Phonological Context Approximation and Homophone Treatment for NEWS 2009 English-Chinese Transliteration Shared Task

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

This paper describes our systems participating in the NEWS 2009 Machine Transliteration Shared Task. Two runs were submitted for the English-Chinese track. The system for the standard run is based on graphemic approximation of local phonological context. The one for the non-standard run is based on parallel modelling of sound and tone patterns for treating homophones in Chinese. Official results show that both systems stand in the mid range amongst all participating systems.

1 Introduction

This paper describes our systems participating in the English-Chinese track of the NEWS 2009 Machine Transliteration Shared Task.

The apparently free combination of Chinese characters in names is not entirely ungoverned. There are no more than a few hundred Chinese characters which are used in names. Moreover, beyond linguistic and phonetic properties, many social and cognitive factors are simultaneously influencing the naming process and superimposing on the surface graphemic correspondence.

Our systems in the standard and non-standard runs aim at addressing two issues in English-Chinese forward transliteration (referred to as *E2C* hereafter), namely graphemic ambiguity and homophones in Chinese respectively.

By graphemic ambiguity, we refer to the multiple mappings between English segments and Chinese segments. For example, the English segment "ty" could be rendered as 蒂 *di4* as in Chris**ty** 克里斯蒂 *ke4-li3-si1-di4*, or 太 *tai4* as in S**ty**le 斯太尔 *si1-tai4-er3*[1]. Although direct orthographic mapping (e.g. Li *et al*., 2004) has been shown to work even more effectively than phoneme-based methods (e.g. Virga and Khudanpur, 2003), it is observed that phonological context plays an important role in resolving graphemic ambiguity. In the absence of an explicit phonemic representation of the source names, our GAP system, to be described in Section 4.1, attempts to approximate the local phonological context for a given segment by means of surface graphemic properties.

An English name could be acceptably transliterated in various ways, e.g. 希拉里 *xi1-la1-li3*, 希拉利 *xi1-la1-li4*, 希拉莉 *xi1-la1-li4*, as well as 希拉蕊 *xi1-la1-rui3* are all possible transliterations for Hilary. Homophones are abundant in Chinese, as evident from the first three alternatives above. However, conventional transliteration models often rely heavily on the distribution of the training data, which might preclude infrequent but similarly acceptable transliteration candidates. Also, Chinese is a typical tonal language. The sound-tone combination is important in names. Names which sound "nice" are often preferred to those which sound "monotonous". Our SoToP system to be described in Section 4.2 thus attempts to model sound and tone patterns in parallel, to deal with homophones more reasonably despite possible skewed prior distributions.

Related work will be briefly reviewed in Section 2, and the datasets will be described in Section 3. The systems for both runs and their performance will be reported in Section 4, followed by future work and conclusion in Section 5.

2 Related Work

There are basically two categories of work on machine transliteration. First, various alignment models are used for acquiring transliteration

[1] The transcriptions in this paper are in Hanyu Pinyin.

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 76–79,
Suntec, Singapore, 7 August 2009. ©2009 ACL and AFNLP

lexicons from parallel corpora and other resources (e.g. Kuo and Li, 2008). Second, statistical models are built for transliteration. These models could be phoneme-based (e.g. Knight and Graehl, 1998), grapheme-based (e.g. Li *et al.*, 2004), hybrid (Oh and Choi, 2005), or based on phonetic (e.g. Tao *et al.*, 2006) and semantic (e.g. Li *et al.*, 2007) features.

The core of our systems is based on Li *et al.*'s (2004) Joint Source-Channel Model under the direct orthographic mapping framework, which skips the middle phonemic representation in conventional phoneme-based methods and models the segmentation and alignment preferences by means of contextual n-grams of the transliteration segment pairs (or token pairs in their terminology). A bigram model under their framework is thus 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.

3 Datasets

The current study used the English-Chinese (EnCh) data provided by the shared task organisers. There are 31,961 English-Chinese name pairs in the training set, 2,896 English-Chinese name pairs in the development set, and another 2,896 English names in the test set. The Chinese transliterations basically correspond to Mandarin Chinese pronunciations of the English names, as used by media in Mainland China (Xinhua News Agency, 1992).

The training and development data were manually cleaned up and aligned with respect to the correspondence between English segments and Chinese segments, e.g. Aa/l/to 阿/尔/托, and the pronunciations for the Chinese characters were automatically looked up.

