Name Origin Recognition Using Maximum Entropy Model and Diverse Features

Min Zhang[1], Chengjie Sun[2], Haizhou Li[1], Aiti Aw[1], Chew Lim Tan[3], Xiaolong Wang[2]

[1]Institute for Infocomm Research, Singapore
{mzhang,hli,aaiti}@i2r.a-star.edu.sg

[2]Harbin Institute of Technology, China
{cjsun,wangxl}@insun.hit.edu.cn

[3]National University of Singapore, Singapore
tancl@comp.nus.edu.sg

Abstract

Name origin recognition is to identify the source language of a personal or location name. Some early work used either rule-based or statistical methods with single knowledge source. In this paper, we cast the name origin recognition as a multi-class classification problem and approach the problem using Maximum Entropy method. In doing so, we investigate the use of different features, including phonetic rules, *n*-gram statistics and character position information for name origin recognition. Experiments on a publicly available personal name database show that the proposed approach achieves an overall accuracy of 98.44% for names written in English and 98.10% for names written in Chinese, which are significantly and consistently better than those in reported work.

1 Introduction

Many technical terms and proper names, such as personal, location and organization names, are translated from one language into another with approximate phonetic equivalents. The phonetic translation practice is referred to as *transliteration*; conversely, the process of recovering a word in its native language from a transliteration is called as *back-transliteration* (Zhang et al, 2004; Knight and Graehl, 1998). For example, English name "Smith" and "史密斯 (*Pinyin*[1]: Shi-Mi-Si)" in Chinese form a pair of transliteration and *back-transliteration*. In many natural language processing tasks, such as machine translation and cross-lingual information retrieval, automatic name transliteration has become an indispensable component.

Name origin refers to the source language of a name where it originates from. For example, the origin of the English name "Smith" and its Chinese transliteration "史密斯 (Shi-Mi-Si)" is English, while both "Tokyo" and "东京 (Dong-Jing)" are of Japanese origin. Following are examples of different origins of a collection of English-Chinese transliterations.

English:	Richard-理查德 (Li-Cha-De)
	Hackensack-哈肯萨克(Ha-Ken-Sa-Ke)
Chinese:	Wen JiaBao-温家宝(Wen-Jia-Bao)
	ShenZhen–深圳(Shen-Zhen)
Japanese:	Matsumoto-松本 (Song-Ben)
	Hokkaido-北海道(Bei-Hai-Dao)
Korean:	Roh MooHyun-卢武铉(Lu-Wu-Xuan)
	Taejon-大田(Da-Tian)
Vietnamese:	Phan Van Khai-潘文凯(Pan-Wen-Kai)
	Hanoi-河内(He-Nei)

In the case of machine transliteration, the name origins dictate the way we re-write a foreign word. For example, given a name written in English or Chinese for which we do not have a translation in

[1] *Hanyu Pinyin*, or *Pinyin* in short, is the standard romanization system of Chinese. In this paper, *Pinyin* is given next to Chinese characters in round brackets for ease of reading.

a English-Chinese dictionary, we first have to decide whether the name is of Chinese, Japanese, Korean or some European/English origins. Then we follow the transliteration rules implied by the origin of the source name. Although all English personal names are rendered in 26 letters, they may come from different romanization systems. Each romanization system has its own rewriting rules. English name "Smith" could be directly transliterated into Chinese as "史密斯(Shi-Mi-Si)" since it follows the English phonetic rules, while the Chinese translation of Japanese name "Koizumi" becomes "小泉(Xiao-Quan)" following the Japanese phonetic rules. The name origins are equally important in back-transliteration practice. Li et al. (2007) incorporated name origin recognition to improve the performance of personal name transliteration. Besides multilingual processing, the name origin also provides useful semantic information (regional and language information) for common NLP tasks, such as co-reference resolution and name entity recognition.

Unfortunately, little attention has been given to name origin recognition (NOR) so far in the literature. In this paper, we are interested in two kinds of name origin recognition: the origin of names written in English (**ENOR**) and the origin of names written in Chinese (**CNOR**). For ENOR, the origins include English (Eng), Japanese (Jap), Chinese Mandarin *Pinyin* (Man) and Chinese Cantonese *Jyutping* (Can). For CNOR, they include three origins: Chinese (Chi, for both Mandarin and Cantonese), Japanese and English (refer to Latin-scripted language).

