2009

Are You Finding the Right Person? A Name Translation System Towards Web 2.0

Yilu Zhou
George Washington University, yzhou@gwu.edu

ARE YOU FINDING THE RIGHT PERSON? A NAME TRANSLATION SYSTEM TOWARDS WEB 2.0

Completed Research Paper

Yilu Zhou
George Washington University
2201 G St. NW, Suite 515, Washington DC 20052
yzhou@gwu.edu

Abstract

In a multilingual world, information available in global information systems is increasing rapidly. Searching for proper names in foreign language becomes an important task in multilingual search and knowledge discovery. However, these names are the most difficult to handle because they are often unknown words that cannot be found in a translation dictionary and even human experts cannot handle the variation generated during translation. Furthermore, existing research on name translation have focused on translation algorithms. However, user experience during name translation and name search are often ignored. With the Web technology moving towards Web 2.0, creating a platform that allow easier distributed collaboration and information sharing, we seek methods to incorporate Web 2.0 technologies into a name translation system. In this research, we review challenges in name translation and propose an interactive name translation and search system: NameTran. *This system takes English names and translates them into Chinese using a combined hybrid Hidden Markov Model-based (HMM-based) transliteration approach and a web mining approach. Evaluation results showed that web mining consistently boosted the performance of a pure HMM approach. Our system achieved top-1 accuracy of 0.64 and top-8 accuracy of 0.96. To cope with changing popularity and variation in name translations, we demonstrated the feasibility of allowing users to rank translations and the new ranking serves as feedback to the original trained HMM model. We believe that such user input will significantly improve system usability.*

Keywords: Name translation, name variation, user feedback, transliteration, web mining, Hidden Markov Model

1 Introduction

Multilingual information available in global information systems is increasing rapidly. However, valuable information generated from non-English-speaking regions is often neglected by English-speaking users due to language barrier. There are various circumstances when one needs information in foreign documents. An intelligence officer needs to track a suspect's activity records in a foreign database system; a financial analyst wants to look for foreign news to understand the global equity market; a researcher seeks for foreign publications; or a regular customer searches international products. With the Internet reaching 1 billion non-English-speaking users, non-English Web content and data sources become ubiquitous. Being able to search in multilingual documents becomes a pressing issue.

Web search engines, such as Google, Yahoo! and MSN Live Search have added multilingual capabilities. However, they are a collection of monolingual search engines rather than a truly worldwide multilingual search engine. When an English query is sent to these commercial search engines and a request for Chinese documents is specified, only a small portion of Chinese Web sites happened to have the original English query term in them are returned. The majority of the relevant Chinese Web sites remain missed. This issue is often addressed by translating the original query into target document languages before search (Oard, 1996).

Proper names, such as organizations, company names, product names, and person names play a most important role in search queries and knowledge discovery. It was reported that 67.8%, 83.4%, and 38.8% of queries to Wall Street Journal, Los Angeles Times, and Washington Post respectively involved name searching (Thompson and Dozier 1997). Being able to handle proper name translation will significantly improve performance of searching in a multilingual world and is especially important in today's global environment. Name translation is straightforward between language pairs employing the same alphabets (e.g., English/Spanish). The translation simply stays the same as the origin. However, for language pairs employing different alphabets (e.g., English/Arabic), proper names are translated phonetically, referred to as *transliteration.* For example, President "Obama" is transliterated into Chinese as "奥巴马" and the company name "SONY" is transliterated into Arabic as "سوني."

However, little research has been devoted to name translation and search. It is in general difficult for human to translate unfamiliar personal names, place names and names of organizations (Lee et al., 2006). One reason is the variability in name translation. In many situations, there is more than one correct translation for the same name. In some languages, such as Arabic, it can go up to as many as forty (Arbabi et al., 1994). Even professional translators find it difficult to identify all variations. For example, when translating "Phelps" into Chinese, there are at least 5 different translations as shown in Table 1. All five translations are phonetically correct translations. Since the central media in China uses the first translation for Michael Phelps and followed up by other media press, it becomes the most widely used one. Besides the variation in translation, the same translation can represent multiple person entities which increase the complexity of multilingual name search.

Table 1: Chinese translations of "Phelps"

English Name	Chinese Translation	Chinese Pinyin	Comment
Phelps	菲尔普斯	Fei Er Pu Si	Most widely used translation of swimmer "Michael Phelps"
	费尔普斯	*Fei* Er Pu Si	Most widely used translation of the famous economist "Edmund S. Phelps." The second most popular translation of "Michael Phelps" by media.
	弗尔普斯	*Fu* Er Pu Si	Correct translation, used as a character name in a novel
	菲尔普思	Fei Er Pu *Si*	Correct translation, not widely used
	菲尔普丝	*Fei* Er Pu *Si*	Correct translation, often used as girl's name

Existing research on name translation have focused on translation algorithms. However, user experience during name translation and name search are often ignored. With the Web technology moving towards Web 2.0, creating a platform that allow easier distributed collaboration and information sharing, we seek methods to incorporate Web 2.0 technologies into a name translation system. The term Web 2.0 is often associated with social software, weblogs, wikis and RSS. In fact, the Web 2.0 core concept is that collective individual effort can surpass professionals (Ankolekar et al., 2007). In a complex situation such as name translation, where professionals find it difficult to identify all the possible translations, adopting Web 2.0 technology is desired.

In this research, we propose a framework for name translation and search system using a transliteration approach. The system incorporates a probability and Web mining based translation module along with an interactive user feedback module which contributes to the probability translation module. Our work intend to demonstrate the feasibility of improving users' name translation and name search experience with user feedback through a system development approach. The rest of the paper is structured as follows. Section 2 reviews related research. In Section 3 we describe our research questions and in Section 4 we propose our name translation and search system: NameTran. Section 5 presents our experimental design and results and shows a sample user session of such interactive name search system. Finally, in Section 6 we conclude our work and suggest some future directions.

2 Related Work

2.1 Comparison of translation approaches

Research in translating proper names has focused on two strategies: One is a direct translation approach (Chen and Zong, 2008). The second approach is to mine translation pairs from bilingual online resources or corpora (Lee et. al, 2006).

The first approach, direct translation, is often done by transliteration because proper names are usually out-of-vocabulary (OOV) terms that cannot be found in a dictionary. Transliteration is the representation of a word or phrase in the closest corresponding letters or characters of a language with different alphabet so that the pronunciation is as close as possible to the original word or phrase (AbdulJaleel and Larkey, 2003). Unlike mining-based approach, transliteration can deal with low-frequency proper names, but may generate ill-formed translations.

The second approach is based on the assumption that the two name equivalents should share similar relevant context words in their languages. Correct transliteration is then extracted from the closest matching proper nouns. This approach is limited by the coverage of corpus, known as data sparsity problem. It becomes less effective when dealing with low-frequency unknown proper names. Resources on the Web have been used to complement insufficient corpora (Goto et al., 2001, Cao and Li 2002, Lu et al., 2004, Zhou et al., 2008).

