Maximum N-gram HMM-based Name Transliteration: Experiment in NEWS 2009 on English-Chinese Corpus

Yilu Zhou
George Washington University
yzhou@gwu.edu

Abstract

We propose an English-Chinese name transliteration system using a maximum N-gram Hidden Markov Model. 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. Our model improves traditional HMM by assigning the longest prior translation sequence of syllables the largest weight. In our non-standard runs, we use a Web-mining module to boost the performance by adding online popularity information of candidate translations. The entire model does not rely on any dictionaries and the probability tables are derived merely from training corpus. In participation of NEWS 2009 experiment, our model achieved 0.462 Top-1 accuracy and 0.764 Mean F-score.

1 Introduction

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 ways to translate this name: "费尔普斯," "菲尔普斯," "弗尔普斯," "菲尔普思," and "菲尔普丝," with some more popular than others.

The variability in translation implies the complexity in name translation that can hardly be addressed in typical machine translation systems. Machine translation systems are often black boxes where only one translation is provided, which do not offer a solution to variability issue. The accuracy of a machine translation system, whether statistical or example-based, largely depends on sentence context information. This context information is often not available with name translation. Furthermore, emerging names are difficult to capture in regular machine translation systems if they have not been included in training corpus or translation dictionary. Thus, being able to translate proper names not only has its own application area, it will also enhance the performance of current machine translation systems.

In our previous English-Arabic name transliteration work (Zhou et al., 2008), we proposed a framework for name transliteration using a 2-gram and a 3-gram Hidden Markov Model (HMM). In this research, we extend our 2-gram and 3-gram HMM to an N-gram HMM where N is the maximum number of prior translation mapping sequence that can be identified in the training corpus. In our non-standard runs, we also integrated a Web mining module. The rest of the paper is structured as follows. Section 2 reviews related work; Section 3 describes our algorithm; Section 4 discusses implementation and evaluation results are provided in Section 5. Section 6 concludes our work.

2 Related Work

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

The first 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. The second approach, direct translation, is often done by transliteration. 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.

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 128–131,
Suntec, Singapore, 7 August 2009. ©2009 ACL and AFNLP

Transliteration models can be categorized into rule-based 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). 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. 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).

3 Challenges with Chinese Language

There are several challenges in transliterating English names into Chinese. 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 Proposed Maximum N-gram HMM

Figure 1 illustrates our name translation framework. The framework consists of three major components: 1) Training, 2) Hidden Markov Model-based Transliteration, and 3) Web Mining-enhanced ranking.

4.1 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. Pinyin is not provided in the training corpus, but is easy to obtain from a Chinese Pinyin table.

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 can be calculated. Using this occurrence information, we can derive probabilities under various situations to support probability models.

We use the Hidden Markov Model which 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). In transliteration, traditional HMM can be viewed as a 2-gram model where the current mapping selection depends on the previous mapping pair. We expand it to an N-gram model and use the combination of 1-gram, 2-gram, ... , (N-1)-gram and N-gram HMM where N is the maximum number of mapping sequence that can be found in training corpus.

The goal of our model is to find the candidate transliteration with the highest transliteration probabilities:

(1) $$\operatorname{argmax} P(t \mid s) = \operatorname{argmax} P(t_1 t_2 . t_n \mid s_1 s_2 .. s_n)$$

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

In a simple statistical model, or a **1-gram** model, transliteration probability is estimated as:

