A List-searching based approach for Language Identification in Bilingual Text: Shared Task report by Asterisk

Abhinav Raj  
M. S. Ramaiah Institute of Technology  
Bangalore, India  
e-mail: abhinav.raj@live.com

Sankha Karfa  
M. S. Ramaiah Institute of Technology  
Bangalore, India  
e-mail: karfa.sankha@gmail.com

Abstract

In this paper, we describe a List Searching based system for word-level language identification of mixed text. Our method uses List searching and minimum edit distance, therefore, can easily be implemented on most languages. Its performance is carried on the test sets provided by the shared task on language Identification for English Hindi (En-Hi) Pair. The experimental results show a consistent performance with with high precision.

Keywords: Back-Transliteration, Language Labelling, Mixed-Text

1 Introduction

Most of the languages are written using indigenous scripts, i.e. Hindi is written in Devanagari. However, often the websites and the user generated content (such as tweets and blogs) in these languages are written using Roman script after transliteration. Transliteration, the process of converting words into Roman script, is used abundantly on the Web not only for documents, and user queries that are used to search for these documents. In this paper we are proposing the solution of language identification in bilingual text (Hindi and English), and back-transliteration of Hindi Words. e.g Desh ki population’ this is an Hindi English mixed message, where Desh ki is in Hindi and the population is in English. We are also targeting Named Entity in Indian Languages by marking proper nouns. A challenge faced while processing transliterated queries is because of extensive spelling variation. For instance, the word Dhanyavad (“thank you” in Hindi and many other Indian languages) can be written in Roman script as dhanyavad, dhanyvad, danyavad, danyavad, dhanyavada, dhanyabad and so on. We tackle this unique situation prevalent in Web search for users of many languages around the world because this important problem that has received very little attention till date. We are checking individual word in list to label the words.

2 Approach

It’s a dictionary based approach. We started with Hindi English Mixed language. English and Hindi words were labelled separately along with some of the names and known places. Different list are made for English corpus, Hindi corpus, Person’s name, Location and map items, common organisations. Each word is checked against the list and if found, corresponding label is marked with it. If a ambiguous word is encountered(which is found in both Hindi and English), we check the neighbouring words in sentence to make decision. If a new word(unknown) is found, its marked as Hindi Word.

Our model is able to identify the Hindi and English words written in Roman language if they are written properly, without missing letters and with different variations. Ambiguous words are labelled on basis of neighbouring words. If both neighbouring word are belongs to language X, the ambiguous word is labelled X, else we we go to second neighbour and continue.

3 Experiments

Implementation

We are checking each word in named list followed by in English and Hindi list. If the word is found, we are mapping it to respective language. If its a named entity, we are denoting it by [WORD]\P for person and [WORD]\L for location. Hindi and English words are marked as WORD\H and WORD\E. The organisations are labelled as (\O).

The model is implemented using python. Each line is read, and then tokens are separated with blank space (" "), Using RE the words are separated from sentence and newlines are also marked .

\texttt{str=re.split\’,’;line)}

A regular expression is built for regex Search and find the words in the list and ignore in-text special characters while processing . Care has been taken to avoid sub-strings result return by checking the position of word (beginning and ending ) or neighbouring white spaces. For example search for Cat can return Cat and Concatatnation. The words are first
searched in English list followed by Hindi list and named entity list. The words which aren’t found in any list are labelled as Hindi. The named entities are expressed in square brackets. The identified Hindi words written in Roman Script are back transliterated to its own script using Google Transliteration API. The result is encoded UTF-8 format and the returned in output file.

Training Experiments

The new word is trained for the first time and then its automatically added in corresponding list. We have constructed ambiguous word dataset.

Datasets

Datasets for English words are taken various websites including from Oxford Dictionary database, with each source having 3000 to 100,000 words in them. The compiled English dataset contains 452 abbreviations and 504543 unique words including medical and scientific terms. Data sources include SIL International: (www-01.sil.org/linguistics/wordlists/english/).

Some words are taken from Leslie Foster, Dept. of Mathematics, SAN JOS STATE UNIVERSITY, (http://www.math.sjsu.edu/ foster/dictionary.txt) and http://www.nicklea.com/articles/wordlist.txt.


