Statistical Machine Transliteration with Multi-to-Multi Joint Source Channel Model

Yu Chen, Rui Wang, Yi Zhang

Language Technology Lab

German Research Center for Artificial intelligence (DFKI)

Saarbrücken, Germany

{firstname.lastname}@dfki.de

Abstract

This paper describes DFKI's participation in the NEWS2011 shared task on machine transliteration. Our primary system participated in the evaluation for English-Chinese and Chinese-English language pairs. We extended the joint source-channel model on the transliteration task into a multi-to-multi joint source-channel model, which allows alignments between substrings of arbitrary lengths in both source and target strings. When the model is integrated into a modified phrase-based statistical machine translation system, around 20% of improvement is observed. The primary system achieved 0.320 on English-Chinese and 0.133 on Chinese-English in terms of top-1 accuracy.

1 Introduction

Machine transliteration has drawn a lot of attention in the previous years. In particular, the previous two shared tasks (Li et al., 2009; Li et al., 2010) attracted more than 30 participants. This year's task only focuses on the transliteration generation task. As our first attempt in this area, we participated in English-to-Chinese transliteration (En-Ch) and Chinese-to-English back transliteration (Ch-En) tasks.

For En-Ch and Ch-En transliterations, there was a discussion on whether to use the intermediate phonemic interpretation, i.e., Pinyin. Li et al. (2004) showed empirically that by skipping the intermediate phonemic interpretation (denoted as grapheme-based methods), the transliteration error rate was reduced significantly, since the mapping between Pinyin and Chinese characters was not trivial. Oh et al. (2009) had a more generalized version of Li et al. (2004)'s system as well as other previous work (e.g., (Knight and Graehl, 1998), denoted as phoneme-based methods) and showed that incorporating Pinyin as one of the features did help the transliteration performance finally. Li et al. (2007) included two other useful features, language of origin and the gender association. This is our first participation of this shared task, instead of considering the "best" setting, we aim at a basic but extensible architecture at first.

2 Systems

Transliteration can be viewed as a special case of the translation task, namely translation at a character level. State-of-the-art statistical machine translation systems were reported as being able to deliver satisfactory results for the transliteration task without additional knowledge on the languages (Knight and Graehl, 1998). However, general statistical machine translation systems do not consider the key features of the transliteration task, which, on the other hand, have been emphasized by the joint source channel models.

Our primary system is a standard phrase-based statistical machine translation (PBSMT) system with a modification based on the Multi-to-Multi Joint Source Channel model. We hope the combination could benefit from the simplicity of a joint source channel model without losing the flexibility of the PBSMT system.

2.1 Phrase-based SMT

The basic architecture of a phrase-based SMT system is an instance of the noisy-channel approaches (Brown et al., 1993). In the context of transliteration, the term "phrase" in phrase-based SMT would refer to a sequence of characters chosen by its statistical rather then any grammatical properties. The transliteration of a name s in the source language into a name t in the target language is modeled as:

Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages 101–105,
Chiang Mai, Thailand, November 12, 2011.

$$\arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{s}) = \arg\max_{\mathbf{t}}(P(\mathbf{t})P(\mathbf{s}|\mathbf{t}));$$

The system involves a *phrase table*, a list of character sequences identified in a source name together with potential transliterations. These sequences derived from the source names may overlap and also have several correspondences in the target language. The process of searching for the target names starts with selecting a subset of the entries in the table. The members of the selected subset must then be arranged in a specific order to give a translation. These operations are determined by statistical properties of the target language enshrined in the so-called *language model*.

The segments in the source name and their counterparts in the target language should always be exactly in the same order, which is clearly not the case for general machine translation tasks. In addition to ordering, there are many other strict rules such that the transliteration task is relatively more deterministic than the translation process. For instance, although it is common that many Chinese characters have the same pronunciation, only a small set of Chinese characters can be used in the transliterated western names. Accordingly, for each source name, there are only a limited set of candidate transliterations, unlike the infinite target set for the general translation task.

It is critical to take into account these characteristics mentioned above when utilizing an SMT system for transliteration. First, the distortion model, one of the major components in a standard PBSMT system, is redundant for transliteration. Including the unnecessary model expands the search space and makes it more difficult to find the good candidates. Second, the word alignment model (Och and Ney, 2004) in a PBSMT system also assumes flexible ordering of correspondence to some extent. This could introduce additional noise to the translation models if applied directly to transliteration tasks without any modifications.