Unlike previous work (Qu and Grefenstette, 2004; Li et al., 2006; Li et al., 2007) where NOR was formulated with a generative model, we regard the NOR task as a classification problem. We further propose using a discriminative learning algorithm (Maximum Entropy model: MaxEnt) to solve the problem. To draw direct comparison, we conduct experiments on the same personal name corpora as that in the previous work by Li *et al.* (2006). We show that the MaxEnt method effectively incorporates diverse features and outperforms previous methods consistently across all test cases.

The rest of the paper is organized as follows: in section 2, we review the previous work. Section 3 elaborates our proposed approach and the features. Section 4 presents our experimental setup and reports our experimental results. Finally, we conclude the work in section 5.

2 Related Work

Most of previous work focuses mainly on ENOR although same methods can be extended to CNOR. We notice that there are two informative clues that used in previous work in ENOR. One is the lexical structure of a romanization system, for example, Hanyu *Pinyin*, Mandarin Wade-Giles, Japanese Hepbrun or Korean Yale, each has a finite set of syllable inventory (Li et al., 2006). Another is the phonetic and phonotactic structure of a language, such as phonetic composition, syllable structure. For example, English has unique consonant clusters such as /str/ and /ks/ which Chinese, Japanese and Korean (CJK) do not have. Considering the NOR solutions by the use of these two clues, we can roughly group them into two categories: rule-based methods (for solutions based on lexical structures) and statistical methods (for solutions based on phonotactic structures).

Rule-based Method

Kuo and Yang (2004) proposed using a rule-based method to recognize different romanization system for Chinese only. The left-to-right longest match-based lexical segmentation was used to parse a test word. The romanization system is confirmed if it gives rise to a successful parse of the test word. This kind of approach (Qu and Grefenstette, 2004) is suitable for romanization systems that have a finite set of *discriminative* syllable inventory, such as *Pinyin* for Chinese Mandarin. For the general tasks of identifying the language origin and romanization system, rule based approach sounds less attractive because not all languages have a finite set of *discriminative* syllable inventory.

Statistical Method

1) N-gram Sum Method (SUM): Qu and Grefenstette (2004) proposed a NOR identifier using a trigram language model (Cavnar and Trenkle, 1994) to distinguish personal names of three language origins, namely Chinese, Japanese and English. In their work, the training set includes 11,416 Chinese name entries, 83,295 Japanese name entries and 88,000 English name entries. However, the trigram is defined as the joint probabil-

ity $p(c_i c_{i-1} c_{i-2})$ for 3-character $c_i c_{i-1} c_{i-2}$ rather than the commonly used conditional probability $p(c_i \mid c_{i-1} c_{i-2})$. Therefore, the so-called trigram in Qu and Grefenstette (2004) is basically a substring unigram probability, which we refer to as the n-gram (n-character) sum model (SUM) in this paper. Suppose that we have the unigram count $C(c_i c_{i-1} c_{i-2})$ for character substring $c_i c_{i-1} c_{i-2}$, the unigram is then computed as:

$$p(c_i c_{i-1} c_{i-2}) = \frac{C(c_i c_{i-1} c_{i-2})}{\sum_{i, c_i c_{i-1} c_{i-2}} C(c_i c_{i-1} c_{i-2})} \quad (1)$$

which is the count of character substring $c_i c_{i-1} c_{i-2}$ normalized by the sum of all 3-character string counts in the name list for the language of interest. For origin recognition of Japanese names, this method works well with an accuracy of 92%. However, for English and Chinese, the results are far behind with a reported accuracy of 87% and 70% respectively.

