A Hybrid Transliteration Model for Chinese/English Named Entities —BJTU-NLP Report for the 5th Named Entities Workshop

Dandan Wang, Xiaohui Yang, Jinan Xu, Yufeng Chen, Nan Wang, Bojia Liu, Jian Yang, Yujie Zhang
School of Computer and Information Technology
Beijing Jiaotong University
{13120427, xhyang, jaxu, chenyf, 14120428, 14125181, 13120441, yjzhang}@bjtu.edu.cn

Abstract

This paper presents our system (BJTU-NLP system) for the NEWS2015 evaluation task of Chinese-to-English and English-to-Chinese named entity transliteration. Our system adopts a hybrid machine transliteration approach, which combines several features. To further improve the result, we adopt external data extracted from wikipeda to expand the training set. In addition, pre-processing and post-processing rules are utilized to further improve the performance. The final performance on the test corpus shows that our system achieves comparable results with other state-of-the-art systems.

1 Introduction

Machine transliteration transforms the script of a word from a source language to a target language automatically. Knight(1998) proposes a phoneme-based approach to solve the transliteration between English names and Japanese katakana. The phoneme-based approach needs a pronunciation dictionary for one or two languages. These dictionaries usually do not exist or can't cover all the names. Jia(2009) views machine transliteration as a special example of machine translation and uses the phrase-based machine translation model to solve it. However, using the English letters and Chinese characters as basic mapping units will make ambiguity in the alignment and translation step. Huang(2011) proposes a novel nonparametric Bayesian using synchronous adaptor grammars to model the grapheme-based transliteration.

This paper describes a machine transliteration system and data measures for participating NEWS2015 evaluation, which is abbreviated as BJTU-NLP. We participated in two transliteration masks: Chinese-to-English and English-to-Chinese named entity transliteration task. This report briefly introduces the implementation framework of our machine transliteration system, and analyzes the experimental results over the evaluation data.

The following parts are organized as follows: Section 2 briefly introduces the implementation framework of the transliteration system. Section 3 introduces the details of the experiment and data processing in brief. In Section 4, experimental results are given and the results of the experiment are analyzed. Section 5 is our conclusion and future work.

2 System Description

By treating transliteration as a translation problem, BJTU-NLP has realized a machine transliteration system based on the combination of multiple features by a log-linear model, to complete the corresponding experiments with English-Chinese and Chinese-English name pairs The description of the whole transliteration system is as follows.

2.1 A Log-linear Machine Transliteration Model

In this evaluation, a tool is used in our machine transliteration system based on the fusion multiple features. In this system, we introduce a linear log model for transliteration (Koehn et al., 2007), using combination features in it. The process of transliteration can be described as follows: for a given source language name $\boldsymbol{s}$ find the optimal result $\hat{\boldsymbol{e}}$ from all possible results $\boldsymbol{e}$, which is computed by:

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{e},\boldsymbol{s})\right)}{\sum_{e'} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{e}',\boldsymbol{s})\right)} \quad (1)$$

Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP, pages 67–71,
Beijing, China, July 26-31, 2015.

Where M is the number of used features, $h_m(\boldsymbol{e},\boldsymbol{s})$ is the *mth* transliteration feature, and λ_m is the weight of the *mth* transliteration feature.

2.2 Features

In the transliteration process, the source name is transformed from left to right in the order, lexical reordering problem does not exist, therefore, the transliteration model does not require replacement model features, and because "phrase translation pair" does not exist lexical correspondence (between English letters and correspondence Chinese characters), forward/reverse phrase lexicalization probability are not used in our transliteration model. In the final, the features we used are as follow:

1. Forward phrase translation probability, $P(e|s)$ is the probability of translating into English name ***e*** from Chinese name ***s***, the formula is as follows.
$$P(e|s) = \frac{count(e,s)}{count_{\bar{e}}(e,s)} \quad (2)$$
2. Reverse phrase translation probability, $P(s|e)$ is the probability of translating into Chinese name ***s*** from English name ***e*** , the formula is as follows.
$$P(s|e) = \frac{count(s,e)}{count_{\bar{f}}(e,s)} \quad (3)$$
3. The length of name
4. The normalized length deviation after transforming the length of the other language into the reference language， $I(e|s)$, $I(s|e)$ are computed as follows.
$$I(e|s) = \frac{|len(s)-len(e)|}{len(s)} \quad (4)$$
$$I(s|e) = \frac{|len(e)-len(s)|}{len(e)} \quad (5)$$
Where *len(s)* is the number of characters this source name contains, *len(e)* is the number of segments target name contains.
5. Language model score, $lm(c)$. In the translation model based on phrase, each source phrase fragments can be translated without considering the source language phrase fragments which are in front of it. Each source language phrases are independent in transliteration, the transliteration between source language phrase and target language phrase only rely on the language model of the target language.