2.2 M2M Jonit Source-Channel Model

The joint source-channel machine transliteration model (Li et al., 2004) calculates the n-gram transliteration probability. More specifically, for a source name s, a target transliteration t, and an alignment α between the source and the target, we have the transliteration probability defined as:

$$P(s,t,\alpha) = \sum_{k=1}^{K} P(< e,c >_k \mid < e,c >_{k-n+1}^{k-1}) \tag{1}$$

where $< e,c >_k$ is the k^{th} aligned pair of translation units. Therefore, forward and backward transliteration can be uniformly obtained by (2) and (3).

$$\bar{t} \;=\; \underset{s,\alpha}{\mathrm{argmax}}\, P(s,t,\alpha) \tag{2}$$

$$\bar{s} \;=\; \underset{t,\alpha}{\mathrm{argmax}}\, P(s,t,\alpha) \tag{3}$$

The alignment statistics can be obtained with an Expectation-Maximization procedure over the training corpus.

For English-Chinese bidirectional transliteration, Li et al. (2004) assumed that each Chinese character aligns with a sequence of one or more letters in English. This assumption drastically reduces the number of possible alignments. For a English source s and a Chinese target t, the number of possible alignment under this assumption is

$$\binom{|s|-1}{|t|-1} = \frac{(|s|-1)!}{(|t|-1)!(|s|-|t|)!}$$

While the assumption holds true in most of the cases, several obvious limitations arise. First, it is assumed that the source string is at least as long as the target which is not necessary true. Second, and more importantly, in some cases multiple Chinese characters should align with one single English letter (for example ‘X’), and in others, multiple Chinese characters constitute one single transliteration unit. Therefore, instead of adopting the “one Chinese character per unit” assumption, we allow alignments between substrings of arbitrary lengths in both the source and the target. We call this a Multi-to-Multi Joint Source-Channel model (M2M-JSC). This constitutes a much larger model, with more possible transliteration units on the Chinese side. To simplify the calculation, we use the 1-gram model for the calculation of the transliteration probability, and hope that the larger transliteration units to compensate for the Markovian effect of mutual dependencies between alignment pairs. We use the similar Expectation-Maximization procedure to train the model on the corpus. One slight variation from Li et al. (2004) is that instead of choosing a random segmentation in the initialization step, we generate all possible multi-to-multi alignment hypotheses, and normalize the counts by the number of hypotheses of each transliteration pair. The segmentation alignment obtained is significantly different from the original Joint Source-Channel model. Table 1 shows some examples of the M2M-JSC alignment.

English	Chinese
A/JA/X	埃/甲/克斯
A/BA/STE/NIA	阿/巴/斯蒂/尼亚
AHL/BERG	阿尔/伯格

Table 1: Examples of M2M Joint Source-Channel Alignment Result

2.3 Combined system

In order to benefit from both previous described components, the M2MJSC model is integrated into the PBSMT system as a substitute of the translation model. Figure 1 illustrates the structure of the combined system.

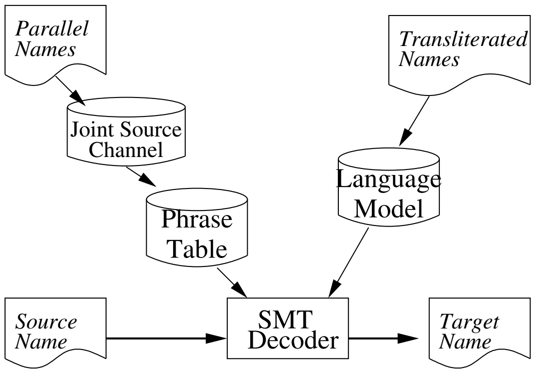

Figure 1: Phrase-based Transliteration System with Joint Source Channel Model

M2MJSC is first applied to the training set to divide each source name in parallel with the corresponding target name into the same number of segments. These segments are then considered as words that are one-to-one aligned. The PBSMT system takes multiple segments, namely phrases, as translation units. The phrase extraction follows the heuristic that starts with the given word alignment and expands to the adjacent alignment points (Koehn et al., 2003). The translation probabilities of the extracted phrases are estimated accordingly.

As the last step, we split all the segments in the translation model into characters to allow more straightforward integration into the original PBSMT system that relies on character based inputs.

3 Experiment setup

3.1 Preprocessing

We worked with the English data only in the uppercase form as provided in the training set. The names are tokenized into characters, but we did not perform any further phonetic mapping for both languages as the phonetic mapping requires additional knowledge which was not available in the training data.