2) N-gram Perplexity Method (PP): Li *et al.* (2006) proposed using n-gram character perplexity PP_c to identify the origin of a Latin-scripted name. Using bigram, the PP_c is defined as:

$$PP_c = 2^{-\frac{1}{N_c} \sum_{i=1}^{N_c} \log p(c_i \mid c_{i-1})} \quad (2)$$

where N_c is the total number of characters in the test name, c_i is the i^{th} character in the test name. $p(c_i \mid c_{i-1})$ is the bigram probability which is learned from each name list respectively. As a function of model, PP_c measures how good the model matches the test data. Therefore, PP_c can be used to measure how good a test name matches a training set. A test name is identified to belong to a language if the language model gives rise to the minimum perplexity. Li *et al.* (2006) shown that the PP method gives much better performance than the SUM method. This may be due to the fact that the PP measures the normalized conditional probability rather than the sum of joint probability. Thus, the PP method has a clearer mathematical interpretation than the SUM method.

The statistical methods attempt to overcome the shortcoming of rule-based method, but they suffer from data sparseness, especially when dealing with a large character set, such as in Chinese (our experiments will demonstrate this point empirically). In this paper, we propose using Maximum Entropy (MaxEnt) model as a general framework for both ENOR and CNOR. We explore and integrate multiple features into the discriminative classifier and use a common dataset for benchmarking. Experimental results show that the MaxEnt model effectively incorporates diverse features to demonstrate competitive performance.

3 MaxEnt Model and Features

3.1 MaxEnt Model for NOR

The principle of maximum entropy (MaxEnt) model is that given a collection of facts, choose a model consistent with all the facts, but otherwise as uniform as possible (Berger *et al.*, 1996). MaxEnt model is known to easily combine diverse features. For this reason, it has been widely adopted in many natural language processing tasks. The MaxEnt model is defined as:

$$p(c_i \mid x) = \frac{1}{Z} \prod_{j=1}^{K} \alpha_j^{f_j(c_i, x)} \quad (3)$$

$$Z = \sum_{i=1}^{N} p(c_i \mid x) = \sum_{i=1}^{N} \prod_{j=1}^{K} \alpha_j^{f_j(c_i, x)} \quad (4)$$

where c_i is the outcome label, x is the given observation, also referred to as an instance. Z is a normalization factor. N is the number of outcome labels, the number of language origins in our case. $f_1, f_2, \cdots, f_K$ are feature functions and $\alpha_1, \alpha_2, \cdots, \alpha_K$ are the model parameters. Each parameter corresponds to exactly one feature and can be viewed as a "weight" for the corresponding feature.

In the NOR task, c is the name origin label; x is a personal name, f_i is a feature function. All features used in the MaxEnt model in this paper are binary. For example:

$$f_j(c, x) = \begin{cases} 1, & \text{if } c = \text{"Eng"} \,\&\, x \text{ contains("str")} \\ 0, & \text{otherwise} \end{cases}$$

In our implementation, we used Zhang's maximum entropy package[2].

3.2 Features

Let us use English name "Smith" to illustrate the features that we define. All characters in a name

[2] http://homepages.inf.ed.ac.uk/s0450736/maxent.html

are first converted into upper case for ENOR before feature extraction.

N-gram Features: N-gram features are designed to capture both phonetic and orthographic structure information for ENOR and orthographic information only for CNOR. This is motivated by the facts that: 1) names written in English but from non-English origins follow different phonetic rules from the English one; they also manifest different character usage in orthographic form; 2) names written in Chinese follows the same pronunciation rules (*Pinyin*), but the usage of Chinese characters is distinguishable between different language origins as reported in Table 2 of (Li *et al.*, 2007). The N-gram related features include:

1) FUni: character unigram <*S, M, I, T, H*>
2) FBi: character bigram <*SM, MI, IT, TH*>
3) FTri: character trigram <*SMI, MIT, ITH* >

Position Specific n-gram Features: We include position information into the n-gram features. This is mainly to differentiate surname from given name in recognizing the origin of CJK personal names written in Chinese. For example, the position specific n-gram features of a Chinese name "温家宝(Wen-Jia-Bao)" are as follows:

1) FPUni: position specific unigram <*0 温*(Wen), *1 家*(Jia), *2 宝*(Bao)>
2) FPBi: position specific bigram <*0 温家*(Wen-Jia), *1 家宝*(Jia-Bao)>
3) FPTri: position specific trigram <*0 温家宝*(Wen-Jia-Bao)>

Phonetic Rule-based Features: These features are inspired by the rule-based methods (Kuo and Yang, 2004; Qu and Grefenstette, 2004) that check whether an English name is a sequence of syllables of CJK languages in ENOR task. We use the following two features in ENOR task as well.

