Graphemic Approximation of Phonological Context for English-Chinese Transliteration

Oi Yee Kwong
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
Olivia.Kwong@cityu.edu.hk

Abstract

Although direct orthographic mapping has been shown to outperform phoneme-based methods in English-to-Chinese (*E2C*) transliteration, it is observed that phonological context plays an important role in resolving graphemic ambiguity. In this paper, we investigate the use of surface graphemic features to approximate local phonological context for *E2C*. In the absence of an explicit phonemic representation of the English source names, experiments show that the previous and next character of a given English segment could effectively capture the local context affecting its expected pronunciation, and thus its rendition in Chinese.

1 Introduction

Proper names including personal names, place names, and organization names, make up a considerable part of naturally occurring texts. Personal names, in particular, do not only play an important role in identifying an individual, but also carry the family history, parental expectation, as well as other information about a person. In natural language processing, the proper rendition of personal names, especially between dissimilar languages such as Chinese and English, often contributes significantly to machine translation accuracy and intelligibility, and cross-lingual information retrieval. This paper addresses the problem of automatic English-Chinese forward transliteration (referred to as *E2C* hereafter) of personal names.

Unlike many other languages, Chinese names are characteristic in their relatively free choice and combination of characters, particularly for given names. Such apparent flexibility does not only account for the virtually infinite number of authentic Chinese names, but also leads to a considerable sample space when foreign names are transliterated into Chinese. Underlying the large sample space, however, is not entirely a random distribution. On the one hand, there are no more than a few hundred Chinese characters which are used in names (e.g. Sproat *et al.*, 1996). On the other hand, beyond linguistic and phonetic properties, many other social and cognitive factors such as dialect, gender, domain, meaning, and perception, are simultaneously influencing the naming process and superimposing on the surface graphemic correspondence.

As the state-of-the-art approach, direct orthographic mapping (e.g. Li *et al.*, 2004), making use of graphemic correspondence between English and Chinese directly, has been shown to outperform phoneme-based methods (e.g. Virga and Khudanpur, 2003). In fact, transliteration of foreign names into Chinese is often based on the surface orthographic forms, as exemplified in the transliteration of Beckham, where the supposedly silent h in "ham" is taken as pronounced, resulting in 汉姆 *han4-mu3* in Mandarin Chinese and 咸 *haam4* in Cantonese[1].

However, as we have observed, there is considerable graphemic ambiguity in *E2C*, where an English segment might correspond to different Chinese segments. Such multiple mappings, to a large extent, is associated with the phonological context embedding the English segment, thus affecting its expected pronunciation. Hence, if such phonological context could be considered in

[1] Mandarin names are shown in simplified Chinese characters and transcribed in Hanyu Pinyin, while Cantonese names are shown in traditional Chinese characters and transcribed in Jyutping published by the Linguistic Society of Hong Kong.

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 186–193,
Suntec, Singapore, 7 August 2009.

the transliteration model, some of the graphemic ambiguity could be resolved. However, instead of going for an explicit phonemic representation, which might introduce an extra step for error propagation, in the current study we investigate the usefulness of surface graphemic features for approximating the local phonological context in *E2C*. Experiments show that the previous and next character of a given segment could effectively capture the local phonological context and improve transliteration accuracy.

A short note on terminology before we move on: We use "segment" to refer to a minimal graphemic transliteration unit in the names. For instance, in the data, the name Amyx is transliterated as 阿米克斯 *a1-mi3-ke4-si1*, the grapheme pairs are <a, 阿>, <my, 米>, and <x, 克斯>. There are three English segments: "a", "my" and "x"; and three Chinese segments: 阿, 米 and 克斯. A segment may or may not correspond to exactly a syllable, although it often does.

In Section 2, we will briefly review some related work. In Section 3, we will discuss some observations on graphemic ambiguity in *E2C*. The proposed method will be presented in Section 4. Experiments will be reported in Section 5, with results discussed in Section 6, followed by a conclusion in Section 7.

