An EMD Based Approach to Transliteration Unit Alignment between English and Chinese

Muyun Yang¹, Shujie Liu¹, Lei Wang¹, Sheng Li¹, Jufeng Li¹, Tiejun Zhao¹, Haoliang Qi²

¹*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001*
{ymy; shujie; lwang; jfli; tjzhao}@mtlab.hit.edu.cn; lisheng@hit.edu.cn;

²*Department of Computer, Heilongjiang Institute of Technology, Harbin, 150001*
hoaliang.qi@gmail.com

ABSTRACT

Machine transliteration is to translate the proper nouns in the source language according to its pronunciation into the target language. Recent orthographical based approach has improved the performance of machine translation significantly. Focusing on the transliteration unit alignment that provides a fundamental parameter for the model, this paper adopts a semi-supervised EMD approach—applying discriminative model over the original EM results. The experiment results prove it is substantial to the improvement of the transliteration performance.

Index Terms—Transliteration, alignment, EMD

1. INTRODUCTION

With the growing trend of globalization, more and more importance is attached to the cross-language information processing technology, such as machine translation (MT), cross-language information retrieval (CLIR) etc., to overcome language barriers. Machine transliteration is one of the non-trivial issues because it solves a substantial part of translation of named entity (NE), which carries the fundamental factual information in human languages.

Basically, the transliteration is a method to translate the person names, place names and other proper nouns by transcribing the source pronunciations into the target language phonemes. For example, the English name “Britain” is transliterated into “不列颠” (/ bu lie dian /) as a phonetic equivalent in Chinese. This process is usually known as “machine transliteration” in natural language processing circle, in which the phonetic feature governs the transformation from source language into target language.

According to the unit of processing, current transliteration (short for machine transliteration hereafter) technologies can be divided into “phoneme-

based” method and “grapheme-based” method. The phoneme-based method usually includes the following process: (a) convert the words into their pronunciation transcriptions (e.g. phoneme); (b) transliterate the phoneme of the source language into the phoneme of the target language; (c) convert phoneme of the target language into the words [1]. In contrast to the “word-phoneme-word” conversion of the phoneme-based method, the grapheme-based method directly models the mapping from the source language word to the target language. Because it avoids the information loss during the phoneme conversions, the grapheme-based method out-performs grapheme method in most recent researches [2, 3].

A key technology in the grapheme based transliteration is the EM based orthographical unit alignment between source and target language. The best performance of the grapheme-based method at present is 69.3% in word-level and 89.8% in the transliteration unit level [4]. Among the factors that influence the performance of the machine transliteration, the quality of transliteration unit alignment is an extremely important factor, owing to its decision on the correctness of the selected transliteration unit in the process of decoding.

To further improve the performance of the grapheme based transliteration model, this paper introduces the discriminative learning into the original EM based transliteration unit alignment. We first label a small part of data with transliteration unit alignment manually. With this data as development dataset, we improve the transliteration unit alignment performance by discriminative learning. The experiments show that, compared to the EM algorithm, this algorithm can significantly improve the quality of transliteration unit alignment, and accordingly, improve the performance of English/Chinese grapheme transliteration on the joint-source channel model.

The rest of this paper is arranged as following. Section 2 describes the whole frame of EMD approach to

transliteration unit alignment. Section 3 presets the experimental results, which proves the validness of the proposed approach. Section 4 finally concludes the paper.

2. EMD APPROACH TO TRANSLITERATION UNIT ALIGNMENT

2.1 EMD: discriminative frame over EM transliteration unit alignment

In the EM algorithm, E steps compute the probability of all the transliteration unit alignment pairs, and M step establish a new estimation model based on the probability distribution. Ideally, the E step will list all the possible alignment and mark them with probability, but this is a very difficult issue. In the M steps, we want to compute all the possible alignment for every transliteration pair based on the probability estimated in the E steps.

Here we expand the EM algorithm to form a new training method by incorporating the minimum error (or maximum likeness) rate to further improve the EM results. The motivation is to take the advantages of both the EM algorithm to estimate the parameters without supervision and the discriminative learning to minimize the error with a small proportion of annotated data. This thought is actually formalized as so-called EMD (EM plus discriminative training) algorithm to improve the performance of word alignment [5]. In this algorithm, the EM model is responsible for initializing large-scale parameters with certain accuracy, and the discriminative training is used for controlling the contribution of each parameter by minimizing the error in the annotated data.

In the framework of EMD as shown in Fig. 1, we use a small part of transliteration pairs aligned by the EM algorithm, which is about 1% in later experiment, to form a manually labeled corpus. In the process of training of the model, we will feed this manually labeled data to the discriminative module, adjusting the parameters for the transliteration unit alignment in a semi-supervised style.