2.1.1 Direct Translation Approach: Transliteration

Previous transliteration models can be categorized into rule-based approach, machine learning approach, and statistical approach. A rule-based approach maps each letter or a set of letters in the source language to the closest sounding letter or letters in the target language according to pre-defined rules or mapping tables. It relies on manual identification of all transliteration rules and heuristics, which can be very complex and time consuming to build (Darwish et al., 2001). The transliteration accuracy depends on the completeness of the rules. Due to the ambiguity of some rules, noise is often introduced. Moreover, this approach is not expandable to different languages pairs. A machine learning approach improves rule-based mapping by filtering out unreliable translations learned from target language patterns. Although some bad transliterations can be removed, good transliterations can be removed as well. A statistical approach is the most promising approach. Instead of relying on a large set of language heuristics, a statistical approach obtains translation probabilities from a training corpus: pairs of transliterated words. When new words come, the statistical approach picks the transliteration candidate with the highest transliteration probabilities generated as the correct transliteration.

Most statistical-based research used phoneme-based transliteration, relying on a pronunciation dictionary. This approach fails when such a dictionary is not available. Al-Onaizan and Knight showed that a grapheme-based approach out-performed a phoneme-based approach in Arabic-English transliteration (Al-Onaizan and Knight,

2002). Zhou et al. (2008) developed a grapheme-based English-Arabic transliteration system which applies a hybrid probability model and a web mining model. The system demonstrated superior performance with web mining boosting effect.

2.1.2 Indirect Translation Approach: Mining Name Entities from the Web

The mining-based approach takes a very different view of the transliteration problem. Web mining is defined as the discovery and analysis of useful information from the WWW. It can be categorized into Web content mining, Web structure mining and Web usage mining (Pal et al., 2002). Web content mining deals with web page content in a form of text and documents. Structure mining copes with hyperlinks between websites. Web usage mining utilizes data generated by users' interactions with the Web, such as server logs and user profiles (Chen & Chau, 2004). Unlike transliteration approach, Web mining-based approach does not rely on transliteration heuristics or probability models. Instead, it searches the Web for transliteration using relevant context words of the source name. The assumption here was that the two name equivalents should share similar relevant context words in their languages. Correct transliteration is then extracted from the closest matching proper nouns.

Goto et al. (2001) proposed such an Internet-based technique for finding English equivalents for Japanese names. They first searched the Internet for relevant context words of the original name, and then used the translated context words as a query to obtain relevant Web documents. Similarly, Lu, Chien, and Lee (2004) presented an approach to finding translation equivalents of query terms and constructing multilingual lexicons through the mining of Web anchor texts and link structures, which was shown to be effective on English-Chinese Web documents. Sproat et al. (2006) studied Chinese-English name transliteration using comparable corpora using temporal distribution of candidate pairs. They achieved improvements when combining mining-based approach with direct transliteration.

The Web mining approach is applicable to any pairs of languages. No rules, dictionaries, or training corpora are needed. However, the performance depends on the ability to identify proper names and accuracy in translating relevant context words. This approach works well for hotspots in news articles, but not normal names.

2.2 Name Translation Directions and Categories

Transliteration can happen in two directions: forward transliteration and back transliteration (Lin & Chen 2002). Given a pair (s, t) where s is the original proper name in the source language and t is the transliterated word in the target language. Forward transliteration is the process of phonetically converting s into t. Back transliteration is the process of correctly finding or recovering s given t. Forward transliteration is a one-to-many mapping. Some transliterations might be more popular than others, but it is difficult to define one "correct" transliteration. On the other hand, back transliteration is often a many-to-one mapping between both letter-based languages has been identified as a more difficult task than forward transliteration for some language pairs (Stalls & Knight, 1998). However, when letter-based and character-based languages are both involved, such as English-Chinese, it becomes a more complex problem and often involves two-step transliteration (Virga and Khudampur, 2003).

Table 2 illustrates examples in both transliteration directions. We use English-Arabic pair to illustrate the transformation between both alphabet-based languages, and English-Chinese pair to demonstrate translation between character-based and alphabet-based languages. For each language pair, there are four types of transliteration which are listed in the table. Based on our review and native speakers' comment, we also access the difficulty and variation level of each type. Previous research studying each direction is included as well.

For English-Arabic pair, forward transliteration has a high variability because there are many different ways you can translate a name. The existence of multiple correct translations also makes the difficulty level medium, since translation only need to generate one of the correct translations. Although being able to identify one correct translation is not hard, being able to identify all or most correct translation is still a difficult task. Back transliteration has high difficulty level because there is often only one correct origin. A simple rule-based mechanism will generate many transliteration candidates but fails to identify the correct one.

For English-Chinese pair, similarly forward transliteration has high variation. We assign a "high" difficulty level to English to Chinese forward translation, because they use very different systems, alphabet-based and character-based. Their forward transliteration cannot be performed directly and often relies on some type of phonetic dictionary to generate an intermediate phonetic translation. Thus, forward translation from English to Chinese is a difficult problem. The case of Chinese to English "forward" transliteration is a special case. Both variation and difficulty levels are low. There are three types of Chinese names. Simplified Chinese is being used in Mainland China, and Traditional Chinese is used in both Taiwan and Hong Kong. Mainland China and Taiwan use similar phonetic system while Hong Kong uses Cantonese phonetic system. For simplified Chinese and traditional Chinese used in Taiwan, there are standard ways to translate Chinese names to English. For traditional Chinese used in Hong Kong, no standard translations are adopted. However, the variation is still low. The difference among three Chinese language systems is not the focus of this paper which we address in a different paper. Our examples are in simplified Chinese used in Mainland China.

Table 2: Taxonomy of Transliteration Directions and Previous Work

Direction	**Forward transliteration**		**Back transliteration**	
Process	Phonetically convert to a foreign language		Recover the original name	
Feature	One-to-many		Many-to-one, sometimes many-to-many	
	English -> Arabic	**Arabic -> English**	**English -> Arabic**	**Arabic -> English**
Examples	Edison → ادسن / ا د يـسن / اديســون / أديســون	محمد → Muhammed / Mohammed / Muhammad	Muhammed / Mohammed / Muhammad → محمد	ادسن / اديســن / اديســون / أديســون → Edison (Eddison)
Difficulty	Medium (to identify one) High (to identify all)	Medium (to identify one) High (to identify all)	High	High
Variation	High	High	Medium	Medium
Examples	**English -> Chinese**	**Chinese-> English**	**English -> Chinese**	**Chinese -> English**
	Edison → 爱迪生 / 埃迪森 / 埃蒂森	姚明 → Yao Ming	Lijun → 力骏 / 丽君 / 莉军	爱迪生 / 埃迪森 / 埃蒂森 → Edison (Eddison)
Difficulty	High (to identify one) High (to identify all)	Low	High (even impossible for human)	High
Variation	High	Low	High	Medium
Previous Research	**Arabic->English** Arbabi et al. (1994) **English->Arabic** AbdulJaleel & Larkey (2002) Darwish et al. (2001) Al-Onaizan & Knight (2002) **English->Chinese** Wan & Verspoor (1998) Virga & Khudanpur (2003)		**Arabic->English** Stalls & Knight (1998) **Thai->English** Kawtrakul et al. (1998) **Japanese->English** Knight & Graehl (1997) Goto et al. (2001) **Chinese->English** Lin & Chen (2002)	

2.3 User Name Searching Experience and Web 2.0 Technologies

When searching for a person's name, there are typical two scenarios: the user is familiar with that person and they know what to explore, or they are unfamiliar with the person so they just want to learn the basics of the person (Lee et al, 2005). The first scenario is called navigational and the second is called informational. In either scenario, finding the right person at a first try is not easy in a monolingual setting. It is even more difficult in a multilingual world where one name can generate multiple translations and even experts cannot identify all correct translations.

