, Semantic Knowledge-based Model, Transliteration of Person Name

1. Introduction

Transliteration is the process of transforming graphemes in source language into target language [1]. Generally, name entities like person name or place name are not seen in the dictionary and those out of vocabulary words are a common source of errors in many natural language processing applications such as machine translation (MT) [2], multilingual text-to-speech synthesis [3] and information retrieval [4]. However, most of the current Uyghur texts from news domain that are used as a main resource for Uyghur-Chinese MT system contain the large number of person names which are of Uyghur and Chinese origin, written in Uyghur writing system. Often, those person names left untranslated, since majority of them are out of vocabulary words and name of Chinese origin that are usually phonetically imported to Uyghur.

Transliteration of person name might be considered trivial between many languages in the same families, e.g., a person name Bill Gates is always Bill Gates in many languages like English or French

Manuscript received April 21, 2016; fisrt revision November 8, 2016; second revision January 31, 2017; accepted January 31, 2017.

Corresponding Author: Yating Yang (yangyt@ms.xjb.ac.cn)

* Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi, China ({alim.murat, yangyt, zhouxi, wanglei, xiaoli}@ms.xjb.ac.cn, turghunjan@sina.com)

** Xinjiang Key Laboratory of Minority Speech and Language Information Processing, Urumqi, China ({alim.murat, yangyt, zhouxi, wanglei, xiaoli}@ms.xjb.ac.cn, turghunjan@sina.com)

*** University of Chinese Academy of Science, Beijing, China (alim.murat@ms.xjb.ac.cn, turghunjan@sina.com)

orthographically and phonetically. However, automatic transliteration across languages with completely different alphabets and phonological structure such as Uyghur and Chinese are far beyond the description by simple phonetic rules. For example, the same Uyghur person name (UPN) ئايشەمگۈل *Ayshëmgül* (Latin-script Uyghur (LSU), hereafter used for ease representation; https://en.wikipedia.org/wiki/Uyghur_Latin_alphabet) is translated into different Chinese variations, which are 阿依夏木古丽, 阿依夏木姑丽, 阿依先古丽 and 阿依先姑丽. As shown in Table 1, the same Uyghur syllables are transliterated into different Chinese characters (underlined in the aligned parts). Furthermore, we randomly extract 100 Uyghur syllables out of 1,000 most frequent UPN and discover that all the syllables have more than one transliterations, and each syllable in UPN gains about average 5.8 equivalent Chinese characters.

Table 1. Uyghur-Chinese name transliteration variations

Uyghur	LSU	Chinese	Gender
ئايشەمگۈل	*Ayshëmgü**l***	阿依夏木古丽	Female
ئادىل	*Adi**l***	阿迪力	Male
يالقۇن	***Ya**lkun*	亚力坤	Male
ياسمەن	*Yasiman*	娅斯曼	Female
نىياز	***Ni**yaz*	尼亚孜	Male
ئاينىگار	*Ay**ni**gar*	阿依妮尕尔	Female

According to the problem stated above, there are still some challenges in the transliteration of person name between Uyghur and Chinese: (1) there has been less work on transliteration of person names in Uyghur-Chinese MT system; (2) the problem of character selection in the resulting transliteration variations. As we observed and analyzed, transliteration task is unlike any other traditional statistical MT which basically involves alignment and reordering. But in many cases, the person name transliteration between Uyghur and Chinese is related to semantic information which includes name's gender and origin. In this paper, we attempt to add the semantic information to general transliteration model and gain the best transliteration output in character selection by using semantic knowledge. However, we propose a semantic knowledge-based transliteration model, which could make use of related semantic instances in the training data to obtain the semantic information of person names such as gender and language origin. In addition, we devise a method to detect the gender and language origin of given UPN, which can favor the selection of the best candidate character on target Chinese side where their semantic meaning is very close to input UPN. Noted that semantic information is not learned in a straight-forward way but is adopted as a bridge to link the input and output in transliteration.

Experimental results have proved that the proposed semantic knowledge-based scheme found to be very advantageous in optimizing the existing traditional transliteration approach which only utilizes a general model to transcribe a source person name into the target person with the limited rules and or distributions.

