Name Origin Recognition in Chinese Texts Based on Conditional Random Fields

Jing Zhang[1,a], Jian Xu[2,b] and Yujie Zhang[3,c]

[1,2,3]School of Computer and Information Technology, Beijing Jiaotong University

[a]cindy90.china@gmail.com, [b]jaxu@bjtu.edu.cn, [c]yjzhang@bjtu.edu.cn

Keywords: Name Origin Recognition, Conditional Random Fields, Natural Language Processing

Abstract. Name origin recognition is to identify the origin of a name. In natural language processing, information of name origin is an important feature for name entity translation and question answering. Language identification of the origins of names can help to know what language-specific transliteration approaches to use. While some early work used two main methods, which are based on rules and statistics. In this paper, we use the conditional random fields (CRFs) model and view the task as a labeling problem on a sequence of words, taking advantage of the ability of using arbitrary features as input in CRFs under the character-based framework. Experimental results show that CRFs model is effective in recognizing origins of personal names in Chinese texts.

Introduction

We can often correctly identify origins of personal names mentioned in books, newspapers or websites, even though we do not really know or speak the original language. Knowing where we are from is very meaningful and has an important historical and cultural significance. Given a name in Chinese for which we do not have a translation in a bilingual English-Chinese dictionary, we first have to decide whether the name if of Japanese, Chinese, or some European origin. Name origin recognition is to identify the source language of a name, which can be used to decide the language of the name origin.

Chinese texts are mixed with Chinese personal names and Chinese transliteration of foreign personal names from different nations. Therefore, name origin recognition has become an important sub-task in Chinese-English machine translation, and plays a key role in many cross-language applications such as information retrieval and question answering.

Unlike previous work, in this paper, we use the conditional random fields to determine the origin of names in the Chinese texts, and view the name origin recognition as a sequence labeling problem. Following Pervouchine et al. [1], we divide the names into three origins: Chinese, Japanese and English, where "English" is a rather broad category that includes names of Europe-American area written natively in Latin script.

In our experiment, we make classification annotations and part-of-speech tagging for each word in the training corpus, then extract its character, part of speech, whether the character is in the surname lists or a commonly-used character of names, and the context information as characteristic attributes to set feature template. Finally, we establish the training collection based on that.

Related Works

Rule-based Methods. Kuo et al. [2] proposed using a rule-based method to recognize 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.

N-gram Statistics Methods. Chen et al. [3] used the *n*-gram method on the basis of letter and letter cluster to recognize name origin of four languages: English, French, German and Portuguese. They divided the names written in English into a plurality of letter clusters by using the information of syllable, and then calculated the likelihood of a name origin from a language. The language with the

highest likelihood was considered to be the final decision. Llitjos et al. [4] used the letter as the basic processing unit, made use of *n*-gram model to calculate the co-occurrence probability between the adjacent letters in the word, thus classified the different origins of foreign translated names.

Classification Methods. Zhang et al. [5] cast the name origin recognition problem as a multi-class classification task, by using a discriminative classification approach and extracting features from the names. They used Maximum Entropy model and a number of features based on *n*-grams, character positions, word length as well as some rule-based phonetic features. Pervouchine et al. [1] followed the discriminating classification approach, but added features on the context of a personal name. They used a list of names out of context but with known origin for the bootstrapping of the learning, and then used expectation-maximization algorithm to further train the model on a large corpus of names of unknown origin but with context features.

Model and Method

Conditional Random Fields. Conditional random fields (CRFs) [6] are undirected graphical models, a special case of which correspond to conditional trained probabilistic finite state automata. Being conditionally trained, CRFs can easily incorporate a large number of arbitrary, non-independent features while still having efficient procedures for non-greedy finite-state inference and training. CRFs have shown success in various sequence modeling tasks.

Name origin recognition can be seen as a sequence segmentation problem: each word is a token in a sequence to be assigned a label. CRFs bring together the best of generative and classification models. Like classification models, they can accommodate many statistically correlated features of the inputs, and they are trained discriminatively. But like generative models, they can trade off decisions at different sequence positions to obtain a globally optimal labeling. Such models are well suited to sequence analysis.

