Mining Translations of Chinese Names from Web Corpora Using a Query Expansion Technique and Support Vector Machine

Kai-Hsiang Yang[*], Wei-Da Chen[#], Hahn-Ming Lee[*,#], Jan-Ming Ho[*]

 # Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology,Taipei 106, Taiwan
{M9415021, hmlee} @mail.ntust.edu.tw

** Institute of Information Science, Academia Sinica Taipei 115, Taiwan*
{khyang, hmlee, hoho}@iis.sinica.edu.tw

Abstract

Chinese name translation is a special case of the problem of named entity translation. It is a very challenging problem because there exist many kinds of Romanization systems and some people like to add additional words into their English names. Translating a scholar's name to its corresponding English name could help find information about his academic achievements. In this paper, we provide a classification for Chinese names, and propose a novel approach to mining Chinese name translations from Web corpora. Our approach is based on three kinds of features, namely the phonetic similarity, the smallest distance, and the number of appearances in the neighborhood, to extract name translation candidates by using a query expansion technique and Support Vector Machine (SVM). Experimental results show that our approach can correctly translate the majority of Chinese names.

1. Introduction

In recent year, Researchers usually create their homepages on the Internet for various reasons, such as describing their research and contributions, or providing material for their new courses. However, scholars usually publish their work under their English names. When we want to find the state-of-the-art research of a scholar, obtaining his/her English name is an essential step in collecting related information; we focus on the "Chinese name translation" problem in this paper.

Chinese name translation is a special case of the problem of named entity translation. It is a very challenging problem because of the following two properties. First, there exist many kinds of Romanization systems, such as Tongyong Pinyin [6] and Hanyu Pinyin [4], so individuals can choose any one of them to translate their names. For example, the Chinese surname "林" may be translated into "Lin" or "Lam". Second, a translation of a Chinese name sometimes contains additional words that are not related to any word in the Chinese name of the person. For example, many Chinese people add Western first names into their English names. We had analyzed many Chinese names and provide a classification of Chinese name in Table 1, which contains eight different types of names.

Table 1. Chinese name classification.

Name format	Example
Type-1. (Chinese given name) (Surname) or (Surname), (Chinese given name)	劉豐哲 (Fon-Che Liu)
Type-2. (Merged Chinese given name) (Surname)	吳德琪 (Derchyi Wu)
Type-3. (Western first name) (Surname)	趙蓮菊(Anne Chao)
Type-4. (Chinese given name) (Western first name) (Surname)	黃光明 (Kwang-Ming Frank Hwan)
Type-5. (Abbreviated Chinese given name) (Surname)	張秀瑜 (S. Y. Chang)
Type-6. (Western first name) (Abbre. Chinese given name) (Surname)	李昭勝(Jack-C. Lee)
Type-7. (Chinese given name) (Abbreviated Chinese given name) (Surname)	蔡桂紅 (Gwei-Hung H. Tsai)
Type-8. (Chinese given name) (Unpredictable Surname)	張韻詩 (Jane Win-Shih Liu)

Most approaches that deal with the translation of named entities are based on the techniques of parallel corpora comparison [9][12]and Web mining [7][8][10]. The parallel corpora comparison approach is ineffective when there are no parallel documents related to the named entity, and the Web mining methods do not consider the order of the terms in Chinese names and the phonetic similarity. Moreover, Chinese names are not translated semantically; sometimes they are translated according personal

preferences. Therefore, calculating the semantic similarity may not resolve this problem.

In this paper, we propose a Chinese name translation system to find the English translation of a Chinese name from Web corpora, for a given Chinese name *N,* we first get its translation of the surname *T* by looking up surname dictionary, and then use the pair (*N*,*T*) as a query to a search engine. Our system will extract candidates from the returned snippets, where a snippet is a short paragraph to describe each searched results from search engines.

