Overview and Datasets of FIRE 2013 Track on Transliterated Search

Rishiraj Saha Roy  
IIT Kharagpur

Monojit Choudhury  
Microsoft Research India

Prasenjit Majumder  
DAIICIT Gandhinagar

Komal Agarwal  
DAIICIT Gandhinagar

Forum for Information Retrieval Evaluation 2013 (FIRE '13)  
New Delhi, India
चाहूँ में या ना - Chahun Main Ya Na
Arijit Singh, Palak Muchchal

Movie/Album: आशिकी २ (2013)
Music By: जीत गांगुली
Lyrics By: इरशाद क़मिल
Performed By: अरिजीत सिंह, पलक मुंछाल

तू ही ये मुझको बता दे, चाहूँ में या ना
अपने तू दिल का पता दे, चाहूँ में या ना
इतना बता दूँ तुझको, चाहते हैं अपनी मुझको
यूँ तो नहीं इक्खित्यार
फिर भी ये सोचा दिल ने, अब जो तुम हाँ मिलने
पूर्व पुछ चक्र एक बार
tू ही ये मुझको बता दे...

Tips For Searching Your Song
To search for songs, you can search through Labels mentioned at the bottom of this blog. Best is to use (Ctrl+F).
This weekend's cooking experiment was to make a traditional kerala dish called "Kappa Puzhukku" (Tapioca Pudding). Grand success.

Facebook and Twitter
Marich aur Subahu ka vadh –

Doosare din brahm muhurt men uth kar tatha nityakram aur sandh
Lakshman guru Vishwamintra ke pas ja kar bole, “Gurudev! kripa
yagya men vighna dalane ke liye kis samay aate hain? Yah ham is
na ho ki hamare anjane men hi ve aakar upadrav machane lagen.

Dashrath ke veer putron ke in utsahbhare vachanon ko sun kar wa
atyant prasanna huye aur bole, “He rughukulbhusan rajkumaron!
se chhah dinon tatha ratriyon tak puran roop se savdhan mudra m
dinon men Vishvamitra ji maun hokar yagya karenge. Is samay bh
denge kyonki ve yagya ki diksha le chuke hain.”
Floated Tasks

- (Pilot) Track in first year
- Focused on basics required for search in transliterated space
- Subtask 1
  - Query word labeling
- Subtask 2
  - Multi-script ad hoc retrieval
Subtask 1: Query Word Labeling

- Label words of a query as English or L
- Subtask presented for three language pairs
  - English-Hindi
  - English-Bangla
  - English-Gujarati
- If labeled as L, generate transliteration in native script
- Process of back transliteration
- Evaluation excludes OOV named entities
Subtask 1: Examples

- **Input**
  - *door ke dhol song lyrics*
  - *electric tar best company ki*
  - *shu tame mane prem karo*

- **Output**
  - *door H=दूर ke H=के dhol H=ढोऱ song E lyrics E*
  - *electric E tar B=তার best E company E ki B=কি*
  - *shu G=શુ tame G=તમે mane G=મને prem G=પ્રેમ karo G=કરો*
Subtask 2: Multi-script Retrieval

- Retrieve top ten relevant documents for a query
- Query in Roman script
  - Bollywood song text
- Large corpus of mixed script Documents
  - Roman/Devanagari/Both
  - Documents contain song lyrics
Subtask 2: Example

- Query: geeto ki rut aur rangon ki barkha
- Document

कोई जो मिला तो मुझे ऐसा लगता था जैसे मेरी सारी दुनिया में गीतों की रुत और रंगों की बरखा है

Khushboo ki andhee hai

Mehki huee si ab saree fizayein hain
Datasets

- General purpose
- Specific to Subtask 1
- Specific to Subtask 2
- Info on datasets at

http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/index.html
Datasets: General Purpose

- Word frequency lists: English, Hindi, Gujarati
- Word transliteration pairs
  - Hindi: Alignment of song lyrics [Gupta et al., 2012]
  - Bangla: Annotations collected from chat, dictation setups [Sowmya et al. 2010]
  - Gujarati: Toy set, processed from FIRE 2013 data
- Large language corpora (Leipzig)
- ITRANS to UTF-8 converter
Datasets: Subtask 1

- Hindi
  - 1000 queries – 500 development set, 500 test set
- Bangla
  - 200 queries – 100 development set, 100 test set
- Gujarati
  - 300 queries – 150 development set, 150 test set

- ~1000, ~300, ~500 translit pairs in dev sets
- Not all entries technically search “queries”
Carefully crafted with instances of language words with valid English dictionary entries
- door, tan, man (Hindi), tar, pore, ache (Bangla); tame, mane, mate (Guajrati)

Created and annotated by respective native speakers

Future plans
- Enrich and expand with more quality control
- Looking for partners for more languages!!
Datasets: Subtask 2