2. Previous Work

Existing transliteration approaches are usually divided into three categories: grapheme-based, phoneme-based and a hybrid of phoneme and grapheme [5]. In this paper, we process the transliteration by means of Grapheme-based modeling framework. So hereby we describe some brief transliteration literatures of the grapheme-based approach in the following.

To our knowledge, there is no effort has been made to the person name transliteration between Uyghur and Chinese, but many studies have been devoted to the related English-Chinese transliteration. In [6], a novel nonparametric Bayesian is presented by using synchronous adaptor grammars to model the grapheme-based transliteration. Zhang et al. [7] introduced the pivot transliteration model with grapheme-based method. As another example of grapheme-based approach, Jia et al. [8] views machine transliteration as a special example of machine translation and used the phrase-based machine translation model to tackle the problem.

Another work has been done on machine transliteration of Japanese-English that uses phrase-based SMT techniques [9]. Rama and Gali [10] performed similar research for the Hindi-English language pair. Finch and Sumita [9] extended their earlier work using joint multilingual model to generate the n-best list of transliteration and presented language independence of their approach and obtained similar results across eight different language pairs. Jia et al. [8] also suggested grapheme-based joint Source-Channel Model within the Direct Orthographic Mapping framework. Models based on characters [11], syllable [12], as well as hybrid units [13], are also studied by researchers.

Regarding semantic transliteration, although there have been some works addressing semantic transliteration is considered as a good tradition in translation practices [14], unfortunately, it has not been adequately stated computationally in the literature. Also, a few work in [15] has tried introducing a probabilistic framework for Chinese character selection in phonetic transliteration. However, any analytical results to the semantic-specific transliteration is not being reported.

3. Semantic Knowledge Description

It is indicated in [14] that semantic information can produce the meaning that can be able to fill the semantic defects neglected by phonetic transliteration.

3.1 Language Origin

Identifying the person name origin is critical to the transliteration's success. Usually, a person name from Uyghur language resources in news domain is not of an absolute Uyghur origin (e.g., ئايشەمگۈل *Ayshëmgül*), besides, many person names that have been originated from Chinese (e.g., جاڭ چۈنشيەن *Jang chünshën*, 张春贤) are also seen in the resources. It is note that in resources scares languages like Uyghur that requires to manually annotate the origin of a source person name, and this affects the characters to be used in transliteration. There is a strict rule that if a person name translates into Chinese, it will need to at best preserve its Uyghur identity in Chinese version.

Thus, we decide to select proper character depending on the person name origin in training example. We present a language origin module which guesses the origin of the UPN and generates Chinese candidate transliterations from the mapping model based on the name origin. The performance of this module will be given in the experiment section.

3.2 Gender

Each UPN typically indicates an obvious gender information, which can be very beneficial in character selection. Therefore, the resulted transliteration of UPN must be very gender-specific by selecting adequate Chinese character that conveys the concept of particular gender of person name. For example, ئالىم (*Alim*) and ئالىيە (*Aliyë*) are transliterated into Chinese, the results are 阿力木, 阿里木 and 阿丽娅, 阿利亚, 阿丽雅, respectively. In this case, transliteration of UPNs are phonetically correct, but semantically inappropriate due to an incorrect gender identity in the resulted transliteration. However, in this instances, 阿力木 only shows strong masculine characteristics and 阿丽娅 implies clear feminine characteristics. The performance of this module will detail in the experiment section.

3.3 Viability of Semantic Information in Transliteration

UPN consists of syllables that formed by letters, while transliteration of UPN consists of Chinese characters. In transliteration, each Chinese character not only aligns to a Uyghur syllable but also potentially express the semantic meaning of person name. Usually, a good transliteration adequately shows semantic attribution while an inadequate one may result in undesirable transliteration.

