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

Abstract

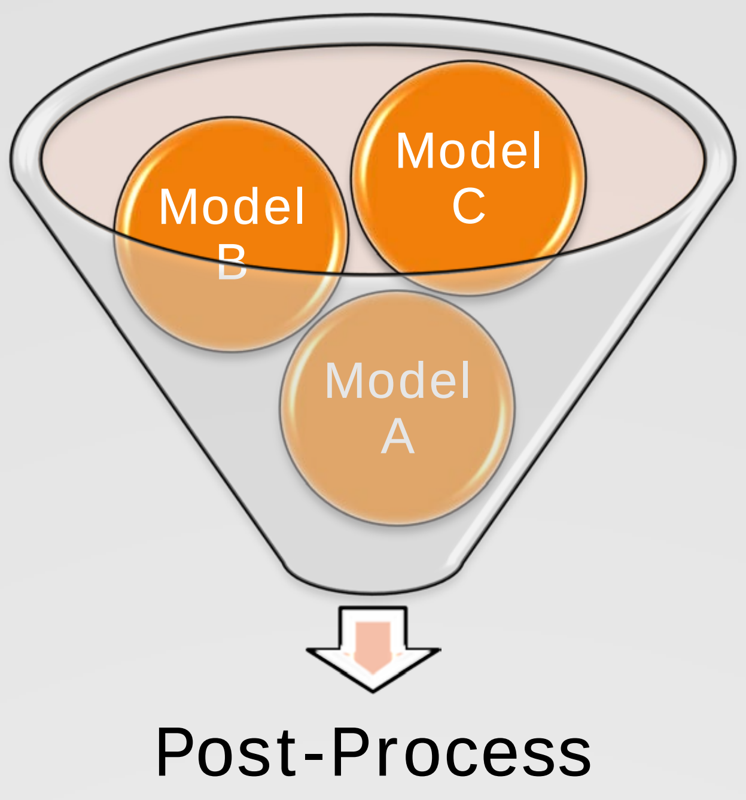
This paper reports about our work in the NEWS 2009 Machine Transliteration Shared Task held as part of ACL-IJCNLP 2009. We submitted one standard run and two non-standard runs for English to Hindi transliteration. The modified joint source channel model has been used along with a number of alternatives. 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.

Fi | r | dau| si
फ़ि| र |दौ| सी

Transliteration Unit

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 linguistic features/knowledge.



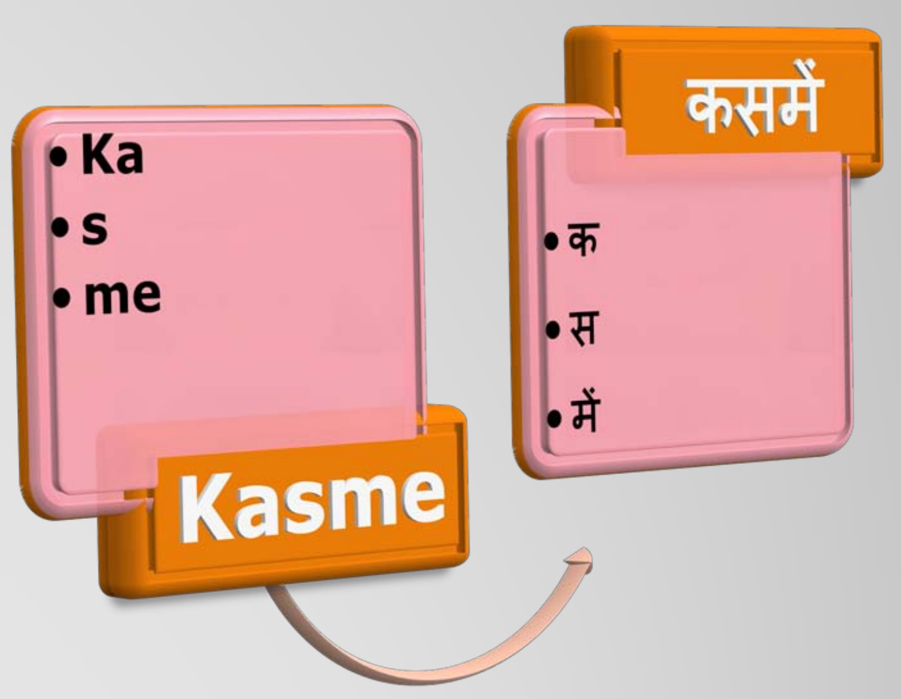
System Architecture

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.