1) FMan: a Boolean feature to indicate whether a name is a sequence of Chinese Mandarin *Pinyin*.
2) FCan: a Boolean feature to indicate whether a name is a sequence of Cantonese *Jyutping*.

Other Features:

1) FLen: the number of Chinese characters in a given name. This feature is for CNOR only. The numbers of Chinese characters in personal names vary with their origins. For example, Chinese and Korean names usually consist of 2 to 3 Chinese characters while Japanese names can have up to 4 or 5 Chinese characters
2) FFre: the frequency of n-gram in a given name. This feature is for ENOR only. In CJK names, some consonants or vowels usually repeat in a name as the result of the regular syllable structure. For example, in the Chinese name "Zhang Wanxiang", the bigram "an" appears three times

Please note that the trigram and position specific trigram features are not used in CNOR due to anticipated data sparseness in CNOR[3].

4 Experiments

We conduct the experiments to validate the effectiveness of the proposed method for both ENOR and CNOR tasks.

4.1 Experimental Setting

Origin	# entries	Romanization System
Eng[4]	88,799	English
Man[5]	115,879	*Pinyin*
Can	115,739	*Jyutping*
Jap[6]	123,239	Hepburn

Table 1: D_E: Latin-scripted personal name corpus for ENOR

Origin	# entries
Eng[7]	37,644
Chi[8]	29,795
Jap[9]	33,897

Table 2: D_C: Personal name corpus written in Chinese characters for CNOR

[3] In the test set of CNOR, 1080 out of 2980 names of Chinese origin do not consist of any bigrams learnt from training data, while 2888 out of 2980 names do not consist of any learnt trigrams. This is not surprising as most of Chinese names only have two or three Chinese characters and in our open testing, the train set is exclusive of all entries in the test set.

[4] http://www.census.gov/genealogy/names/

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

[6] http://www.csse.monash.edu.au/~jwb/enamdict_doc.html

[7] Xinhua News Agency (1992)

[8] http://www.ldc.upenn.edu LDC2005T34

[9] www.cjk.org

Datasets: We prepare two data sets which are collected from publicly accessible sources: D_E and D_C for the ENOR and CNOR experiment respectively. D_E is the one used in (Li *et al.*, 2006), consisting of personal names of Japanese (Jap), Chinese (Man), Cantonese (Can) and English (Eng) origins. D_C consists of personal names of Japanese (Jap), Chinese (Chi, including both Mandarin and Cantonese) and English (Eng) origins. Table 1 and Table 2 list their details. In the experiments, 90% of entries in Table 1 (D_E) and Table 2 (D_C) are randomly selected for training and the remaining 10% are kept for testing for each language origin. Columns 2 and 3 in Tables 7 and 8 list the numbers of entries in the training and test sets.

Evaluation Methods: Accuracy is usually used to evaluate the recognition performance (Qu and Gregory, 2004; Li *et al.*, 2006; Li *et al.*, 2007). However, as we know, the individual accuracy used before only reflects the performance of *recall* and does not give a whole picture about a multi-class classification task. Instead, we use *precision* (*P*), *recall* (*R*) and F-measure (*F*) to evaluate the performance of each origin. In addition, an overall accuracy (*Acc*) is also given to describe the whole performance. The *P*, *R*, *F* and *Acc* are calculated as following:

$$P = \frac{\#\ correctly\ recognized\ entries\ of\ the\ given\ origin}{\#\ entries\ recognized\ as\ the\ given\ origin\ by\ the\ system}$$

$$R = \frac{\#\ correctly\ recognized\ entries\ of\ the\ given\ origin}{\#\ entries\ of\ the\ given\ origin}$$

$$F = \frac{2PR}{P+R} \qquad Acc = \frac{\#\ all\ correctly\ recognized\ entries}{\#\ all\ entries}$$

4.2 Experimental Results and Analysis

Table 3 reports the experimental results of ENOR. It shows that the MaxEnt approach achieves the best result of 98.44% in overall accuracy when combining all the diverse features as listed in Subsection 3.2. Table 3 also measures the contributions of different features for ENOR by gradually incorporating the feature set. It shows that:

1) All individual features are useful since the performance increases consistently when more features are being introduced.
2) Bigram feature presents the most informative feature that gives rise to the highest performance gain, while the trigram feature further boosts performance too.
3) MaxEnt method can integrate the advantages of previous rule-based and statistical methods and easily integrate other features.