2 Related Work

There are basically two categories of work on machine transliteration. On the one hand, various alignment models are used for acquiring transliteration lexicons from parallel corpora and other resources (e.g. Lee *et al.*, 2006; Jin *et al.*, 2008; Kuo and Li, 2008). On the other hand, statistical transliteration models are built for transliterating personal names and other proper names, such as by means of noisy channel models or direct models amongst others, phoneme-based (e.g. Knight and Graehl, 1998; Virga and Khudanpur, 2003), or grapheme-based (e.g. Li *et al.*, 2004), or a combination of them (Oh and Choi, 2005), or based on phonetic (e.g. Tao *et al.*, 2006; Yoon *et al.*, 2007) and semantic (e.g. Li *et al.*, 2007) features.

Li *et al.* (2004), for instance, used a Joint Source-Channel Model under the direct orthographic mapping (DOM) framework, skipping the middle phonemic representation in conventional phoneme-based methods, and modelling the segmentation and alignment preferences by means of contextual n-grams of the transliteration units. Their method was shown to outperform phoneme-based methods and those based on the noisy channel model.

The n-gram model used in Li *et al*. (2004) was based on previous local context of grapheme pairs. However, as we are going to show in Section 3, contexts on both sides of a segment are important in determining the actual rendition of it in Chinese. In addition, graphemic ambiguity could in part be resolved by means of the phonological context embedding the segment. Hence in the current study, we propose a method modified from the Joint Source-Channel Model to take into account contexts on both sides of a segment, and to approximate local phonological context by means of surface graphemic features.

3 Some Observations

In this section, we will quantitatively analyse some properties of *E2C* based on our data, and show the importance of considering neighbouring context on both sides of a certain segment, as well as the possibility of approximating phonological properties graphemically.

3.1 Dataset

The data used in the current study are based on the English-Chinese (EnCh) training and development data provided by the organisers of the NEWS 2009 Machine Transliteration Shared Task. There are 31,961 English-Chinese name pairs in the training set, and 2,896 English-Chinese name pairs in the development set. The data were manually cleaned up and aligned with respect to the correspondence between English and Chinese segments, e.g. Aa/l/to 阿/尔/托. The analysis in this section is based on the training set.

The Chinese transliterations in the data basically correspond to Mandarin Chinese pronunciations of the English names, as used by media in Mainland China (Xinhua News Agency, 1992). Note that transliterations for English names could differ considerably in Chinese, depending on the dialect in question. Names transliterated according to Mandarin Chinese pronunciations are very different from those according to Cantonese pronunciations, for instance. Transliterations used in Mainland China are also different from those used in Taiwan region, despite both are based on Mandarin Chinese. A well cited example is a syllable initial /d/ may surface as in Baghdad 巴格达 *ba1-ge2-da2*, but the syllable final /d/ is not represented. This is true for transliteration based on Mandarin Chinese pronuncia-

tions. For Cantonese, however, it is different since ending stops like –p, –t and –k are allowed in Cantonese syllables. Hence the syllable final /d/ in Baghdad is already captured in the last syllable of巴格達 *baa1-gaak3-daat6* in Cantonese.

Such phonological properties of Mandarin Chinese might also account for the observation that extra syllables are often introduced for certain consonant segments in the middle of an English name, as in Hamilton, transliterated as 汉密尔顿 *han4-mi4-er3-dun4* in Mandarin Chinese (c.f. 咸美頓 *haam4-mei5-deon6* in Cantonese); and Beckham, transliterated as 贝克汉姆 *bei4-ke4-han4-mu3* in Mandarin Chinese (c.f. 碧咸 *bik1-haam4* in Cantonese).

3.2 Graphemic Ambiguity

Table 1 quantitatively describes the training data. On average each English name has around 3.14 segments, or transliteration units. On average each English segment has around 1.7 different renditions in Chinese. On the other hand, although the number of unique Chinese segments is just a few hundred, on average one Chinese segment could correspond to about 10 different English segments. This suggests that English-Chinese graphemic segment correspondence could be quite ambiguous. Further analysis is therefore needed to see if any systematic patterns could be found among such ambiguity.

