Machine Transliteration of Names from Different Language Origins into Chinese

Yan Shao, Jörg Tiedemann

Department of Linguistics and Philology
Uppsala University
yan.shao.9735@student.uu.se, jorg.tiedemann@lingfil.uu.se

1. Introduction

Machine transliteration is the process of automatically transforming the script of a word from a source language to a target language based on its pronunciation, which is an effective approach to complement the overall performance of machine translation system. It also plays an important role in cross-language information retrieval. Chinese as a major Asian language is often the target language in many machine transliteration tasks. The syllabic representation of Chinese characters is very special, so it is challenging to build a transliteration model which is universal and efficient as well as capable of achieving satisfying precision in transliterating named entities from different languages into Chinese.

In this research, focusing on the problem of transliterating names from different source languages into Chinese, we build an HMM based machine transliteration system and apply it on different language data sets.

2. Background

There is a number of publications on machine transliteration. In many cases, the source language and target language are English and an Asian language, for examples Chinese, Korean or Japanese. Similar to machine translation, machine transliteration also started with the attempt of creating rule-based systems. However, the research of machine transliteration greatly developed with the utilization of statistical models. Various phonetic-based models as well as orthographic-based ones were built devoted to different languages (Karimi et al., 2011). A large number of approaches particularly focus on the transliteration in which Chinese is used as the target language. Furthermore, there are also standard metrics to evaluate the quality of automatic transliteration system, such as ACC (word accuracy, complete match) and Mean F-score (edit distance based evaluation) (Li et al., 2010b).

3. Data and Tools

We use the *Translation Dictionary for Foreign Names* (Xia, 1993) as the main dataset for investigating how different language origins of the names influence the quality of machine transliteration system. Meanwhile, the dataset of NEWS 2010 shared task (Li et al., 2010b), specifically English to Chinese generation task is also used to build and test our basic transliteration model in order to compare it with the state-of-the-art transliteration systems evaluated in that conference. We also use a Chinese character-pinyin dictionary to obtain the pinyin of transliterations. Pinyin is the most widely used romanization of Chinese characters, which is applied as intermediate representation of our transliteration system. Additionally, M2M Aligner (Jiampojamarn et al., 2007) is used as the crucial tool for retrieving the segmentations of the source string and for obtaining the alignment substring pairs.

4. Transliteration Model

4.1 Baseline System

Considering the transliteration as a labelling process, we construct an HMM based machine transliteration system. We modify the classical HMM model by adding pinyin as the intermediate phonetic representation in the system. The M2M Aligner is used to segment and align the training data via unsupervised learning. We assume that a single Chinese character is an independent phonetic unit, so we set the maximum length of substrings on the source side as five and on the target side as one when applying M2M Aligner. We use maximum likelihood estimation with the result returned by M2M Aligner to estimate the parameters of our transliteration model. The Viterbi Algorithm is implemented in the decoder to guarantee efficient decoding. Our baseline yields competitive results with ACC of 0.336 and mean F-score of 0.691 on the test data of the NEWS 2010 workshop.

4.2 Modified System

Additionally, based on the experiments on the development set, we use several approaches to modify the baseline system.

First, we pre-contract letter combinations for the M2M Aligner. In our transliteration task, there are some letters on the source side that are pronounced as one single phoneme and they are never transliterated into two Chinese characters. If we pre-contract those letter combinations and regard them as single units, the precision of the M2M Aligner significantly improves therefore the entire transliteration system performs better.

We also assume that the low frequency alignments in the output of the M2M Aligner are likely to be incorrect which has a negative impact when used as training instances. Therefore, we trim those terms whose frequencies are one.

Additionally, in order to compensate for the negative effect of the unigram model that we use in the transliteration model, we add a penalty score to modify the probability of substrings to make the length of substrings have less impact on the overall estimation.