We carry out an investigation on the viability of the semantic knowledge in order to improve the transliteration performance. Thus we conduct a quick statistic on our person name dataset shown in Table 2. The U-C data, which is a parallel corpus containing person name pairs between Uyghur to Chinese, whose input UPN is Uyghur origin. The UC-C data is almost the same as the U-C data, but whose input UPN is Chinese origin (Chinese person name written in Uyghur). The dataset provided by the key lab for network security and public opinion analysis in Xinjiang Normal University is collected from 20 different government websites in Xinjiang Uyghur Autonomous Region and manually inspected. To highlight the gender-specific information in transliteration, all entries in the data set are classified into male and female category.

Table 2. Person name pairs in datasets

	U-C	UC-C
Male	18,545	767
Female	7,320	4,817
Total	25,865	5,584

In [16], it is stated that Chinese has over 400 unique characters that can nearly be transcribed in other language. In grapheme mapping, different Uyghur syllable or sub-syllable may render the same Chinese character and construct a set of homonyms. Among the homonyms, characters arousing positive meaning can be used for person names. As presented in [17], out of several thousand common Chinese characters, merely a small set of few hundred characters are used overwhelmingly for transliterating English names to Chinese. Here, we conduct a statistics on the whole Chinese character usage in the dataset, and it shows that each semantic attribute has a certain connection with some unique characters (Table 3).

Table 3. Chinese character coverage in datasets

	U-C	UC-C
Male	404	240
Female	379	218
Total	652	380

From the statistics of Chinese character above, it is noticed that the semantic information of person names are determined by the character selection in Chinese. If the semantic information is given in advance, we can directly model these information as a feature. On the contrary, in the case where the semantic information is not available beforehand, we are not able to model the prior knowledge of semantic information. Therefore, dealing with semantic information of UPN is needed before the transliteration.

4. General Transliteration Model

Our general transliteration model is similar to [18], in which they worked on translation among Chinese and English names based on syllable and determined the best candidate using the statistical n-gram model. However, we propose a semantic knowledge-based model by modifying the previous general model. The overall structure of our model is shown in Fig. 1.

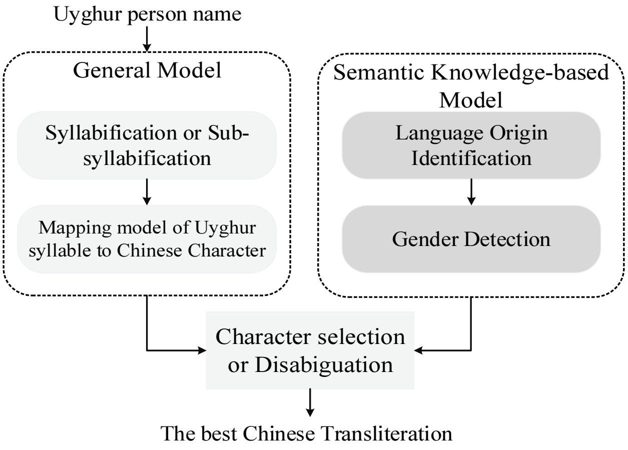
Uyghur person name
General Model
Syllabification or Sub-syllabification
Mapping model of Uyghur syllable to Chinese Character
Semantic Knowledge-based Model
Language Origin Identification
Gender Detection
Character selection or Disabiguation
The best Chinese Transliteration

Fig. 1. The overall system architecture.

We intuitively assume that given a source Uyghur name $U = U_1, \dots, U_k, \dots, U_K$, which consists of K Uyghur syllables, and mean to search for target Chinese transliteration $C = C_1, \dots, C_k, \dots, C_K$, which composed of K Chinese characters, and that of the highest probability. Here, it is believed that a Person Name is literally transliterated without any insertion or deletion, only by using orthographic grapheme mapping within a probabilistic framework. In this framework, a translation system produces the optimum target Chinese person name, C^*, which outputs the highest posterior probability given the source UPN.

$$C^* = argmax_{C \in \Phi_C} P(C|U)$$

where Φ_C is the set of all possible transliteration candidates for the input UPN. In order to incorporate our grapheme mappings which can ensure more exact transliteration, $P(C|U)$ is rewritten as:

$$P(C|U) = \sum_M P(C,M|U)$$

$$\cong max_M P(C,M|U)$$

where M is orthographic grapheme mapping for U and C. In the training phase, a set of U2C grapheme mappings are generated in this form $M = M_1^k = M_1, \ldots, M_k, \ldots, M_K$ in which M_k is defined as a mapping pair [mu_k,mc_k]. In this paper, mu_k is a Uyghur syllable or sub-syllable of UPN, while mc_k while mc_k is the equivalent Chinese character of the transliterated person name. The mappings are illustrated in the following form:

[يا,亚], [سىن,森], [جان,江], or [ئا,阿], [ي,依], [شه,夏] [م,木], [گۈ,古], [ل,丽].