Unique English names	31,822
Total English segments	99,930
Unique English segments	2,822
Unique Chinese segments	458
Unique grapheme pairs	4,750

Table 1. Quantitative Aspects of the Data

Assume transliteration pair mappings are in the form $<e_k, \{c_{k1},c_{k2},\ldots,c_{kn}\}>$, where e_k stands for the kth unique English segment, and $\{c_{k1},c_{k2},\ldots,c_{kn}\}$ for the set of n unique Chinese segments observed for it in the data. It was found in the training data that n varies from 1 to 15, while 32.2% of the distinct English segments have multiple grapheme correspondence. Table 2 shows the degree of graphemic ambiguity with illustrative examples. Some of the ambiguity, however, is the result of homophones. The effect of homophones (whether or not tones are taken into account) in *E2C* transliteration is worth more in-depth investigation, but it is beyond the scope of the current study.

n	**Proportion**	**Examples**			
		English Segment	**Chinese Segments**	**Source Name**	**Transliteration**
≥5	4.8%	na	内 *nei4*	Aber**na**thy	阿伯**内**西
			娜 *na4*	Adami**na**	阿达米**娜**
			尼 *ni2*	Cra**na**ge	克拉**尼**奇
			拿 *na2*	Buo**na**parte	波**拿**巴
			瑙 *nao3*	K**ena**ll	克**瑙**尔
			纳 *na4*	Stra**na**han	斯特拉**纳**汉
			诺 *nuo4*	Wid**na**ll	威德**诺**尔
4	2.9%	tain	丹 *dan1*	Lafon**tain**	拉方**丹**
			坦 *tan3*	S**tain**ton	斯**坦**顿
			廷 *ting2*	Sar**tain**	沙**廷**
			顿 *dun4*	Chas**tain**	查斯**顿**
3	7.3%	ran	兰 *lan2*	G**ran**berg	格**兰**伯格
			朗 *lang3*	F**ran**cine	佛**朗**辛
			伦 *lun2*	Kar**ran**	卡**伦**
2	17.2%	ty	蒂 *di4*	Chris**ty**	克里斯**蒂**
			太 *tai4*	S**ty**le	斯**太**尔
1	67.8%	gie	吉 *ji2*	An**gie**	安**吉**
				Cow**gie**l	考**吉**尔

Table 2. Graphemic Ambiguity of the Data

The other multiple correspondences are nevertheless genuine ambiguity. The same English graphemic segment, depending on its pronunciation within the name, could be rendered in various Chinese segments of very different pronunciations. To determine the expected pronunciation of the ambiguous English segment, however, the phonological context embedding the segment has an important role to play. For instance, the graphemic segment "na", when appearing at the end of a name, is often pronounced as /na/ and rendered as 娜 *na4*, especially for female names. But when it is in the middle of a name, and especially before "th", it is often pronounced as /nei/ and rendered as 内 *nei4*. Similarly, the segment "ty" is often pronounced as /ti/ at the end of a name and transliterated as 蒂 *di4*. On the other hand, if it is in the middle of a name, after an "s" or in front of "le" or "re", it is often pronounced as /tai/ and therefore transliterated as 太 *tai4*.

Take another segment "le" as an example. It is found to correspond to as many as 15 different Chinese segments, including 利 *li4*, 勒 *le4*, 历 *li4*, 尔 *er3*, 莱 *lai2*, 里 *li3*, etc. When "le" appears at the end of a name, all but a few cases are pronounced as /l/ and rendered as 尔 *er3*, particularly when it follows "a", e.g. Dale 戴尔 *dai4-er3* and Dipasquale 迪帕斯奎尔 *di2-pa4-si1-kui2-er3*. Exceptions are when "le" at the end of a name follows "r", where it is often rendered as 利 *li4* instead. On the other hand, when "le" appears at the beginning of a name where the vowel is often prominently pronounced, it is usually rendered as 勒 *le4* or 莱 *lai2*, e.g. Lepke 莱普克 *lai2-pu3-ke4*, except when it is followed by the vowel "o", where it is then often transliterated as 利 *li4*, e.g. Leonor 利奥诺 *li4-ao4-nuo4*. When "le" appears in the middle of a name, the transliteration is nevertheless more variable. Still it is remarkable that "le" is transliterated as 历 *li4* when it is followed by "c" or "x", e.g. Alex 阿历克斯 *a4-li4-ke4-si1*.

Such observations thus suggest two important points for *E2C*. First, contexts on both sides of a given segment do play a role in determining its likely rendition in Chinese. Second, the phonological context is important for determining the expected pronunciation of an English segment given its position in a name. Hence we propose a method, making use of contexts on both sides of a segment, to approximate the local phonological context of a segment via surface graphemic features.

