# Improving Name Origin Recognition with Context Features and Unlabelled Data

**Vladimir Pervouchine, Min Zhang, Ming Liu and Haizhou Li**
Institute for Infocomm Research, A-STAR
vpervouchine@gmail.com,{mzhang,mliu,hli}@i2r.a-star.edu.sg

## Abstract

We demonstrate the use of context features, namely, names of places, and unlabelled data for the detection of personal name language of origin.

While some early work used either rule-based methods or n-gram statistical models to determine the name language of origin, we use the discriminative classification maximum entropy model and view the task as a classification task. We perform bootstrapping of the learning using list of names out of context but with known origin and then using expectation-maximisation algorithm to further train the model on a large corpus of names of unknown origin but with context features. Using a relatively small unlabelled corpus we improve the accuracy of name origin recognition for names written in Chinese from 82.7% to 85.8%, a significant reduction in the error rate. The improvement in $F$-score for infrequent Japanese names is even greater: from 77.4% without context features to 82.8% with context features.

## 1 Introduction

Transliteration is a process of rewriting a word from a source language to a target language in a different writing system using the word's phonological equivalent. Many technical terms and proper nouns, such as personal names, names of places and organisations are transliterated during translation of a text from one language to another. A process reverse to the transliteration, which is recovering a word in its native language from its transliteration in a foreign language, is called back-transliteration (Knight and Graehl, 1998). In many natural language processing (NLP) tasks such as machine translation and cross-lingual information retrieval, transliteration is an important component.

Name origin refers to the language of origin of a name. For example, the origin of 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.

For machine transliteration the name origins dictate the way we re-write a foreign name. For example, given a name written in Chinese for which we do not have a translation in an English-Chinese dictionary, we first have to decide whether the name is of Chinese, Japanese, Korean, English or another origin. Then we follow the transliteration rules implied by the origin of the name. Although all English personal names are rendered in 26 letters, they may come from different romanization systems. Each romanisation sys-



*Coling 2010: Poster Volume*, pages 972–978,
Beijing, August 2010

tem 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. Li et al. (2007b) demonstrated that incorporating name origin recognition (NOR) into a transliteration system greatly improves 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, not much attention has been given to name origin recognition (NOR) so far in the literature. In this paper, we are interested in recognition of the origins of names written in Chinese, which names can be of three origins: Chinese, Japanese or English, where "English" is a rather broad category that includes other West European and American names written natively in Latin script.

Unlike previous work (Qu and Grefenstette, 2004; Li et al., 2007a; Li et al., 2007b), where NOR was formulated with a generative model, we follow the approach of Zhang et al. (2008) and regard the NOR task as a classification problem, using a discriminative learning algorithm for classification. Furthermore, in the training data with names labelled with their origin is rather limited, whereas there is vast data from news articles that contains many personal names without any labels of their origins. In this research we propose a method to harness the power of the unlabelled noisy news data by bootstrapping the learning process with labelled data and then using the *personal name context* in the unlabelled data to improve the NOR model. We achieve that by using the maximum entropy model and the expectation-maximisation training, and demonstrate that our method can significantly improve the accuracy of NOR compared to the baseline model trained only from the labelled data.

The rest of the paper is organised as follows: in Section 2 we review the previous research. In Section 3 we present our approach, and in Section 4 we describe our experimental setup, the data used and the evaluation method. We conclude in Section 5.

## 2 Related research

Most the research up to date focuses primarily on recognition of origin of names written in Latin script, called English NOR (ENOR), although the same methods can be extended to names in Chinese script (CNOR). We notice that there are two informative clues that used in previous work in ENOR. One is the lexical structure of a romanisation system, for example, Hanyu Pinyin, Mandarin Wade-Giles, Japanese Hepbrun or Korean Yale, each has a finite set of syllable inventory (Li et al., 2007a). 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 et al. (2007) proposed using a rule-based method to recognise different romanisation system for Chinese only. The left-to-right longest match-based lexical segmentation was used to parse a test word. The romanisation 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 romanisation 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 romanisation system, rule based approach sounds less attractive because not all languages have a finite set of discriminative syllable inventory.