And thus,

$$M = [mu_k, mc_k]_1^K = \{[mu_1, mc_1], \ldots [mu_k, mc_k], \ldots, [mu_K, mc_K]\}.$$

As we presented above, our general transliteration model incorporates orthographic grapheme-based mapping, in order to search the best probable mappings between Uyghur syllable and Chinese character, and then produces the final transliteration. Then, we describe mapping model of syllable to Chinese character.

4.1 Mapping Model of Syllable to Chinese Character

Since the Chinese character used in person names are limited, most of the conversions from syllable to character are fixed. However, some Uyghur syllables or sub-syllables still have several equivalent characters in Chinese, and this phenomenon generates the one-to-many mapping relation between them. For which, we should make an appropriate character selection in the transliterated Chinese person name using orthographic grapheme mapping table. We illustrate the mappings partially in Tables 4 and 5.

Table 4. Uyghur to Chinese single character mapping

Vowel	Chinese character	Consonant	Chinese character
ئا	阿	ت	特|提
ئە	艾	ر	尔
ئو	乌	ن	尼|妮
ئې	艾	ل	力|丽
ئى	依	ي	依
ئا	阿	ت	特|提
ئە	艾	ر	尔

Table 5. Uyghur to Chinese multiple character mapping

C/V	ت	ر	ن	ل	ي
ئا	塔	热|拉	纳|娜	拉	亚|娅
ئە	太	热|拉	乃	莱	亚|娅
ئو	托	如	诺	罗	尧
ئې	特	热	涅	列	叶
ئى	提	热|日	尼	力	依|黑
ئا	塔	热|拉	纳|娜	拉	亚|娅
ئە	太	热|拉	乃	莱	亚|娅

C=consonant, V=vowel.

In order to achieve the best character selection which can precisely reflect the UPN convention, we a model manifesting the usage of each Chinese character at different position in the person name, and build a 2-gram language model with absolute discounting according to the character positions varies.

A resulting Chinese person name is represented as C_1, C_2,…, C_l, C_i $(1 \leq i \leq L)$. The character appeared at the first position of person name is referred to as FW, while the one which appeared at the last position is referred to as LW. Generally, each Chinese character obtains different frequencies at this two different position in person name, e.g., the most used Uyghur syllable in UPN را (ra) can be mapped to two characters in Chinese including 热 (re) and 拉 (la), which are shown in Table 6. However, this implies that the most suitable character which will be selected in transliteration is subject to their frequencies at different position.

Table 6. Character frequencies (%) at two distinctly different positions

	FW	LW
热	0.62	0.31
拉	0.21	0.59

From Table 6, it can be seen that character 热 gains considerably high probability at FW and gets lower frequency at LW. On the contrary, character 拉 shows comparatively high frequency at LW and relatively lower at FW in the training data. Nonetheless, some characters positioned in lower frequency basically are also correct in certain context. In this case, it results in ambiguities in character selection. As for ambiguities, we use a character sequence model (CSM) with bigram language modeling.

CSM for language modeling is a probability distribution over a sequence of characters within a word [19]. Let $W = \langle C_1, C_2, \dots, C_n \rangle$ be Chinese person name. Then with the N-order Markov Assumption, the probability distribution of output Chinese person name in transliteration is formulated in the following form.

$$P(W) = \prod_{i=1}^{n} P(C_i | C_{i-1}, \dots, C_{i-N})$$

In this model, CSM learns the character sequence patterns of Chinese person name in training data. If a Uyghur syllable can be mapped to more than on Chinese characters, the probability distribution $P(W)$ indicates the most possible Chinese character to be selected.

