English-to-Chinese Machine Transliteration using Accessor Variety Features of Source Graphemes

Mike Tian-Jian Jiang
Department of Computer Science, National Tsing Hua University
Institute of Information Science, Academia Sinica
tmjiang@iis.sinica.edu.tw

Chan-Hung Kuo
Institute of Information Science,
Academia Sinica
laybow@iis.sinica.edu.tw

Wen-Lian Hsu
Institute of Information Science,
Academia Sinica
hsu@iis.sinica.edu.tw

Abstract

This work presents a grapheme-based approach of English-to-Chinese (E2C) transliteration, which consists of many-to-many (M2M) alignment and conditional random fields (CRF) using accessor variety (AV) as an additional feature to approximate local context of source graphemes. Experiment results show that the AV of a given English named entity generally improves effectiveness of E2C transliteration.

1 Introduction

Transliteration is a subfield of computation linguistics, and is defined as the phonetic translation of names across languages. Transliteration of named entities is essential in numerous applications, such as machine translation, corpus alignment, cross-language information retrieval, information extraction, and automatic lexicon acquisition. The transliteration modeling approaches can be classified as phoneme-based, grapheme-based, and a hybrid of phoneme and grapheme.

Numerous studies focus on the phoneme-based approach (Knight and Graehl, 1998; Virga and Khudanpur, 2003). Suppose that *E* is an English name and *C* is its Chinese transliteration, the phoneme-based approach first converts *E* into an intermediate phonemic representation *p*, and then converts *p* into its Chinese counterpart *C*. The idea is to transform both the source and target names into comparable phonemes so that the phonetic similarity between the two names can be measured easily. The grapheme-based approach, which treats the transliteration as a statistical machine translation problem under monotonic constraint, has also attracted much attention (Li *et al.*, 2004). This approach aims to obtain the bilingual orthographical correspondence directly to reduce the possible errors introduced in multiple conversions. The hybrid approach attempts to utilize both phoneme and grapheme information for transliteration. Oh and Choi (2006) proposed a strategy to include both phoneme and grapheme features in a single learning process.

This work presents a grapheme-based approach of English-to-Chinese (E2C) transliteration using many-to-many alignment (M2M-aligner) (Jiampojamarn *et al.*, 2007) and conditional random fields (CRF) (Lafferty *et al.*, 2001) with additional features of accessor variety (AV) (Feng *et al.*, 2004). The remainder of this article is organized as follows. Section 2 briefly introduces related works involving M2M-aligner, CRF, and AV. The concept of this work for transliteration using M2M-aligner, CRF, and AV are explained in Section 3. Section 4 describes the experiment results and discussion. Finally, the conclusion is presented in Section 5.

2 Related Works

2.1 CRF-based Transliteration

Yang *et al.* (2009) proposed a two-step CRF model for direct orthographical mapping (DOM) machine transliteration, in which the first CRF segments a source word into chunks and the second CRF maps the chunks to a word in the target language. Reddy and Waxmonsky (2009) presented a phrase-based translation system that characters are grouped into substrings to be mapped atomically into the target language, which showed how substring representation can be incorporated into a CRF model with local context and phonemic information. Shishtla *et al.* (2009) adopted a statistical transliteration technique that consists of alignment model of GIZA++ (Och and Ney, 2003) and CRF model.

Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages 86–90,
Chiang Mai, Thailand, November 12, 2011.

The approach of this work is similar to the technique of Shishtla *et al*., yet this work focuses on the additional AV feature of CRF and uses M2M-aligner, which will be described in Section 2.2, instead of GIZA++.

