Research for Tibetan-Chinese Name Transliteration Based on Multi-granularity

Chong Shao[1], Peng Sun[1], Xiaobing Zhao[1,2], and Zhijuan Wang[1,2]

[1] School of Information Engineering Minzu University of China,Beijing,China 100081
[2] Minority Languages Branch,National Language Resource and Monitoring Research Center
wangzj_muc@126.com

Abstract. In order to solve the problem of data sparseness caused by less training corpus in Tibetan-Chinese transliteration, this paper analyzes the alignment granularity of Tibetan-Chinese names as the research object and uses the pronunciation feature to reduce the corresponding relationships. The method of transliteration of Tibetan and Chinese names and the design of related experiments is comparable with traditional methods and improve the top-1 accuracy of transliteration of Tibetan and Chinese names to 65.72%. The experimental results show that the method can improve the accuracy of Tibetan-Chinese name transliteration.

Keywords: Transliteration· Segmentation Granularity· Tibetan-Chinese.

1 Introduction

In the various tasks of cross-lingual natural language processing, the problem of translation of Out-of-Vocabulary (OOV) is often encountered. Usually, we can use a transliteration method to translate OOV, and the accuracy of OOV transliteration results can directly affect the actual Application [10]. Due to the difference of language feature, when transliterating the person name, the transliteration unit (ie, segmentation) of both source language and the target language should be appropriately adjusted. Therefore, translation granularity has always been one of the key points of transliteration research [12].

C. M. Verspoor proposes a forward-transliteration method based on four-step rules from English to Chinese [11]. By artificially establishing a conversion table for English syllables to Chinese Pinyin, and then establishing a conversion table for Chinese Pinyin to Chinese characters, better performance has been obtained. Lin and Chen used the English-Chinese unified phonetic table IPA (International Phonetic Alphabet) to realize the conversion of English to phonetic symbols and phonetic symbols to Chinese [6]. Knight and Graehl used the similarity between English phonemes and Japanese phonemes to achieve English to English phonemes, English phonemes to Japanese phonemes, Japanese phonemes to Japanese conversion [4]. Li et al. proposed a joint source-channel model in English-to-Chinese transliteration tasks, directly aligning English and

Chinese, and achieved good transliteration result [3][8]. Kunchukuttan and Bhattacharyya finished transliteration of eight languages pairs with three granularities: alphabet, word, mixed alphabet and the word [5]. Zhou and Zhao regard the transliteration of person name as sentence pair in statistical machine translation. Each transliteration unit is regarded as a word in a sentence, and the machine learning method is used to transliterate names and reached a great translation effect [1]. Yu et al. proposed a method of using word graphs to fuse multiple granularities to achieve transliteration of English-Chinese names [10]. Liu et al. used the combination of statistics and rules to analyze the characteristics of English pronunciation, constructed the rules of fine division of transliteration units, and proposed a method of transliteration unit based on glyphs and speech [7]. The above methods have achieved good results in segmentation granularity of transliteration from different views, but each method has its own inadequacies, mainly in the following aspects:

1) Rule-based methods require a large number of artificially constructed transliteration rules between specific language pairs, these rules are language-dependent and less robust.

2) The method of transliteration granularity by letter or token is language-independent, so the robustness is better. But because the information of person names pronunciation is not used, a large number of errors will be generated in the alignment stage, thereby affects the final transliteration effect.

3) Through the method of phoneme or unified phonetic table, the information of person names pronunciation can be used to generate the alignment with higher accuracy, but because of more conversion steps, the probability error or errors accumulation also causes the final transliteration accuracy.

4) The direct alignment method reduces the error caused by too many conversion steps and improves the accuracy of transliteration. However, this method skips the speech step, information loss is inevitable.

2 Methodology

The main pipeline of person name transliteration consists of four phases: preprocessing, transliteration model, decoding and post-processing, as Figure 1 displays. Preprocessing is the main topic that this paper focused on.

Step one, both Tibetan and Chinese corpus are sliced into three granularities for the experimental settings, so we get three corpora including same transliteration units. Step two, the transliteration model is trained with the aligned parallel corpus. Step three, a decoding experiment is performed on the source language test corpus that has been divided into transliteration units. Step four, because it doesnt need to reorder the result of transliteration, the output is mainly combined into pinyin, and then converted into corresponding Chinese.