5. Semantic Transliteration Model

We described general probabilistic transliteration framework in Section 4, as shown in Eq. (1).

$$C^* = argmax_{C \in \Phi_C} P(C|U) \tag{1}$$

We can see that any external information between U and C is not taken into consideration in Eq. (1). In general transliteration model, $P(C|U)$, probability of the hypothesized transliteration, C and given input Uyghur person name U, is directly modeled without concerning any form of semantic information. However, in this paper, we propose a novel transliteration approach based on semantic knowledge which includes language origin and gender identity, in order to incorporate the possible semantic meaning of UPN in transliteration. To this end, $P(C|U)$ is reformulated in the following form.

$$P(C|U) = \sum_{L \in l, G \in g} P(C, L, G|U) \tag{2}$$

$$= \sum_{L \in l, G \in g} P(C|U, L, G) P(L, G|U) \tag{3}$$

where $P(C|U, L, G)$ is the transliteration probability from U to C, given the language origin (L) and gender (G) identity. l and g represent the sets of language origins and genders respectively. $P(L, G|U)$, is the probability of language origin and gender identity of given UPN.

In this approach, the UPN and Chinese person name in training example are aligned using the mapping model, and this alignment produces possible mappings between Uyghur syllable and Chinese character. During testing, the best character selection is processed based on the semantic information. Once the resulted transliteration candidates are created, CSM ranks each transliteration output according to the scores obtained.

As indicated in Eq. (3), the semantic transliteration greatly depends on the predefined semantic information, language origin and gender identity. If a semantic information is not available for particular UPN, a uniform probability distribution is assumed. By changing $P(L, G|U)$ as follows:

$$P(L, G|U) = P(G|L, U) P(L|U) \tag{4}$$

It is quite obvious that person name has semantic features. In the case in which the semantic knowledge is not presented, we try to learn semantic features from the person name using Bayes Theorem which express $P(L, G|U)$ in the following form.

$$P(L, G|U) = \frac{P(U|L, G) P(L|G)}{P(U)} \tag{5}$$

$P(U|L,G)$ can be modeled using an N-gram language model for the syllables of all UPN in training data. $P(L|G)$ is typically uniform. If we fail to model the semantic information of UPN, then a general model will be used for transliteration by discarding the reliance on those information that is not available. When both the language origin and gender identity are unknown, the transliteration uses the general model as a baseline.

6. Experiments

In this section, we present the experiments of transliteration performed on datasets of person name (see Section 3.3). In addition to transliteration, we conduct the experiments on language origin and gender to analyze how the semantic information affects transliteration results. The datasets and evaluation criteria are also described in this section.

6.1 Datasets

The transliteration task uses datasets of two language origins (Uyghur and Chinese), which are shown in Table 7 and also contain gender information (male and female). Each dataset is randomly divided such that 3,000 name pairs out of 25,865 are used for the testing in the U-C, and that another 600 name pairs out of 5,584 are used for the testing in the UC-C. The remaining is used for training the model. Note that we have no overlapping name pairs between training and testing samples.

Table 7. Datasets used in the experiments

Language origin	Data set	Male	Female	All
U-C	Train	17,640	5,225	22,865
	Test	2,500	500	3,000
UC-C	Train	4,367	617	4,984
	Test	450	150	600

Statistics on datasets are shown in Table 7. All Uyghur-Chinese transliteration pairs are distinct, while some Uyghur names share same Chinese transliteration. Therefore, the total number of unique Chinese names is less than that of Uyghur names. The Chinese character used in transliteration are fixed, for this reason, there are comparatively less amount of unique Chinese character in both training examples (Table 8).

Table 8. Numbers of unique entries in training sample

	FW	LW
Number of unique name pairs	25,865	4,984
Number of unique Uyghur syllables	1,966	1,172
Number of unique Chinese characters	652	380
Average Uyghur syllables for per name (%)	6.38	4.20
Average Chinese character for per name (%)	6.82	3.58

6.2 Language Origin Identification

Identifying the language origin of UPN is critical to the semantic transliteration. N-gram language model have been used for the Language origin detection in [20]. We also use the same model where we learn two different CSM classes according to each language origin, and the model assigns the input UPN to the class which produces the higher probability. We also use the same training datasets that are manually annotated with language origin information.