Let ***Error! Reference source not found.*** $o = \langle o_1, o_2, \ldots, o_n \rangle$ be a sequence of observed words of length n. Let S be a set of states in a finite state machine, each corresponding to a label $l \in$ ***Error! Reference source not found.***. Let $s = \langle s_1, s_2, \ldots, s_n \rangle$ be the sequence of states in S **Error! Reference source not found.**that correspond to the labels assigned to words in the input sequence ***Error! Reference source not found.***. Linear-chain CRFs define the conditional probability of a state given an input sequence to be:

$$P(s \mid o) = \frac{1}{Z_0} exp(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(s_{i-1}, s_i, o, i)) , \quad (1)$$

where Z_0 is a normalization factor of all state sequences, $f_j(s_{i-1}, s_i, o, i))$ ***Error! Reference source not found.***is one of m functions that describes a feature, and λ_j is a learned weight for each such feature function. Feature functions could ask arbitrary questions about two consecutive states, any part of the observation sequence and the current position. Their values may range between $-\infty$ and $+\infty$, but typically they are binary. To make all conditional probabilities sum up to 1, we must calculate the normalization factor

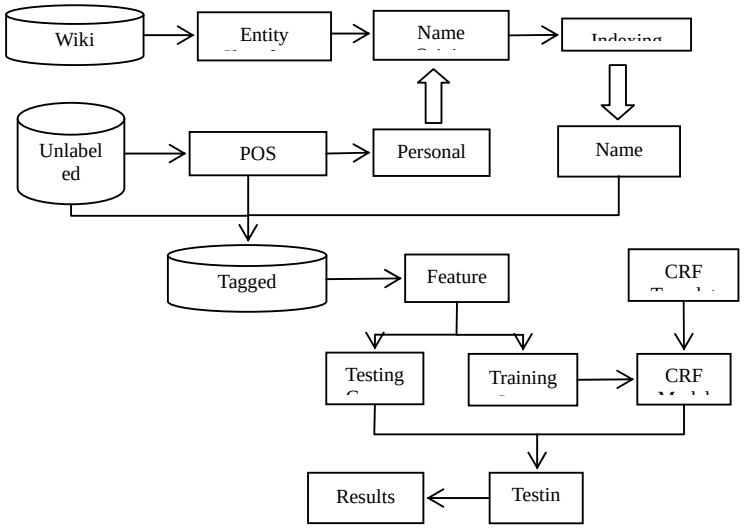
$$Z_0 = \sum_s exp(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(s_{i-1}, s_i, o, i)) . \quad (2)$$

Corpus Acquisition and Labeling. The rapidly growing Web is one of the largest distributed databases in the world. In this article, we propose an approach that extracts corpus and identify the personal names origin from the Web. Wikipedia is an open, collaborative encyclopedia on the Web, and has a rapid and successful growth in recent years. Wikipedia provides a variety of data resources for NER and other NLP research. Most articles in Wikipedia are about named entities, including the names of famous person, and they are more structured than raw texts.

We use Wikipedia database on the Chinese version as the initial corpus of the entire system. An article in Wikipedia is identified by a unique name. By using the first paragraph of the entity as the

classified information, we map the Wikipedia entities to categories (LOC, PER, ORG). On the basis of the classification result, we can get the name corpus if the classified result is PER. In the name corpus, the name itself is the Wiki entry, and according to the first paragraph of the text we can get the national information of the name.

We establish the region matching information for Chinese, Japanese and Europe-American area, and we can map the places and the organization names into the country names. But for some personal names, the first paragraph of the article may appear some different geographical names of places or



organizations. In this case, we extract all names of places and organizations; use them as features, then map this information to the country names. In our experiment, the closer a place or organization name is to the personal name, the more likely it belongs to the origin of the corresponding country. In this way, we find out the origin of the personal name by using the articles in Wikipedia.

Fig. 1 Workflow of our approach

Proposed Approach. Based on the above resources, we propose an approach, which relies on three models: recognition model which gives the part-of-speech tagging and judges the word whether a name or not; linking model which determines the origin of a name; and training model, which does the training and testing with CRFs. The workflow of the approach for the name origin recognition is shown in Fig.1.

IOB Tagging Label. We use “B-X”, “I-X”, “O” tags, where “B”, “I”, “O” mean the beginning of a person name correspondingly, the inside of a person name, and the outside of a person name. Suffix X represents the origin of a name. For different countries, namely Chinese, Japanese and Europe-American area, are tagged by the letters “CH”, “JA”, “EN”, respectively. According to IOB tagging label, the tag set in our system is L, where L = {B-CH, I-CH, B-EN, I-EN, B-JA, I-JA, O}. Thus, identifying name origin in Chinese texts with CRFs is to make classification for each word in the texts.

Features. We design three kinds of features to represent each sentence for CRFs.

*Content (n-gram)***:** We use features for sentence contents represented by: i) words, ii) word bigrams, namely **Error! Reference source not found.**W_n (n = -2,-1,0,1,2) and **Error! Reference source not found.Error! Reference source not found.**W_nW_{n+1} (n = -2,-1,0,1), where W refers to a Chinese character while **Error! Reference source not found.**W_0 denotes the current character and W_n (W_{-n})**Error! Reference source not found.** denotes the character n positions to the right (left) of the current character.