Web2.0 technologies allow for an easier distributed collaboration and information sharing. The idea is individual contributor can surpass experts. Many e-commerce websites have implemented Web2.0 to allow richer user experience (O'Reilly, 2005). In e-commerce applications where customers find it difficult to search for desired products among enormous offerings, Web 2.0 technologies are used to alleviate this problem by allowing users to provide their feedback. Recommendation system is another form that has been widely adopted to help customers locate products. For example, Netflix allows users to rank each movie and then make recommendations according to this rank and users' preference. Because name search can be triggered by major events, we can imagine users may share similar search goals. Thus, it is important to incorporate user feedback during a name search.

3 Research Questions

Based on our review, several research needs and gaps have been identified. *First*, name translation is an important component in multilingual Web search and has many application areas. It is a complex problem because they do not appear in lexicon and are of high variability. *Second*, most previous studies have focused on either direct name translation approach, often a statistical approach, or a mining-based indirect translation approach. Although some pioneer works have tried to incorporate both, this area is still not fully explored. *Third*, most previous research has focused on translation algorithms, but little has been investigated in developing a practical real-time name translation and search system. The problem is not only translating the names, but using the translated name to find the relevant person and the right information. *Fourth*, with Web 2.0 technologies more widely available in e-commerce applications such as movie review and customer feedbacks that allows individual contributor to have a collective impact on the overall product ranking, similar techniques are desired in a name translation system that might support users to better find the person they are searching for. Thus, we propose the following research questions.

1. How can we build a name translation system that incorporates direct and indirect translation approach?

2. Does such system demonstrate superior performance over a pure statistical-based transliteration system?

3. What does the translation variation distribution look like? How does this distribution guide our system design?

We address these questions in the next sections.

4 Proposed Framework: An Interactive Name Search System

4.1 Challenges with Chinese Language

We chose to work on English-Chinese translation because this language pair is unique and post special challenges as alphabet and character-based language pair. First, written Chinese is a logogram language. Thus, a phonetic representation of Chinese characters, Pinyin, is used as an intermediate Romanization. Our process of translating an English name into Chinese consists of two steps: translating English word into Pinyin and then mapping Pinyin into Chinese characters. Second, Chinese is not only monosyllabic, but the pronunciation of each Chinese character is always composed of one (or none) Consonant unit and one Vowel unit with the Consonant always appears at the beginning. For example, /EKS/ is one syllable in English but is three syllables in Chinese (/E/ + /KE/ + /SI/). English syllables need to be processed in a way that can be mapped to Chinese Pinyin..

4.2 Proposed N-gram Hidden Markov Model

We describe a name translation framework as shown in Figure 1. The framework consists of four major modules: 1) Training, 2) Hidden Markov Model-based Transliteration, 3) Web Mining enhanced ranking and 4) User feedback models. We will explain each module in this section.

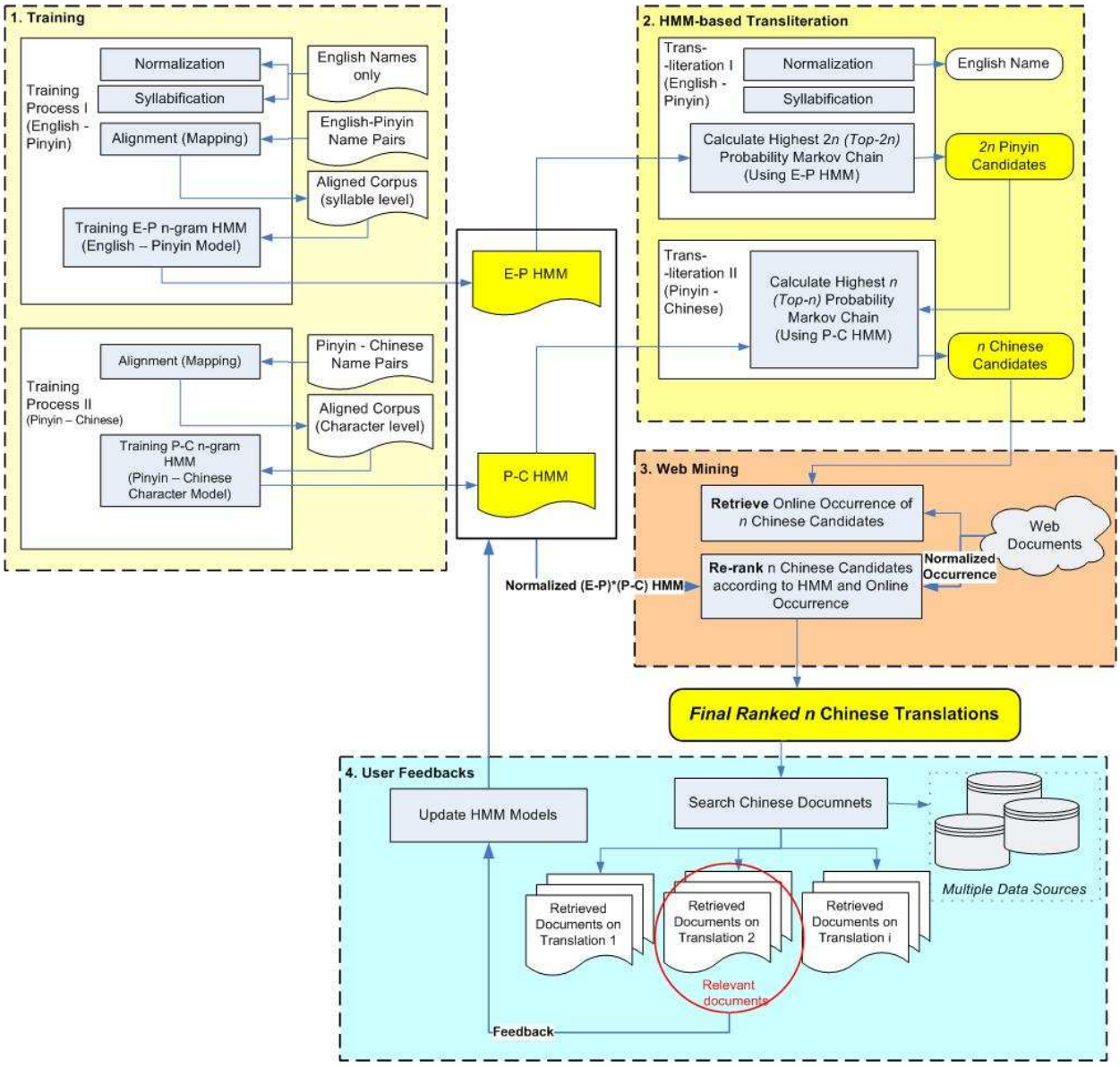


Figure 1: An Interactive Proper Name Translation System

4.2.1 Module 1: Training

The training process (Figure 1 Module 1) generates two transliteration probability tables based on a training corpus of English-Pinyin pair and Pinyin-Chinese name pairs. These two probability tables are then used in HMM transliteration.