Features	Origin	*P(%)*	*R(%)*	*F*	*Acc(%)*
FUni	Eng	91.40	80.76	85.75	85.29
	Man	83.05	81.90	82.47	
	Can	81.13	82.76	81.94	
	Jap	87.31	94.11	90.58	
+FBi	Eng	97.54	91.10	94.21	96.72
	Man	97.51	98.10	97.81	
	Can	97.68	98.05	97.86	
	Jap	94.62	98.24	96.39	
+FTri	Eng	97.71	93.79	95.71	97.97
	Man	98.94	99.37	99.16	
	Can	99.12	99.19	99.15	
	Jap	96.19	98.52	97.34	
+FPUni	Eng	97.53	94.64	96.06	98.16
	Man	99.21	99.43	99.32	
	Can	99.41	99.24	99.33	
	Jap	96.48	98.49	97.47	
+FPBi	Eng	97.68	94.98	96.31	98.28
	Man	99.32	99.50	99.41	
	Can	99.53	99.34	99.44	
	Jap	96.59	98.52	97.55	
+FPTri	Eng	97.62	94.97	96.27	98.30
	Man	99.34	99.58	99.46	
	Can	99.63	99.37	99.50	
	Jap	96.61	98.45	97.52	
+FFre	Eng	97.74	95.06	96.38	98.35
	Man	99.37	99.59	99.48	
	Can	99.61	99.41	99.51	
	Jap	96.66	98.56	97.60	
+ FMan + FCan	Eng	97.82	95.11	96.45	98.44
	Man	99.52	99.68	99.60	
	Can	99.71	99.59	99.65	
	Jap	96.69	98.59	97.63	

Table 3: Contribution of each feature for ENOR

Features	Eng	Jap	Man	Can
FMan	-0.357	0.069	0.072	-0.709
FCan	-0.424	-0.062	-0.775	0.066

Table 4: Features weights in ENOR task.

Feature	Origin	*P*(%)	*R*(%)	*F*	*Acc*(%)
FUni	Eng	97.89	98.43	98.16	96.97
	Chi	95.80	95.03	95.42	
	Jap	96.96	97.05	97.00	
+FBi	Eng	96.99	98.27	97.63	96.28
	Chi	96.86	92.11	94.43	
	Jap	95.04	97.73	96.36	
+FLen	Eng	97.35	98.38	97.86	97.14
	Chi	97.29	95.00	96.13	
	Jap	96.78	97.64	97.21	
+FPUni	Eng	97.74	98.65	98.19	97.77
	Chi	97.65	96.34	96.99	
	Jap	97.91	98.05	97.98	
+FPBi	Eng	97.50	98.43	97.96	97.56
	Chi	97.61	96.04	96.82	
	Jap	97.59	97.94	97.76	
FUni +FLen + FPUni	Eng	98.08	99.04	98.56	98.10
	Chi	97.57	96.88	97.22	
	Jap	98.58	98.11	98.34	

Table 5: Contribution of each feature for CNOR

Table 4 reports the feature weights of two features "FMan" and "FCan" with regard to different origins in ENOR task. It shows that "FCan" has positive weight only for origin "Can" while "FMan" has positive weights for both origins "Man" and "Jap", although the weight for "Man" is higher. This agrees with our observation that the two features favor origins "Man" or "Can". The feature weights also reflect the fact that some Japanese names can be successfully parsed by the Chinese Mandarin *Pinyin* system due to their similar syllable structure. For example, the Japanese name "Tanaka Miho" is also a sequence of Chinese *Pinyin*: "Ta-na-ka Mi-ho".

