Chinese-to-English Backward Machine Transliteration

Guo Yuqing, Wang Haifeng
Toshiba (China) Research and Development Center
5/F, Tower W2, Oriental Plaza, No.1, East Chang An Ave., Dong Cheng District
Beijing, 100738
{guoyuqing, wanghaifeng}@rdc.toshiba.com.cn

Abstract

It is challenging to transliterate named entities across languages. It is even more challenging to backward transliterate the transliterated term into its original form. This paper addresses the problem of backward translating person name from Chinese to its English counterpart. We propose a statistical backward transliteration method. Our method uses English sub-syllable and Chinese syllable as the basic units for the Chinese-English transliteration model and the English language model. Experiments show that our method is promising.

1 Introduction

Transliteration is the process of replacing words in source language with their approximate phonetic or spelling equivalents in target language. Commonly, transliteration is used to translate named entities across languages. Although the bilingual lexicons may be updated from time to time, new named entities out of the lexicons still appear frequently. Automatic transliteration is helpful for many applications, such as Machine Translation (MT), Cross Language Information Retrieval (CLIR) and Information Extraction (IE), etc.

There are two directions of transliteration: forward and backward. To forward transliterate an original word into target language, several candidates may be all valid. But it is less forgiving in the reverse direction. For example, the English word "Bianco" can be transliterated into "比安科 (BI AN KE)", "比安克 (BI AN KE)", and "比昂科 (BI ANG KE)" etc. But for backward transliteration, neither "Biank" nor "Byanco" would be accepted when we mean "Bianco". So backward transliteration is often more challenging.

In previous works, both directions of machine transliteration for several language pairs have been studied. Knight and Greahl (1998) described a generative model implemented by finite state transducers for backward transliteration from Japanese to English. This model was extended to Arabic to English backward transliteration by Stalls and Knight (1998). Jung et al. (2000) used an extending HMM to forward transliterate English to Korean. There is also some research on Chinese. For forward transliterate English into Chinese, Wan and Verspoor (1998) used some heuristic rules, Virga and Khudanpur (2003) employed the statistical MT techniques. For backward transliteration from Chinese to English, a similarity-based framework was proposed. Chen et al. (1998) used a manually assigned scoring matrix for measuring similarities between English and Chinese names on the grapheme level. Lin and Chen (2002) used a scoring matrix automatically learned from a training corpus to measure similarities on the phoneme level.

In the forementioned work on Chinese to English backward transliteration, the process of mate-matching was conducted, that is, a candidate list had to be provided by hand or by an extra named entity extraction system. In this paper, we present an automatic method to backward transliterate person names from Chinese to English without any candidate list. We propose a statistical transliteration model in which English sub-syllable and Chinese syllable are used as the basic units.

2 Chinese-to-English Backward Machine Transliteration

2.1 Transliteration System Description

Transliteration from Chinese to English is a rather difficult task due largely to the distinct differences

in their phonetics and alphabets. English is a language written by phonograms (words), whereas, Chinese language is written as a string of ideograms (characters). There is no straightforward way of mapping between English words and Chinese characters. A feasible solution is to represent the pronunciation of Chinese character by a process of romanization. Hanyu Pinyin is an official romanization system which is commonly used in Mainland China and also gains popularity in other countries and regions. We use Pinyin as a medium to connect Chinese characters with English words.

However, it is still difficult to find the obvious relationships between Chinese Pinyin and English word because English words are mostly multisyllabic but Chinese characters are mostly monosyllabic. For example, the English word "Christopher" has three syllables "Chris", "to" and "pher". On the Chinese side, the transliterated word "克里斯托弗" consists of five characters represented as "KE", "LI", "SI", "TUO" and "FU" by Pinyin, which each is a syllable. So we need to segment the English word into appropriate units to be a counterpart of Chinese Pinyin.

Then the backward transliteration process is broken down into various steps depicted as follows:

(1) Converting Chinese characters to Pinyins;
(2) Translating Pinyins to corresponding English sub-syllables;
(3) Combining English sub-syllables to word.

Steps (1) is almost deterministic, while step (2) and (3) are accomplished using statistical methods.

2.2 English Sub-syllable Segmentation

English syllable is the basic unit to segment English word. However, mapping between English syllable and Chinese Pinyin isn't a trivial work. As shown above, the English syllable "Chris" maps to three Chinese syllables "KE", "LI" and "SI". The reason is that English permits initial and final consonant clusters in syllable. Mandarin Chinese, in contrast, primarily has a syllable structure of initial part and final part, the initial part is a consonant or null, and the final part is a simple or compound vowel with a nasal consonant ending optionally. When English consonant cluster is transliterated into Chinese, it should be reduced to several single phonemes.