Languages	Language Specific Systems					Generic System			
	Size of	Baseline		Modified		Baseline		Modified	
	training set	ACC	MFS	ACC	MFS	ACC	MFS	ACC	MFS
Czech	32,189	0.547	0.849	0.582	0.864	0.345	0.746	0.423	0.798
English	45,239	0.331	0.688	0.364	0.710	0.204*	0.585	0.211*	0.605
Finnish	16,458	0.647	0.876	0.663	0.878	0.467	0.801	0.507	0.819
French	75,568	0.422*	0.770	0.424*	0.768	0.233	0.619	0.197	0.607
German	46,118	0.507	0.825	0.573	0.854	0.269	0.675	0.325	0.719
Hungarian	25,600	0.398	0.747	0.426	0.760	0.203	0.584	0.235	0.618
Italian	55,057	0.598	0.859	0.623	0.887	0.441	0.779	0.478	0.797
Portuguese	9,641	0.465*	0.784	0.454*	0.773	0.346*	0.695	0.361*	0.709
Romanian	26,950	0.548	0.833	0.560	0.837	0.413	0.776	0.483	0.807
Russian	45,249	0.512	0.833	0.530	0.837	0.329	0.751	0.357	0.765
Serbian	31,548	0.585	0.860	0.591	0.861	0.440	0.794	0.479	0.810
Spanish	27,600	0.579	0.856	0.586	0.858	0.255	0.661	0.274	0.682
Swedish	27,674	0.616*	0.867	0.625*	0.870	0.382	0.749	0.428	0.782
Turkish	13,609	0.650	0.865	0.632	0.865	0.324	0.720	0.396	0.754

Table 1: Performances of language specific systems and generic system on different language sets

Languages	Size of	Swedish Language Specific System				Generic System			
		Baseline		Modified		Baseline		Modified	
	testing set	ACC	MFS	ACC	MFS	ACC	MFS	ACC	MFS
Danish	851	0.286	0.684	0.294	0.696	0.208	0.631	0.228	0.664
Norwegian	2,189	0.303	0.717	0.301	0.718	0.222	0.652	0.255	0.687

Table 2: Performances of the language specific system for Swedish and the generic system on Danish and Norwegian

"x" is the only letter that sometimes needs to be transliterated into two Chinese characters. Our system cannot transliterate a single letter on the source side into multiple Chinese characters. In this research, we use a simple solution to the problem. We replace the letter "x" by "ks" in the source names before the decoding system transliterates them.

The integrated modified system achieves significant improvements. The ACC is increased to 0.370 and Mean F-score reaches 0.714. The precision of the modified transliteration system is comparable to state-of-the-art machine transliteration systems that are submitted in NEWS 2010 (Li et al., 2010a).

5. Experiments on Different Language Sets

Both our baseline system and modified system are applied on 14 different language sets. Using the different training data, we build distinct language specific systems which are dedicated to certain languages. Meanwhile we train a generic system using the assembled training data to compare with the language specific systems. The detailed results are shown in Table 1. In general, the language specific systems significantly outperform the generic systems. The language specific systems are particularly effective on some languages such as German, Spanish, Italian and Russian. Compared to the baseline transliteration system, our modified system performs better on most language sets.

In order to further investigate the significance of differences between the precisions of these different machine transliteration systems, we performed the Wilcoxon signed-rank test (Wilcoxon, 1945) on the ACC scores that are evaluated on our testing data. The test result shows that the differences between the language specific systems and generic system are all highly statistically significant with p-values far below any critical value. The majority of the differences between the baseline system and modified system are significant as well. We mark those pairs that fail to reject H_0 when $p = 0.05$ with "*" in Table 1

Note that the language specific systems are also more efficient because of the smaller search space for the decoder. We also notice that some languages are more difficult for the transliteration to process, for examples English, French and Hungarian probably due to language-specific properties of orthography and pronunciation.

On the other hand, even though the language specific systems show considerable superiorities, they also have some disadvantages. The language specific systems are limited on certain language sets, and therefore sometimes a highly accurate classifier is required to determine the language origin first in practical transliteration problems. Another experiment on Danish and Norwegian using the Swedish transliteration system indicates that systems for closely related languages are reasonable alternatives in case of insufficient training data. The results are shown in Table 2.

6. Summary

Overall, the experimental results on our machine transliteration models have reached our initial expectation. We have successfully built the transliteration system and made it applicable on different languages with satisfying precisions. Our research indicates that specifying language origins in machine transliteration is significant.

References

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April. Association for Computational Linguistics.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Computing Surveys (CSUR)*, 43, April.

Haizhou Li, A. Kumaranz, Min Zhang, and Vladimir Pervouchine. 2010a. Report of NEWS 2010 transliteration generation shared task. In *NEWS '10 Proceedings of the 2010 Named Entities Workshop: Shared Task on Transliteration*, Uppsala, Sweden.

Haizhou Li, A. Kumaranz, Min Zhang, and Vladimir Pervouchine. 2010b. Whitepaper of NEWS 2010 shared task on transliteration generation. In *NEWS '10 Proceedings of the 2010 Named Entities Workshop: Shared Task on Transliteration*, Uppsala, Sweden.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December.

Defu Xia. 1993. *Translation Dictionary for Foreign Names*. China Translation and Publishing Corporation, Beijing, China, October.