2.2 M2M-Aligner

Jiampojamarn *et al.* (2007) argued that previous work has generally assumed one-to-one alignment for simplicity, but letter strings and phoneme strings are not typically in the same length, so null phonemes or null letters must be introduced to make one-to-one-alignments possible. Furthermore, two letters frequently combine to produce a single phoneme (double letters), and a single letter can sometimes produce two phonemes (double phonemes). For example, the English word "ABERT" with its Chinese transliteration "阿贝特", which Jaimpojamarn *et al*. referred as "phonemes", is aligned as:

A	BE	RT
|	|	|
阿	贝	特

The letters "BE" are an example of the double letter problem which mapping to the single phoneme "贝." These alignments provide more accurate grapheme-to-phoneme relationships for a phoneme prediction model. Hence the M2M-aligner is for alignments between substrings of various lengths and based on the expectation maximization (EM) algorithm. For more details of the algorithm, readers are encouraged to explore previous works of Ristad and Yianilos (1998), and Jiampojamarn *et al*. (2007).

Despite ambiguity between Chinese transliteration and phoneme, the above paragraph of the opinion of Jaimpojamarn *et al*. indicates a particular problem of E2C transliteration, that the training data comprised pairs of names written in source and target scripts lacks explicit grapheme-level alignment. This work uses M2M-aligner as an unsupervised method for generating alignments of the training data, which provide hypotheses of DOM without null graphemes.

2.3 Accessor Variety

Feng *et al.* (2004) proposed accessor variety (AV) to measure how likely a character substring is a Chinese word. Another similar measurement of English and Chinese words called boundary entropy or branching entropy (BE) was used in several works (Tung and Lee, 1994; Chang and Su, 1997; Cohen and Adams, 2001; Cohen *et al.*, 2002; Huang and Powers, 2003; Tanaka-Ishii, 2005; Jin and Tanaka-Ishii, 2006; Cohen *et al.*, 2007). The basic idea behind these measurements is closely related to one particular perspective of *n*-gram and information theory of cross entropy or perplexity. Zhao and Kit (2007) induced that AV and BE both assume that the border of a potential word is located where the uncertainty of successive characters increases, where AV and BE are regarded as the discrete and continuous versions, respectively, of the fundamental work of Harris (1970), and then chose to adopt AV as the additional feature of CRF-based Chinese Word Segmentation (CWS). The AV of a string *s* is defined as:

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\} \quad (1)$$

In Eq. (1), L_{av}(s) and R_{av}(s) are defined as the number of distinct preceding and succeeding characters, except when the adjacent character is absent due to a sentence boundary, and then the pseudo-character of the beginning or end of a sentence is accumulated indistinctly. Feng *et al.* (2004) also developed more heuristic rules to remove strings that contain known words or adhesive characters. For the strict meaning of unsupervised features and for simplicity, this study does not include those additional rules.

The necessity of AV is primarily on the demand for semi-supervised learning. Since AV can be extracted from large corpora without any manual segmentation or annotation, hidden variables underlying frequent surface patterns of languages may be captured via an inexpensive and unsupervised algorithm such as suffix array. Unsupervised feature selection of AV or similar features has generally improved effectiveness of supervised CWS on cross-domain and unlabeled data (Jiang *et al*., 2010), and this work consequently considers that AV of un-segmented English names from training, development, and test data might help enhancing E2C transliteration.

3 Transliteration using EM and CRF

3.1 CRF Alignment Labeling

In the work, M2M-aligner first maximizes the probability of the observed source-target word pairs using the EM algorithm and subsequently sets the grapheme alignments via maximum a posteriori estimation. CRF is then conditioned on the grapheme alignments to produce globally

optimal solutions. However, the performance of the EM algorithm is frequently affected by the initialization. To obtain better alignment results of M2M-aligner, this work empirically sets the "maxX" parameter for the maximum size of sub-alignments in the source side to 8, and sets the "maxY" parameter for the maximum size of sub-alignments in the target side to 1 (denoted as X8Y1 in short), since one of the well known *a priori* of Chinese is that almost all Chinese characters are monosyllabic, which reflects the situation of "double phoneme" mentioned in Section 2.2. Notably, this work follows the definition of grapheme described by Oh and Choi (2005) to prevent from confusion of phoneme, grapheme, character, and letter, that graphemes refer to the basic units (or the smallest contrastive units) of written language: for example, English has 26 graphemes or letters or characters, Korean has 24, and German has 30. Table 1 is an example of M2M-aligner results. With aligned training data, a transliteration model can be then trained by CRF to generate names in the target language from names in the source language. This work uses Wapiti (Lavergne *et al.*, 2010) as CRF toolkit. Table 2 is an example of training data for a CRF alignment labeling, where the tags *B* and *I* indicate whether the grapheme is in the starting position of the sub-alignment.

