# Query word labeling and Back Transliteration for Indian Languages: Shared task system description

**Spandana Gella**, **Jatin Sharma**, **Kalika Bali**
**Microsoft Research India**
{t-spgell, jatinsha, kalikab}@microsoft.com

**Abstract**

In this paper, we target the problem of word level language identification for Indian languages written in roman script and mixed with English language. In addition to the language identification we also handle transliteration of Indian language words into the native Indic scripts. To address these problems we consider a supervised approach of building a classifier with monolingual samples together with a context-switching probability from Indian Language (IL) to English (Eng). The proposed system is submitted to the FIRE-2013 shared task on Transliterated Search. Our submitted system showed the best performing results.

## 1 Introduction

Code-mixing and transliteration are common phenomena observed on computer mediated communication in languages that use non-roman script especially those from the Arabic, Asian and African regions. It presents serious challenges to translation, understanding, search etc. Therefore, we require a back-transliteration [1] tool that can handle spelling variations, errors, homophonic words and canonical forms. Such a tool would be a helpful as a native language input method tool for QWERTY keyboards [2]. The FIRE shared task on transliterated search targets query labeling i.e. labeling the individual words with their original language and then to back-transliterate non-English words to their native scripts. A typical task is illustrated in Table 1. In this paper we propose two different ways of query word labeling using i) character n-grams approach, and ii) coupling character n-grams approach with context switch probability.

The paper is organized as follows - we describe the dataset for training, development and evaluation in Section 2, our method and experimental setup in Section 3, we present our results in Section 4 and error analysis in Section 5. Finally, we conclude with listing possibility of future work in Section 6.

## 2 Dataset

To build the query word labeling system we used the word files along with frequency in Hindi, Gujarati and Bangla in their native scripts collected from a monolingual corpus, Roman transliterations provided to us as part of the FIRE-2013 shared task[1]. We have experimented with different number of training samples in each of the languages.

[1] http://cse.iitkgp.ac.in/resgrp/cnerg/qa/firetranslit/#subtask1

| Input | Query Labeling | Back-Transliteration |
|---|---|---|
| sachin tendulkar number of centuries | sachin\H tendulkar\H number\E of\E centuries\E | सचिन तेंदुलकर number of centuries |
| palak paneer recipe | palak\H paneer\H recipe\E | पालक पनीर recipe |
| mungeri lal ke haseen sapney | mungeri\H lal\H ke\H haseen\H sapney\H | मुंगेरी लाल के हसीन सपने |
| iguazu water fall argentina | iguazu\E water\E fall\E argentina\E | iguazu water fall argentina |

Table 1: Shared Task description in two seperate steps of query labeling and back translieration

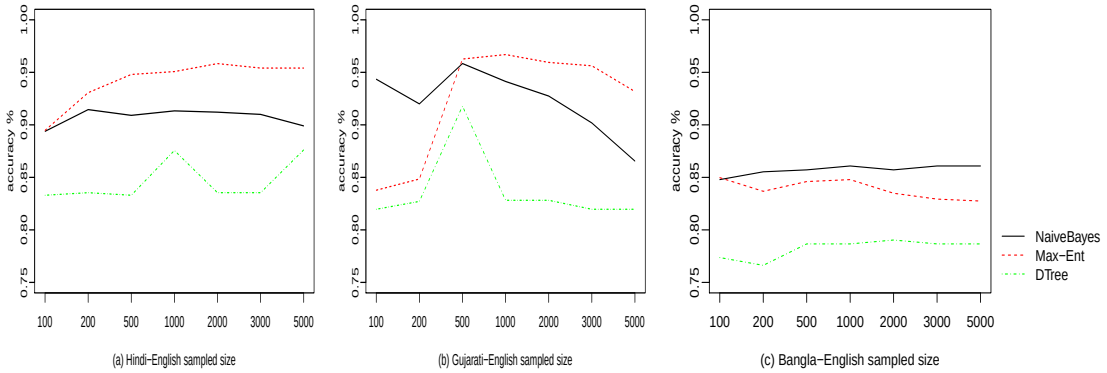



Figure 1: Learning curves for NaiveBayes, MaxEnt and DecisionTree on word labeling for Hindi, Gujarati and Bangla language on development data