Chinese Pinyin has 409 legal symbols. Each of them is represented by one or more Roman letters. In order to perform mapping from English names to Chinese Pinyin, an English name is divided into sub-syllables and this process is called **Syllabification**. Although many English syllabification algorithms have been proposed, they need to be adjusted. During syllabification, light vowels are inserted between two continuous consonants and silent letters are deleted. We use a finite state machine to implement the syllabification process. For example, "Phelps" becomes {/ph/ /e/ /l/ /@/ /p/ /@/ /s/ /@/} with "@" being inserted light vowels. **Alignment** process maps each sub-syllable in an English name to target Pinyin. The accuracy of alignment process largely depends on the accuracy of Syllabification.

Pinyin to Chinese character alignment is more straightforward where each Pinyin syllable (consonant + vowel) is mapped to the corresponding Chinese character. Once the alignment is done, occurrence of each translation pair in the training data can be calculated. Using this occurrence information, we can derive probabilities under various situations to support probability models. The output of training module: two probability tables of English-Pinyin and Pinyin-Chinese are used as input of HMM.

4.2.2 Module 2: A Two-step N-gram Hidden Markov Model

To address the out-of-vocabulary problem, we adopted a direct name transliteration model. In the translation phase, a grapheme-based Hidden Markov Model (HMM) is used to obtain probability distribution as shown in Figure 1 Module 2. The smallest unit in a grapheme approach is a letter, while the smallest unit in a phoneme approach is a syllable. A grapheme-based approach is used instead of a phoneme-based approach for easier portability to other languages since it is easier to port to other language pairs and no phonetic dictionary is needed here. Although we do not keep the entire syllable, phonetic integrity is captured in Hidden Markov Chain. Most previous research used either a rule-based or a simple statistical approach with independent probability estimation which lose the phonetic context. Correct transliteration is dependent on both languages' phonetic sequences. Hidden Markov Model is one of the most popular probability models and has been used in speech recognition, the human genome project, consumer decision modeling, etc. (Rabiner, 1989). Traditional HMM can be viewed as a bigram model where the current mapping selection depends on the previous mapping pair. We expand it to an hybrid N-gram model and use the combination of unigram, bigram, and trigram HMM. The goal of our model is to find the candidate transliteration with the highest transliteration probabilities:

(1) $\arg\max P(t \mid s) = \arg\max P(t_1 t_2 ..t_n \mid s_1 s_2 ....s_m)$

Where s is the source name to be transliterated, which contains letter string $s_1 s_2 ... s_i$; t is the target name, which contains letter string $t_1 t_2 ... t_i$.

In a simple statistical model, or a **unigram** model, transliteration probability is estimated as:

(2) $P(t_1, t_2, t_3, ..., t_n \mid s_1, s_2, s_3, ..., s_n) = P(t_1 \mid s_1) P(t_2 \mid s_2) ..... P(t_n \mid s_n)$

Where $$P(t_i \mid s_i) = \frac{\#of\ times\ s_i\ translates\ to\ t_i\ in\ corpus}{\#of\ times\ s_i\ appears\ in\ corpus}$$

The **bigram** HMM improves the simple statistical model in that it incorporates context information into a probability calculation. The transliteration of the current letter is dependent on the transliteration of **ONE** previous letter (one previous state in HMM). Transliteration probability is estimated as:

(3) $$P(t_1, t_2, t_3, ..., t_n \mid s_1, s_2, s_3, ..., s_n)$$
$$= P(t_1 \mid s_1) P(t_2 \mid s_2, t_1)(t_3 \mid s_3, t_2) ..... p(t_n \mid s_n, t_{n-1})$$

Where $$P(t_i \mid s_i) = \frac{\#of\ times\ s_i\ translates\ \ to\ t_i}{\#of\ times\ s_i\ occurs}$$

and

$$P(t_i \mid s_i, t_{i-1}) = \frac{\#\ of\ times\ s_i\ translates\ to\ t_i\ given\ s_{i-1} -> t_{i-1}}{\#\ of\ times\ s_{i-1}\ translates\ to\ t_{i-1}}$$

The **trigram** HMM intends to capture even more context information by translating the current letter dependent on the **TWO** previous letters. Transliteration probability is estimated as:

(4) $$P(t_1, t_2, t_3, ..., t_n \mid s_1, s_2, s_3, ..., s_n)$$
$$= P(t_1 \mid s_1) p(t_2 \mid s_2, t_1) P(t_3 \mid s_3, t_2, t_1) ..... p(t_n \mid s_n, t_{n-1}, t_{n-2})$$

Where

$$P(t_i \mid s_i) = \frac{\#\ of\ times\ \ s_i\ translates\ \ \ to\ t_i}{\#\ of\ times\ \ s_i\ occurs}$$

$$P(t_i \mid s_i, t_{i-1}) = \frac{\#\ of\ times\ s_i\ translates\ to\ t_i\ given\ \ s_{i-1} -> t_{i-1}}{\#\ of\ times\ s_{i-1}\ translates\ to\ t_{i-1}}$$

and $$P(t_i \mid s_i, t_{i-1}, t_{i-2}) = \frac{\#\ of\ times\ \ s_i\ translates\ to\ t_i\ given\ \ s_{i-1} -> t_{i-1}\ and\ s_{i-2} -> t_{i-2}}{\#\ of\ times\ s_{i-1}\ translates\ \ to\ \ t_{i-1} and\ s_{i-2}\ translates\ \ to\ \ t_{i-2}}$$

Finally, we also proposed to use a hybrid probability. It is a weighted combination of all N-grams:

$$Final\ Transliter\ ation\ Score\ = \alpha_1(Unigram\ \ HMM\ ) + \alpha_2(Bigram\ \ HMM\ ) + \alpha_3(Trigram\ \ HMM\ )$$

we applied α_1=1, α_2=2, α_3=3 such that longer matched sequence has a larger contribution in the final probability. The rationale is that the longer the prior sequence identified in training data, the higher probability that the translation sequence is the correct tone. These α parameters can be tuned in the future. We call this approach hybrid HMM. The same process is conducted for Pinyin to Chinese character translation as shown in the lower part of Figure 1 Module 2.