4 Proposed Method

The Joint Source-Channel Model in Li *et al*. (2004) making use of direct orthographic mapping and a bigram language model for the segment pairs (or token pairs in their terms) is as follows:

$$\begin{aligned} P(E,C) &= P(e_1, e_2, ..., e_k, c_1, c_2, ..., c_k) \\ &= P(<e_1,c_1>, <e_2,c_2>, ..., <e_k,c_k>) \\ &\approx \prod_{k=1}^{K} P(<e_k,c_k> | <e_{k-1},c_{k-1}>) \end{aligned}$$

where E refers to the English source name and C refers to the transliterated Chinese name. With K segments aligned between E and C, e_k and c_k refer to the kth English segment and its corresponding Chinese segment respectively.

While we have grounds for orthographic mapping as mentioned in the introduction, there is some modification we hope to make to the above model. As pointed out in the last section, local contexts on both sides of a given segment should be important and useful for modelling the context embedding the segment, which in turn could help determine its expected pronunciation. In addition, the phonological environment might be sufficiently represented by a neighbouring phoneme instead of even a syllable. Thus we take the last character from the previous segment and the first character of the next segment (instead of the whole neighbouring segment) into account, irrespective of their corresponding Chinese segments. This could be considered an attempt to approximate the local phonological context of a given segment by means of surface graphemic features, even if we do not go for an explicit phonemic representation of the source name.

Hence we propose to make use of bigrams in both directions with equal weighting, and assign a score, *Score(E,C)*, to a transliteration candidate as below:

$$\prod_{k=1}^{K} P(<e_k,c_k> | lc(e_{k-1}))\ P(<e_k,c_k> | fc(e_{k+1}))$$

where $lc(e_{k-1})$ refers to the last character of the previous English segment, and $fc(e_{k+1})$ refers to the first character of the next English segment.

In the rest of this paper, we will refer to this method as GAP, which stands for Graphemic Approximation of Phonological context.

5 Experiments

The 31,961 English-Chinese name pairs from the NEWS shared task training set were used for training, and the 2,896 names in the development set were used for testing. The data were first manually cleaned up and aligned with respect to the correspondence between English segments and Chinese segments.

5.1 Segmentation of Test Names

Each test name was first segmented. All possible segmentations were obtained based on the unique English segments obtained from the manual alignment above.

The graphemic units are made case-insensitive. When finding all possible graphemic segmentations of the English source names, segments with length 1 are only allowed if no longer segment with that initial letter followed by a vowel is possible. For example, while "a", "k", "l", "o", "v", "s" and "y" are all observed segments in the training data, when computing the transliteration for the test name Akalovsky, only two of the possible segmentations, A/ka/lo/v/s/ky and A/kal/o/v/s/ky, were considered while the rest involving more single-letter segments were ignored. This is justified by three reasons. First, the more alternative segmentations, the more alternative transliteration candidates are to be evaluated. This is computationally expensive, and many alternatives are in fact quite unlikely. Second, single-letter segments are redundant if a longer segment is possible. On the one hand, transliterations are usually based on a consonant-vowel combination as a unit. A consonant will only be on its own as a segment if it occurs among a consonant cluster, which has no direct syllable correspondence in Chinese. For example, it is useless to single out the second "k" in Akalovsky as the longer segment "ka" is pronounceable anyway, unlike in names with consonant clusters like Akst. On the other hand, in the cases of doubling consonants like Ross, both "s" and "ss" will correspond to similar sounds. Third, the n-gram models favour transliterations with fewer segments anyway, so the segmentations with more single-letter segments will be less probable in any case.

The possible segmentations obtained were then ranked by a method similar to GAP. The score for each segmentation candidate S, $Score(S)$, is computed by:

$$\prod_{k=1}^{K} P(s_k \mid lc(s_{k-1}))\, P(s_k \mid fc(s_{k+1}))$$

where s_k is the kth segment in a name, $lc(s_{k-1})$ is the last character of the previous segment and $fc(s_{k+1})$ is the first character of the next segment.

In the experiments, we selected the top N segmentation candidates for use in subsequent steps, where N was varied from 1 to 3.