Even though it is possible to combine the training sets for both English-to-Chinese and Chinese-to-English, we restrained ourselves to the set that are designated for the particular direction. In other words, the Chinese-to-English training set was not included for training of all the components of our English-to-Chinese system and vice versa.

3.2 SMT system for transliteration

3.2.1 Statistical models

Our system consists the following major statistical components:

- An n-gram language model;
- A translation model, including two phrase translation probabilities (both directions), two lexical weightings (both directions) induced from word translation probabilities, and a phrase penalty. This model is further decomposed into phrases;
- Word penalty used to penalize longer hypotheses.

The n-gram language model is estimated using the SRILM toolkit (Stolcke, 2002). The translation model is built from the character alignments given the M2MJSC model and we did not construct any distortion models.

3.2.2 Moses decoder

We used the open-source SMT decoder Moses (Koehn et al., 2007). Moses allows a log-linear model to combine various models and implements an efficient beam search algorithm that quickly finds the best translation among the large number of hypotheses. In order to adapt the SMT decoder to the transliteration task, we not only supplied the decoder with no reordering models, but also constrained the decoder in a monotone manner by setting distortion limit to 0.

Tasks	System	ACC	Mean F	MRR	Map_ref
English-to-Chinese	M2MJC+PBSMT	0.320	0.674	0.397	0.308
English-to-Chinese	M2MJC	0.260	0.638	0.340	0.251
Chinese-to-English	M2MJC+PBSMT	0.133	0.746	0.210	0.133
Chinese-to-English	M2MJC	0.117	0.731	0.177	0.117

Table 2: Official results

3.2.3 Parameter tuning

The system integrates all the models into a more complex discriminative model in a log linear formulation. The weights for the individual models can be optimized on development data so that the system outputs are as close as possible to correct candidates. Minimum error rate training (MERT) (Och, 2003) is one of the common method for balancing between features on different bases. We used Z-MERT (Zaidan, 2009) to search for the set of feature weights that maximizes the official f-score evaluation metric on the development set.

Moreover, we extracted a small development set of 500 names randomly from the official development set. The rest of the official development set served as a development test set, so we could run additional experiments on the provided data set apart from our submission. The feature weights we used for our submission are obtained from the complete development set.

4 Results

We participated in English-to-Chinese and Chinese-to-English transliteration tasks in NEWS2011. Table 2 lists the official evaluation scores for our submission to these two tracks. Our contrast system is the stand-alone M2MJSC system. It is clear that the final combined system has outperformed the M2MJSC system by around 20% for both directions.

We notice that there is a group of multi-word names in the development set that are particularly difficult for our system to transliterate correctly. Most of these names consists of parts that should be translated by the meanings instead of transliterated by the phonemes, for example, "DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA". To handle such cases, we need to include additional recognition and translation modules that clearly require knowledge beyond the provided training data set.

5 Conclusion

We successfully participated in this year's En-Ch and Ch-En machine transliteration shared tasks. We extended the original joint source-channel model proposed by Li et al. (2004) by allowing more possible transliteration units than single characters (in Chinese) and single letters (in English). When the M2M-JSC model is integrated into a modified phrase-based SMT system, around 20% of improvement is observed. In the future, we will further explore the M2M-JSC model with richer feature sets as well as the integration of other SMT approaches.

Acknowledgments

The first two authors were supported by the EuroMatrixPlus project (IST-231720) funded by the European Community under the Seventh Framework Programme for Research and Technological Development. The last was supported by the German Excellence Cluster of Multimodal Computing and Interaction.

References

Peter Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Comput. Linguist.*, 24:599–612, December.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses:

Open Source Toolkit for Statistical Machine Translation. In *the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, June.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *The 42nd Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Barcelona, Spain, July.

Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong. 2007. Semantic transliteration of personal names. In *The 45th Annual Meeting of Association for Computational Linguistics*, pages 120–127, Prague, Czech Republic, June.

Haizhou Li, A Kumaran, Vladimir Pervouchine, and Min Zhang. 2009. Report of news 2009 machine transliteration shared task. In *Proceedings of the 2009 Named Entities Workshop*, pages 1–18, Singapore, Singapore, August.

Haizhou Li, A Kumaran, Min Zhang, and Vladimir Pervouchine. 2010. Report of news 2010 transliteration generation shared task. In *Proceedings of the 2010 Named Entities Workshop*, pages 1–11, Uppsala, Sweden, July.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Can chinese phonemes improve machine transliteration?:a comparative study of english-to-chinese transliteration models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 658–667, Singapore, Singapore, August.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado.

Omar F. Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.