Moreover, our system uses a query expansion technique to reduce the required number of web documents during the mining process. Huang *et al.* [8] demonstrated how a query expansion improves the performance of retrieving useful information. During the process, there is no constant rule to determine whether a chain of English terms in a snippet is someone's English name or not. Therefore, it is hard to define heuristic rules to help make the decision. However, the performance could be improved by a well trained machine learning algorithm. Murata *et al.* [11] conducted experiments on several machine learning methods and showed that SVM achieves best performance. Hence, our system employs a query expansion technique and the SVM technique to select appropriate name candidates for a given person's Chinese name. Our approach uses three kinds of features, namely the phonetic similarity, the smallest distance, and the number of appearances in the neighborhood. Experiment results show that our system can correctly translate the majority of Chinese names.

The rest of paper is organized as follows. In Section 2, we present the system architecture of our proposed system, and demonstrated experimental results in Section 3. Finally, in Section 4, we provide our conclusions and discuss our future work.

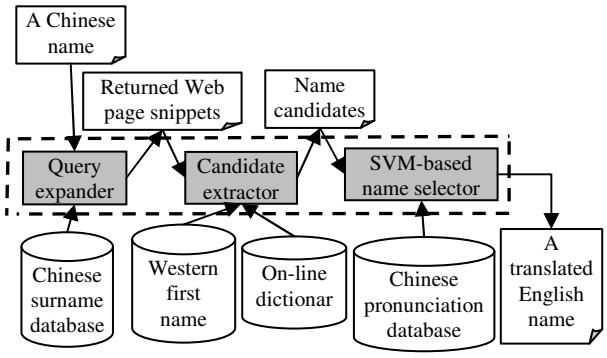

Figure 1. System architecture.

2. System architecture

Figure 1 shows the system architecture, which contains three modules: a Query expander, a Candidate extractor, and a SVM-based name selector. In addition, the system utilizes an on-line dictionary and three databases, namely a Chinese surname database, a Western first name database, and a Chinese word pronunciation database. These modules and databases are briefly described in the following. The Query expander retrieves Web page snippets containing the inputted Chinese name and the translation of its surname. Then, the Candidate extractor extracts translation candidates using some rule-based methods. Finally, the SVM-based name selector extracts each candidate's features and then utilizes SVM to determine whether the candidate is the correct translation of the inputted Chinese name.

2.1. SVM-based name selector

After the first two modules find out candidates from snippets, the SVM-based name selector uses three features of each candidate to determine whether the candidate is the translation of the inputted Chinese name. These three features include the "phonetic similarity", the "smallest distance", and the "number of appearances in the neighborhood".

2.1.1. Phonetic similarity

The famous Soundex technique [5] is used in our system for calculating the phonetic similarity of terms. The system first translates each word in a Chinese name by looking up a Chinese word pronunciation database, and then calculates the Soundex codes for the translated name and each candidate. Because a code *c* has 8 codes, say c_1, c_2 …c_8, and, the phonetic similarity between code *c* and *p* is calculated as follows:

$$F_{PhSim}(c \mid p) = \sum_{i=1}^{8} f(c_i \mid p_i), \qquad (1)$$

where $f(c_i|p_i)$ is a value that indicates whether c_i is equal to p_i or not. When i = 1 or 5, if c_i is equal to p_i, then $f(c_i|p_i) = 2$. When i = 2, 3, 4 or 6, 7, 8, if c_i is equal to p_i, then $f(c_i|p_i) = 1$. If c_i is not equal to p_i, then $f(c_i|p_i) = 0$.

2.1.2. Smallest distance

The second feature we choose is the minimal distance between the inputted Chinese name and each candidate in the retrieved snippets. Because there may have many occurrences for each candidate, we only

concern the shortest distance between the inputted Chinese name and all occurrences of each candidate.

2.1.3. Number of appearance in neighborhood

The third feature is the number of appearances of each candidate in the neighborhood of the inputted Chinese name. The neighborhood is defined by the distance between the Chinese name and the candidate in the Web page snippets, where the distance is smaller than a given threshold.