5.2 Transliteration Candidates

With the top N segmentation candidates, the transliteration candidates were generated by looking up the grapheme pairs obtained from manual alignment with frequency over a certain threshold f. We tested with $f \geq 3$ and $f \geq 5$. If there is no grapheme pair for a certain segment above the threshold, all pairs below the threshold would be considered. All combinations obtained were then subject to ranking by the GAP transliteration method.

5.3 Testing

The transliteration candidates were evaluated and ranked by the GAP method. For comparison, we also run the Joint Source-Channel Model (JSCM) described in Li *et al*. (2004) on the test data. In addition, we also tested a variation of GAP, called GAP-s, where the neighbouring characters are replaced by the neighbouring segments in the computation of the scores, that is, $lc(e_{k-1})$ is replace by $<e_{k-1},c_{k-1}>$ and $fc(e_{k+1})$ is replaced by $<e_{k+1},c_{k+1}>$. Note that similar changes were applied to the ranking of the source name segmentations for both methods accordingly.

System performance was measured by the Mean Reciprocal Rank (MRR) (Kantor and Voorhees, 2000), as well as the Word Accuracy in Top-1 (ACC) and Fuzziness in Top-1 (Mean F-score) used in the NEWS shared task. Only the top 10 transliteration candidates produced by the systems were considered.

6 Results and Discussion

6.1 Candidates Filtering

As mentioned in the last section, candidates were filtered in two stages. First, when the source English name was segmented, only the top N segmentation candidates were retained for subsequent processes. Second, when transliteration candidates were generated, only those grapheme pairs with frequency $\geq f$, where applicable, were considered for the candidates. Table 3 shows the

results of GAP with various combinations of N and f.

	$f \setminus N$	1	2	3
ACC	**3**	0.6357	0.6443	**0.6450**
Mean F		0.8558	0.8600	**0.8598**
MRR		0.6961	0.7279	**0.7319**
ACC	**5**	0.6336	0.6423	0.6430
Mean F		0.8547	0.8597	0.8595
MRR		0.6910	0.7233	0.7280

Table 3. Performance of GAP

As seen in Table 3, although the top 1 segmentation candidate could already achieve a certain performance level, taking the top 3 segmentation candidates could nevertheless considerably improve the MRR. This apparently suggests that the source name segmentation step could have significantly affected the overall performance of transliteration. Taking more segmentation candidates into account could help raise some correct transliterations to a higher rank, but there was not much improvement in terms of the accuracy at the top 1 position.

In terms of the grapheme pair frequency, setting the threshold at 3 gave only slightly better results than setting it at 5. A possible reason is that about 70% of all unique grapheme pairs have frequency below 5, and out of these over 47% only have single correspondence. In other words, there are a lot of grapheme pairs of low frequency, and for those ambiguous English segments, the distribution of their corresponding Chinese segments could be relatively uneven.

Hence the following comparison between various transliteration methods was based on the combination of N=3 and $f \geq 3$.

6.2 System Performance

To show the effectiveness of our proposed method, GAP was compared with JSCM and GAP-s. Table 4 shows the results of the three methods.

	JSCM	GAP-s	GAP
ACC	0.5760	0.6174	**0.6450**
Mean F	0.8309	0.8507	**0.8598**
MRR	0.6881	0.7175	**0.7319**

Table 4. System Performance Comparison

As evident from Table 4, system GAP-s outperformed JSCM. The accuracy at top 1 position is much improved, thus boosting the MRR too. This improvement therefore supports our hypothesis that contexts on both sides of a given segment are important for determining its rendition in Chinese, where part of the graphemic ambiguity could be successfully resolved. Meanwhile, system GAP further improves the results from GAP-s, bringing ACC up to 0.6450 and MRR to 0.7319. This shows that the phonological context could be better captured, though only approximately, by means of the last character of the previous segment and the first character of the next segment, instead of the whole neighbouring segments. This is because the phonological context is often most closely related to the neighbouring phonemes instead of a whole syllable.

6.3 Examples

In this section we show two examples from the experimental outcomes to illustrate the usefulness of the GAP method.

The name Abercromby, according to the gold standard, should be transliterated as 阿伯克龙比 *a4-bo2-ke4-long2-bi3*. This transliteration came third in the JSCM system, whose top first and second candidates were 阿伯克罗姆比 *a4-bo2-ke4-luo2-mu3-bi3* and 阿贝尔克罗姆比 *a4-bei4-er3-ke4-luo2-mu3-bi3* respectively. On the contrary, the expected transliteration came first in the GAP system.