Based on all the unique English segments resulting from manual alignment, all possible segmentations of a test name were first obtained, and they were then ranked using a probabilistic score computed by:

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

where S is a segmentation sequence with K segments, s_k is the kth segment in S, $lc(s_{k-1})$ is the last character of segment s_{k-1} and $fc(s_{k+1})$ is the first character of segment s_{k+1}.

4 System Description

4.1 Standard Run – GAP

Our system for the standard run is called GAP, which stands for Graphemic Approximation of Phonological context.

Although direct orthographic mapping has been shown to be an effective method, it is nevertheless observed that phonological context significantly contributes to the resolution of some graphemic ambiguity. For example, the English segment "le" was found to correspond to as many as 15 Chinese segments in the data, 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 rendered as 尔 *er3*, e.g. Dale 戴尔 *dai4-er3* and Dipasquale 迪帕斯奎尔 *di2-pa4-si1-kui2-er3*. This is especially true when the previous character is "a". On the contrary, when "le" appears at the end of a name following an "r", it is more often rendered as 利 *li4* instead, e.g. Berle 伯利 *bo2-li4*. On the other hand, "le" at the beginning of name is often 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*. Such observation thus indicates two important points for *E2C*. First, the phonological context is useful as English graphemic segments could be ambiguous in terms of pronunciation, and the actual pronunciation often determines which Chinese segment is to be used. Second, local contexts on both sides are important as they indicate the environment in which the segment is embedded, which might affect the way it is pronounced.

GAP thus attempts to approximate local phonological context by means of surface graphemic properties, making use of bigrams in both directions. Since the phonological environment might be sufficiently represented by a neighbouring phoneme instead of a whole syllable, we approximate the phonological context with one character on both sides of a given English segment, irrespective of their corresponding Chinese

segments. Using single characters on both sides could also ensure that a small and consistent parameter space is maintained. Hence, weighting the context on both sides equally, GAP assigns a score *Score(E,C)* to a transliteration candidate with *K* segment pairs as follows:

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

where $<e_k,c_k>$ is the *k*th English-Chinese segment pair, $lc(e_{k-1})$ is the last character of segment e_{k-1} and $fc(e_{k+1})$ is the first character of segment e_{k+1}.

Taking the top 3 segmentation candidates, the transliteration candidates were generated by looking up the grapheme pairs obtained from manual alignment with frequency $f \geq 3$. If there is no grapheme pair above the threshold, all pairs below the threshold would be considered. All combinations obtained were then subject to ranking with *Score(E,C)* above.

4.2 Non-standard Run – SoToP

The homophone problem is notorious in Chinese. As far as personal name transliteration is concerned, unless there are standardised principles prescribed, the "correctness" of transliterated names is not clear-cut at all. As a tonal language, how a combination of characters sounds is also important in naming. As in the example given in Section 1, one cannot really say any of the transliterations for Hilary is "right" or "wrong", but perhaps only "better" or "worse". Hence naming is more of an art than a science, and automatic transliteration should avoid over-reliance on the training data and thus missing unlikely but good alternative candidates.

Our system for the non-standard run, SoToP, thus aims at addressing this cognitive or perceptual aspect of transliteration beyond its linguistic and phonetic properties. Instead of direct orthographic mapping, we use a Sound model (SoM) and a Tone model (ToM) in Parallel. The SoToP architecture is shown in Figure 1.

SoM basically assembles the homophones and captures the sound patterns in terms of a grapheme-phoneme mapping. The operation of SoM is like GAP above, except that the $<e_k,c_k>$ pairs are replaced by $<e_k,so_k>$ pairs, where so_k refers to the phonetic transcription in Hanyu Pinyin (without tone) for the *k*th Chinese segment in a candidate.

ToM, on the other hand, captures the tone patterns of transliteration, irrespective of the sound and the character choice. Although English does not have tones, the intonation and stress of a syllable may prompt for the usage of a Chinese character of a certain tone. Chinese, on the other hand, is a tonal language. The tone patterns are more cognitive in nature, as some combinations may just sound awkward for no apparent reason. Moreover, some sound-tone combinations might result in undesirable homophones, which are also avoided in names in general. The operation of ToM is also like GAP, except that the $<e_k,c_k>$ pairs are replaced by $<e_k,to_k>$ pairs, where to_k refers to the tone for the *k*th Chinese segment in a candidate.