As is illustrated in Fig. 1, the transliteration units are first aligned through EM algorithm without supervision in the EMD framework. Then the corpus is divided into two parts: a small portion of data is manually labeled as the development set and the most of the data are left untouched. Then the discriminative model updates the EM results by achieving a best alignment performance over the development sets. It should be noted that during the discriminative learning process, the training corpus is continuously realigned with each updated parameters acquired. Therefore, when the EMD method stops, the training corpus has

actually been re-aligned several times so as to get the best transliteration unit alignment results in the development corpus.

Fig.1 Framework of EMD training of the transliteration unit alignment

2.2 EM based transliteration unit alignment

The first step in EMD is to initialize transliteration unit alignment by the EM algorithm. Basically, EM algorithm is a typical example of unsupervised algorithm including two steps: E step calculates the probability of the hidden variable relative to the sample based on the given model parameters; M step recalculates model parameters on the basis of the probability distribution obtained from the former step. After certain iterations of these two steps, the EM algorithm gradually improves the model performance and finally converges at a maximum point.

We use EM algorithm to automatically align the grapheme unit for English-Chinese personal name transliteration. Given the whole corpus W with m transliteration pairs, the task of transliteration can be regarded as to find a best alignment sequence $\gamma^m (= \gamma_1, \gamma_2, \dots, \gamma_m)$, in which γ_i is the transliteration unit alignment of i^{th} transliteration pair. Specifically, the process of using EM algorithm to carry out the auto alignment in the transliteration unit level on the English-Chinese word pair includes:

1) Establish n -gram, transliteration model. A bi-gram grapheme transliteration model is:

$$\hat{\gamma}^n \equiv \underset{\gamma^n}{\operatorname{argmax}} \prod_{i=1}^n p(\langle e, c \rangle_i | \langle e, c \rangle_{i-1})$$

in which $\langle e, c \rangle_i$ is the i^{th} transliteration unit alignment aligned in a transliteration pair.

2) For the corpus that hasn't been aligned, we shall carry out all possible forms of segmentation as a learning knowledge base. For example, given the word pair "史密斯/smith", we shall use the segmentation algorithm to get all the possible segmentation:

"史密斯/s m ith" ;
 "史密斯/s mi th" ;
 "史密斯/s mit h" ;
 "史密斯/sm i th" ;
 "史密斯/sm it h" ;
 "史密斯/smi t h" .

3) Initialize the parameter values by averagely distribute the probability for each $P(<e, c>_i | <e, c>_{i-1})$;

4) E-step: for all the possible segmentations, calculate the expectation of $N_{<e, c>_i}$ and $N_{(<e, c>_i, <e, c>_{i-1})}$ with the current parameter estimation.

$E(N_{<e, c>_i})$ as the expectation of a specific transliteration pair $<e, c>_i$ is calculated as:

$$E(N_{<e, c>_i}) = \sum_{\gamma^n} count\langle e, c \rangle_i \times P(\gamma^n | W, \Lambda)$$

$E(N_{(<e, c>_i, <e, c>_{i-1})})$ as the expectation from $<e, c>_{i-1}$ to $<e, c>_i$, is calculated by:

$$E(N_{(<e, c>_i, <e, c>_{i-1})}) = \sum_{\gamma^n} C_{count}(\langle e, c \rangle_i, \langle e, c \rangle_{i-1}) \times P(\gamma^n | W, \Lambda)$$

5) M-step: use MLE to update the probability of $P(<e, c>_i | <e, c>_{i-1})$:

$$P(\langle e, c \rangle_i | \langle e, c \rangle_{i-1}) = \frac{E[N_{\langle e, c \rangle_i, \langle e, c \rangle_{i-1}}]}{E[N_{\langle e, c \rangle_i}]}$$

6) Repeat Step 4 and Step 5 until the parameters converge;

2.3 Features in the discriminative alignment model

When the EM algorithm stops, it generates the conditional probability between each transliteration unit pairs. Then with the discriminate learning module, we try to optimize these probabilities by minimize the errors in align the development transliteration pairs.

Based on the experience of word alignment, under the discriminative learning model in this paper, we initially select 3 features according to the characteristics of transliteration unit alignment:

1) Probability of transliteration unit, which is easily obtained from statistics from the initial alignment, is calculated as:

$$P_p(e, f) = \frac{Count(e, f)}{\sum_{e \in E, f \in F} Count(e, f)}$$

in which $Count(e, f)$ represents the times of the alignment between the transliteration unit of the source language e and the target language f . E is the set of the words of the source language, and F is the set of the target language words.