This work tests several combinations of conventional CRF features along with their abbreviated notations for E2C transliteration, as shown in Table 3, where C_i represents the input graphemes bound individually to the prediction label at its current position *i*. Take Table 2 as an example, if the current position is at the label "*B* 迪", features generated by C_{-1}, C_0 and C_1 are "A" "D" and "I" respectively. Note that a prediction label may either comprise a positioning tag and a Chinese grapheme, or just be the positioning tag itself.

Source	Target	M2M-Aligner Result	
ABBADIE	阿巴迪	A:B|B:A|D:I:E|	阿|巴|迪|

Table 1. An Example of M2M Alignment

Character	Label
A	*B*阿
B	*I*
B	*B*巴
A	*I*
D	*B*迪
I	*I*
E	*I*

Table 2. Example of a CRF labeling format for E2C transliteration

	Context		
Function	C_0, C_{-1}, C_1, C_0C_1, $C_{-1}C_0$,	C_0, C_{-1}, C_1, C_{-2}, C_2 C_0C_1, $C_{-1}C_0$, $C_{-2}C_1$, C_1C_2	C_0, C_{-1}, C_1, C_{-2}, C_2 C_{-3}, C_3 C_0C_1, $C_{-1}C_0$, $C_{-2}C_1$, C_1C_2 $C_{-3}C_{-2}$, C_2C_3
Notation	1UB	2UB	3UB
Positioning Tag of Prediction Label			
Function	*B, I*		*B, I, E*
Notation	P_{BI}		P_{BIE}
Chinese Grapheme of Prediction Label			
Function	On *B* only		On *B* and *I*
Notation	G_B		G_{BI}

Table 3. Conventional CRF Features

3.2 CRF with AV

This work extends the work of Zhao and Kit (2008) into a unified representation for AV features of English graphemes. The representation accommodates both the position of a string and the string's likelihood ranking by the logarithm. Formally, the ranking function for a string, *s,* with a score, *x,* counted by AV is defined as:

$$f(s) = r, if\ 2^r \le x < 2^{r+1} \quad (2)$$

.

The logarithm ranking mechanism in Eq. (2) is inspired by Zipf's law to alleviate the potential data sparseness of infrequent strings. The rank *r* and the corresponding positions of a string are then concatenated as feature tokens. To provide readers with a clearer picture of the appearance of feature tokens, a sample representation for AV is presented and explained in Table 4.

For example, considering strings with two graphemes, one of the strings "AB" is ranked *r* = 3; therefore, the column of di-grapheme feature tokens has "A" denoted as *3B* and "B" denoted as *3E*. If another di-grapheme string, "BA,"

Input	AV Feature					Label
	1 char	2 char	3 char	4 char	5 char	
A	*7S*	*3B*	*2B*	*0B*	*1B*	*B*阿
B	*5S*	*3E*	*2B*	*0B*	*1B*	*I*
B	*5S*	*3B*	*2B*	*0B*	*1B*	*B*巴
A	*7S*	*4B*	*2B*	*1B*	*1B*	*I*
D	*7S*	*4E*	*3B*	$1B_1$	*1E*	*B*迪
I	*5S*	*4E*	$3B_1$	$1B_2$	*0E*	*I*
E	*7S*	*3E*	*3E*	*1E*	*0E*	*I*

Table 4. Example of AV features

competes with "AD" at the position of "A" with a higher rank of r = 4, then *4B* is selected for feature representation of the token at a certain position. Notably, when the string "AD" conflicts with the string "DI" at the position of "D" with the same rank of r = 4, the corresponding position with the ranking of the leftmost string, which is *4E* in this case, is applied arbitrarily.

