# FIRE@ISM-2013 Transliterated Search Task

Dinesh Kumar Prabhakar
Sukomal Pal

Department of Computer Science & Engineering
Indian School of Mines, Dhanbad, India

# Contents

- Introduction
- FIRE Task
- Solution
- Approaches
- Result
- Analysis
- Conclusion
- References

# Introduction

**Transliteration:**

A process of writing a term/phrase/sentence of one language (e.g. Hindi) using script of another language (e.g Roman script as used in English)

( e.g.- *yaaron sab dua karo <---> यारों सब दुआ करो)*

Two categories

**Forward**: Phonetic presentation of terms in non-native script

(*e.g. Hindi written using Roman script)*

**Backward:** Conversion of terms from non-native script to its native script

(*e.g. Converting a Hindi phrase written using Roman script back to Devnagari script)*

# FIRE Task

- Task 1: **Query Word Labeling**
  - palak paneer recipe (i/p)
  - palak\H=पालक paneer\H=पनीर recipe\E (o/p)

- Task 2: **Multi-script Ad hoc retrieval for Hindi Song Lyrics**
  - Iss pyar ko main kya naam doon
  - List of song

# Solution

**Query Word Labeling**

Phase 1: Classification

- Dictionary based Classification
- ML-based Classifier (MaxEnt)

Phase 2: Transliteration

- List-based

# Approach-I

- Preprocessing
  - Assuming English wordlist contains sufficient data
  - Created 26 different text file (e.g.- a.txt, b.txt, ..., z.txt)
- Phase 1: Classification
  - List-based
- Phase 2: Transliteration
  - List-based

# Algorithm

1. Input *term* from Test Document
2. Check first letter of *term {A-Z,a-z}*
3. Match *term* in corresponding Document
4. *if* match found
4.1. { Match *term* in E-H pair Document
4.2. *if* found
4.2.1. {Print *term*,\H, word's native script from E-H pair}
4.3. *else*
4.3.1 {Print *term*,\E}}
5. *else*
5.1. {Match term in E-H pair Document
5.2. *if* found
5.2.1. {Print *term* ,\H,=, native script from E-H pair}
5.3. *else*
5.3.1. {Print *term*, \H}}
6. *end*

# Results

- Exact query match fraction (EQMF) = #(Quer. for which lang. labels and translits. match exactly)/#(All queries)
- Transliteration precision (TP) = #(Correct transliterations)/#(Generated transliterations)
- Transliteration recall (TR) = #(Correct transliterations)/#(Reference transliterations)
- Transliteration F−score (TF) = 2 × TP × TR/(TP + TR)
- Labelling accuracy (LA) = #(Correct label pairs)/(#(Correct label pairs) + #(Incorrect label pairs))

# Results

| Language Stats | Metric | ISMDhanbad | Maximum Score | Median Score |
|---|---|---|---|---|
| Hindi | Exact query match fraction | **0.0860** | 0.1980 | 0.0290 |
| 10 runs | Exact transliteration pairs match | 1584/2117 | N. A. | N. A. |
| 5 teams | Transliteration-precision | **0.7253** | 0.8135 | 0.4486 |
| #(True \H) = 2444 | Transliteration-recall | **0.6484** | 0.8125 | 0.4300 |
| #(True \E) = 777 | Transliteration-Fscore | **0.6847** | 0.8130 | 0.4260 |
| #(\N) = 232 | Labelling accuracy | 0.8780 | 0.9848 | 0.9540 |
| N = Names | Eng-precision | 0.6853 | 0.9667 | 0.9302 |
| and ambiguities | Eng-recall | 0.9138 | 0.9755 | 0.9640 |
| excluded from | Eng-Fscore | 0.7832 | 0.9685 | 0.9019 |
| analysis | L-precision | 0.9693 | 0.9906 | 0.9883 |
| | L-recall | 0.8666 | 0.9894 | 0.9791 |
| | L-Fscore | 0.9151 | 0.9900 | 0.9700 |

# Analysis

- English wordlist in corpus is considerably high
- Out-of- Dictionary word will be treated as hindi word
  - (e.g.-peenekeliye\H)
- Named entity may come with correspoding transliterated word if it is in E-H pair file
  - (e.g.- khusbu  khusbu\H=खुश्बू)
  - Why?
  - Since term is there in E-H pair document
    - NER technique used X
    - Context consider X

# Approach-II

- Preprocessing
  - Annotate “**E**” to english words and “**H**” to hindi term of E-H pair words

    (e.g.-tera H, khushboo H and good E, apple E)
  - Train the classifier with these annotion
- Phase 1: Classification
  - Using this classifier, terms are classified
- Phase 2: Transliteration
  - List-based

# Algorithm

1. Input *term* from Test Document

2. Classify terms into E\H

3. *if term* is of E class

3.1. {Print *term*, “\”,class}

4. else

4.1. { match *term* in E-H pair Document

4.2. *if* found

4.2.1. {Print *term*,class, term's native script from E-H pair}

4.3. *else*

4.3.1 {Print *term*,\,class}}

5. *end*

# Analysis

bibi\H=बीबी ka\H=का maqbara\E paryatak\H guide\E

- maqbara\E wrongly classified
  - Why?
  - Less no of hindi term in training data ( in E-H pair document)
- paryatak\H equivalent transliteration is not here.
  - Why?
  - Out-of-dictionary (E-H pair document)

# Conclusion

- Backward transliteration technique
- Our system has performed better for some of the metrics
  - (e.g.- EQMF, TP,TR and TF)
  - Why?
  - Equvalent transliterations was there
- There are some limitations of this  system
  - (e.g. Named-entity may not be identifiable)
  - Why?
  - We haven't used any NER technique
- System may give unwanted transliteration for few *terms (e.g.- koee koee/H=के)*
  - Why?
  - Since it is there in E-H pair document

# References


1. King, B., Abney, S.: Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods In Proceedings of NAACL-HLT-2013, Atlanta, Georgia (2013) 1110-1119

2. Gupta, K., Choudhury, M., and Bali, K.: Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics, In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12), Istanbul, Turkey (2012) 2459-2465

3. Sowmya, V.B., Choudhury, M., Bali K., Dasgupta, T. and Basu, A.: Resource Creation for Training and Testing of Transliteration Systems for Indian Languages, LREC (2010)

4. Karimi, S., Scholer F., and Turpin, A.: Machine Transliteration Survey. In ACM Computing Surveys (CSUR), Volume 43 Issue 3, New York, USA (2011) 17:1-46

5. Dale, R.: Language Technology. Slides of HCSNet Summer School Course. Sydney (2007)

6. Stanford Classifier v3.2.0 – 2013-06-19 classification tool from Stanford University

# THANK YOU