2.3 Parameter Tuning and Decoding

The system adopts GIZA++, which is a word alignment model to extract transliteration phrases pairs. In order to get the best weight of features and the best name transliteration model, the process of parameter tuning is as follows:

1. The weights of five features mentioned in the previous section are initialized to 1.
2. Using the log-linear model on the development set, we can obtain the NBest transliteration candidate, then merge with the original NBest candidate to form new candidate results.
3. According to the new NBest candidate results obtained, in order to get the best BLEU value, each feature weight is adjusted with the ZMERT (Zaidan et al.2009) toolkit for a better log-linear model.
4. Repeat steps 2, 3 until the model reaches convergence, finally we obtain the optimal weight of each feature. Then decode given names, using phrase table formed in training stage and transliteration model with optimal weight.

3 Rule-based Adaptation

3.1 External Dictionary

In this evaluation, in addition to the official data sets, we proposed to import the Wikipedia data set as an external dictionary. After obtaining the data from the Wikipedia database, we use clustering and iterative methods to obtain named entity pairs. We did data cleansing, de-noising and de-emphasis for the obtained name entity pairs. For the reserved data, it need to comply with the following requirements:

1. Retain only the English and Chinese name transliterations.
2. For some English names contains a modified letter, for example Áá, Àà, Ăă, Ắắ, we would replace the letter with its corresponding ordinary alphabet letters.
3. Cannot have duplicate transliteration results (including given official data sets).

After the above steps, we got about 37,151 available named entity pairs. During the expanded training of non-standardized methods, we need to add the above corpus into English-to-Chinese and Chinese-to-English training set respectively, and then do the de-emphasis operation to ensure the uniqueness of each named entity pair.

3.2 Chinese-to-English preprocessing

For Chinese corpus, our preprocessing rules are as follows:

1. Simplified Chinese representation

2. Chinese word segmentation method

Segmentation

During the segmentation stage, we take the given word as a sequence of characters. Then combined with the characteristics of Chinese grammar, we take particular rule to Chinese word segmentation as divide the Chinese word by space.

Word Alignment

Word alignment here accurately refers to the alignment of segmentation result of the above step result. Word alignment tool we used is the GIZA ++ (Och et al., 2003). Since the corpus is named entity pairs, we took the result of GIZA ++ as the final word alignments.

Language Model

After several times comparison test, the two systems involved in this evaluation adopt the 3-gram language model.

3.3 English-to-Chinese preprocessing

For English corpus, our preprocessing rules are as follows:

1. Capitalization representation
2. English word segmentation method

Segmentation

During this segmentation stage, we also take the given word as a sequence of characters. Then we take particular rules to English word segmentation as divide these words by syllable.

Word Alignment

Word alignment here uses the same tool as above.

Language Model

The two systems involved in this evaluation also adopt the 3-gram language model.

3.4 Corpus usage

The evaluation directions of our participation are Chinese-to-English and English-to-Chinese named entity transliteration direction. And all evaluation corpus we used for this evaluation (including the training sets, development sets, test sets and reference sets) are as follows:

	Training Set	Dev Set	Test Set
English-to-Chinese	37,753	2,802	1008
Chinese-to-English	28,678	2,719	1019

Table 1 standardized methods of data list

	Training Set	Dev Set	Test Set
English-to-Chinese	74,904	2,802	1008
Chinese-to-English	65,829	2,719	1019

Table 2 Non-standardized methods of data list

4 Experiments

4.1 Data Sets

The standard training set of English-Chinese transliteration track contains 37753 pairs of names. We pick up 37151 pairs of names extracted from Wikipedia to merge into the training set. 2802 pairs are treated as the final dev set to tune the weights of system features. For the Chinese-English back transliteration track, the final training and test sets are formed in the same way. The official dev set is used directly.

The Srilm (Stolcke et al., 2002) toolkit is used to count n-gram on the target of the training set. Here we use a 3-gram language model. In the transliteration model training step, the Giza++ (Och et al., 2003) generates the alignment with the grow-diag-and-final heuristic, while other setup is default. The following 4 metrics are used to measure the quality of the transliteration results (Li et al., 2009a): Word Accuracy in Top-1 (ACC), Fuzziness in Top-1 (Mean F-score), Mean Reciprocal Rank (MRR), MAPref.

