English to Hindi Machine Transliteration System at NEWS 2009

Amitava Das, Asif Ekbal, Tapabrata Mandal and Sivaji Bandyopadhyay
Computer Science and Engineering Department
Jadavpur University, Kolkata-700032, India
amitava.research@gmail.com, asif.ekbal@gmail.com, tapabratamondal@gmail.com, sivaji_cse_ju@yahoo.com

Abstract

This paper reports about our work in the NEWS 2009 Machine Transliteration Shared Task held as part of ACL-IJCNLP 2009. We submitted one standard run and two non-standard runs for English to Hindi transliteration. The modified joint source-channel model has been used along with a number of alternatives. The system has been trained on the NEWS 2009 Machine Transliteration Shared Task datasets. For standard run, the system demonstrated an accuracy of 0.471 and the mean F-Score of 0.861. The non-standard runs yielded the accuracy and mean F-scores of 0.389 and 0.831 respectively in the first one and 0.384 and 0.828 respectively in the second one. The non-standard runs resulted in substantially worse performance than the standard run. The reasons for this are the ranking algorithm used for the output and the types of tokens present in the test set.

1 Introduction

Technical terms and named entities (NEs) constitute the bulk of the Out Of Vocabulary (OOV) words. Named entities are usually not found in bilingual dictionaries and are very generative in nature. Proper identification, classification and translation of Named entities (NEs) are very important in many Natural Language Processing (NLP) applications. Translation of NEs involves both translation and transliteration. Transliteration is the method of translating into another language by expressing the original foreign word using characters of the target language preserving the pronunciation in their source language. Thus, the central problem in transliteration is predicting the pronunciation of the original word. Transliteration between two languages that use the same set of alphabets is trivial: the word is left as it is. However, for languages those use different alphabet sets the names must be transliterated or rendered in the target language alphabets. Transliteration of NEs is necessary in many applications, such as machine translation, corpus alignment, cross-language Information Retrieval, information extraction and automatic lexicon acquisition. In the literature, a number of transliteration algorithms are available involving English (Li et al., 2004; Vigra and Khudanpur, 2003; Goto et al., 2003), European languages (Marino et al., 2005) and some of the Asian languages, namely Chinese (Li et al., 2004; Vigra and Khudanpur, 2003), Japanese (Goto et al., 2003; Knight and Graehl, 1998), Korean (Jung et al., 2000) and Arabic (Al-Onaizan and Knight, 2002a; Al-Onaizan and Knight, 2002c). Recently, some works have been initiated involving Indian languages (Ekbal et al., 2006; Ekbal et al., 2007; Surana and Singh, 2008).

2 Machine Transliteration Systems

Three transliteration models have been used that can generate the Hindi transliteration from an English named entity (NE). An English NE is divided into Transliteration Units (TUs) with patterns C*V*, where C represents a consonant and V represents a vowel. The Hindi NE is divided into TUs with patterns C+M?, where C represents a consonant or a vowel or a conjunct and M represents the vowel modifier or matra. The TUs are the lexical units for machine transliteration. The system considers the English and Hindi contextual information in the form of collocated TUs simultaneously to calculate the plausibility of transliteration from each English TU to various Hindi candidate TUs and chooses the one with maximum probability. This is equivalent to choosing the most appropriate sense of a word in the source language to identify its representation in the target language. The system learns the mappings automatically from the bilingual NEWS training set being guided by lin-

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 80–83,
Suntec, Singapore, 7 August 2009.

guistic features/knowledge. The system considers the linguistic knowledge in the form of conjuncts and/or diphthongs in English and their possible transliteration in Hindi. The output of the mapping process is a decision-list classifier with collocated TUs in the source language and their equivalent TUs in collocation in the target language along with the probability of each decision obtained from the training set. Linguistic knowledge is used in order to make the number of TUs in both the source and target sides equal. A Direct example base has been maintained that contains the bilingual training examples that do not result in the equal number of TUs in both the source and target sides during alignment. The Direct example base is checked first during machine transliteration of the input English word. If no match is obtained, the system uses direct orthographic mapping by identifying the equivalent Hindi TU for each English TU in the input and then placing the Hindi TUs in order. The transliteration models are described below in which S and T denotes the source and the target words respectively:

● Model A

This is essentially the joint source-channel model (Hazhiou et al., 2004) where the previous TUs with reference to the current TUs in both the source (s) and the target sides (t) are considered as the context.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid < s,t >_{k-1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