Table 5 reports the contributions of different features in CNOR task by gradually incorporating the feature set. It shows that:

1) Unigram features are the most informative
2) Bigram features degrade performance. This is largely due to the data sparseness problem as discussed in Section 3.2.
3) FLen is also useful that confirms our intuition about name length.

Finally the combination of the above three useful features achieves the best performance of 98.10% in overall accuracy for CNOR as in the last row of Table 5.

In Tables 3 and 5, the effectiveness of each feature may be affected by the order in which the features are incorporated, i.e., the features that are added at a later stage may be underestimated. Thus, we conduct another experiment using "all-but-one" strategy to further examine the effectiveness of each kind of features. Each time, one type of the n-gram (n=1, 2, 3) features (including orthographic n-gram, position-specific and n-gram frequency features) is removed from the whole feature set. The results are shown in Table 6.

Features	Origin	*P*(%)	*R*(%)	*F*	*Acc*(%)
w/o Unigram	Eng	97.81	95.01	96.39	98.34
	Man	99.41	99.58	99.49	
	Can	99.53	99.48	99.50	
	Jap	96.63	98.52	97.57	
w/o Bigram	Eng	97.34	95.17	96.24	98.26
	Man	99.30	99.48	99.39	
	Can	99.54	99.33	99.43	
	Jap	96.73	98.32	97.52	
w/o Trigram	Eng	97.57	94.10	95.80	97.94
	Man	98.98	99.23	99.10	
	Can	99.20	99.08	99.14	
	Jap	96.06	98.42	97.23	

Table 6: Effect of n-gram feature for ENOR

Table 6 reveals that removing trigram features affects the performance most. This suggests that trigram features are much more effective for ENOR than other two types of features. It also shows that trigram features in ENOR does not suffer from the data sparseness issue.

As observed in Table 5, in CNOR task, 93.96%

accuracy is obtained when removing unigram features, which is much lower than 98.10% when bigram features are removed. This suggests that unigram features are very useful in CNOR, which is mainly due to the data sparseness problem that bigram features may have encountered.

4.3 Model Complexity and Data Sparseness

Table 7 (ENOR) and Table 8 (CNOR) compare our MaxEnt model with the SUM model (Qu and Gregory, 2004) and the PP model (Li *et al.*, 2006). All the experiments are conducted on the same data sets as described in section 4.1. Tables 7 and 8 show that the proposed MaxEnt model outperforms other models. The results are statistically significant (χ^2 test with $p<0.01$) and consistent across all tests.

Model Complexity:

We look into the complexity of the models and their effects. Tables 7 and 8 summarize the overall accuracy of three models. Table 9 reports the numbers of parameters in each of the models. We are especially interested in a comparison between the MaxEnt and PP models because their performance is close. We observe that, using trigram features, the MaxEnt model has many more parameters than the PP model does. Therefore, it is not surprising if the MaxEnt model outperforms when more training data are available. However, the experiment results also show that the MaxEnt model consistently outperforms the PP model even with the same size of training data. This is largely attributed to the fact that MaxEnt incorporates more robust features than the PP model does, such as rule-based, length of names features.

One also notices that PP clearly outperforms SUM by using the same number of parameters in ENOR and shows comparable performance in CNOR tasks. Note that SUM and PP are different in two areas: one is the PP model employs word length normalization while SUM doesn't; another that the PP model uses n-gram conditional probability while SUM uses n-character joint probability. We believe that the improved performance of PP model can be attributed to the effect of usage of conditional probability, rather than length normalization since length normalization does not change the order of probabilities.

Data Sparesness:

We understand that we can only assess the effectiveness of a feature when sufficient statistics is available. In CNOR (see Table 8), we note that the Chinese transliterations of English origin use only 377 Chinese characters, so data sparseness is not a big issue. Therefore, bigram SUM and bigram PP methods easily achieve good performance for English origin. However, for Japanese origin (represented by 1413 Chinese characters) and Chinese origin (represented by 2319 Chinese characters), the data sparseness becomes acute and causes performance degradation in SUM and PP models. We are glad to find that MaxEnt still maintains a good performance benefiting from other robust features.

Table 10 compares the overall accuracy of the three methods using unigram and bigram features in CNOR task, respectively. It shows that the MaxEnt method achieves best performance. Another interesting finding is that unigram features perform better than bigram features for PP and MaxEnt models, which shows that data sparseness remains an issue even for MaxEnt model.