3. Experiments

We evaluate our approach on two datasets. Dataset I is the one we used to train the training vector, and it contains 78 pairs of Chinese and English names. For creating a larger database, we developed a program to collect the expert database in the National Science Council; the database contains 1,117 pairs of Chinese and English name of researchers and professors in Taiwan. We classify the data according to our classification (8 types). Table 2 shows the distribution of the dataset.

Table 2. Distribution of the two datasets

Name Type	Dataset I		Dataset II	
	#	%	#	%
Type-1.	19	24.3%	1000	89.5%
Type-2.	10	12.8%	42	3.8%
Type-3.	9	11.5%	9	0.8%
Type-4.	14	17.9%	50	4.5%
Type-5.	3	3.8%	0	0%
Type-6.	8	10.3%	9	0.8%
Type-7.	3	3.8%	3	0.3%
Type-8.	12	15.4%	4	0.4%

In Dataset I, the data is selected manually in order to reduce the differences in the number of each name type. Dataset II is collected from a real world scenario. Thus, we can observe the true distribution of the Chinese name types in Dataset II. Most people translate their names as Type-1; only a few people translate their names into Type-2 and Type-4. Hence, most of people do not add extra terms when translating their names into English.

We use the alignment accuracy proposed in [8] as our measurement. It is defined as the probability of selecting the correct answers when the searched snippets contain the correct answers. It is a conditional probability. We measured the performance of our approach by the top-1 to top-5 alignment accuracy.

To train a SVM model, we selected 78 names as the training set to produce the training vectors for the three features. Moreover, if more than two similarity values of features of a candidate are better than the answer, the candidate will be labeled as +1, otherwise is marked as -1. By using this kind of labeling method, the over-fitting problem can be greatly resolved.

Each training vector contains four features: a label, the phonetic similarity, the smallest distance and the number of appearances in the neighborhood. We use the training function in LIBSVM [3] to train the model.

3.3. Experimental results

We first evaluate our system performance on Dataset I and II. Figure 2 shows the translation accuracy of our system on each dataset. We can obviously see that our approach achieved 70.5% of Top-1 accuracy and 90.1% of Top-5 accuracy on Dataset I; and 57.9% of Top-1 accuracy and 86.2% of Top-5 accuracy on Dataset II.

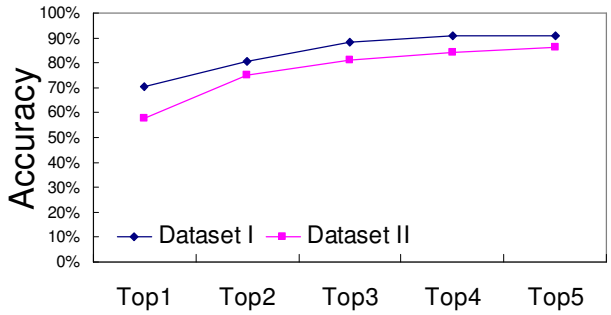

Figure 2. Translation accuracy on Dataset I and II.

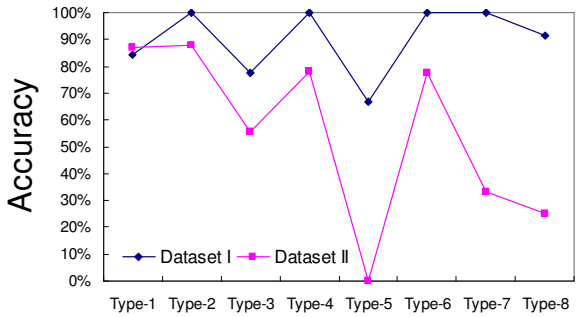

Figure 3. Translation accuracy of each name type on Dataset I and II.