**N-gram statistics methods**

**N-gram sum method** 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, 83,295 Japanese and 88,000 English name entries. However, the trigram is defined as the joint probability $p(c_i c_{i-1} c_{i-2})$ rather than the commonly used conditional probability $p(c_i|c_{i-1}c_{i-2})$. Therefore it is basically a substring unigram probability. 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.

**N-gram perplexity method** Li et al. (2007a) proposed a method of NOR using n-gram character perplexity $PP_c$ to identify the origin of names written in Latin script. Using bigrams, the perplexity is defined as

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

where $N_c$ is the total number of characters in a given name, $c_i$ is the $i$-th character in the name and $p(c_i|c_{i-1})$ is the bigram probability learned from a list of names of the same origin. Therefore, $PP_c$ can be used to measure how well a new name fits the model learned from the training set of names. The origin is assigned according to the model that gives the lowest perplexity value. Li et al. (2007a) demonstrated that using $PP_c$ gives much better performance than with the substring unigram method.

**Classification method** Zhang et al. (2008) proposed using a discriminative classification approach and extract features from the names. They use Maximum Entropy (MaxEnt) model and a number of features based on n-grams, character positions, word length as well as some rule-based phonetic features. They performed both ENOR and CNOR and demonstrated that their method indeed leads to better performance in name origin recognition then the n-gram statistics method. They attribute that to the fact their model incorporates more robust features than the n-gram statistics based models.

In this paper we too follow the discriminating classification approach, but we add features based on the context of a personal name. These features require the original text with the names to be available. Our approach closely models the real-life situation when large corpora of articles with personal names is readily available in the Web, yet the origins of the names are unknown.

## 3 Model and training methods

### 3.1 Maximum entropy model for NOR

The principle of maximum entropy is that given a collection of facts we should choose a model that is consistent with all the facts but otherwise as uniform as possible (Berger et al., 1996). maximum entropy model (MaxEnt) is known to easily combine diverse features and

has been used widely in natural language processing research. Given an observation $x$ the probability of outcome label $c_i$, $i = 1 \ldots N$ given $x$ is given by

$$p(c_i|x) = \frac{1}{Z} \exp\left(\sum_{j=1}^{n} \lambda_j f_j\,(x, c_i)\right) \quad (1)$$

where $N$ is the number of the outcome labels, which is the number of name origins in our case, $n$ is the number of features, $f_j$ are the feature functions and $\lambda_j$ are the model parameters. Each parameter corresponds to exactly one feature and can be viewed as a "weight" for the corresponding feature. $Z$ is the normalisation factor given by

$$Z = \sum_{i=1}^{N} p(c_i|x) \quad (2)$$

In the problem at hand $x$ is a personal name and all the features are binary. The features, also known as contextual predicates, are in the form

$$f_i(x, c) = \begin{cases} 1 & \text{if } c = c_i \text{ and } cp(x) = true \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $cp$ is the contextual predicate that maps a pair $(c_i, x)$ to $\{true, false\}$.

In our experiments we use Zhang's maximum entropy library[1].

### 3.2 Initial training with labelled data and n-gram features

For the initial training of MaxEnt model we use labelled data: personal names of Chinese, Japanese or English origin written in Chinese. The origin of each name is known. Following paper by Zhang et al. (2008) and their findings regarding the contribution value of each feature that they studied, we extract unigram, positional unigram and word length features. For example, Chinese name "温家宝" has the following features:

温 家 宝 (温,0) (家,1) (宝,2) 3

We restrict the n-gram features to unigram only to avoid the data sparseness, because our data contains a number of Chinese surnames and given names, which have a length of one or two characters.

### 3.3 Further training with unlabelled data and context features

For further training of MaxEnt model we use unlabelled data collected from news articles. The name origin is not known but each personal name is in a context and is often surrounded by names of places that may give a hint about the personal name origin. For each personal name we extract all names of places in the same paragraph and use them as features. If a place name is repeated many times in the same paragraph we only include it once in the feature list.