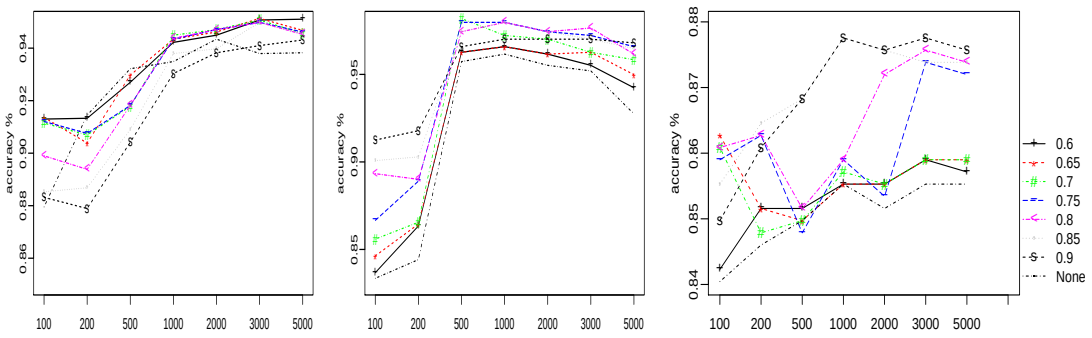
## 2.1 Features

We trained our classifier with combinations of character unigram, bigram, trigram, 4-grams, 5-gram and full word as features. This was inspired from the (King&Abney-2013)[3] work on word level language classification. The classifiers using full set of features listed above performed well when compared to different combination of features, for brevity we describe the systems using full set of features. The Figure 1 shows the learning curves for all classifiers as the number of training samples varied from $100 - 5000$.

## 2.2 Methodology

Using all available features, we compared three classifiers: Naive Bayes, Maximum Entropy and Decision Tree using MALLET [4] toolkit. We observed that a combination of all character 1-5grams performed well compared to individual feature sets. Learning curves for each of the classifier with the size of the training sample is presented in Figure 1. The figure shows that maximum entropy classifiers performed well on Hindi and Gujarati whereas for Bangla Naive Bayes classifier showed the best results.

We further explored the influence of context switching probability on the learning curves for MaxEnt (Hindi and Gujarati) and NaiveBayes (Bangla) classifiers. As depicted by Figure 2 use of a context switching probablity on top of n-grams improves the performance. From this curve we also obtain the optimal values of context switch probabilities for all three languages which are listed in Table 2.




(a) Hindi - MaxEnt (b) Gujarati - MaxEnt (c) Bangla - NaiveBayes

Figure 2: Learning curves for Hindi, Gujarati and Bangla on MaxEnt and NaiveBayes with varying context switch probabilities and sampling size on development data

# 3 System Description

In this section we present the details of our feature set, the training (classification) algorithms, tools used and assumptions made executing the experiments. Three official runs `MSRI-1`, `MSRI-2` and `MSRI-3` were submitted for 'FIRE SharedTask on Transliterated Search subtask-1' based on the performance on development set. The systems `MSRI-2` and `MSRI-3` use the context switch probability, monolingual frequency factor on top of classifier output to assign a language label to a word whereas `MSRI-1` uses the classifier output + frequency factor. These three systems are chosen for each of the language based on the best results achieved on the development set. The development results are presented in Table 2.

We used a hash function based MSRI Name Search [5] tool trained on our Hindi word-list to back-transliterate Hindi words. As all three languages belong to Indo-Aryan family and both Gujarati and Bangla show pronunciation similarity to Hindi, we used the same tool for both the languages in `MSRI-1` and `MSRI-2` to find a Hindi transliteration which was then converted to Gujarati or Bangla using Indic character mapping. To further improve the accuracy of the tool for Bangla in `MSRI-3` we converted Bangla word-list into Hindi using Indic character mapping and appended it to training word-list of the tool. This showed a significant improvement in Bangla transliteration. As for Gujarati we had very few words available we didn't apply this technique.