In language origin identification, how to tune the ideal N-gram count is a significant factor for achieving the best accuracy. Surana and Singh [21] proposed 5-grams for detecting the origin of words written in the Roman script. Whereas, we discover that 4-grams are found to be the most effective way for identifying the language origin of input UPN. The main reason we use 4-gram in identification is that many syllables in written Uyghur include an inherent "ئـ" Hamza which is neither vowel nor consonant.

Table 9. The result of language origin identification

N-gram size	Uyghur origin (%)	Chinese origin (%)
3	62.7	89.50
4	94.26	85.90
5	70.7	53.20

As shown in Table 9, the detection of the UPN which is of Uyghur origin gains the highest accuracy with 4-gram, while the UPN of Chinese origin obtains bit lower accuracy. The slight decrease in accuracy of the Chinese origin together with the same N-gram count is attributed to the fact that Chinese person names in our datasets consist of 3 to 4 characters. However, the 4-gram still shows the highest accuracy for both language origins in average.

6.3 Gender Detection

Unlike language origin, we use a separate tactic to recognize the specific gender of UPN. For the UPN of Uyghur origin, we use a rule bank (Table 9) proposed by [22], which contains 106 suffixes that are very useful for indicating the gender information. In the rule bank, the suffix is annotated based on the position, whether the suffix is in the beginning (B) or in the end (E) of UPN. The few part of rule bank shown in Table 10 for reference.

Table 10. A snippet of rule bank for gender detection

Suffix	Gender (M/F)	Position (B/E)
پەرى	F	B
قىز	F	E
ئابدۇ	M	B
جان	M	E
گۈل	F	B,E
يار	M	B,E

The result of gender detection shows that the rule bank works so well for the UPN of Uyghur origin. Nonetheless, this rule bank is not applicable to other language. Therefore, we devise another approach to gender detecting for the UPN of Chinese origin. We assume that if the language origin of UPN S is already known, the gender information can be expressed based on maximum likelihood approach in this form $G_s = argmax_{G\in g}P(S|L_S, G)$, where Ls is the detected language. The results of gender detection are presented based on two different language origins.

Table 11. The result of gender detection for both language origins

Language	Male (%)	Female (%)	Average (%)
Uyghur origin	96.45	90.78	93.7
Chinese origin	63.28	70.93	67.1

As shown in Table 11, the average detection accuracies are 93.7% and 67.1%, respectively for the UPN of Uyghur origin and Chinese origin. Among the language origins, gender detection for UPN of Chinese origin performs relatively poor compared to Uyghur origin. The main reason is that the Chinese characters are shared very much across genders.

6.4 Transliteration Results and Analysis

The transliteration performance was measured using mean reciprocal rank (MRR) metric [23], a measure that tells approximately the average rank of the correct result, assuming that there is only one correct output. If MRR score close to 1 it means that the correct answer is often included in the top of the n-best list.

Firstly, we performed the experiment using the general transliteration model (GTM) as the baseline without incorporating any semantic information, in order to analyze and highlight how these semantic knowledge makes transliteration more effective. Then, we performed the experiment using the semantic knowledge-based model (SKM) which is based on all possible semantic information of UPN. Table 12 shows the transliteration performance of 4 different scenarios in the experiment. In the experiment, L and G stands for language and gender; M and F stand represent male and female.

Table 12. Overall transliteration performance

Method	U-C		UC-C	
	Male	Female	Male	Female
General model	0.57	0.48	0.43	0.37
GM+L	0.63	0.54	0.58	0.50
GM+G	0.80	0.62	0.56	0.47
Semantic model	0.89	0.83	0.71	0.69

The comparison between semantic model and the baseline with two separate semantic information for the transliteration performance are analyzed.