For example, paragraph containing passage *"The U.S. President Barack Obama ..."* will result in two personal names "Barack" and "Obama" having "U.S." as their context feature. Due to the diversity of place names we also attempt to map the names of the places into the country names. In this case, features like "U.S.", "USA", "America" are manually substituted with "USA". In our experiments we also try to narrow the place name extraction to windows of different sizes surrounding the personal name. The rationale here is that the closer a place name is to the personal name, the more likely it has a connection to the origin of the personal name.

In summary, our algorithm includes two steps.

[1] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

First, we use the boostrap data and n-gram, positional n-gram and name length features to do the initial training (the 0-th iteration) of MaxEnt model with L-BFGS method (Byrd et al., 1995). After that we use the model to assign origin labels to names of the training set of the unlabelled data.

Next, we use both the bootstrap data and the training set of the unlabelled data, labelled in the previous step, and add the context features to the already used n-gram, positional n-gram and name length features. Since there is no context available for the bootstrap data, the context features for it are missing, which can be handled by the MaxEnt model. We perform the Expectation-Maximisation (EM) iterations by using the mixed data to train the $i$-th iteration of the MaxEnt model, then use the model to re-label the training set of the unlabelled data and repeat the training of the model for the $(i + 1)$-st iteration. We stop the iterations when the ratio of patterns that change the origin labels becomes less than 0.01%.

## 4 Experiments

### 4.1 Corpora

The corpora consists of two datasets. One dataset, called the "bootstrap data", is a set of Chinese, Japanese and English names written in Chinese following the respective transliteration rules according to the name origins. The names are a mixture of full names, first (given) names and surnames. Table 1 shows the number of names of each origin. This is the labelled data; the origin of each name is known. The data is used to start the MaxEnt model training.

The second dataset, called the "unlabelled data", is Chinese, Japanese and English personal names written in Chinese, which have been extracted from the news articles collected over 6 months from Xinhua news website. The articles have been processed by an automatic part-of-speech (POS) tagger, after which personal names and names of places have been manually identified (the latter for extracting the context features). Normally the first (given) name and surnames are identified as two separate personal names. The data is split into a training set of 27,882 names with unknown origin and a testing set of 1,476 names whose origin was manually assigned. We split data in such a way that there is no overlap between patterns in the training and testing sets, although there may be overlap between names. For example, if a name may be present in both training and testing sets but in a different context, making the two names two distinct patterns. The number of names of each origin in the testing set is shown in Table 2. As seen from the table, the number of Chinese names exceeds the number of English or Japanese names. This is an expected consequence of using articles from a Chinese news agency because many of the articles are reporting on local affairs. We have manually removed a number of Chinese name patterns from the testing set, since the original percentage of Chinese names in the articles is about 83%.

| Origin | Number of names |
|---|---|
| Chinese | 52,342 |
| Japanese | 26,171 |
| English | 26,171 |

Table 1: Number of names of each origin in the bootstrap dataset.

| Origin | Number of names |
|---|---|
| Chinese | 738 |
| Japanese | 369 |
| English | 422 |

Table 2: Number of names of each origin in the testing dataset.

### 4.2 Evaluation method

Following Zhang et al. (2008) to make our results comparable to theirs, we evaluate our system using precision $P_o$, recall $R_o$ and $F$-score $F_o$ for each origin $o \in \{\text{“}Chinese\text{”} \text{“}Japanese\text{”} \text{“}English\text{”}\}$. Let the number of correctly recognised names of a given origin $o$ be $k_o$, and the total number of names recognised as being of origin $o$ be $m_o$, while the actual number of names of origin $o$ be $n_o$. Then the precision, recall and $F$-score are given as:

$$P_o = \frac{k_o}{m_o}$$

$$R_o = \frac{k_o}{n_o}$$

$$F_o = \frac{2 \times P_o \times R_o}{P_o + R_o}$$

We also report the overall accuracy of the system (or, rather the overall recall), which is the ratio of the total number of correctly recognised names to the number of all names:

$$Acc = \frac{k_{Chinese} + k_{Japanese} + k_{English}}{n_{Chinese} + n_{Japanese} + n_{English}}$$

### 4.3 Results

After each iteration of our MaxEnt-based EM algorithm, we record the number of patterns in the training set that changed their origin labels, as well as calculate the precision, recall and $F$-score for each origin as well as the overall accuracy. The results are reported in Tables 3 and 4, where for the sake of brevity the origin subscripts are “C”, “J” and “W” for Chinese, Japanese and English name origin respectively.