Once the HMM Model is obtained, new incoming name is translated by obtaining a letter sequence that maximizes the overall probability through the HMM. This step uses a modified Viterbi's search algorithm (Viterbi, 1967). The original Viterbi's algorithm only keeps the most optimal path, in other words, only generates one translation. To cope with name translation variations, we keep the top-2n optimal paths for further analysis, n being the final number of desired translation candidates.

4.2.3 Module 3: Web Mining

To boost the transliteration performance we propose to use the Web mining approach, which Analyzes Candidates' Occurrence on the Web. Each one of the top-n transliterations obtained from the previous step is sent to a Web search engine using a meta-search program which records the number of documents retrieved, referred to as Web frequency. By examining the popularity of all possible transliterations on the Internet, bad transliterations can be filtered and their online popularity can serve as an indicator of transliteration correctness. The popularity is estimated by acquiring the number of documents returned from a search engine using the translation candidate as query. We used occurrence count of target translation instead of co-occurrence information of both source and target translation, because frequency of co-occurrence will be low for unpopular names. This final rank of transliterations is derived from a *weighted* score of the *normalized* Web frequency and the probability score.

$$Final\ score = \alpha * normalized\ probability\ score + \beta * normalized\ Web\ frequency,$$

$$\text{s.t. } \alpha + \beta = 1.$$

On the one hand, even though we are using Web mining for disambiguation, we do not want to treat all the top-N transliteration candidates equally. Instead, we retain information from the probability model. In this way, if two transliterations have a similar Web frequency score (e.g. 128,000 vs. 128,001) their probability scores will play a major role in selecting the best transliteration. On the other hand, we still want to distinguish between different Web frequency counts if the difference is big enough. In transliteration the occurrence difference between 128,000 and 1 should have a much bigger effect than the difference between 128,000 and 127,000, in which case the Web frequency score will play a more important role in the final ranking score.

In our framework, we used linear normalization. During each transliteration, Web frequency and probability score for each candidate were divided by the highest ones achieved among all candidates. We chose α=0.5and ß=0.5 to generate the final score. This setting gave same weights to the probability model and the Web mining model. Other settings of α and ß were not tested in this work. We have interest in testing the effect of different α and ß settings in the future. All the transliteration candidates are then ranked by their final scores.

4.2.4 Module 4: User Feedback

Previous research showed that the larger the training dataset (or the corpus), the higher the accuracy. However, the initial training dataset is often not complete because of limited availability. Name translation changes over time and varies in regions. On the other hand, a translation can become popular and dominant in a short time if the name has high coverage in media. This dynamic development of translation can hardly be captured in direct transliteration, because training dataset are determined in advance.

With the fast development of Web 2.0 which encourages more interactive user experience, we propose an interactive user feedback module which allows real-time updates to the original training dataset. The idea is similar to Netflix movie rating and Amazon's product ranking where user generated ranking data is added to the backend database. Every time a translation is rated as a "good" translation, that translation is added to our training data and the entire probability model is re-calculated. If that translation pair already exists in our training data, the entire translation sequence will be reinforced by increasing the occurrence frequency by 1 in the training data. This model addresses the limitation of training dataset and dynamics of translation.

5 Experiments

In order to answer our three research questions, we conducted three experiments. Experiment 1 studied the overall system performance using a standard top-n translation accuracy measure. Experiment 2 looked at variations in translation and Experiment 3 shows a sample user experience of NameTran system and demonstrates how user feedback can be captured.

5.1 Experiment 1: Top-n Translation Accuracy

5.1.1 Measures

Traditional transliteration studies have used "accuracy" to measure performance which is defined as:

$$\text{Accuracy} = \frac{\textit{Number of Correct Transliterations}}{\textit{Total Number of Transliterations}}$$

However, it is in general difficult to judge correct transliteration during forward transliteration, which is a one-to-many mapping. One English name could have more than five Chinese transliterations that are acceptable to human. In this case, there are two ways to define "correct transliterations" (Al-Onaizan and Knight, 2002). For evaluation purpose, we use the notion of *rigid accuracy*, assuming that there is one and only one correct transliteration. A transliteration is considered correct only if it matches the pre-defined gold-standard. Gold-standard can be translations extracted from a dictionary. Or it can be judged by human experts as the most widely used or most acceptable transliteration. On the other hand, in a *relaxed accuracy*, a transliteration is considered correct if it is acceptable to human experts. A relaxed accuracy allows multiple correct answers. However, it is often difficult to generate all the possible translations in advance. Relaxed accuracy can be done by judging the machine-generated

transliterations afterwards. For example, President Clinton is transliterated as "克林顿" in most Chinese news articles, which is considered a gold-standard transliteration. The transliteration system could generate a different name "柯林顿", which has the same pronunciation as the gold-standard, and is actually used in many contexts as well, yet not as dominant. Our expert could decide that it was acceptable. In a rigid accuracy judgment, the machine-generated "柯林顿" will not be considered as a correct transliteration, while in a relaxed accuracy it will be treated as a correct one.

Since rigid accuracy is a more stringent test which avoids human judgment variations, we chose to use rigid accuracy for our measure. Note that a relaxed accuracy is always higher than a rigid accuracy, and a rigid accuracy can be viewed as the worst case scenario of system accuracy.

Besides measuring the accuracy for the highest ranked transliteration, identifying a set of top-n transliteration candidates are of interest. Top *n* accuracy is defined as the percentage of names whose selected top *n* transliterations include correct transliterations. Among all the ranked transliteration candidates, top-n accuracy is defined as

$$\text{Top-n Accuracy} = \frac{\textit{Number of Times Correct Transliterations appeared in the first n Candidates}}{\textit{Total Number of Transliterations Performed}}$$

To stay consistent with previous research, we chose to examine n=1, 2, 4, 8.

5.1.2 Dataset and Methodology

Our English-Chinese dataset is a list of 2000 unique English names and their transliterations extracted from a bilingual dictionary. These dictionary-provided translation are considered to be gold-standard in our experiments. Both datasets are unaligned.

We used the 10-fold cross validation method to test system accuracy. 10-fold cross validation is a most common method used in testing data mining algorithms and models. We first divided the data into 10 subsets of equal size randomly. We trained the model 10 times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the system top-n accuracy. Accuracy scores obtained from each subset were then averaged.

5.1.3 Results

Table 3 summarizes average accuracy achieved and paired t-test results of comparing probability model alone and a combined probability and Web mining model. The best performance was achieved using a combined hybrid HMM and Web mining model (column 9), a 0.64 top-1 accuracy and a 0.96 top-8 accuracy. Bigram HMM, Trigram HMM, hybrid HMM, and a combined hybrid HMM and Web mining model enhanced a simple statistical approach by 244.08%, 325.81%, 414.24% and 415.01% respectively for top-1 accuracy. Improvement in top-2, top-4 and top-8 categories were not as tremendous as that of top-1, yielded from 69.58% (top-8 accuracy for bigram) to 142.43% (top-8 accuracy for Web mining enhanced).