4 Results and Discussions

4.1 E2C Transliteration Results

In the interest of brevity, only the 3rd and the 4th standard runs that exceed 0.3 in terms of top-1 accuracy (ACC) are listed in Table 5. Numerous models of pilot tests have been trained using both the training set and the development set, and then evaluated on the development set for optimizing CRF feature combinations, as shown in Table 6.

4.2 Error Analysis and Discussions

Based on observations of the pilot tests, there is a clear trend that AV features improve performances significantly. However, improvements on the test set are not as good as expected. After carefully investigating NEWS-2011 data, one particular phenomenon has been noticed: only the development set contains phrasal named entities. Furthermore, some E2C word pairs are not pure transliterations and aligned in very different character lengths, such as the word pair of "COMMONWEALTH OF THE BAHAMAS" and "巴哈马联邦," and this phenomenon is noted as "semi-semantic transliteration" for convenience. In fact, the M2M parameter "maxX" of this work has been designed for these phrasal structure to be relatively larger and less symmetrical to the parameter "maxY" than previous works that usually set both X and Y to 2 as default values. Since the M2M and the CRF models might over-fit the development set, phrasal structure and semi-semantic transliterations that only appeared in the development set probably became noises according to the test set.

ID	Configuration	ACC	Mean F-score
4	X8Y1, 3UB, P_{BIE}, G_B, AV	0.327	0.688
3	X8Y1, 2UB, P_{BI}, G_{BI}, AV	0.303	0.675

Table 5. Selected E2C standard runs

Configuration	ACC	Mean F-score
X8Y1, 1UB, P_{BI}, G_B	0.001	0.151
X8Y1, 1UB, P_{BI}, G_B, AV	0.000	0.078
X8Y1, 2UB, P_{BI}, G_B	0.001	0.122
X8Y1, 2UB, P_{BI}, G_B, AV	0.000	0.064
X8Y1, 3UB, P_{BI}, G_B, AV	0.569	0.860
X8Y1, 1UB, P_{BI}, G_{BI}	0.454	0.762
X8Y1, 1UB, P_{BI}, G_{BI}, AV	0.547	0.813
X8Y1, 2UB, P_{BI}, G_{BI}	0.547	0.814
X8Y1, 2UB, P_{BI}, G_{BI}, AV	0.753	0.910
X8Y1, 1UB, P_{BIE}, G_B	0.182	0.586
X8Y1, 1UB, P_{BIE}, G_B, AV	0.273	0.656
X8Y1, 2UB, P_{BIE}, G_B	0.347	0.708
X8Y1, 2UB, P_{BIE}, G_B, AV	0.483	0.800
X8Y1, 3UB, P_{BIE}, G_B	0.449	0.771
X8Y1, 3UB, P_{BIE}, G_B, AV	0.597	0.857

Table 6. Selected E2C pilot tests

To analyze semi-semantic transliterations, NEWS-2011 Chinese-to-English (C2E) back-transliteration corpus have been acquired, and the corresponding standard runs have been submitted owing to the policy of NEWS shared task. The C2E experiments, however, encountered a serious problem of CRF L-BFGS training requirement on space complexity, therefore the submitted results are actually incomplete and erroneous, since C2E transliteration using the proposed approach produces too many labels and features to train a CRF model with the whole training set. In authors' experiences, even a workstation with 24GB memory spaces is insufficient for such training. Notably, the similar hardware constraint makes the 4th standard run of E2C, which is the primary one, to regress to the simpler Chinese grapheme labeling strategy, namely G_B, while introducing deeper contexts and more specific positioning tags, to trade efficiency of CRF training phases.

5 Conclusion and Future Work

This work proposes to use AV of source grapheme for E2C transliteration. Experiments indicate the AV features generally improve the performance in terms of ACC. Recommended future investigations would be features of target graphemes or source-channel models (Li *et al*., 2004) that are efficient and capable of recognizing semi-semantic transliteration.