Compared to the 0-th iteration there is an significant improvement in accuracy, particularly in recognition of Japanese names, which are relatively infrequent compared to Chinese and English ones in the unlabelled training data. This clearly shows the effectiveness of our proposed method.

| Iteration | $P_C$ | $P_J$ | $P_W$ | $R_C$ | $R_J$ | $R_W$ |
|---|---|---|---|---|---|---|
| 0 | 0.887 | 0.724 | 0.857 | 0.823 | 0.911 | 0.761 |
| 1 | 0.914 | 0.736 | 0.875 | 0.823 | 0.968 | 0.775 |
| 2 | 0.910 | 0.736 | 0.874 | 0.823 | 0.968 | 0.767 |
| 3 | 0.914 | 0.737 | 0.874 | 0.824 | 0.973 | 0.767 |
| 4 | 0.913 | 0.742 | 0.875 | 0.825 | 0.968 | 0.778 |

Table 3: Results of running EM iterations, original names of the places are kept.

| Iteration | $Acc$ | $F_C$ | $F_J$ | $F_W$ |
|---|---|---|---|---|
| 0 | 0.829 | 0.854 | 0.807 | 0.806 |
| 1 | 0.847 | 0.866 | 0.836 | 0.822 |
| 2 | 0.845 | 0.864 | 0.836 | 0.817 |
| 3 | 0.847 | 0.867 | 0.839 | 0.817 |
| 4 | 0.849 | 0.867 | 0.840 | 0.824 |

Table 4: Results of running EM iterations, original names of the places are kept.

## 5 Conclusions

We propose extension of MaxEnt model for NOR task by using two types of data for training: origin-labelled names alone and origin-unlabelled names in their context surrounding. We show how to apply a simple EM method to make use of the contextual words as features, and improve the NOR accuracy from 82.9% to 84.9% overall, while for rare names such as Japanese the effect of using unlabelled data with context features is even greater.

The purpose of this research is to demonstrate how the unlabelled data can be used. In the future we hope to investigate the use of other context features, as well as to study the effect of data size on the NOR accuracy improvement.

The feature of names' places normally exhibit great variation: one country name may be spelled in many different ways, and often there are names of cities etc that surround personal names. We will explore to normalise names of places by substituting each name with name of the country where the place is in the future work.

## References


[Berger et al.1996] Berger, A., Stephen A. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

[Byrd et al.1995] Byrd, R. H., P. Lu, and J. Nocedal. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific and Statistical Computing*, 16(5):1190–1208.

[Cavnar and Trenkle1994] Cavnar, William B. and John M. Trenkle. 1994. Ngram based text categorization. In *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 275–282.

[Knight and Graehl1998] Knight, Kevin and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4).

[Kuo et al.2007] Kuo, Jin-Shea, Haizhou Li, and Ying-Kuei Yang. 2007. A phonetic similarity model for automatic extraction of transliteration pairs. *ACM Transactions on Asian Language Information Processing*, 6(2).

[Li et al.2007a] Li, Haizhou, Shuanhu Bai, and Jin-Shea Kuo. 2007a. Transliteration. In *Advances in Chinese Spoken Language Processing*, chapter 15, pages 341–364. World Scientific.

[Li et al.2007b] Li, Haizhou, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong. 2007b. Semantic transliteration of personal names. In *Proc. 45th Annual Meeting of the ACL*, pages 120–127.

[Qu and Grefenstette2004] Qu, Yan and Gregory Grefenstette. 2004. Finding ideographic representations of Japanese names written in Latin script via language identification and corpus validation. In *Proc. 42nd ACL Annual Meeting*, pages 183–190, Barcelona, Spain.

[Zhang et al.2008] Zhang, Min, Chengjie Sun, Haizhou Li, Aiti Aw, Chew Lim Tan, and Xiaolong Wang. 2008. Name origin recognition using maximum entropy model and diverse features. In *Proc. 3rd Int'l Conf. NLP*, pages 56–63.