● Model B

This is basically the trigram model where the previous and the next source TUs are considered as the context.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid s_{k-1}, s_{k+1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

● Model C

In this model, the previous and the next TUs in the source and the previous target TU are considered as the context. This is the improved modified joint source-channel model.

$$P(S \mid T) = \prod_{k=1}^{K} P(< s,t >_k \mid < s,t >_{k-1,}\ s_{k+1})$$

$$S \rightarrow T(S) = \arg\max_T \{P(T) \times P(S \mid T)\}$$

For NE transliteration, P(*T*), i.e., the probability of transliteration in the target language, is calculated from a English-Hindi bilingual database of approximately 961,890 English person names, collected from the web[1]. If, *T* is not found in the dictionary, then a very small value is assigned to P(*T*). These models have been desribed in details in Ekbal et al. (2007).

● Post-Processing

Depending upon the nature of errors involved in the results, we have devised a set of transliteration rules. A few rules have been devised to produce more spelling variations. Some examples are given below.

Spelling variation rules

Badlapur बदलापुर | वदलापुर

Shree | Shri श्री

3 Experimental Results

We have trained our transliteration models using the English-Hindi datasets obtained from the NEWS 2009 Machine Transliteration Shared Task (Li et al., 2009). A brief statistics of the datasets are presented in Table 1. Out of 9975 English-Hindi parallel examples in the training set, 4009 are multi-words. During training, we have split these multi-words into collections of single word transliterations. It was observed that the number of tokens in the source and target sides mismatched in 22 multi-words and these cases were not considered further. Following are some examples:

Paris Charles de Gaulle पेरिस रॉसे चार्ल्स डे ग्यूले

South Arlington Church of Christ साउथ अर्लिंग्टन

In the training set, some multi-words were partly translated and not transliterated. Such examples were dropped from the training set. Finally, the training set consists of 15905 single word English-Hindi parallel examples.

[1]http://www.eci.gov.in/DevForum/Fullname.asp

Set	Number of examples
Training	9975
Development	974
Test	1000

Table 1. Statistics of Dataset

The output of the modified joint source-channel model is given more priority during output ranking followed by the trigram and the joint source-channel model. During testing, the *Direct example base* is searched first to find the transliteration. Experimental results on the development set yielded the accuracy of 0.442 and mean F-score of 0.829. Depending upon the nature of errors involved in the results, we have devised a set of transliteration rules. The use of these transliteration rules increased the accuracy and mean F-score values up to 0.489 and 0.881 respectively.

The system has been evaluated for the test set and the detailed reports are available in Li et al. (2009). There are 88.88% unknown examples in the test set. We submitted one standard run in which the outputs are provided for the modified joint source-channel model (Model C), trigram model (Model B) and joint source-channel model (Model A). The same ranking procedure (i.e., Model C, Model B and Model A) has been followed as that of the development set. The output of each transliteration model has been post-processed with the set of transliteration rules. For each word, three different outputs are provided in a ranked order. If the outputs of any two models are same for any word then only two outputs are provided for that particular word. Post-processing rules generate more number of possible transliteration output. Evaluation results of the standard run are shown in Table 2.

Parameters	Accuracy
Accuracy in top-1	0.471
Mean F-score	0.861
Mean Reciprocal Rank (MRR)	0.519
Mean Average Precision $(MAP)_{ref}$	0.463
MAP_{10}	0.162
MAP_{sys}	0.383

Table 2. Results of the standard run

The results of the two non-standard runs are presented in Table 3 and Table 4 respectively.

Parameters	Accuracy
Accuracy in top-1	0.389
Mean F-score	0.831
Mean Reciprocal Rank (MRR)	0.487
Mean Average Precision $(MAP)_{ref}$	0.385
MAP_{10}	0.16
MAP_{sys}	0.328

Table 3. Results of the non-standard run 1

Parameters	Accuracy
Accuracy in top-1	0.384
Mean F-score	0.823
Mean Reciprocal Rank (MRR)	0.485
Mean Average Precision $(MAP)_{ref}$	0.380
MAP_{10}	0.16
MAP_{sys}	0.325