Acknowledgements
This research was supported in part by the National Science Council under grant NSC 100-2631-S-001-001, and the research center for Humanities and Social Sciences under grant IIS-50-23. Wallace Academic Editing service is appreciated for their editorial assistance. The authors would like to thank anonymous reviewers for their constructive criticisms.

References

Jing-Shin Chang and Keh-Yih Su. 1997. An unsupervised iterative method for Chinese new lexicon extraction. *Computation Linguistics and Chinese language Processing*, 2(2):97-148.

Paul Cohen and Niall Adams. 2001. An Algorithm for Segmenting Categorical Time Series into Meaningful Episodes. *Advances in Intelligent Data Analysis*, 198-207.

Paul Cohen, Niall Adams and Brent Heeringa. 2007. Voting Experts: An Unsupervised Algorithm for Segmenting Sequences. *Intelligent Data Analysis*, 11(6):607-625.

Paul R Cohen, B Heeringa and Niall M Adams. 2002. An Unsupervised Algorithm for Segmenting Categorical Timeseries into Episodes. *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*, 49-62.

Haodi Feng, Kang Chen, Xiaotie Deng, and Wiemin Zheng. 2004. Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1):75-93.

Zellig Sabbetai Harris. 1970. Morpheme boundaries within words. *Papers in Structural and Transformational Linguistics*, 68-77.

Jin Hu Huang and David Powers. 2003. Chinese Word Segmentation based on contextual entropy. *Proceedings of the 17th Asian Pacific Conference on Language, Information and Computation*, 152-158.

Sittichai Jiampojamarn, Grzegorz Kondrak and Tarek Sherif. 2007. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 372-379.

Tian-Jian Jiang, Shih-Hung Liu, Cheng-Lung Sung and Wen-Lian Hsu. 2010. Term Contributed Boundary Tagging by Conditional Random Fields for SIGHAN 2010 Chinese Word Segmentation Bakeoff. *Proceeding of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 266-269.

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

John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of ICML*, 591-598.

Thomas Lavergne, Oliver Cappé and François Yvon. 2010. Practical Very Large Scale CRFs. *Proceedings the 48th ACL*, 504-513.

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

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51.

J. H. Oh and K. S. Choi. 2006. An Ensemble of Transliteration Models for Information Retrieval. *Information Processing and Management*, 42:980-1002.

Eric Sven Ristad and Peter N. Yianilos. 1998. Learning String Edit Distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 20(5):522-532.

Sravana Reddy and Sonjia Waxmonsky. 2009. Substring-based transliteration with conditional random fields. *Proceedings of the 2009 Named Entities Workshop,* 92-95.

Praneeth Shishtla, V. Surya Ganesh, Sethuramalingam Subramaniam and Vasudeva Varma. 2009. A language-independent transliteration schema using character aligned models at NEWS 2009. *Proceedings of the 2009 Named Entities Workshop,* 40-43.

Kumiko Tanaka-Ishii. 2005. Entropy as an Indicator of Context Boundaries: An Experiment Using a Web Search Engine. *Proceedings of International Joint Conference on Natural Language Processing* , 93-105.

Cheng-Huang Tung and His-Jian Lee. 1994. Identification of unknown words from corpus. *Computational Proceedings of Chinese and Oriental Languages*, 131-145.

P. Virga and S. Khudanpur. 2003. Transliteration of Proper Names in Cross-lingual Information Retrieval. In the *Proceedings of the ACL Workshop on Multi-lingual Named Entity Recognition*.

Dong Yang, Paul Dixon, Yi-Cheng Pan, Tasuku Oonishi, Masanobu Nakamura, Sadaoki Furui. 2009. Combining a two-step conditional random field model and a joint source channel model for machine transliteration. *Proceedings of the 2009 Named Entities Workshop,* 72-75.

Hai Zhao and Chunyu Kit. 2007. Incorporating Global Information into Supervised Learning for Chinese Word Segmentation. *Proceedings of the 10th Conference of the Pacific Association for Computation Linguistics,* 66-74.

Hai Zhao and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.