The Candidate Generator combines the top *M* candidates from ToM and top *N* candidates from SoM to generate character combinations by looking up a pronunciation table. The lookup table lists the homophones for each sound-tone combination found in the data. In the current study, both *M* and *N* were set to 3. The generated candidates were then ranked by a simple bigram model based on the bigram probabilities of the Chinese segments.

4.3 System Testing

The two systems were tested on the NEWS development data, containing 2,896 English names. System performance was measured by the following evaluation metrics: Word Accuracy in Top-1 (ACC), Fuzziness in Top-1 (Mean F-score), Mean Reciprocal Rank (MRR), MAP_{ref}, MAP_{10}, and MAP_{sys}. Detailed description of these metrics can be found in the NEWS shared task whitepaper (Li *et al*., 2009).

Table 1 shows the system testing results on the development data. The standard run, GAP, in general gives better results than the non-standard run, SoToP. One possible reason is apart from the source name segmentation step, SoToP has more steps allowing error propagation as the mapping was done separately with sound and tone, whereas GAP directly maps English segments to Chinese segments at the graphemic level.

Metric	GAP	SoToP
ACC	0.645	0.597
Mean F-score	0.860	0.836
MRR	0.732	0.674
MAP_{ref}	0.645	0.597
MAP_{10}	0.223	0.206
MAP_{sys}	0.225	0.335

Table 1. System Testing Results

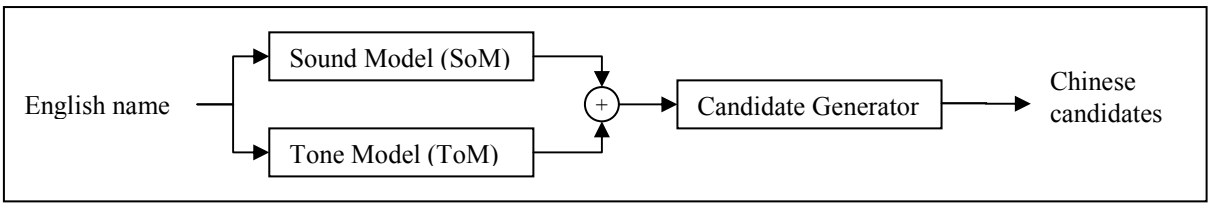

Figure 1. The SoToP Architecture for *E2C* Transliteration

4.4 Official Results

The two systems were trained on both the training data and development data together, and run on the test data. The official results are shown in Table 2. The performance of the two systems is in the mid range amongst all participating systems, including standard and non-standard runs. Despite the shortcoming and lower performance of SoToP, modelling the sound and tone patterns has its merits for handling homophones. For example, the expected transliteration for Mcgiveran, 麦吉弗伦 *mai4-ji2-fu2-lun2*, was ranked 6th by GAP but 1st by SoToP. The segment "ve" is much more likely rendered as 夫 *fu1* than as 弗 *fu2*, but ToM in SoToP was able to capture the preferred tone pattern 4-2-2-2 in this case.

Metric	GAP	SoToP
ACC	0.621	0.587
Mean F-score	0.852	0.834
MRR	0.718	0.665
MAP_{ref}	0.621	0.587
MAP_{10}	0.220	0.203
MAP_{sys}	0.222	0.330

Table 2. Official Results on Test Data

5 Future Work and Conclusion

Thus we have reported on the two systems participating in the NEWS shared task. The standard run, GAP, relies on direct orthographic mapping and approximates local phonological context with neighbouring graphemes to help resolve graphemic ambiguity. The non-standard run, SoToP, attempts to address the homophone issues in Chinese, by modelling the sound and tone patterns in parallel, and subsequently combining them to generate transliteration candidates. In general GAP gives better results than SoToP, while both are in the mid range amongst all participating systems. Future work includes more error analysis and improving the accuracy of individual steps to minimise error propagation. The possible combination of the two methods is also worth further investigation.

Acknowledgements

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

References

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 SIGHAN-6*, Hyderabad, India, pp.16-23.

Li, H., Zhang, M. and Su, J. (2004) A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of the 42nd Annual Meeting of ACL*, 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 45th Annual Meeting of ACL*, Prague, Czech Republic, pp.120-127.

Li, H., Kumaran, A., Zhang, M. and Pervouchine, V. (2009) Whitepaper of NEWS 2009 Machine Transliteration Shared Task. In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop (NEWS 2009)*, Singapore.

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

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 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.