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

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



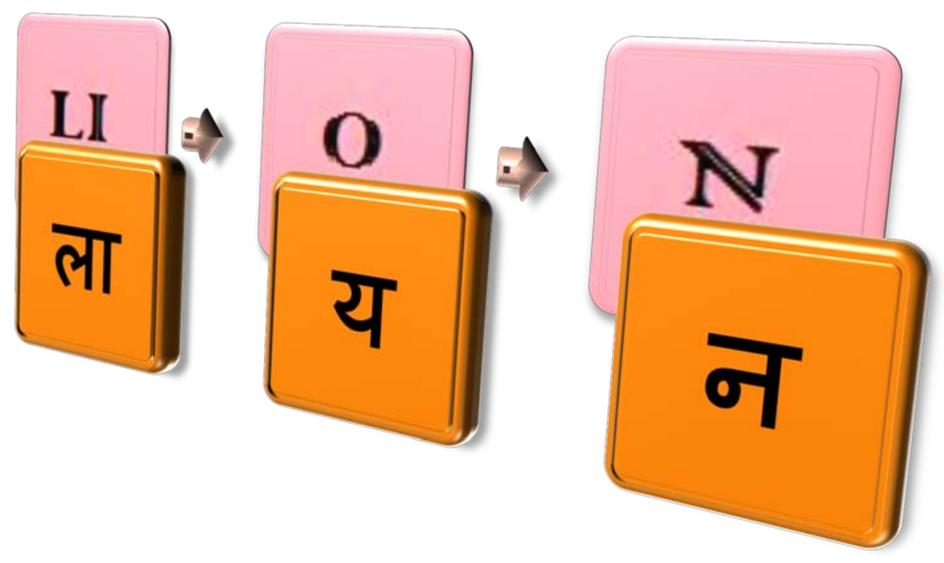
Joint Source Channel Model

Model B

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

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

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



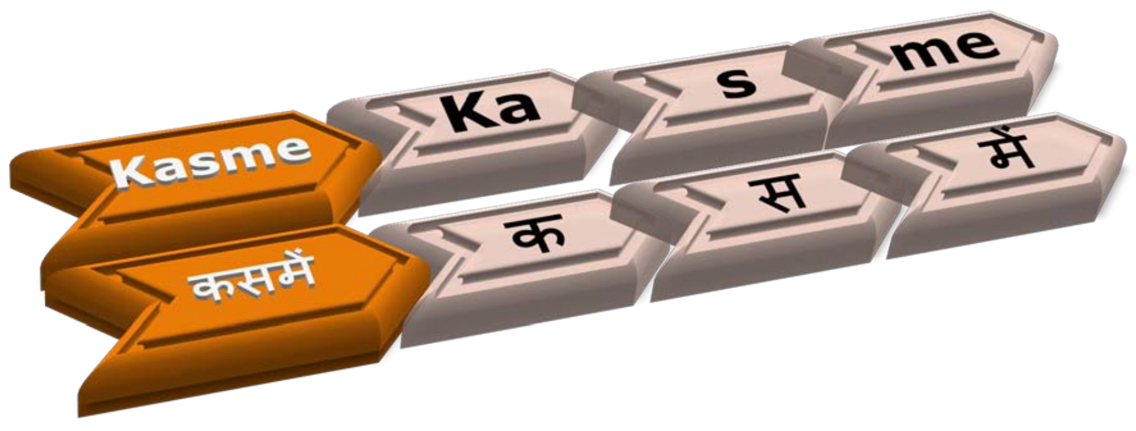
Model B

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. 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 . 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).

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

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



Model C

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:.

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

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 non-standard run 1

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 non-standard run 2

Post-Process

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

Post-Process Rules

Conclusion

But the non-standard runs resulted in substantially worse performance than the stan-dard 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 dataset 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).

Authors	Email
Amitava Das	amitava.research@gmail.com
Asif Ekbal	asif.ekbal@gmail.com
Tapabrata Mandal	tapabratamondal@gmail.com
Sivaji Bandyopadhyay	sivaji_cse_ju@yahoo.com