We then analyzed the translation accuracy for each name type on Dataset I and II, as shown in Figure 3. It

is obvious from the result that our approach achieves good performance, especially on Type-1, Type-2, Type-4, and Type-6 name types. Note that Type-1, Type-2 and Type-4 are commonly used Chinese name types. Our system also achieves a good performance in translating the Type-6 name type. However, the translations we found for the Type-6 Chinese names are not exactly the same as the answers; furthermore, most of these translations belong to Type-1, Type-2 or Type-4. This shows that it is not easy to mine translations corresponding to Type-6 name types. But we can easily observe that, in Dataset II, few people translate their Chinese names into the Type-3, Type-5, Type-6, Type-7, and Type-8.

4. Conclusion

We propose a novel approach for mining the Chinese name translations from Web corpora. Our approach extracts name translation candidates by using a query expansion technique and selects appropriate name candidates as the English translation of the inputted Chinese name using SVM. We also provide a classification of Chinese names. Experiment results show that our approach can correctly translate the majority of Chinese names (providing 90% translation accuracy for Dataset I and 86% for Dataset II at the top-5 results).

When mining for personal information, the name ambiguity problem often arises [13][14][15]. In this paper, we have not dealt with the person name disambiguation problem, but we will investigate it in our future work. Thus, users can provide information about a person of interest to obtain the name translation from our system. Then we can translate the name precisely according to the information.

Acknowledgement

This work was supported by the National Digital Archive Program(NDAP, Taiwan), the National Science Council of Taiwan under grants NSC 95-2422-H-001-024, NSC 95-2218-E-001-001, NSC 95-2218-E-011-015, NSC 95-3114-P-001-002-Y02, NSC95-3114-P-001-001-Y02 and NSC 95-2221-E-001-021-MY3.

References

[1] Directory of Division of Computer Science of National Science Council. http://cs.nsc.ncku.edu.tw/news

[2] Directory of scholars of Institute of Mathematics, Academia Sinica. http://www.math.sinica.edu.tw/addbook/default.jsp

[3] LIBSVM -- A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[4] Pinyin - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Pinyin

[5] Soundex - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Soundex

[6] Tongyong Pinyin - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Tongyong_Pinyin

[7] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu and L.-F. Chien, "Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval," *Special Interest Group on Information Retrieval 2004 (SIGIR'04)*, Sheffield, South Yorkshire, UK, July, 2004.

[8] F. Huang, Y. Zhang and S. Vogel, "Mining Key Phrase Translations from Web Corpora," *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, October, 2005.

[9] T. Kumano, H. Kashioka, H. Tanaka and T. Fukusima, "Acquiring Bilingual Named Entity Translations from Content-Aligned Corpora," *International Joint Conference on Natural Language Processing (IJCNLP)*, China, March, 2004.

[10] W.-H. Lu, L.-F. Chien and H.-J. Lee, "Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach," *ACM Transactions on Information Systems*, Vol. 22, No. 2, pages 242–269, 2004.

[11] M. Murata, K. Uchimoto, Q. Ma and H. Isahara, "Using a Support-Vector Machine for Japanese-to-English Translation of Tense, Aspect, and Modality," *Annual Meeting of the ACL archive Proceedings of the workshop on Data-driven methods in machine translation (WDDMT)*, France, July, 2001.

[12] M.-S. Shia, J.-H. Lin, S. Yu and W.-H. Lu, "A Web-based Unsupervised Algorithm for Learning Transliteration Model to Improve Translation of Low-Frequency Proper Names," *Natural Language Processing and Knowledge Engineering, 2005. IEEE (NLP-KE '05)*, China, October, 2005.

[13] Y.-C. Wei, M.-S. Lin and H.-H. Chen, "Name Disambiguation in Person Information Mining," *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*, Hong Kong, December, 2006.

[14] K.-H. Yang, K.-Y. Chiou, H.-M. Lee and J.-M. Ho, "Web Appearance Disambiguation of Personal Names Based on Network Motif," *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*, Hong Kong, December, 2006.

[15] K.-H. Yang, J.-Y. Jiang, H.-M. Lee and J.-M. Ho, "Extracting Citation Relationships from Web Documents for Author Disambiguation," *Technical Report (TR-IIS-06-017)*, Institute of Information Science, Academia Sinica, 2006.