2) Reverse fertility probability, which is the likelihood of n source words to be aligned to one target word, is computed as:

$$P_n(e, n) = \frac{Count(e, n)}{\sum_{e \in E, m \in \{1, 2, \dots\}} Count(e, m)}$$

in which $Count(e, n)$ is the number of source word e can be aligned to the word of the target language, and E is the set of the source word.

3) Context information, which is to describe the relation between the adjacent syllables of target language, is calculated by a character based bi-gram Markov model. The context information used in this paper is shown with a few examples in the following table.

Table 1. Negative log probability of context feature

$C_i \ C_{i-1}$	$-\log P(c_i c_{i-1})$
勒_海	4.25
普_汤	2.08
夫_措	0.69

The discriminative process is used to train the weight of the above-mentioned feature. In practice, a one-dimension error-minimized training method suggested in [6] to minimize the alignment errors in the development corpus.

5. EXPERIMENTS

The data used in this paper is obtained from a bilingual dictionary named "Chinese Transliteration of Foreign Name", containing 37,668 entries of the English names and their Chinese translations [7].

First, we use the EM algorithm to obtain the initial alignment results of the 37,688 pairs of English and Chinese names. Then we select 292 transliteration pairs and manually correct the alignment errors of translation units among them. This small part of data is used as development set, and all the 37,688 pairs are used as training dataset. We train the data with EMD method to obtain new alignment results of transliteration unit.

With the two methods of EM and EMD, we actually get two results of the English-Chinese

transliteration unit alignment. With these two sets of data, we can establish two grapheme based transliteration models. Here we take the final transliteration model performance as an indicator of transliteration unit alignment accuracy because it the final transliteration results that matters. To test the performance of the models, we adopt a ten-fold cross validation on the data. Firstly, we divide the training data into ten parts. Then we use one of them for the open test, and the left nine are reserved for training data.

Two kinds of evaluation methods are employed in the experiment for transliteration performance. The first measure is BLEU, which we often used for MT evaluation. Since the calculation unit is Chinese character, BLEU score roughly represents the accuracy of the transliteration in grapheme unit level. Then we treat the transliterated words as a whole to check whether they comply with the standard answer. And this evaluation method reflects the transliteration performance in the word level.

The result of this experiment is shown in Table 2. According to the table, we can see that after introducing the discriminative learning, the transliteration performance has improved markedly. The BLEU-4 score increases 3.4%, while the accuracy of the entire vocabulary level increases 4.0%.

Table 2. Performances of EMD approach to transliteration

		Open Test
BLEU Score	EM	70.2%
	EMD	72.6%
Word Accuracy	EM	66.8%
	EMD	69.5%

After analyzing the errors occurred in the results, we discover that besides the transliteration unit alignment errors, there are problems of data sparseness. In the future work, we will explore how the proposed method will perform in a larger data scale: if a larger data will bring a better performance for the EMD strategy. In addition, we will compare the application of the EMD approach for transliteration unit alignment and the discriminative learning at the transliteration model level.

6. CONCLUSION

This paper has explored the technology of enhancing the grapheme based English-Chinese transliteration by transliteration unit alignment based on the semi-supervised discriminative learning method. First, we label a small part of data with transliteration unit alignment manually, and we use this data as training dataset to improve the transliteration unit alignment

quality based on the discriminative learning method. Compared to the unsupervised EM training method, this method overcomes the local optimal error of the EM method with a small price of small scale of data labeled manually. In relation to the general discriminative learning method, this method uses a part of the training dataset as measurement dataset and improves the quality of the parameters getting from the training dataset. The experiment has shown that using the results of the transliteration unit alignment can significantly improve the performance of the entire English/Chinese transliteration model.

ACKNOWLEDGEMENT

This work is supported by Natural Science Foundation of China (Grant No. 60773066, 60736044)

REFERENCES

- [1] J. S. Lee and K. S. Cho, "English to Korean Statistical Transliteration for Information Retrieval", *Computer Processing of Oriental Languages*, pp. 17-37, 1998, 12(1)
- [2] H. Li, M. Zhang and J. Su, "A Joint Source-Channel Model for Machine Transliteration", *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004
- [3] M. Zhang, H. Li and J. Su, "Direct Orthographical Mapping for Machine Transliteration", *Proc. of 20th International Conference on Computational Linguistics*, Aug, 2004
- [4] A. Ekbal, S. K. Naskar and S. Bandyopadhyay, "A Modified Joint Source-Channel Model for Transliteration", *Proc. of the COLING/ACL 2006*, pp. 191-198, 2006
- [5] A. Fraser and D. Marcu, "Semi-Supervised Training for Statistical Word Alignment", *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL 2005*, pp. 769-776, 2005
- [6] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", *Proc. of ACL 2003*, 2003
- [7] Xinhua News Agency, *Chinese Transliteration of Foreign Name*, Business Press, 1992