We used some frequency based decisions over the output of classifier C(L, E, w) as depicted below. Also the words containing special characters (e.g. &), numerals (e.g. 3D ) and all capitals (e.g. MBA) were strictly considered as English.

$$C'(L,E,w) = \begin{cases} L & \text{if } C(E,L,w) = E \ \&\ conf(E,w) \geq max(0.98p, 0.98) \ \&\ freq(w) < 20 \\ L & \text{if } C(E,L,w) = E \ \&\ length = 2 \ \&\ freq(w) < 50 \\ E & \text{if } w \in S \cup N \cup A \\ C(L,E,w) & \text{otherwise} \end{cases}$$

where, $C(L,E,w)$: Trained classifier, $C'(L,E,w)$: Updated classifier, $conf(E,w)$: Confidnce of word $w$ being English , $S$: Set of special characters, $N$: Set of numerals, $A$: Set of all capital letters, $E$: English label and $L$: Other language label

# 4 Results

## 4.1 Experiments on Development Data

| System | Hindi | | | | Gujarati | | | | Bangla | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LA | S | Algo | Prob | LA | S | Algo | Prob | LA | S | Algo | Prob |
| MSRI-1 | 0.9434 | 2000 | maxent | - | 0.9615 | 1000 | maxent | - | 0.8420 | 1000 | maxent | - |
| MSRI-2 | 0.9513 | 3000 | maxent | 0.65 | 0.9818 | 500 | maxent | 0.7 | 0.8775 | 1000 | naive | 0.85 |
| MSRI-3 | 0.9207 | 5000 | maxent | 0.6 | 0.9765 | 3000 | maxent | 0.8 | 0.8738 | 3000 | naive | 0.75 |

Table 2: Language labeling analysis on submitted runs on development data in all three languages. LA - Labeling Accuracy, Algo - Classifier, Prob - Context probability parameter

## 4.2 Test set results

| System | Hindi | | | Gujarati | | | Bangla | | |
|---|---|---|---|---|---|---|---|---|---|
| | LA | TF | TQM | LA | TF | TQM | LA | TF | TQM |
| MSRI-1 | 0.9823 | 0.8127 | 0.1940 | 0.9614 | 0.4711 | **0.0800** | 0.9259 | 0.4914 | **0.0100** |
| MSRI-2 | **0.9848** | **0.8130** | **0.1980** | **0.9755** | **0.4803** | 0.0733 | **0.9499** | 0.5033 | **0.0100** |
| MSRI-3 | 0.9826 | 0.8101 | 0.1860 | 0.9661 | 0.4748 | 0.0667 | 0.9459 | **0.5137** | **0.0100** |
| Maximum | **0.9848** | **0.8130** | **0.1980** | **0.9755** | **0.4803** | **0.0800** | **0.9499** | **0.5137** | **0.0100** |
| Median | 0.9540 | 0.4160 | 0.0290 | 0.9661 | 0.4748 | 0.0733 | 0.9359 | 0.4973 | 0.0100 |

Table 3: Language labeling analysis on submitted runs in all three languages, along with maximum and median scores. Our runs which had maximum scores are presented in **bold**. LA - Labeling Accuracy, TF- Transliteration F-score, TQM - % of queries that had exact labeling and transliteration

# 5 Error Analysis

System MSRI-2 works best out of the three systems submitted as it considers a context-switch probability along with the character n-grams. For Bangla MSRI-3 improves upon the transliteration accuracy due to inclusion of Bangla words while training.

The most frequent labeling and transliteration errors are categorized and listed in Table 4 and Table 5 respectively. Short words, ambiguous words and erroneous words fail to provide enough n-gram information for a confident classification. Treating all mixed-numerals words as English caused a few Indic words to be wrongly classified.