In the experiment on the U-C data, both GTM+L and GTM+G have improved the baseline and presented the MRR scores of 0.63 and 0.80, respectively, and also there have been a significant differences between them. Given that the knowledge of gender identity of UPN has a great superiority, while the language origin has a less effect on the U-C data with a slightly poor performance. As is shown in the result, those separate semantic knowledge combinations seem to help the transliteration, however, the SKM still gained the better MRR performance than those two (GTM+L and GTM+G) based on the integrated semantic knowledge. However, it is also surprising to find that SKM achieved 0.32 and 0.35 relative significant improvement in MRR.

Another experiment on the UC-C data, on the contrary that both separate models have showed uniformly a poorer performance in MRR with scores of 0.58 and 0.56, respectively compared to the case on the U-C. Due to the poorer detection accuracy for the UPN of Chinese origin, it was shown to have a reduction in MRR performance to some degree. Noted that the GTM+G has produced relatively lower performance than the GTM+L in this dataset, since most of the character used in Chinese person name shares the same one across the different gender, for which, we discover that these separate semantic model presented not ideal performance on this data. However, the SKM contributed a better improvement in MRR performance with the increase of 0.32 and 0.35, respectively.

In this work, the most significant finding is that the general transliteration model may be greatly enhanced by integrating semantic knowledge such as the language origin and gender information into transliteration. Another important finding is that semantic transliteration model is more adaptable and attains best performance in U-C data. In particular, the results of this study indicate that knowing the semantic information of UPN has proved that semantic knowledge-based transliteration can also be applicable to named entity, and it could further adopt more different semantic information.

7. Conclusions

In this paper, we presented a transliteration model based on semantic information (i.e. language origin and gender identity) which is directly learnt from the person name, to transliterate the person name of Uyghur into Chinese. Two kinds of information including the person names of Uyghur origin and Chinese origin are used in experiments. Transliteration results show that the proposed semantic knowledge based model outperforms the general transliteration model. On the other hand, results also have revealed that our model performs greatly on the person name data of Uyghur origin.

This paper has contributed a successful attempt toward transliteration by using person name as a case study and built a mathematical framework that incorporates language and gender information into the transliteration work.

In the future work, it is expected that the proposed semantic transliteration model for person name could be further extended to machine transliteration of other named entities. In addition, we will apply our model to other languages, like Kazak, Kyrgyz, Uzbek, and Turkish for a further research topic.

Acknowledgement

This paper is supported by Natural Science Foundation of Xinjiang (Grant No. 2015211B034).

References

[1] R. E. Banchs, M. Zhang, X. Duan, H. Li, and A. Kumaran, "Report of NEWS 2015 machine transliteration shared task," in *Proceedings of NEWS 2015: The 5th Named Entities Workshop*, Beijing, China, 2015, pp. 10-23.

[2] W. J. Hutchins and H. L. Somers, *An Introduction to Machine Translation (Vol. 362)*. London: Academic Press, 1992.

[3] R. W. Sproat, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Dordrecht: Kluwer Academic Publishers, 1997.

[4] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

[5] K. Kaur and P. Singh, "Review of machine transliteration techniques," *International Journal of Computer Applications*, vol. 107, no. 20, pp. 13-16, 2014.

[6] Y. Huang, M. Zhang, and C. L. Tan, "Nonparametric Bayesian machine transliteration with synchronous adaptor grammars," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Portland, OR, 2011, pp. 534-539.

[7] M. Zhang, X. Duan, V. Pervouchine, and H. Li, "Machine transliteration: leveraging on third languages," in *Proceedings of the 23rd International Conference on Computational Linguistics*: *Posters*, Beijing, China, 2010, pp. 1444-1452.

[8] Y. Jia, D. Zhu, and S. Yu, "A noisy channel model for grapheme-based machine transliteration," in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore, 2009, pp. 88-91.

[9] A. Finch and F. Sumita, "Phrase-based machine transliteration," in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, Hyderabad, India, 2008, pp. 13-18.

[10] T. Rama and K. Gali, "Modeling machine transliteration as a phrase based statistical machine translation problem," in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore, 2009, pp. 124-127.