4.2 Experimental results

Our transliteration systems' outputs have following format problems:

1. English-to-Chinese outputs: the Chinese output words are still separated by spaces
2. Chinese-to-English outputs: English output words are still divided by syllable

To solve these problems, we make the following amendments to the outputs:

1. Remove the spaces between character and character, syllable and syllable
2. The English results are expressed as: initial capital letters, other letter lowercase

We adopt Niutrans (Xiao et al., 2012) to realize our log-linear model to combining several features. By comparing the experiment, we found that segmentation by syllable of English words is more effective and segmentation by Pinyin and syllable of Chinese words performs better. We adopt the above standard and non-standard training set to evaluate the official test set, and use official development set to adjust parameters. The evaluation results of the standard and non-

standard training set and corresponding analysis are shown as follows.

4.2.1 Evaluation Results and Analysis of Standard Training Set

We evaluated the four official test sets respectively. We calculated the four parameter values, ACC, F-score, MRR and MAP_ref, according to the four official evaluation standards. The experimental results are shown in Table 3.

Test Sets	ACC	F-score	MRR	MAP_ref
ChEn_2266	0.151	0.766	0.151	0.151
ChEn_1019	0.157	0.732	0.157	0.151
EnCh_2000	0.225	0.620	0.225	0.212
EnCh_1008	0.204	0.605	0.204	0.195

Table 3 Standard training set evaluation results

In Table 3, we found that the effect of English-Chinese transliteration is better than the Chinese-English transliteration. The effect of English-Chinese transliteration is better than the Chinese-English transliteration, which shows that segmentation of syllable is more reasonable for preprocessing when the source language is English, and preprocessing method of Chinese needs to be improved.

4.2.2 Evaluation Results and Analysis of non-Standard Training Set

We added the English-Chinese and Chinese-English named entities drawn from the Wikipedia to the training set, and evaluate the official test sets by the expanded training set as non-Standard training set. We calculated the four official parameter values likewise and experimental results are shown in Table 4.

Test Sets	ACC	F-score	MRR	MAP_ref
ChEn_2266	0.105	0.746	0.105	0.105
ChEn_1019	0.157	0.732	0.157	0.151
EnCh_2000	0.224	0.629	0.224	0.212
EnCh_1008	0.193	0.605	0.193	0.182

Table 4 non-Standard training set evaluation results

We can conclude from Table 4 that the results of the evaluation on the non-Standard training set have promotion over that on the Standard training set. This suggests that increasing the training set has a positive influence on improving the evaluating results.

5 Conclusions and Future Work

This paper mainly describes the machine transliteration system and data measures for participating NEWS2015 evaluation of BJTU-NLP. We adopt a hybrid transliteration model to realize named entities transliteration. In the process of training, we added the preprocessing of training corpus, modified related parameters of Niutrans system and the compared results of the experiment with different parameters. Related post-processing is also added according to the transliteration results. Simultaneously, we expand the training set with the help of Wikipedia in the named entities. The experimental results show that after joining in the named entities to Wikipedia, the evaluating results have a certain increase.

As to future work, we plan to conduct in-depth research and discussion in the preprocessing of named entities transliteration, post-processing and machine transliteration model, etc.

Acknowledgement

The research work has been supported by the National Nature Science Foundation of China under grant no. 61370130, and 61473294, the Fundamental Research Funds for the Central Universities (2014RC040), and also the International Science & Technology Cooperation Program of China under grant No. 2014DFA11350.

Reference

Keven Knight, Jonathan Graehl. 1998. Machine Transliteration. Computational Linguistics, Vol. 24, No. 4, pp. 599-612.

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. 2012. NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation. In Proc. of ACL, demonstration session.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In Proc. Of ICSLP, Denver, USA.

Franz Josef Och, Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Comput.Linguistics 29, 1, 19-51.

Yun Huang, Min Zhang and Chewlim Tan. 2011. Nonparametric Bayesian Machine Transliteration with Synchronous Adaptor Grammars. In Proceedings of ACL-HLT 2011: Short Papers, Portland, Oregon, pp.534-539.

Yuxiang Jia, Danqing Zhu, and Shiwen Y. 2009. A Noisy Channel Model for Grapheme-based Machine Transliteration. In the Proceedings of the 2009 Named Entities Workshop, 2009, pp. 88-91.

Koehn P, Och F J, Marcu D. Statistical Phrase-Based Translation[J]. Statistical Phrase-Based Translation, 2002, (5):127--133.

Zaidan O. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems[C]. //Prague Bulletin of Mathematical Linguistics. 2010:2009.