The top 3 source name segmentation candidates for both methods are shown in Table 5. The expected segmentation has already been identified as the best candidate in GAP, while it came third in JSCM.

Top	JSCM	GAP
1	a/ber/c/ro/m/by	a/ber/c/rom/by
2	a/be/r/c/ro/m/by	a/ber/c/ro/m/by
3	a/ber/c/rom/by	a/be/r/c/rom/by

Table 5. Segmentations for Abercromby

When it comes to the evaluation of the transliteration candidates, the longer candidates could even score higher than the expected outcome in JSCM. The statistical data show that the bigram c/克+ro/罗 is far more likely than c/克+rom/龙, but $P(<e_k,c_k>=<\text{rom}, 龙> \mid fc(e_{k+1})=\text{b})$ is much stronger than $P(<e_k,c_k>=<\text{m}, 姆> \mid fc(e_{k+1})=\text{b})$. Hence, taking the character on both sides of a segment, GAP managed to rank 阿伯克龙比 highest.

Another example is the name Regelson, which is transliterated as 里格尔森 *li3-ge2-er3-sen1* in the gold standard. The expected transliteration is

ranked 8th in JSCM and 2nd in GAP. Although $P(<e_k,c_k>=<ge,杰> \mid <e_{k-1},c_{k-1}>=<re,里>)$ is much higher than $P(<e_k,c_k>=<ge,格> \mid <e_{k-1},c_{k-1}>=<re,里>)$, when taking the next segment <l,尔> into account, the likelihood of <ge,杰> is lowered. Hence the expected transliteration is ranked higher in GAP.

6.4 Error Analysis

As the proposed method stands, errors could have been propagated from two steps. The first is the source name segmentation step. If it happens that the top segmentation candidates are already wrong to start with, there is no way to reach the expected transliteration at all. Hence it is even more important to maintain a high accuracy for the segmentation step. The other error-propagation step is certainly when transliteration candidates are evaluated. The results for this step often heavily rely on the training data. If it happens that the grapheme pair distributions are somewhat skewed, particular Chinese segments would be preferred irrespective of relevant linguistic or other factors. On the other hand, if many homophones are used for a particular English segment, the chance of reaching the expected transliteration with one of the homophones is again loosened. More on this will be discussed in the next section.

For the latter error-propagation step, our attempt to make use of contexts on both sides of a segment has been shown to be able to improve the results. To see how much of the errors is attributable to the segmentation step, we roughly made an estimation by comparing the length of the top 1 candidates given in JSCM and GAP with the gold standard. It was found that 17.8% and 14.2% of the first candidates in JSCM and GAP respectively do not match the length of the gold standard. More detailed analysis of the segmentation results is in progress.

6.5 Current Limitations and Future Work

Our current treatment of neighbouring context and graphemic approximation of phonological context is shown to outperform pure DOM based on previous context only. Nevertheless, there are several directions of work which would require more investigation to further improve *E2C* performance.

First, the source name segmentation step needs further improvement to minimise error propagation from an early step. Phonological knowledge is obviously important in this regard as how a given English name should be segmented and pronounced is determined by its phonological context. Even without an explicit phonemic representation of the source names, more could be done in terms of modelling the phonological context via the surface graphemes.

Second, relating to the above, foreign names of different origins often have very different phonological properties leading to different pronunciations for the same orthographic forms. The silent h in Beckham mentioned earlier is one example, even though Chinese transliterations are often based on surface orthographic properties. Other problematic cases could be from languages like Russian and German where there are relatively more consonant clusters. For instance, the segment "scho" is often transliterated as one syllable (e.g. 绍 *shao4*, 肖 *xiao4*, or 舍 *she4*) but the segment "stro" often leads to three syllables (e.g. 斯特罗 *si1-te4-luo2*). It is therefore important to incorporate more phonological knowledge into the transliteration model, not only to generate more reliable and acceptable transliteration candidates, but also to reduce effort in evaluating phonologically invalid segmentation candidates and syllable structures, thus making the task computationally less expensive.