*POS-tag***:** The Part-Of-Speech tagging is the task of assigning to each word its linguistic category. The point of using POS-tags relies mainly on that determine the beginning and the end of a personal name and thus help the classifier to capture the boundaries of the names.

After years of intensive researches, Chinese word segmentation has achieved a quite high performance [7]. Among all of them, the ICTCLAS (developed by Chinese Academy of Sciences) is

currently the best one both in accuracy and speed. This Chinese lexical analysis system combines part-of-speech (POS) tagging, word segmentation and unknown word recognition [8]. Therefore, we use NLPIR [9], namely ICTCLAS 2013, for our POS tagging task. For the character in the sentence, we retain its original POS of the words accordingly.

External Resources (ER): We use the lists of names of different origins in the corpus to analyze the statistical information of the commonly-used words of these names. For different origins, we integrate the common words of surname and the word frequency information for the three origins.

In our experiment, we take characters as the basic processing unit, and collect the statistical information of the commonly-used characters for the names of different origins in the corpus. The statistical results show that there exists great difference in commonly-used characters and surnames between the three origins.

Experiments

Data and Setting. The acquisition of corpus in our experiments refers to the above section, and the corpus contains about 120,000 sentences. Table 1 lists the numbers of personal names of the three origins in the corpus. In the experiments, 90% of the sentences are randomly selected for training and the remaining 10% are kept for testing for each language origin.

Table 1 Number of names of each origin

Origin	Number of names (duplicate names have been removed)	Number of sentences
Chinese	8282	63058
Japanese	7678	18477
English	14119	33638

Evaluation. We use the**Error! Reference source not found.** $F_{\beta=1}$-measure for evaluation:

$$F_{\beta=1}=\frac{(\beta^2+1)*precision*recall}{\beta^2*(precision+recall)}, \tag{3}$$

where precision is the percentage of names recognized by the system and which are correct, and recall is the percentage of names existing in the corpus and which were found by the system.

Results and Analysis. We use CRF++ (ver.0.53) [10] as the basis of our implementation of CRFs. The baseline model we use is the Maximum Entropy (ME) model. And in order to compare the performance of the Maximum Entropy (ME) with the CRF performance, we keep the same feature-set for the two models. In our implementation, we use Zhang's maximum entropy package [11].

Baseline and CRF Results. Table 2 and 3 report the performances of the baseline and the proposed methods on all the features respectively. The proposed method outperforms the baseline method in all experimental configurations. This suggests that CRFs are more suitable for name origin recognition in Chinese texts. It is noteworthy that the CRF classifier gained higher performance than the ME. The data of the two tables reflect the capability of CRFs to determine the origin of names in the Chinese texts.

Table 2 Baseline Results

Baseline	*P*(%)	*R*(%)	*F*(%)
Chinese	77.73	88.03	82.56
English	91.82	95.25	93.50
Japanese	84.88	90.90	87.79
Overall	82.24	90.34	86.10

Table 3 CRF Results

CRF	*P*(%)	*R*(%)	*F*(%)
Chinese	89.45	91.03	90.23
English	97.97	97.83	97.90
Japanese	96.37	95.98	96.17
Overall	92.67	93.55	93.11

Impact of Features. Table 4 and 5 report the contributions of different features in our experiments by gradually incorporating the feature set. For the sake of brevity, the origin subscripts are "C", "J" and "E" for Chinese, Japanese and English name origin respectively. Table 4 and 5 show that:

1) Unigram features are the most fundamental and informative. Even in the Chinese texts, using CRF model, the unigram features can perform well.

2) Bigram features upgrade the performance on the basis of unigram features.

3) The POS information can greatly improve the performance and the combination of the above three useful features achieves the best performance of 93.15% in overall F-1 as shown in Table 7. This is largely due to the boundary division of the names as discussed before.

4) External Resources (ER) Features degrade the performance. This large due to the data sparseness problem, because the commonly-used surnames and characters is also common characters in Chinese texts. However, compared with "Unigram+Bigram" features, the "Uni+Bi+ER" features increase recall of English origin by 0.38%, and improve the Japanese origin's precision by 0.24%.