In this paper, we segment an English word into sub-syllables based on the Chinese transliteration of the English word. All English words in the training set are segmented into sub-syllables, which are mapped to Chinese Pinyins, as shown in Figure 1. Then the mapping list of English sub-syllable and Pinyin is built, which will be used in the transliteration process. Although the offline segmentation can be performed by hand, we employ a method combining rules and statistics for more efficient segmentation.

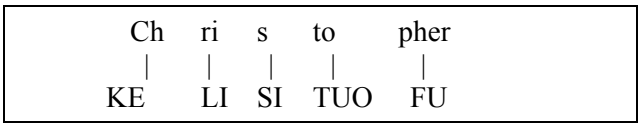

Figure 1. Mapping between English sub-syllable and Chinese Pinyin

2.3 Transliteration Model

Given a Chinese word C, for all English words E, we minimize our chance of error by choosing the English word $\hat{E}$ for which $P(E \mid C)$ is greatest. Using Bayes's theorem and dropping the constant denominator, we can equivalently perform the following maximization:

$$\hat{E} = \arg\max_{E} P(C \mid E)P(E) \tag{1}$$

In Equation (1), $P(C \mid E)$ is the Chinese-English transliteration model and $P(E)$ is the English language model. If Chinese word $C = (c_1, c_2 ..., c_n)$ has n Chinese characters, and English word $E = (e_1, e_2 ..., e_n)$ has n English sub-syllables, then

$$\begin{aligned} P(C \mid E) &= P(c_1, c_2, ..., c_n \mid e_1, e_2, ..., e_n) \\ &\approx \prod_{i=1}^{n} P(c_i \mid e_i) \end{aligned} \tag{2}$$

$$\begin{aligned} P(E) &= P(e_1, e_2, ..., e_n) \\ &\approx P(e_1) \prod_{i=2}^{n} P(e_i \mid e_{i-1}) \end{aligned} \tag{3}$$

As the limitation of the training data, the model may suffer from the sparseness problem. We use

interpolation for smoothing, which are expressed as follows:

$$P(c_i \mid e_i) = \alpha P(c_i \mid e_i) + (1-\alpha)P(c_i) \quad (4)$$

Where $0 \le \alpha \le 1$

$$P(e_i \mid e_{i-1}) = \beta P(e_i \mid e_{i-1}) + (1-\beta)P(e_i) \quad (5)$$

Where $0 \le \beta \le 1$

For a given Chinese word, e.g. "KE LI SI TUO FU", there will be more than 1,000,000,000 transliterated candidates, which are all the possible combination of sub-syllables, e.g. "ch/ri/s/to/ph/er", "ch/ri/s/tau/fer", "k/rei/th/too/fu" and so on. In practice, we use the n-best search algorithm (Schwartz and Chow, 1990) to choose the most possible transliterated words from all these candidates according to their probabilities described in Equation (1).

3 Experiments

3.1 Training Set and Test Set

We build the training set and test set from the LDC Chinese <-> English Named Entity List[1] of more than 500 thousand person name pairs.

In our experiments, the test set including 500 pairs is randomly chosen from the LDC bilingual corpus. The rest part of the LDC corpus is used as training set except Japanese names and some pairs with formatting errors. Table 2 shows the statistical information about the training set.

English-Chinese word pairs	424,788
Pinyin set (pieces)	405
Sub-syllable set (pieces)	2,387
Sub-syllable bigram pairs	1,174,187
Sub-syllable and Pinyin pairs	1,598,975

Table 2. Size of the training set

3.2 Evaluation

Besides the best answer of the transliterated word, our algorithm also gives other possible candidates sorted by their probabilities. Figure 3 shows an example of the list of the transliterated candidates for "阿特金斯 (A TE JIN SI)".

阿特金斯
Atkins,
Artkins,
Ateckins,
Arteckins,
Hartkins,
...

Figure 3. An example of the transliteration results

In order to evaluate the performance of transliteration results, we define the Precision of transliteration as in Equation (6).

$$Precision = \frac{\#CorrectTransWords}{\#AllWords} \times 100\% \quad (6)$$

Where $\#AllWords$ is the total number of words in the test set (500 pairs), and $\#CorrectTransWords$ is the number of the correct transliterated words occurring in the top n candidates. We choose top 1, 10, 50, 100, 200, and 500 as the evaluation points. Figure 4 shows the experiment results ($\alpha = 0.99$ and $\beta = 0.98$).