Hindi Datasets are obtained from Dicts.info, "FIRE2013 Track on Transliterated Search" and training material provided by FIRE organisers giving us a unique list of 36253 words. The Name of person was obtained through baby-name.org, deron.meranda.us/data/, State CET results, MicrosoftStudentPartner /MicrosoftStudentAssociates Selection List (total 20398 names from both language). The name of places were obtained through encyclopedia and DBpedia (1407 locations). List of Organisation were obtained through wikipedia database with 71 entries.

4 Results

A test data of 1270 lines for Hi-En pair was run for the model, with total 27296 tokens(en-12324, Hindi-13676, NE-1186), and it was evaluated on precision, recall and f-measure for Hindi and English and label accuracy. Final scores were on basis of Exact transliterated pair match.

| EQMF All (No transliteration) | 0.126 |
| EQMF without NE (No transliteration) | 0.223 |
| EQMF without mix (No transliteration) | 0.126 |
| EQMF without NE and Mix (No transliteration) | 0.223 |
| EQMF All | 0.002 |
| EQMF without NE | 0.003 |
| EQMF without Mix | 0.002 |
| EQMF without Mix and NE | 0.003 |
| ETPM | 1605/1936 |

This is the the evaluation Result for run1.

5 Error analysis

Errors are there because of new words, creative words which couldn’t be found in Database and ambiguous words. We are working on minimum edit distance algorithm and then compare the token to predict the language. This will decrease the error rate. More errors could be because of spelling mistakes and containing creative spellings (gr8 for great), word play (goooood for good), abbreviations (OMG for Oh my God!). We are also working to make labelling decision more accurate.

6 Conclusion and Future Work

In this paper, we described a method of labelling and mapping words from a mixed bilingual text, and back transliterate Hindi words into native script using list based searching. This model can be further used in shared task for building new design / product in field of artificial reality as well. The machine can be trained to be more like human and it can give intelligent response. The model is aiming a very important part of language processing which is rarely addressed. We will be making a smarter model soon with improved efficiency and increased accuracy.

We are targeting a predictive model now, which can work on informal words (nt= NOT), as they arent found in dictionary. Also we are working on words which are combination of two or more individual words, for both language. (e.g. Parmeshware= param + eshwara, lejayenge= le+ jayenge). We plan to do this parallel to increase our processing speed. We will be coming with tries model for faster search, organised data-structure, and training for new words. The new word will be asked for correct marking for once, and then it will be added in the list of that language. We are looking for context based model to tackle ambiguous words for higher accuracy in a Smarter model.
### Metrics Definition

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP, LR, LF</td>
<td>Token level precision, recall and F-measure for the Indian language in the language pair</td>
</tr>
<tr>
<td>EP, ER, EF</td>
<td>Token level precision, recall and F-measure for English tokens</td>
</tr>
<tr>
<td>TP, TR, TF</td>
<td>Token level transliteration precision, recall, and F-measure</td>
</tr>
<tr>
<td>LA</td>
<td>Token level language labeling accuracy = correct label pairs/(correct label pairs + incorrect label pairs)</td>
</tr>
<tr>
<td>EQMF</td>
<td>EQMF (Exact query match fraction) as defined in [1]</td>
</tr>
<tr>
<td>EQMF(without transliteration)</td>
<td>EQMF as defined in [1], but only considering language identification</td>
</tr>
<tr>
<td>ETPM</td>
<td>Exact transliterated pair match as defined in [1]</td>
</tr>
</tbody>
</table>

### 7 acknowledgement

We thank Mr. Krishna Raj PM, Assistant Professor, Information Science and Engineering, MSRIT for guidance and support. We also thank Archit Vyas (BMSCE), Prakruthi Deepak -(PESIT), Tanu Rampal (PESIT) and Harshitha Bidappa -(BMSIT) for help in corpus creation.

### 8 References

- FIRE 2013 Shared Task detailed description: FAQ retrieval using noisy queries, [URL verified 19 Nov. 2013]
- Spandana Gella, Jatin Sharma, Kalika Bali Query word labeling and Back Transliteration for Indian Languages: Shared task system description
- Sujoy Das, Aarti Kumar Performance Evaluation of Dictionary Based CLIR Strategies for Cross Language News Story Search