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

Where

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

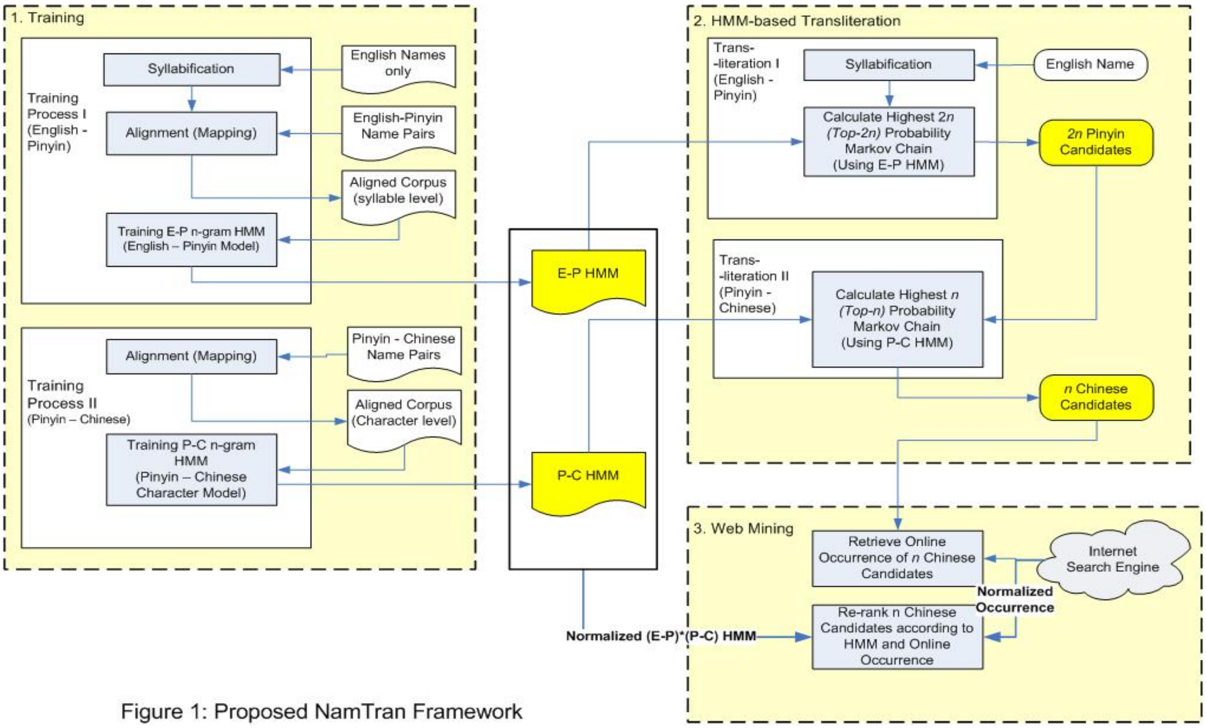
1. Training
Training Process I (English - Pinyin)
Syllabification
English Names only
Alignment (Mapping)
English-Pinyin Name Pairs
Aligned Corpus (syllable level)
Training E-P n-gram HMM (English – Pinyin Model)
Training Process II (Pinyin – Chinese)
Alignment (Mapping)
Pinyin - Chinese Name Pairs
Aligned Corpus (Character level)
Training P-C n-gram HMM (Pinyin – Chinese Character Model)
E-P HMM
P-C HMM
2. HMM-based Transliteration
Trans- -literation I (English - Pinyin)
Syllabification
English Name
Calculate Highest 2n (Top-2n) Probability Markov Chain (Using E-P HMM)
2n Pinyin Candidates
Trans- -literation II (Pinyin - Chinese)
Calculate Highest n (Top-n) Probability Markov Chain (Using P-C HMM)
n Chinese Candidates
3. Web Mining
Retrieve Online Occurrence of n Chinese Candidates
Internet Search Engine
Normalized Occurrence
Re-rank n Chinese Candidates according to HMM and Online Occurrence
Normalized (E-P)*(P-C) HMM

Figure 1: Proposed NamTran Framework

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)

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

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

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)

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

Where

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

$$P(t_i \mid s_i, t_{i-1}) = \frac{\#of\ timess_i\ translatestot_i\ givens_{i-1} -> t_{i-1}}{\#of\ timess_{i-1}\ translatestot_{i-1}}$$

and

$$P(t_i \mid s_i, t_{i-1}, t_{i-2}) =$$
$$\frac{\#\,of\ times\ \ s_3\ translates\ to\, t_i\ given\ \ s_{i-1} -> t_{i-1}\ and\ s_{i-2} -> t_{i-2}}{\#\,of\ times\ s_{i-1}\ translates\ \ to\ \ t_{i-1} and\ s_{i-2}\ translates\ \ to\ \ t_{i-2}}$$