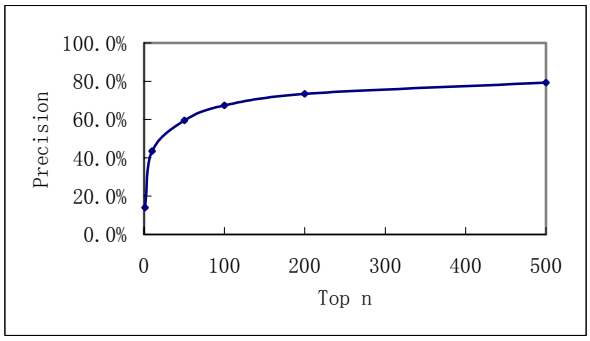

Figure 4. Experiments results of the test set

Satisfactory results (*Precision* = 79.2%) can be got when choosing more candidates (*n* = *500*). Then, we can re-score the top n candidates with other clues, such as the statistical information on Web (Al-Onaizan and Knight, 2002), to find the best transliterated word from the top n candidates with higher efficiency.

Experiments also show that our transliteration process is robust when the noise is produced in the forward transliteration.

For an English word "Bianco", "比安科" is frequently used in the forward transliteration. While "比昂科" is also used for some occasions,

[1] LDC Catalog number: LDC2003E01.

such as the articles on Web. The Pinyin of "比安科" is "BI AN KE" and that of "比昂科" is "BI ANG KE". The two Pinyins are different. Our system still can get the correct transliteration for both "比安科" and "比昂科".

比安科	比昂科
Bianco	Bienko
Bianko	Bianco
Byanco	Bianko
Biank	Bionko
Byanko	Byanco
…	…

Figure 5. An example with noise

When observing the error results, we found many of them are due to loan words from other languages. For example, "贝努瓦(BEI NU WA)" is transliterated from the original word "Benoit", which does not appear in the top 500 candidates from our system. The word "Benoit" comes from French with non-English phonetic features. It is quite hard for modeling this kind of feature. Even for the good human translators, it is also a challenge to give the right translation when the English pronunciation is not in a common way or when the pronunciation of the transliterated word varies greatly from the original one.

4 Conclusion and Discussion

We have proposed a statistical model to backward transliterate person names from Chinese to English. We use English sub-syllable and Chinese syllable as the basic units in the Chinese-English transliteration model and the English language model. Promising results of our experiments suggest our method will be helpful to some applications, such as MT, CLIR, IE, etc.

Further research will include a more sophisticated transliteration model allowing insertion and deletion, and a more powerful language model with larger context and better smoothing. Furthermore, we will analyze the performance of our approach under various training set and test set. We also plan to do more research on the noise robustness of our method. Finally, we will study how to adapt the method to real applications.

Acknowledgements

We would like to thank Dr. He Hongzhao for his contribution to this work.

References

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Linguistics*, 24(4): 599-612.

Bonnie G. Stalls and Kevin Knight. 1998. Translating names and technical terms in Arabic text, In *Proceedings of 17th COLING and 36th ACL Workshop on Computational Approaches to Semitic Languages*, pages 34-41, Montreal, Canada.

R. Schwartz and Y. L. Chow. 1990. The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, pages 81-84, Albuquerque, NM.

SungYoung Jung, SungLim Hong and Eunok Paek, 2000. An English to Korean transliteration model of extending markov window. In *Proceedings of 18th International Conference on Computational Linguistics*, pages 383-389, Saarbrucken, Germany.

Stephen Wan and Cornelia Maria Verspoor. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In *Proceedings of 17th COLING and 36th ACL*, pages 1352-1356. Montreal, Canada.

Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding and Shih-Chung Tsai. 1998. Proper name translation in cross-language information retrieval. In *Proceedings of 17th COLING and 36th ACL*, pages 232-236. Montreal, Canada.

Wei-Hao Lin and Hsin-Hsi Chen. 2002. Backward machine transliteration by learning phonetic similarity. In *Proceedings of 6th Conference on Natural Language Learning*, pages 139-145. Taipei, Taiwan.

Yaser Al-Onaizan and Kevin Knight. 2002. Translating name entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 400-408. Philadelphia, PA.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of Proper Names in Cross-Lingual Information Retrieval. In *Proceedings of 41st ACL Workshop on Multilingual and Mixed-language Named Entity Recognition*, pages 57-64. Sapporo, Japan.