Figure 2 illustrates the improvements obtained from Web mining model. By combining Web mining with three direct translation approaches: simple statistical method (or a unigram model), bigram HMM model and Trigram HMM model, accuracy performances are significantly improved over using direct transliteration approach alone. It is confirmed that Web mining will boosted direct transliteration approach.

5.2 Experiment 2: Translation Variation Distributions

Our second experiment aims to study the translation variation distributions. There are often multiple translations for a given name, and some of them might be more dominant than others, although this popularity might change over time. We capture three distributions from our 2000 English name translations shown in Figure 4.

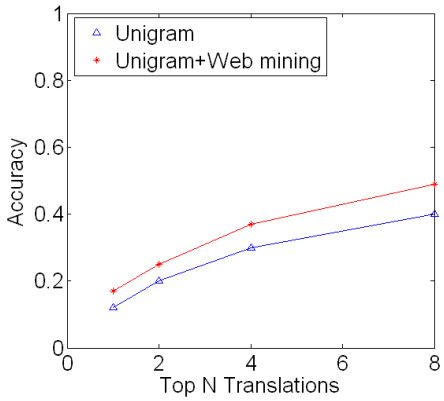
The first diagram (Figure 3 top left) shows the distribution of number of correct Chinese translations from English names. For each English name, we first generated 16 candidate translations from Hidden Markov Model. We then automatically searched for web documents using that candidate translation as a keyword. We put double quotation marks around the translation such that only web pages containing exactly the same translation were returned. As long as the candidate translation retrieved results, that translation was considered a valid translation. A human expert than went through the results to make sure that no odd translations were included. We examined a total of 2000 English names and found that number of valid translations is quite evenly distributed from 0 to 16.

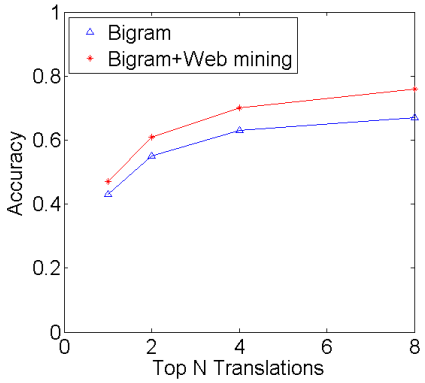
We then studied the distribution of dominant translations shown in Figure 3 top right. We used number of web documents retrieved from each translation as an indicator of popularity. We first ranked all translation candidates by number of documents retrieved, and then applied two criteria: First, if a translation retrieved more than 10,000 web pages, that translation was considered to be a dominant translation. That means if an English name had a total of 3 translations that all retrieved more than 10,000 web pages, it was considered to have 3 dominant translations. Second, if the web pages retrieved by top-n popular translations accounted for 90% of all documents retrieved by all 16 translation candidates, those n translations were considered dominant translations. This distribution is skewed where most English names have less than 5 dominant translations that people actually use. 32% of all names have one dominant translation, 24% has 2 dominant translations and 15% has 3 dominant translations.

The bottom diagram in Figure 3 shows aggregated number of Web pages returned from all name translations of an English name. From this diagram we can see that English name translations either have a high level of online presence (18% returned more than 100,000 web pages) or have very low online frequency (50% returned less than 1000 web pages). The frequently occurred online names are most likely famous people who often appear in news articles.

Table 3: Summary of average accuracy achieved and *t*-test results

	Simple (Unigram)	***Unigram +Web mining***	**Bigram**	***Bigram +Web mining***	**Trigram**	***Trigram +Web mining***	**Hybrid**	***Hybrid +Web mining***
Top 1	0.12	0.17	0.43	0.47	0.53	0.58	0.64	0.64
Top 2	0.20	0.25	0.55	0.61	0.65	0.71	0.75	0.80
Top 4	0.30	0.37	0.63	0.70	0.73	0.81	0.81	0.91
Top 8	0.40	0.49	0.67	0.76	0.77	0.86	0.85	0.96
Paired *t*-test (2 tail, α=0.05)								
P value		2.29E-15		1.38E-15		1.41E-16		4.37E-10





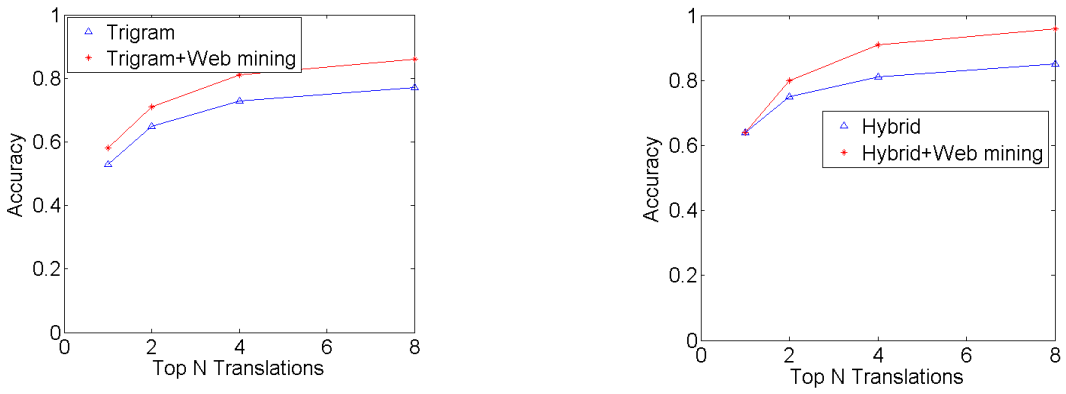


Figure 2: Performance comparison of combined probability and Web mining models (accuracy)