This process is continued until the maximum mapping sequence is found in the transliteration corpus. The final probability estimation is a weighted combination of all N-grams:

$$FinalTranslitertationScore =$$
$$\alpha_1(1 - gramHMM) + \alpha_2(2 - gramHMM) + \ldots\ldots + \alpha_n(N - gramHMM)$$

In our submitted results, we applied α_1=1, α_2=2, …., α_n=N 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 **Maximum N-gram HMM**. The same process is conducted for Pinyin to Chinese character translation as shown in the lower part of Figure 1 Module 1.

4.2 Translation and Ranking

Once the two Maximum N-gram HMM Model are obtained, new incoming names are translated by obtaining a letter sequence that maximizes the overall probability through the HMM (Figure 1 Module 2). This step uses a modified Viterbi's search algorithm (Viterbi 1967). The original Viterbi's algorithm only keeps the most optimal path. To cope with name translation variations, we keep the top-20 optimal paths for further analysis.

4.3 Web Mining Component

To boost the transliteration performance we propose to use the Web mining approach, which analyzes candidates' occurrence on the Web

(Figure 1 Module 3). Each one of the top-20 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. The final rank of transliterations is derived from a weighted score of the normalized Web frequency and the probability score.

5 Evaluation

Named Entity Workshop (NEWS) 2009 Machine Transliteration Shared Task provided a training corpus with 31,961 pairs of English and Chinese name translations and 2,896 testing cases. We submitted one standard run with Maximum N-gram HMM *(N-HMM)* setting, and two non-standard runs with 3-gram HMM *(3-HMM)*, and Maximum N-gram HMM + Web mining *(N-HMM+W)*. There are two other runs that we submitted which contains error in the results and they are not discussed here. We present our evaluation results in Table 1.

	Top-1 Acc	F-score	MRR	MAP (Ref)	MAP (10)
N-HMM	0.456	0.763	0.587	0.456	0.185
N-HMM+W	**0.462**	**0.764**	0.564	**0.462**	0.175
3-HMM	0.458	0.763	**0.602**	0.458	**0.191**

Table 1: Evaluation Results with Top-10 Candidates

It is confirmed that Web-mining module boosted the performance of N-gram HMM in all measure except for $\text{MAP}_{(10)}$. However, the boosting effect is small (1.3%). To our surprise, 3-gram HMM outperformed Maximum N-gram HMM slightly (3% in $\text{MAP}_{(10)}$). Our best Top-1 accuracy is 0.462, and best Mean F-score is 0.764 both achieved by N-gram HMM with Web mining module. We believe this slightly lower performance of Maximum N-gram HMM can be improved with some tuning of weight parameters.

6 Conclusions

We propose an English-Chinese name transliteration system using a maximum N-gram Hidden Markov Model. 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. In participation of NEWS 2009 experiment, our model achieved 0.462 Top-1 accuracy and 0.764 Mean F-score. We plan to conduct further study the impact of Web mining component and find optimal set of parameters. Our model does not rely on any existing dictionary and the translation results are entirely based on learning the corpus data. In the future, this framework can be extended to other language pairs.

Acknowledgment

We thank the data source provider of this shared task from

English-Chinese (EnCh): Haizhou Li, Min Zhang, Jian Su, "A joint source channel model for machine transliteration", Proc. of the 42nd ACL, 2004

References

AbdulJaleel, N., and Larkey, L. S., Statistical transliteration for English-Arabic Cross Language Information Retrieval, in *Proceedings of (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).

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).

Chen, Y., and Zong, C., A Structure-based Model for Chinese Organization Name Translation, *ACM Transactions on ACL*, 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 *TREC*, Gaithersburg, Maryland (2001).

Lee, C.J., Chang, J. S., Jang, J.S.R, Extraction of transliteration pairs from parallel corpora using a statistical transliteration model, Information Sciences, 176(1), 67-90 (2006).

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

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

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).