2.1 Preprocessing

This paper studies the influence of segmentation granularity on the transliteration of Tibetan and Chinese names. Therefore, the same corpus is processed

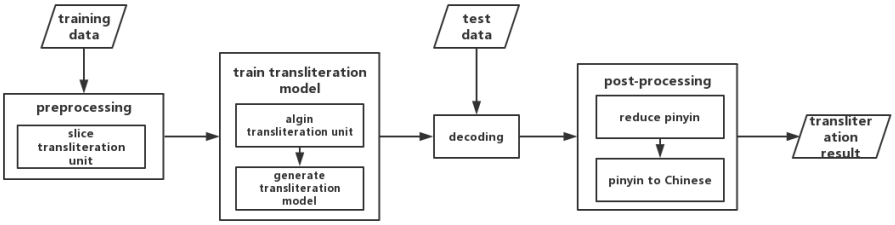

Fig. 1. The pipeline of transliteration.

with different segmentation granularity, and three corpora with the same names but different transliteration units are obtained. Three experiments are carried out to verify the research objectives by comparing the experimental results. In the preprocessing part, the emphasis is on the segmentation of transliteration units between source and target language. According to the different granularity of segmentation and alignment units, we divide them into three parts: direct alignment, further transformation, and further segmentation. We will describe the following methods based on the transliteration example offered by Figure 2.

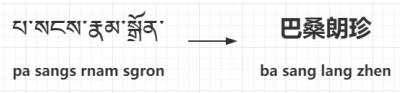

Fig. 2. An example of Tibetan person name and its Chinese transliteration. The following line is transcribed Latin (for Tibetan) or Pinyin (for Chinese).

Direct Alignment Tibetan has a special symbol for segmenting syllables, each syllable corresponds to a character in Chinese. We call the alignment of Tibetan syllables and Chinese characters as direct alignment.

Transcribed Latin to Chinese Pinyin We use Latin Transliteration to get representation Tibetan and Chinese person names as Figure 2 displays.

Fine-grained Segmentation Slice Tibetan Latin with rules ([p a] [s angs] [rn am] [sgr on]) and slice Chinese Pinyin by initial and vowel ([b a] [s ang] [l ang] [zh en]).

2.2 Transliteration Model

Alignment After corpus segmentation, it is necessary to align the segmented corpus so as to align the transliteration units of Tibetan and Chinese bilingual

names. For the three different corpora obtained by three different segmentation methods, we call them corpus one, corpus two and corpus three respectively. In machine transliteration, there is no ordering problem of transliteration units, so for the above three corpora, alignment is to combine the Tibetan list elements into the same number of parts according to the number of Chinese names segmentation, and then align the transliteration units of the same number parts.

In the process of segmentation, there are different numbers of transliteration units between the source language strings and the target language strings. Generally, one to zero often occurs in automatic alignment. For corpus three is constructed by fine-grained segmentation, using common automatic alignment method will cause faults, for example, vowels may appear before vowels. The alignment probably increases the error rate of transliteration. Therefore, we adopt the improved automatic alignment algorithm, the specific steps are as follows:

1) For corpus one and two, if the number of transliteration units of Chinese and Tibetan names is the same after segmentation, they will be aligned directly. That is to say, the segmented corpus will be aligned one by one and form corresponding transliteration pairs.

2) For corpus one and corpus two, if the number of transliteration units of Chinese and Tibetan names is different after segmentation, The maximum expectation algorithm (EM) algorithm is used to align them. The method of dividing the most probabilistic values is taken as the final dividing method, and then the corresponding transliteration pairs are obtained by aligning them in order.

3) For corpus three, we add the boundary symbol "|" to the segmented corpus, so that the whole segregated by the two boundary symbols is a whole before further segmenting, that is, "p a | s angs | rn am | sgr on" and "b a | s ang | l ang | zh en", and then align them with EM algorithm at the initial stage, if the boundary is a | s angs | rnam | sgr on". If the symbol is not at the beginning or end of the alignment result, we think that the alignment scheme is wrong. We eliminate them directly. Then we calculate the initial probability of the remaining schemes and iterate. We take the most probabilistic partition as the final partition method, and then align them in order to get the corresponding transliteration pairs of names.

Train For Tibetan to Chinese person name transliteration, if Chinese name denotes $\alpha = c_1c_2...c_m$, Tibetan name denotes $\beta = b_1b_2...b_m$, c_i represents the minimum segmented unit of Chinese name. So the relationship between Tibetan and Chinese is γ, predicted Tibetan name alignment sequence denotes $\bar{\beta}$:

$$\begin{aligned}\overline{\beta} &= \arg\max_{\beta} P(\beta, \alpha) \\ &= \arg\max_{\beta} \sum_{\gamma} P(\beta, \alpha, \gamma) \\ &\approx \arg\max_{\beta} \left(\arg\max_{\beta} P(\beta, \alpha, \gamma) \right) \\ &= \arg\max_{\beta, \gamma} P(\beta, \alpha, \gamma)\end{aligned} \tag{1}$$

Among, $P(\beta, \alpha, \gamma)$ represents the joint probability of α , β and γ,for K aligned transliteration units, we have

$$\mathrm{P}(\beta, \alpha, \gamma) \approx \prod_{k=1}^{K} P\left(\langle b, c\rangle_k | \langle b, c\rangle_1^{k-1}\right) \tag{2}$$

2.3 Post-processing

For experiment one, we get the corresponding Chinese results after the decoding step. We just need to format the results for our needs. For experiment two and experiment three, we get the corresponding Chinese Pinyin after decoding, so We need to convert it into the corresponding Chinese.