Third, as one of our separate ongoing studies shows, homophones are not only abundant in Chinese language per se, but also in *E2C* transliteration. The situation is particularly salient in Chinese transliterations based on Cantonese pronunciations. For example, while some names might have two transliterations with different pronunciations, like Jackson as 積遜 *zik1-seon3* or 積臣 *zik1-san4*, the same name might also be rendered in two forms with a different character having the same pronunciation, such as Adam as 亞當 or 阿當 (both pronounced as *aa3-dong1* in Cantonese). Two transliterations for the same name might have the same sound but different tones, e.g. Ashley as 艾殊利 *aai6-syu4-lei6* or 艾舒利 *aai6-syu1-lei6*. We therefore attempt to model the English-Chinese segment correspondence via an intermediate representation of the phonetic transcription of the Chinese characters. Preliminary results are reported in Kwong (2009). Although it happens that only one transliteration is given for each name in the gold standard data used in this study, the variability of *E2C* in reality is evident. It is therefore important for systems to be able to accommodate acceptable transliteration alternatives, particularly for transliteration extraction and information retrieval.

Fourth, given that tonal patterns could help distinguish some homophone ambiguity, the effect of the tonal factor and its potential association with the pitch and accent in the English names is worth further investigation.

7 Conclusion

Hence in this paper, we have reported our work on approximating phonological context for *E2C* with surface graphemic features. This is based on the observation that certain graphemic ambiguity is closely associated with the local contexts on both sides of a given segment, the phonological properties of which often determine its expected pronunciation. Experiments have shown that in the absence of an explicit phonemic representation of the English source names, the previous and next character of a given segment could be effectively employed to approximate the local phonological context affecting the rendition of a given segment in Chinese. Our proposed method GAP gives better results than the conventional JSCM which only makes use of previous context, and GAP-s which considers the whole neighbouring segments. Future work includes improving the source name segmentation step to minimise error propagation from an early stage, incorporating other factors like name origin and special phonological properties of different source languages into the transliteration model, as well as effectively handling homophones and tonal patterns in *E2C* transliteration.

Acknowledgements

The work described in this paper was substantially supported by a grant from City University of Hong Kong (Project No. 7002203).

References

Jin, C., Na, S-H., Kim, D-I. and Lee, J-H. (2008) Automatic Extraction of English-Chinese Transliteration Pairs using Dynamic Window and Tokenizer. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, Hyderabad, India, pp.9-15.

Kantor, P.B. and Voorhees, E.M. (2000) The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval, 2(2-3)*: 165-176.

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

Kuo, J-S. and Li, H. (2008) Mining Transliterations from Web Query Results: An Incremental Approach. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, Hyderabad, India, pp.16-23.

Kwong, O.Y. (2009) Homophones and Tonal Patterns in English-Chinese Transliteration. To appear in *Proceedings of ACL-IJCNLP 2009*, Singapore.

Lee, C-J., Chang, J.S. and Jang, J-S. R. (2006) Extraction of transliteration pairs from parallel corpora using a statistical transliteration model. *Information Sciences, 176*:67-90.

Li, H., Zhang, M. and Su, J. (2004) A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, pp.159-166.

Li, H., Sim, K.C., Kuo, J-S. and Dong, M. (2007) Semantic Transliteration of Personal Names. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, pp.120-127.

Oh, J-H. and Choi, K-S. (2005) An Ensemble of Grapheme and Phoneme for Machine Transliteration. In R. Dale, K-F. Wong, J. Su and O.Y. Kwong (Eds.), *Natural Language Processing – IJCNLP 2005*. Springer, LNAI Vol. 3651, pp.451-461.

Sproat, R., Shih, C., Gale, W. and Chang, N. (1996) A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics, 22(3)*: 377-404.

Tao, T., Yoon, S-Y., Fister, A., Sproat, R. and Zhai, C. (2006) Unsupervised Named Entity Transliteration Using Temporal and Phonetic Correlation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australia, pp.250-257.

Virga, P. and Khudanpur, S. (2003) Transliteration of Proper Names in Cross-lingual Information Retrieval. In *Proceedings of the ACL2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*.

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

Yoon, S-Y., Kim, K-Y. and Sproat, R. (2007) Multilingual Transliteration Using Feature based Phonetic Method. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, pp.112-119.