5 Conclusion

We propose using MaxEnt model to explore diverse features for name origin recognition. Experiment results show that our method is more effective than previously reported methods. Our contributions include:

1) Cast the name origin recognition problem as a multi-class classification task and propose a MaxEnt solution to it;
2) Explore and integrate diverse features for name origin recognition and propose the most effective feature sets for ENOR and for CNOR

In the future, we hope to integrate our name origin recognition method with a machine transliteration engine to further improve transliteration performance. We also hope to study the issue of name origin recognition in context of sentence and use contextual words as additional features.

References

Adam L. Berger, Stephen A. Della Pietra and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics. 22(1):39–71.

William B. Cavnar and John M. Trenkle. 1994. Ngram based text categorization. In 3rd Annual Symposium

on Document Analysis and Information Retrieval, 275–282.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. Computational Linguistics. 24(4), 599-612.

Jin-Shea Kuo and Ying-Kuei Yan. 2004. Generating Paired Transliterated-Cognates Using Multiple Pronunciation Characteristics from Web Corpora. PACLIC 18, December 8th-10th, Waseda University, Tokyo, Japan, 275–282.

Haizhou Li, Shuanhu Bai and Jin-Shea Kuo. 2006. Transliteration. Advances in Chinese Spoken Language Processing. World Scientific Publishing Company, USA, 341–364.

Haizhou Li, Khe Chai Sim, Jin-Shea Kuo and Minghui Dong. 2007. Semantic Transliteration of Personal Names. ACL-2007. 120–127.

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

Yan Qu and Gregory Grefenstette. 2004. Finding ideographic representations of Japanese names written in Latin script via language identification and corpus validation. ACL-2004. 183–190.

Min Zhang, Jian Su and Haizhou Li. 2004. Direct Orthographical Mapping for Machine Translation. COLING-2004. 716-722.

Origin	**# training entries**	**# test entries**	**Trigram SUM**			**Trigram PP**			**MaxEnt**		
			***P*(%)**	***R*(%)**	***F***	***P*(%)**	***R*(%)**	***F***	***P*(%)**	***R*(%)**	***F***
Eng	79,920	8,879	94.66	72.50	82.11	95.84	94.72	95.28	97.82	95.11	**96.45**
Man	104,291	11,588	86.79	94.87	90.65	98.99	98.33	98.66	99.52	99.68	**99.60**
Can	104,165	11,574	90.03	93.87	91.91	96.17	99.67	97.89	99.71	99.59	**99.65**
Jap	110,951	12,324	89.17	92.84	90.96	98.20	96.29	97.24	96.69	98.59	**97.63**
Overall *Acc* (%)			89.57			97.39			**98.44**		

Table 7: Benchmarking different methods in ENOR task

Origin	**# training entries**	**# test entries**	**Bigram SUM**			**Bigram PP**			**MaxEnt**		
			***P*(%)**	***R*(%)**	***F***	***P*(%)**	***R*(%)**	***F***	***P*(%)**	***R*(%)**	***F***
Eng	37,644	3,765	95.94	98.65	97.28	97.58	97.61	97.60	98.08	99.04	**98.56**
Chi	29,795	2,980	96.26	87.35	91.59	95.10	87.35	91.06	97.57	96.88	**97.22**
Jap	33,897	3,390	93.01	97.67	95.28	90.94	97.43	94.07	98.58	98.11	**98.34**
Overall *Acc* (%)			95.00			94.53			**98.10**		

Table 8: Benchmarking different methods in CNOR task

Methods	**# of parameters for ENOR**	**# of parameters for CNOR**	
	Trigram	**Unigram**	**Bigram**
MaxEnt	124,692	13,496	182,116
PP	16,851	4,045	86,490
SUM	16,851	4,045	86,490

Table 9: Numbers of parameters used in different methods

	SUM	**PP**	**MaxEnt**
Unigram Features	90.55	97.09	**98.10**
Bigram Features	95.00	94.53	**97.56**

Table 10: Overall accuracy using unigram and bigram features in CNOR task