Table 4 Contribution of Each Feature

Feature	P_C(%)	P_E(%)	P_J (%)	R_C(%)	R_E(%)	R_J(%)	F_C(%)	F_E (%)	F_J(%)
Unigram	77.22	79.86	79.03	67.08	69.94	65.36	71.79	74.57	71.55
+Bigram	86.39	86.43	88.36	82.50	82.45	81.30	84.40	84.39	84.68
+POS	89.46	98.17	96.41	91.07	97.86	95.88	90.26	98.02	96.14
+ER	89.45	97.97	96.37	91.03	97.83	95.58	90.23	97.90	96.17
Uni+Bi+ER	86.14	86.04	88.60	82.44	82.83	81.25	84.25	84.40	84.76

Table 5 Contribution of Each Feature

Feature	*P*(%)	*R*(%)	*F*(%)
Unigram	78.18	67.56	72.48
+Bigram	86.69	82.30	84.44
+POS	92.74	93.56	93.15
+ER	92.67	93.55	93.11
Uni+Bi+ER	86.47	82.36	84.37

The rest of the tables show the results obtained using each of the features individually. When each feature is used individually, the Bigram (Table 6) features show the best performance, whereas compared with the "Unigram+Bigram" features, the latter greatly improves the recall. The POS-tag features get the lowest performance on the recall of the Japanese origin (Table 7), this is largely due to the accuracy of the POS tagging. Using external resources is not very effective (Table 8), as we said above, this is because that the commonly-used surnames and characters is also common characters in Chinese texts, and the most evident data is the performance of names of Chinese origin.

Table 6 Results obtained using the Bigram feature

Bigram	*P*(%)	*R*(%)	*F*(%)
Chinese	87.82	76.39	81.71
English	86.67	76.72	81.39
Japanese	89.49	72.35	80.01
Overall	87.75	75.87	81.38

Table 7 Results obtained using the POS-tag feature

POS	*P*(%)	*R*(%)	*F*(%)
Chinese	62.46	79.55	69.98
English	65.02	58.59	61.63
Japanese	66.02	25.69	36.99
Overall	63.23	65.94	64.56

Table 8 Results obtained using the External Resources feature

ER	*P*(%)	*R*(%)	*F*(%)
Chinese	32.90	8.51	13.53
English	51.01	20.75	29.50
Japanese	48.27	14.03	21.74
Overall	41.48	12.53	19.24

Conclusion and Future Work

In this paper, we present our preliminary experiments which aim at recognizing name origin, our system for Chinese texts, by using CRF model. The results show that with the CRF model we can obtain a performance substantially higher with respect to the ME model.

We propose using CRF model to explore diverse features for name origin recognition in Chinese texts. We make use of the Wikipedia articles to formulate our experimental data, and take the advantage of the structured texts to discriminate the origin of the names. Experiment results show that our method is effective and very promising.

The purpose of this research is to demonstrate that the CRFs can be used to the name origin recognition task. In the next future we hope to increase the size of the data size in order to obtain a higher performance of the system. We also plan to investigate the use of other context features, carry out experiments using different features, and explore the possibility of designing a feature-set for each origin. Furthermore, we plan to conduct a comparative study between many probabilistic models (SVM, HMM, ME, CRF, etc.) and also experiments using a combination of different models.

References

[1] V. Pervouchine, M. Zhang, M. Liu, and H. Li, Improving name origin recognition with context features and unlabeled data, In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, (2010): 972-978.

[2] J. S. Kuo, H. Z. Li, and Y. K. Yang, A phonetic similarity model for automatic extraction of transliteration pairs, In ACM Transactions on Asian Language Information Processing (TALIP), (2007) 6(2): 6.

[3] Y. Chen, J. You, M. Chu, Y. Zhao, and J. Wang, Identifying language origin of person names with N-grams of different units, In Acoustics, Speech and Signal Processing, 2006 (ICASSP 2006 Proceedings). 2006 IEEE International Conference on IEEE, (2006) 1: I-I.

[4] A. F. Llitjós, Improving pronunciation accuracy of proper names with language origin classes, ESSLLI Student Session. (2001) 53.

[5] M. Zhang, C. Sun, H. Li, A. Aw, C. L. Tan and X. Wang, Name Origin Recognition Using Maximum Entropy Model and Diverse Features, In IJCNLP, (2008) 56-63.

[6] J. Lafferty, A. McCallum, and F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, (2001).

[7] C. Huang, H. Zhao, Chinese word segmentation: A decade review, In Journal of Chinese Information Processing, (2007), 21(3): 8-20.

[8] L. Wang, S. Li, D. F. Wong, and L.S. Chao, A Joint Chinese Named Entity Recognition and Disambiguation Sys-tem[C], In The 2nd CIPSSIGHAN Joint Conference on Chinese Language Processing (CLP-2012), (2012).

[9] Institute of Computing Technology, Chinese Academy of Sciences, http://ictclas.nlpir.org/.

[10] CRF++: Yet another CRF toolkit, Software available at http://crfpp.sourceforge.net, (2005).

[11] Software available at http://homepages.inf.ed.ac.uk/s0450736/maxent.html.