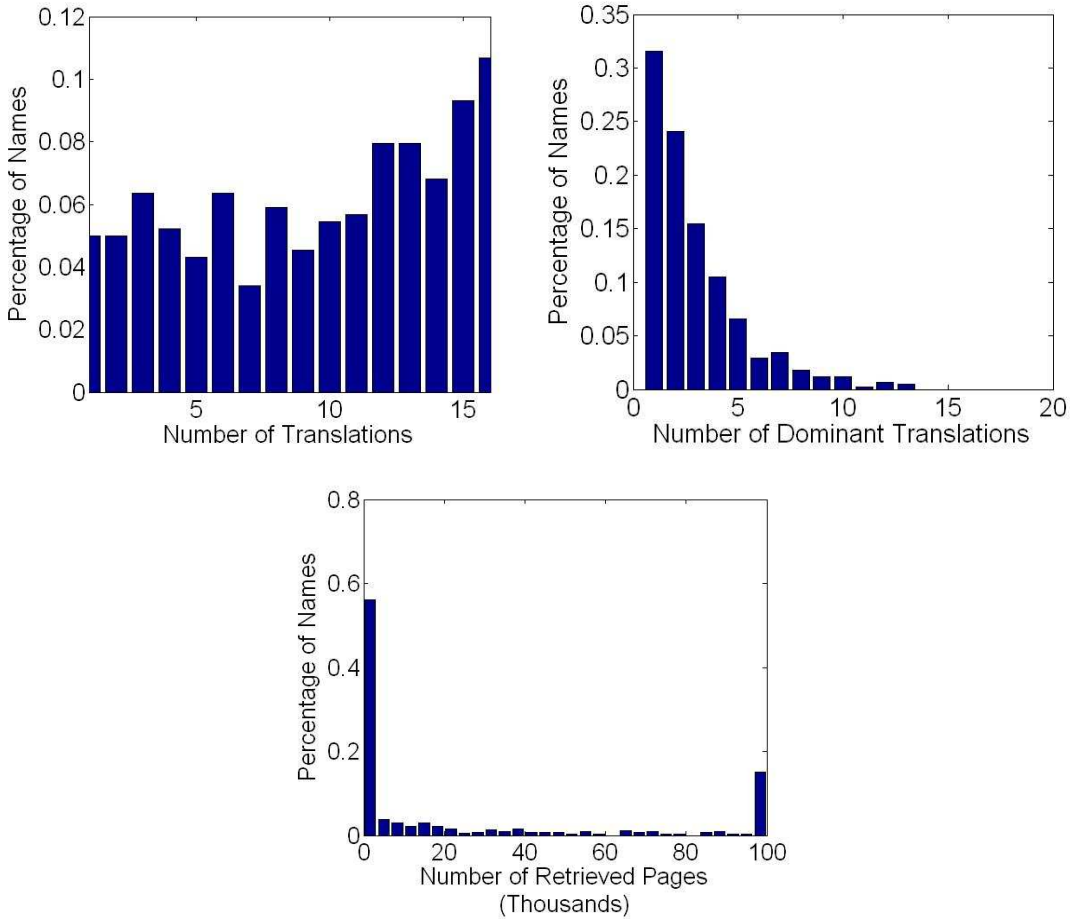


Figure 3: Translation Variation Distributions

5.3 Experiment 3: User Experience with an Interactive Name Search System

Based on our proposed approach, we developed a name search and translation system: *NameTran*. Figure 3 shows the interactive user interface and a sample search result. A user is interested in finding Chinese news on "Phelps" and he enters "Phelps" in search box and hit "Search." It is passed to the system for name translation and eight top ranked translations are displayed on the left with a star rating system. The star rating is based on HMM and Web mining model which indicates the likelihood of being the correct translation. User can switch among different translations by clicking on the eight translations. Once a translation is selected, documents retrieved using that translation is displayed on the right panel. In this example, the first translation "菲尔普斯" retrieves documents mostly about the famous swimmer and the second translation "费尔普斯" retrieves documents mostly about the Economics professor. There are some mixed results in both translations as well. After reading the documents and their translations, the user will be able to tell which translation is the one that he/she is looking for. He/She can choose to rate the translations according to the relevance of the retrieved documents. In the case where the user does not understand Chinese, popular search engines now offer "translation" option where the entire retrieved document can be translated into English. Although such document translation is not perfect, it provides a gist of information which helps the user understand the text.

A user can also provide feedback to the system by clicking on the star rating next to the translations. The rationale is similar to what is adopted in Amazon and Netflix user rating. If a translation is rated by a user as 4 or 5 stars, that particular translation will be added to the training data and all the corresponding translation probability will be updated in the backend system. For 4 stars, each translation probability is discounted by half why applying formula (4). And for 5 stars, the entire translation is used. Ideally each user could have their own profile and the system can capture users' individual preference which we hope to address in our future research.

Such interactive system addresses couple challenges in name translation. First, it compensates the problem of lacking comprehensive training data and linguistic resources. Secondly, it also provides more timely and popular translation results with users' feedback. Besides, unlike most translation systems that provide only one correct translation, multiple translations of the same name are kept and presented.

6 Conclusions and Future Directions

There is a need for searching and understanding multilingual text to supporting e-business in a global market. There are now over 1 billion people whose native language is not English out of the 1.4 billion Internet users (www.internetworldstats.com). Moreover, non-English-speaking Internet population is growing faster than English population, which predicts a faster grow of non-English content.

In this research, we review challenges in name translation and propose an interactive name translation and search system: NameTran. This system takes English names and translates them into Chinese using a hybrid Hidden Markov Model-based (HMM-based) transliteration approach. To handle special challenges with alphabet-based and character-based language pair, we apply a two-phase transliteration model by building two HMM models, one between English and Chinese Pinyin and another between Chinese Pinyin and Chinese characters. To cope with changing popularity and variation in name translations, NameTran displays top-8 translation candidates along with Web search results from each candidate. With machine translation system available with major search engines, a user can select the correct translation by reading the translated foreign documents. A user can also rank translations and the new ranking serves as feedback to the original trained HMM model. We conducted experiments to study the system performance and concluded that a combination of probability and web mining model achieved the highest accuracy. We also studied the distribution of translation variations and showed that although an English name can have many translation variations, for over 80% names there are less than 5 dominant translations. We believe that by tracking user search behavior and capturing user ranking system performance can be improved.

In the future, we plan to conduct a comprehensive user evaluation to understand the practical use of such Name Search systems. We also plan to study the phenomenon that each translation could represent several different persons and different translations can represent the same person. An initial plan is to use document clustering to identify different entities. We also plan to incorporate user profiles in the name search system such that user feedback can be used more effectively.