We use the Conditional Random Field (CRF) model to transform the Chinese of the training corpus into the corresponding Pinyin. Then we input the corresponding pinyin into the CRF as the training corpus and get the corresponding pinyin-to-Chinese model. Then we input the Pinyin from dataset two and dataset three into the CRF. The output of the model is the Chinese result we need.

3 Experiment

This experiment adopts the strategy of determining pronunciation first and then font shape. First, the pronunciation of Chinese names is determined by the transliteration pairs of Tibetan and Chinese names, and then the Chinese names are determined by the Chinese monolingual names. The data used in the experiment include 6405 transliteration pairs of Tibetan and Chinese names. The training set contains 5121 transliteration pairs of Tibetan and Chinese names, and the test set contains 1284 transliteration pairs of Tibetan and Chinese names.The NEWS 2018 Named Entity Transliteration Shared Task [2] has supported a statistical machine transliteration Baselines [9],we followed their work on the Tibetan-Chinese corpus and the result were served as our baseline.

3.1 Evaluate

A precision method is used in this experiment. Only when all parts of a persons name are transliterated correctly, can we think that the result of transliteration is accurate.

$$Precision = \frac{Number\ of\ Correct\ Transliteration}{Total\ Number\ of\ Transliteration} \tag{3}$$

3.2 Result

In order to compare the effects of different segmentation granularity on transliteration results as a whole, we designed three experiments.

Setting One We divide the Chinese name into characters, divide the Tibetan name into syllables, align the segmented corpus with EM algorithm, and train the corresponding transliteration model, then decode the corpus by Viterbi algorithm and output the best result. The result of transliteration serves as a reference.

Setting Two In the preprocessing stage, the Tibetan is converted to corresponding Latin transcription, the Chinese are converted to corresponding Pinyin, and then the other steps refer to setting one remain unchanged.

Setting Three . Preprocess data based on setting two, the Latin transcriptions corresponding to Tibetan are separated according to the five letters of a, e, i, o, u, and the Chinese phonetic alphabets corresponding to Chinese are separated according to the form of initial and vowel, then the other steps refer to setting one unchanged.

After corpus alignment, we extract transliteration parameters from the aligned corpus: logarithmic probability table of binary-order transliteration, probability table of one-order transliteration, and transliteration unit table of target language corresponding to source language transliteration units. In order to better analyze the experimental results, we remove duplication of the transliteration units of the source language and the target language. The statistic of the data set is shown in Table 1.

Table 1. Experimental settings and data statistic.

	Number of Tibetan Transliteration Units	**Number of Chinese Transliteration Units**
Setting One	538	1314
Setting Two	538	323
Setting Three	727	75

We can learn that the types of transliteration units in the target language have been significantly reduced from setting one to setting three, and the total number of corresponding relations between the source language and the target language has been significantly reduced. Under the same corpus, it is helpful to obtain the differentiated probability distribution, thus getting closer to the real value and improving the accuracy of transliteration.

In setting two and three, we adopted the strategy of determining pronunciation first and then font shape. We carried out the statistics of pronunciation accuracy before transliterating Pinyin into Chinese characters. That is, if all pronunciations of a persons name are correctly transliterated, then we think that the pronunciation of that persons name is correct. The overall results of the experiment are as shown in Table 2.

Table 2. Results for the Tibetan to Chinese transliteration task.

Setting	**Precision of Pronunciation(%)**	**Precision(%)**
Baseline	/	21.31
Setting One	/	32.14
Setting Two	84.95	62.95
Setting Three	92.36	65.72

Compared with setting one, setting two and setting three have achieved better performance. The analysis is as follows:

1) Tibetan is a low resource language. It is difficult to obtain enough transliteration pairs of names. In experiment 1, when Tibetan and Chinese are aligned directly and the corresponding relationship is trained, a large number of data sparsity problems arise, which have a great impact on the results, while the other two experiments have almost no problem of data sparsity.

2) There are a large number of homonyms in Chinese. Because of the limited number of transliteration pairs of names, there are a lot of data sparseness problems in the corresponding matrix we get. Even if we use data smoothing technology, the corresponding relationship obtained is far from the real situation. There is no loss conversion between Tibetan and Latin transcriptions. Converting Latin transcriptions to Pinyin first can get better correspondence because the types of Pinyin are far less than the number of Chinese characters, and then the Pinyin can be converted to Chinese characters. Because of the higher accuracy of pronunciation, the final result of this step is better than that of direct transliteration even if some errors.