Table 4. Results of the non-standard run2

In both the non-standard runs, we have used an English-Hindi bilingual database of approximately 961, 890 examples that have been collected from the web[2]. This database contains the (frequency) of the corresponding English-Hindi name pair. Along with the outputs of three models, the output obtained from this bilingual database has been also provided for each English word. In the first non-standard run, only the most frequent transliteration has been considered. But, in the second non-standard run all the possible transliteration have been considered. It is to be noted that in these two non-standard runs, the transliterations obtained from the bilingual database have been kept first in the ranking. Results of the tables show quite similar performance in both the runs. But the non-standard runs resulted in substantially worse performance than the standard run. The reasons for this are the ranking algorithm used for the output and the types of tokens present in the test set. The additional da-

[2]http://www.eci.gov.in/DevForum/Fullname.asp

taset used for the non-standard runs is mainly census data consisting of only Indian person names. The NEWS 2009 Machine Transliteration Shared Task training set is well distributed with foreign names (Ex. Sweden, Warren), common nouns (Mahfuz, Darshanaa) and a few non named entities. Hence the training set for the non-standard runs was biased towards the Indian person name transliteration pattern. Additional training set was quite larger (961, 890) than the shared task training set (9,975). Actually outputs of non-standard runs have more alternative transliteration outputs than the standard set. That means non-standard sets are superset of standard set. Our observation is that the ranking algorithm used for the output and biased training are the main reasons for the worse performance of the non-standard runs.

4 Conclusion

This paper reports about our works as part of the NEWS 2009 Machine Transliteration Shared Task. We have used the modified joint source-channel model along with two other alternatives to generate the Hindi transliteration from an English word (to generate more spelling variations of Hindi names). We have also devised some post-processing rules to remove the errors. During standard run, we have obtained the word accuracy of 0.471 and mean F-score of 0.831. In non-standard rune, we have used a bilingual database obtained from the web. The non-standard runs yielded the word accuracy and mean F-score values of 0.389 and 0.831 respectively in the first run and 0.384 and 0.823 respectively in the second run.

References

Al-Onaizan, Y. and Knight, K. 2002a. Named Entity Translation: Extended Abstract. In *Proceedings of the Human Language Technology Conference,* 122– 124.

Al-Onaizan, Y. and Knight, K. 2002b. Translating Named Entities using Monolingual and Bilingual Resources. In *Proceedings of the 40th Annual Meeting of the ACL*, 400–408, USA.

Ekbal, A. Naskar, S. and Bandyopadhyay, S. 2007. Named Entity Transliteration. *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, Volume (20:4), 289-310, World Scientific Publishing Company, Singapore.

Ekbal, A., Naskar, S. and Bandyopadhyay, S. 2006. A Modified Joint Source Channel Model for Transliteration. In *Proceedings of the COLING-ACL 2006*, 191-198, Australia.

Goto, I., Kato, N., Uratani, N. and Ehara, T. 2003. Transliteration Considering Context Information based on the Maximum Entropy Method. In *Proceeding of the MT-Summit IX*, 125–132, New Orleans, USA.

Jung, Sung Young , Sung Lim Hong and Eunok Paek. 2000. An English to Korean Transliteration Model of Extended Markov Window. In *Proceedings of International Conference on Computational Linguistics (COLING 2000)*, 383-389.

Knight, K. and Graehl, J. 1998. Machine Transliteration, *Computational Linguistics,* Volume (24:4), 599–612.

Kumaran, A. and Tobias Kellner. 2007. A generic framework for machine transliteration. In *Proc. of the 30th SIGIR.*

Li, Haizhou, A Kumaran, Min Zhang and Vladimir Pervouchine. 2009. Whitepaper of NEWS 2009 Machine Transliteration Shared Task. In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop (NEWS 2009)*, Singapore.

Li, Haizhou, A Kumaran, Vladimir Pervouchine and Min Zhang. 2009. Report on NEWS 2009 Machine Transliteration Shared Task. In *Proceedings of ACL-IJCNLP 2009 amed Entities Workshop (NEWS 2009)*, Singapore.

Li, Haizhou, Min Zhang and Su Jian. 2004. A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of the 42nd Annual Meeting of the ACL*, 159-166. Spain.

Marino, J. B., R. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa and M. Ruiz. 2005. Bilingual n-gram Statistical Machine Translation. In *Proceedings of the MT-Summit X,* 275–282.

Surana, Harshit, and Singh, Anil Kumar. 2008. A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing* (*IJCNLP-08*), 64-71, India.

Vigra, Paola and Khudanpur, S. 2003. Transliteration of Proper Names in Cross-Lingual Information Retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition*, 57–60.