- 50 hand crafted queries in Roman script – 25 dev, 25 test
- About 63,000 documents in pure/mixed scripts
- Documents collected by crawling ~15 popular Bollywood lyrics domains like dhingana, musicmaza and hindilyrix
- XML documents parsed and cleaned to contain only lyrics text
- Around 28 relevance judgments per query (6-point scale) after pooling using several baselines
Participation

- Initial show of interest from 17 teams
- 5 teams participated, 25 runs submitted
  - India: ISM Dhanbad, Gujarat University (GU), Microsoft Research India (MSRI)
  - Abroad: TU Valencia (TU-V), NTNU Norway
- MSRI participating but non-competing
Participation

- Subtask 1: ISM, GU, MSRI, TU-V, NTNU (17 runs)
  - Hindi: 10 runs (all 5 teams)
  - Bangla: 4 runs (NTNU, MSRI)
  - Gujarati: 3 runs (MSRI)
- Subtask 2: NTNU, TU-V, GU (8 runs)
Evaluation Metrics – Subtask 1

- **Exact Query Match Fraction** =

\[
\frac{\#(\text{Queries for which lang labels and translit pairs match exactly})}{\#(\text{All queries})}
\]

- **Exact Transliteration Pairs Match** =

\[
\frac{\#(\text{Pairs for which transliterations match exactly})}{\#(\text{Pairs for which both o/p and reference labels are L})}
\]

- Motivation: Exactly one correct answer for back transliteration

- Some cases of normalization have been handled
  - Thanks to Spandana from MSRI!!
Evaluation Metrics – Subtask 1

- **Transliteration Precision (TP)**
  \[
  TP = \frac{\#(Correct \ transliterations)}{\#(Generated \ transliterations)}
  \]

- **Transliteration Recall (TR)**
  \[
  TR = \frac{\#(Correct \ transliterations)}{\#(Reference \ transliterations)}
  \]

- **Transliteration F-Score**
  \[
  F-Score = \frac{2 \cdot TP \cdot TR}{TP + TR}
  \]
Evaluation Metrics – Subtask 1

- **Labeling accuracy** =
  \[
  \frac{\#(\text{Correct label pairs})}{\#(\text{Correct label pairs}) + \#(\text{Incorrect label pairs})}
  \]

- **English Precision (EP)** =
  \[
  \frac{\#(E\text{-}E \text{ pairs})}{\#(E\text{-}E \text{ pairs}) + \#(E\text{-}L \text{ pairs})}
  \]

- **English Recall (ER)** =
  \[
  \frac{\#(E\text{-}E \text{ pairs})}{\#(E\text{-}E \text{ pairs}) + \#(L\text{-}E \text{ pairs})}
  \]

- **English F–Score** =
  \[
  \frac{2 \times EP \times ER}{EP + ER}
  \]

- Similarly LP, LR, and LF are computed
Evaluation Metrics – Subtask 2

- nDCG@5, nDCG@10
  \[ DCG@p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}; \quad nDCG@p = \frac{DCG@p}{IDCG@p} \]

- MAP
  \[ Ave(P) = \frac{\sum_{k=1}^{n} (P(k) \times rel(k))}{\#Rel.docs}; \quad MAP = \frac{\sum_{q=1}^{Q} Ave(P)}{Q} \]

- MRR
  \[ MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \]
Results

- Detailed metric values and approaches coming up soon in participant talks
- Subtask 1:
  - Transliteration F-score (Hindi): 0.8130
  - Transliteration F-score (Bangla): 0.5137
  - Transliteration F-score (Gujarati): 0.4803
- Subtask 2:
  - nDCG@10: 0.8002
Results

- Winners (several very close results!!)
  - Subtask 1 (Hindi): TU-Valencia [Best on 5/12 metrics]
  - Subtask 1 (Bangla): NTNU-Norway [Best on 12/12 metrics]
  - Subtask 1 (Gujarati): None
  - Subtask 2: TU-Valencia [Best on 4/4 metrics]
  - MSRI topped Subtask 1 but was non-competing
  - Congratulations to all!!
Road Ahead

- Encouraging response to task in first year – why the dropouts?
- Metric values reflect room for improvement (grain of salt)
- Extend to at least one non-Indian language (Arabic?)
- Extend to at least Dravidian language (Kannada?)
- Want to enrich datasets in a shared environment – in process
- Plans to create awareness on importance of transliteration for IR like organizing workshops – please visit http://bit.ly/1k7pG55
Acknowledgements

- CMU
  - Rohan Ramanath
- IIT Kharagpur
  - M. Dastagiri Reddy
  - Ranita Biswas
  - Swadhin Pradhan
  - Yogarshi Vyas
- Entire FIRE team for making this track possible!
Feedback

- Overview online at http://www.isical.ac.in/~fire/wn/STTS/2013-translit_search-track_track_overview.pdf
Thank you!!

- Looking forward to increased participation at FIRE 2014!!

- Primary contact: monojitc@microsoft.com