3) Finer-grained segmentation of Latin transcription and Pinyin can further reduce the total number of corresponding relationships, thus increasing the average number of occurrences of each corresponding relationship, making the number of occurrences of each corresponding relationship more different, thus closer to the real value, and improving the effectiveness of transliteration.

After more fine-grained segmentation of corpus, the accuracy of transliteration has been significantly improved, which verifies that finer-grained segmentation not only reduces the problem of data sparsity caused by the method based on the morphology of character but also transfers the problem of corpus size in less-resourced languages to the problem of the unilateral corpus in Chinese. Thus, a finer-grained method of name segmentation can be obtained. It can improve the effect of transliteration.

4 Conclusion

This paper proposes a more fine-grained method of Latin transcription and Pinyin segmentation through the conversion of Tibetan Latin transcription and Chinese Pinyin based on shape segmentation of character. The experimental result shows that the proposed segmentation method can improve the accuracy of transliteration of Tibetan and Chinese names, and it also performs well in solving the problem of scarcity of names in low resource scenarios. This research has the following innovations:

1) The idea of removing duplication of correspondence on the basis of pronunciation is put forward. In previous studies, most of them have focused on expanding the size of the corpus and improving the coverage and complexity of correspondence. Such ideas are often not ideal in the transliteration of languages with low resources. In this study, we first determine the pronunciation, then determine the shape of the characters, and effectively removing duplication of the corresponding relationship on the basis of pronunciation, thus reducing the requirements of the model for the size of Tibetan-Chinese names corpus, and better enhancing the effect of Tibetan-Chinese names transliteration.

2) The problem that bilingual person name pairs in low resource languages are transformed into the problem of Chinese monolingual person-name pairs that can be easily solved. The feasibility of this method is verified by experiment.

However, we only segment Tibetan Latin transcription based on simple rules. In the process of alignment, some wrong segmentation and alignment will occur, which will reduce the accuracy of transliteration. In the future work, we will introduce more rules of language segmentation and alignment to reduce the errors caused by the segmentation and alignment stage, so as to further improve the accuracy of transliteration of Tibetan and Chinese person names.

References

1. BO, Z., ZHAO, J.: Comparison of several english-chinese name transliteration methods. The Fourth National Seminar on Computational Linguistics for Students pp. 24–30 (2008)
2. Chen, N., Banchs, R.E., Zhang, M., Duan, X., Li, H.: Report of news 2018 named entity transliteration shared task. In: Proceedings of the Seventh Named Entities Workshop. pp. 55–73 (2018)

3. Haizhou, L., Min, Z., Jian, S.: A joint source-channel model for machine transliteration. In: Proceedings of the 42nd Annual Meeting on association for Computational Linguistics. p. 159. Association for Computational Linguistics (2004)
4. Knight, K., Graehl, J.: Machine transliteration. Computational linguistics **24**(4), 599–612 (1998)
5. Kunchukuttan, A., Bhattacharyya, P.: Data representation methods and use of mined corpora for indian language transliteration. In: Proceedings of the Fifth Named Entity Workshop. pp. 78–82 (2015)
6. Lin, W.H., Chen, H.H.: Backward machine transliteration by learning phonetic similarity. In: proceedings of the 6th conference on Natural language learning-Volume 20. pp. 1–7. Association for Computational Linguistics (2002)
7. LIU, B., XU, J., Yefeng, C., ZHANG, Y.: Integrating of grapheme-based and phoneme-based transliteration unit alignment method. Acta Scientiarum Naturalium Universitatis Pekinensis pp. 75–80 (2016)
8. Min, Z., Haizhou, L., Jian, S.: Direct orthographical mapping for machine transliteration. In: Proceedings of the 20th international conference on Computational Linguistics. p. 716. Association for Computational Linguistics (2004)
9. Singhania, S., Nguyen, M., Ngo, G.H., Chen, N.: Statistical machine transliteration baselines for news 2018. In: Proceedings of the Seventh Named Entities Workshop. pp. 74–78 (2018)
10. Tingting, L.: Research on Nonparametric Bayesian Based Multi-Language Names Transliteration. Ph.D. thesis, Harbin: Harbin Institute of Technology (2013)
11. Wan, S., Verspoor, C.M.: Automatic english-chinese name transliteration for development of multilingual resources. In: COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics. vol. 2 (1998)
12. YU, H., TU, Z., LIU, Q., LIU, Y.: Lattice-based multi-granularity name-entity machine transliteration. Journal of Chinese Information Processing **27**(4), 16–22 (2013)