The assumption of pronunciation similarity of Gujarati and Bangla holds good in general except for a few cases where these languages exhibit phonological variations compared to Hindi. In Bangla 'a' is frequently pronounced as 'o' and in Gujarati 'na' at the end of a word is sometimes pronounced as 'nna'. In such cases our Hindi phonetics based transliteration method fails. For the words which have already been seen in development set, we picked the reference transliterations. However, development set had a few errors which are propagated as is in the output. Other errors include those caused by phonological variations,

Hindi is inflectional in nature whereas Gujarati and Bangla have some extent of agglutination.

| Type | Romanized | Predicted | Reference |
|---|---|---|---|
| Short Words | i; ve | H; E | E; H |
| Ambiguous Words | the; ate | E; E | H; H |
| Erroneous Words | emosal | H | E |
| Mixed Numerals Words | zara2; duwan2 | E; E | H; H |

Table 4: Annotation Errors

| Type | Romanized | Predicted | Reference |
|---|---|---|---|
| Erroneous Latin Source | hau\H; utari\G; banglae\B | हाउ; उतारी; বংগলে | है; उतारो; বাংলায় |
| Multiple Candidates | kali\H; vidhi\G; par\B | काली; વિધી; পর | कली; વિધિ; পার |
| Multiple Transcriptions | tanhai\H; barbadi\G | तनहाई; બર્બાદી | तन्हाई; બરબાદી |
| Merged Words | gayazamana\H; hradayama\G; saralikaraner\B | घनश्याम; હ્મદદૈ; সর্লেক্ষণ | गयाज़माना; હૃદયમાં; সরলীকরণের |
| Plural Words | neendo\H; mandiro\G | नींदों; મંદિરો | नींदो; મંદિરો |
| Distorted Words | mauja\H | मौज | मौजा |
| Language Specific | paani\G; kolkatar\B | પાની; কোলকাতার | પાણી; কলકાતાર |
| Lexicon Coverage | chaudavi\H | चढ़ाव | चौदवी |
| Vowel Error | Gai\B; bali\B | গাঈ; বালী | গાই; বালি |
| Errors in Training Set | bijuriya\H; nahi\G | बिजुरिय; નહિ | बिजुरिया; નહીં |
| Miscellaneous | bina\H | बिना | सिवा |

Table 5: Transliteration Errors

# 6 Conclusions and Future Work

In this paper we presented a brief overview of our method to classify query words to their native language and back-transliterate non-English words to their original script. We showed that character n-grams features coupled with Maximum Entropy (logistic regression) classifier and context probability is an effective method to label to words in a query. We have also found that the simple heuristics like frequency of a word from a monolingual corpora can be used to increase the performance of a word labeling system. We have executed a thorough error analysis and contributed a normalization script to handle different types of word equivalences in Hindi, Gujarati and Bangla such as canonical Unicodes, homorganic nasals, homophonic ending, non-obligatory use of nukta and anuswar-chandrabindu exchange. As future work we would like to explore on increasing efficiency of our transliteration system and deal with case-sensitive text.

# Acknowledgments

We would like to thank Dr. Monojit Choudhury for his valuable suggestions and helpful comments. We also thank Shaishav Kumar for his help in setting up MSRI Name Search tool for our transliteration system and the shared task organizers for patiently answering our queries and the continuous support.

# References

[1] Sowmya V. B., Monojit Choudhury, Kalika Bali, Tirthankar Dasgupta, Anupam Basu. *Resource Creation for Training and Testing of Transliteration Systems for Indian Languages.*

In Proceedings of LREC, 2010.

[2] Umair Z Ahmed, Kalika Bali, Monojit Choudhury, Sowmya V. B. *Challenges in Designing Input Method Editors for Indian Languages: The Role of Word-Origin and Context.* In Proceedings of IJCNLP-WATIM, 2011.

[3] B King, S Abney. *Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods.* In Proceedings of NAACL-HLT, 2013.

[4] McCallum, Andrew Kachites. *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu, 2002.

[5] Shaishav Kumar, Ragahvendra Udupa. *Learning Hash Functions for Cross-View Similarity Search.* In Proceedings of IJCAI, 2011.