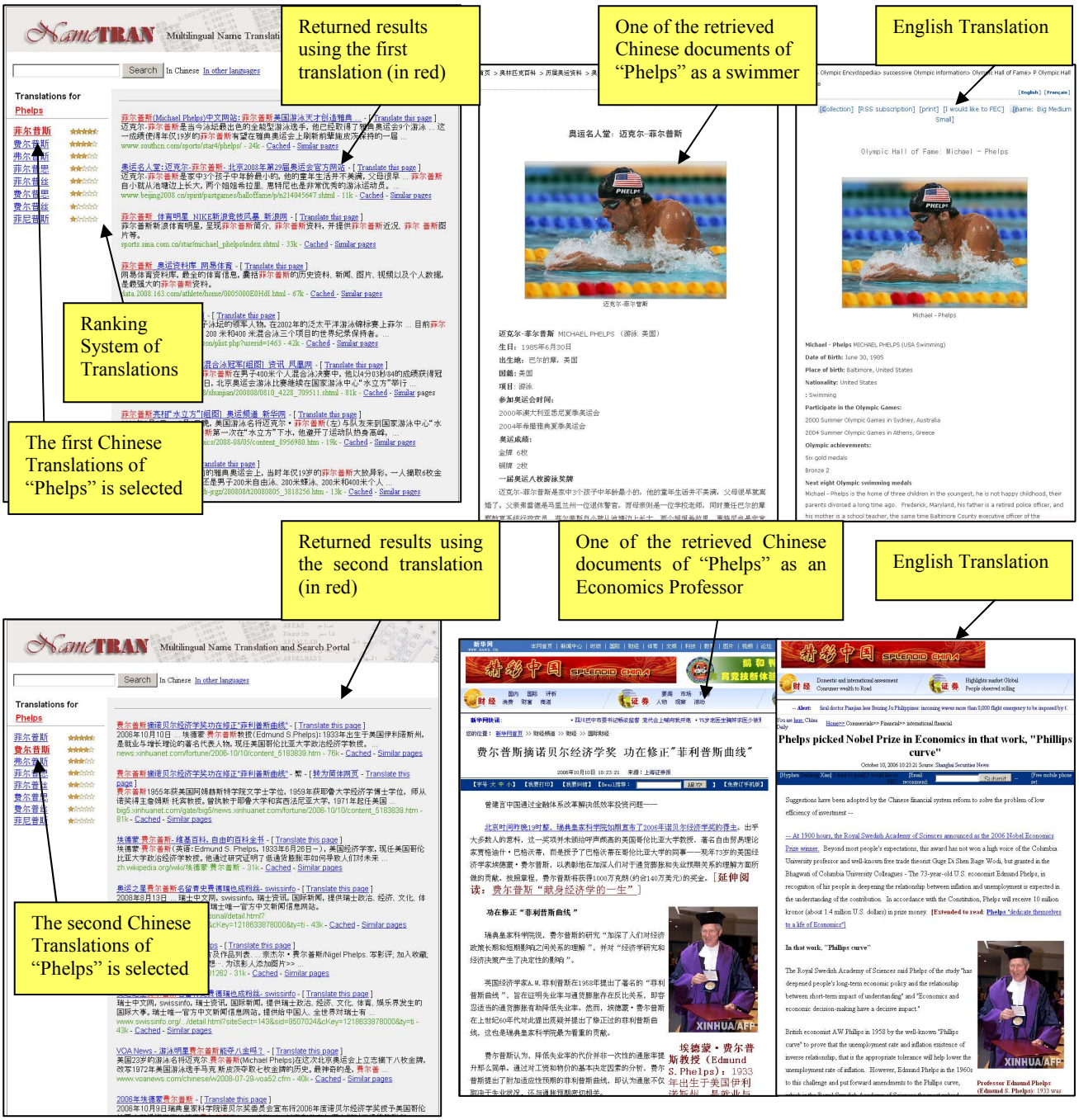


Figure 4: Sample User Session of NameTran

References

AbdulJaleel, N., and Larkey, L. S., Statistical transliteration for English-Arabic Cross Language Information Retrieval, in *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM)* New Orleans, LA, pp. 139 (2003).

Al-Onaizan, Y., and Knight, K., Machine Transliteration of Names in Arabic Text, in *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages* Philadelphia, Pennsylvania pp. 1 (2002).

Ankolekar, A., Krotzsch, M., Tran, T., Vrandecic, D., "The Two Cultures: Mashing up Web 2.0 and the Semantic Web," in *Proceedings of the International World Wide Web Conference,* Banff, Alberta, Canada, May 8-12, 2007, pp. 825-834.

Arbabi, M., Fischthal, S. M., Cheng, V. C., and Bart, E., Algorithms for Arabic Name Transliteration, *IBM Journal of Research and Development*, 38, 183 (1994).

Cao, Y., and Li, H., Base Noun Phrase Translation Using Web Data and teh EM Algorithm, in *COLING 2002*, pp. 127 (2002).

Chen, H. and Chau, M., "Web Mining: Machine Learning for Web Applications," Annual Review of Information Science and Technology (ARIST), 38, 289-329, 2004.

Chen, Y., and Zong, C., A Structure-based Model for Chinese Organization Name Translation, *ACM Transactions on Asian Language Information Processing*, 7, 1 (2008).

Darwish, K., Doermann, D., Jones, R., Oard, D., and Rautiainen, M., TREC-10 Experiments at University of Maryland CLIR and Video in *Text REtrieval Conference*, Gaithersburg, Maryland (2001).

Goto, I., Uratani, N., and Ehara, T., Cross-Language Information Retrieval of Proper Nouns using Context Information, in *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, pp. 571 (2001).

Kawtrakul, A., Deemagarn, A., Thumkanon, C., and Khantonthong, N., Backward Transliteration for Thai Document Retrieval, in *Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAD)*, Chiangmai, Thailand, pp. 563 (1998).

Knight, K., and Graehl, J., Machine Transliteration in *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Somerset, New Jersey, pp. 128 (1997).

Lee, C.-J., Chang, J. S., and Jang, J.-S. R., Alignment of Bilingual Named Entities in Parallel Corpora Using Statistical Models and Multiple Knowledge Sources, *ACM Transactions on Asian Language Information Processing*, 5, 121 (2006).

Lee, U., Liu, Z., Cho, J., "Automatic Identification of User Goals in Web Search," in *Proceedings of the International World Wide Web Conference*, Chiba, Japan, May 10-14, 2005, pp. 391-400.

Lin, W.-H., and Chen, H.-H., Backward Machine Transliteration by Learning Phonetic Similarity in *Proceedings of The 6th Workshop on Computational Language Learning* (CoNLL-2002), Taipei, Taiwan, pp. 139 (2002).

Lu, W.-H., Chien, L.-F., and Lee, H.-J., Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach, *ACM Transactions on Information Systems* (TOIS), 22, 242 (2004).

Oard, D., Cross-language Text Retrieval Research in the USA, in *The 3rd ERCIM DELOS Workshop*, Zurich, Switzerland (1997).

O'Reilly, 2005, "What is Web 2.0," URL http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (found 3.5.2006).

Rabiner, L. R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, 77, 257–286 (1989).

Sproat, R., Tao, T., Zhai, C. X., "Named Entity Transliteration with Comparable Corpora," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 73-80.

Stalls, B. G., and Knight, K., Translating Names and Technical Terms in Arabic Text, in *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, Montreal, Quebec, Canada (1998).

Thompson, P., and Dozier, C. C., Name Searching and Information Retrieval, in *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island (1997).

Virga, P., and Khudanpur, S., Transliteration of Proper Names in Cross-Lingual Information Retrieval, in *Proceedings of the ACL Workshop on Multi-lingual Named Entity Recognition Sapporo*, Japan, pp. 57 (2003).

Viterbi, A. J., Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm, *IEEE Transactions on Information Theory*, 13, 260 (1967).

Wan, S., and Verspoor, C. M., Automatic English-Chinese Name Transliteration for Development of Multilingual Resources, in *Proceedings of the 17th international conference on Computational linguistics*, Montreal, Quebec, Canada pp. 1352 (1998).

Zhang, Y., Huang, F., and Vogel, S., Mining Translations of OOV Terms from the Web through Cross-lingual Query Expansion, in *SIGIR'05*, Salvador, Brazil, pp. 669 (2005).

Zhou, Y., Huang, F., and Chen, H., Combining probability Models and Web Mining Models: A Framework for Proper Name transliteration, *Information Technology and Management*, 9, 91 (2008).