[11] P. Shishtla, V. S. Ganesh, and S. Subramaniam, "A language-independent transliteration schema using character aligned models at NEWS 2009," in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore, 2009, pp. 40-43.

[12] C. Wutiwiwatchai and A. Thangthai, "Syllable-based Thai-English machine transliteration," in *Proceedings of the 2010 Named Entities Workshop*, Uppsala, Sweden, 2010, pp. 66-70.

[13] J. H. Oh and K. S. Choi, "An ensemble of grapheme and phoneme for machine transliteration," in *International Conference on Natural Language Processing*. Heidelberg: Springer, 2005, pp. 450-461.

[14] M. Hagiwara and S. Sekine, "Latent semantic transliteration using Dirichlet mixture," in *Proceedings of the 4th Named Entity Workshop*, Jeju, Korea, 2012, pp. 30-37.

[15] L. Xu, A. Fujii, and T. Ishikawa, "Modeling impression in probabilistic transliteration into Chinese," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006, pp. 242-249.

[16] H. Li, K. C. Sim, J. S. Kuo, and M. Dong, "Semantic transliteration of personal names," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007, pp. 120-127.

[17] X. Duan, R. E. Banchs, M. Zhang, H. Li, and A. Kumaran, "Report of NEWS 2016 machine transliteration shared task," in *Proceedings of NEWS 2016: The 6th Named Entities Workshop*, Berlin, Germany, 2016, pp. 58-72.

[18] X. Jiang, L. Sun, and D. Zhang, "A syllable-based name transliteration system," in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore, 2009, pp. 96-99.

[19] M. K. Chinnakotla and O. P. Damani, "Character sequence modeling for transliteration," in *Proceedings of 7th International Conference on Natural Language Processing (ICON)*, Hyderabad, India, 2009, pp. 1-10.

[20] M. M. Khapra and P. Bhattacharyya, “Improving transliteration accuracy using word-origin detection and lexicon lookup,” in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore, 2009, pp. 84-87.

[21] H. Surana and A. K. Singh, “A more discerning and adaptable multilingual transliteration mechanism for Indian languages,” in *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, Hyderabad, India, 2008, pp. 64-71

[22] A. Murat, A. Yusup, and Y. Abaydulla, “Research and implementation of the uyghur-chinese personal name transliteration based on syllabification,” in *Proceedings of 2013 International Conference on Asian Language Processing*, Urumqi, China, 2013, pp. 71-74.

[23] H. Li, A. Kumaran, M. Zhang, and V. Pervouchine, “Whitepaper of NEWS 2009 machine transliteration shared task,” in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Singapore, 2009, pp. 19-26.

Alim Murat http://orcid.org/0000-0001-8510-7808

He received B.E. and M.S. degrees in School of Computer Science and Technology from Xinjiang Normal University in 2011 and 2014, respectively. Since September 2014, he is with the Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science as a PhD candidate. His current research interests include Natural Language Processing, Machine Translation and Semantic Web.

Turghun Osman

He received M.S. degree in School of Computer Science and Engineering from Xinjiang University in 2009. Since September 2014, he is with the Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science as a PhD candidate. His current research interests include Natural Language Processing and Machine Translation.

Yating Yang

She received her Ph.D. degree in Computer Science from Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science in 2012. Since July 2012, she is with the Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science as an associate researcher. Her current research interests include Natural Language Processing and Machine Translation.

Xi Zhou

He is a research fellow and a head of Multilingual Information Processing Lab in the Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science. His current research interests include Computer Application Technology and Multilingual Information Processing.

Lei Wang

He is a research fellow and a deputy head of Multilingual Information Processing Lab in the Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science. His current research interests include Natural Language Processing and Machine Translation.

Xiao Li

He is a research fellow and a director of Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science. His research interests include Multilingual Information Processing, Artificial Intelligence. He has published more than 70 papers in various journal and refereed conference. He is the recipient of many awards including the 2nd Prize for the National Sci-Tech Award, Outstanding Achievement Award of Chinese Academy of Science and the Extra Prize for Regional Sci-Tech Award.