ISM@FIRE-2015: Mixed Script Information Retrieval

Dinesh Kumar Prabhakar
Indian School of Mines
Dhanbad, Jharkhand
India 826004
dinesh.nitr@gmail.com

Sukomal Pal
Indian School of Mines
Dhanbad, Jharkhand
India 826004
sukomalpal@gmail.com

ABSTRACT

This paper describes the approach we have used for identification of languages for a set of terms written in Roman script and approaches for the retrieval in mixed script domain, in FIRE-2015. The first approach identifies the class (native language of terms and whether a term is any named entity or of any other type) of given terms/words. MaxEnt a supervised classifier has been used for the classification which performed best for *strict f-measure NE* has score is 0.46 and *strict f-measure NE_P* has score 0.24. For the MSIR subtask Divergence from Randomness (DFR) based approach is used and performed better with block indexing and query formulation. Overall scores of our submission on NDCG@10 0.4335, 0.5328, 0.4489 and 0.5369 for ISMD1, ISMD2, ISMD3 and ISMD4 respectively.

.

Keywords

Word classification, Transliteration, Information Retrieval

1. INTRODUCTION

With the development of the Web 2.0, user's count on Social sites are increasingly becoming higher. They write messages (specially blogs and post) on sites (such as Twitter and Facebook) in their own languages preferably using Roman scripts (transformed form). These post might consist terms of Non-English (or terms from user's native) languages, a simple English word, a mixed language term (like gr8, 2moro) or a Named Entity (NE). Identification of such categories play significant role in Natural Language Processing (NLP). It doesn't remains limited to the NLP but also used in other sub-domains of linguistic processing and Information Retrieval (IR).

Since, blog posts contain some important information that opens up the scope of IR in informal texts (in form of posts or massages). Raw blogs data often have some erroneous text. Hence, before applying any IR steps data must be preprocessed using some linguistic processing approaches.

There are huge collection of data on/off the Web for various information needs but the track for adhoc retrieval. For the retrieval, collection has documents written in two scripts: Roman (transliterated form of Hindi terms in Roman script) and Devanagari. In whole corpus, some document has information in Devanagari, some others has in Roman and rest of the document has information in mixed (transliterated and native) scrip one after another. To maximize the number of most relevant documents on the Web (in web retrieval) or from the corpus (in ad-hoc retrieval) it is necessary to retrieve the documents of other language and/or script. It is important to discuss three terms *monolingual*, *multilingual* and *mixed script* retrieval. In IR *monolingual* means query and documents to be retrieved are in single language where as *multilingual* query and documents may be in written different language. But, the *mixed script* retrieval is slightly different than *monolingual* retrieval. In *mixed script* retrieval, system should retrieve the relevant documents of same language written in more than one script.

In FIRE-2015, for the *Mixed Script Information Retrieval* track participant has to design the system for term classification and for the retrieval of relevant documents written in Devanagari script and in Roman script.

We have used query expansion to reformulate the seed (information need) for addressing the mixed script retrieval issues.

Further in Section 2, we discussed the task descriptions. Section 3 shows related work on and Section 4, describes our approaches for annotation and MSIR. In Section 5, we have discussed results and analyzed errors. Section 6, conclude the strategies with the direction of future work.

2. TASK DESCRIPTION

The track, *Shared Task on Mixed Script Information Retrieval (MSIR)* has three subtasks: *Query Word Labeling*, *Mixed Script Ad-hoc Retrieval* and *Mixed-script Question Answering.* We have participated in first two subtask.

Query Word Labeling

Input:- Let Q be the query set containing n query word $w_i (1 \leq i \leq n)$ written in Roman script. The word $w_i \in Q$ $(w_1, w_2, \ldots, w_n)$, could be standard English (en) words or transliterated from another language L = {Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (ka), Malayalam (ml), Marathi (mr), Tamil (ta), Telugu (te)} and some Named Entities (NE). The task is to label the words as En or a member of L depending on whether it an English word, or a transliterated L-language word. Input and expected result for an utterance is given below as an example.

Input:

```
<utterance id="1">
hesitate in to giving is @aapleaks #aapsweep
revenge should this statehood take way bjp
not the #aapstorm best
</utterance>
```

Output:- Result ${w_i}^l$ is corresponding label produced for individual terms.

Output:

```
<utterance id="1">
en en en en en X X en en en en en en NE en
en X en
</utterance>
```

Mixed Script Ad-hoc Retrieval

There are more than 66K documents and 25 queries (seeds). Documents are written in Devanagari script, Roman script or in mixed script. Here mixed script means a document has same content in two scripts one after another. Out of 25 queries seven are in Devanagari and others are in Roman script.

The goal of the task is for a given query system should produce set of relevant documents in ordered where on the most relevant document should should come at first position.

3. RELATED WORK

Subtask-1 accomplished in two phases: *Word Labeling* and *transliteration* of *H* labeled word to its native (Devanagari) script.

3.1 Query Language Labeling

The labeling is concerned with the classification of a given word written in Roman script. Query words w_i can be classified and annotated with corresponding classes manually or using machine learning based classifiers. Various classifiers are there for classification such as Support Vector Machine(SVM), Bayesian networks, Decision Trees, Naive-Bayes, MaxEnt and Neural Networks.

King and Abney started for labeling the languages of words in cross-lingual documents[3]. They have approached this problem in a weakly supervised fashion, as a sequence labeling problem with monolingual text samples for training data. Prabhakar and Pal also attempt in similar fashion using supervised learning algorithm [6].

3.2 Mixed Script Ad-hoc Retrieval

This subtask was introduced in FIRE-2013 [7], continued in FIRE-2014 with more challenges (joint terms need expansion) [1] and in FIRE-2015 (queries are in Devanagari or Roman text along with previous challenges).

Gupta et al. in 2014, approached MSIR using 2-gram *tf-idf* and deep learning based query expansion [2]. The spelling variation in transliterated terms along with mixed script text is the major challenge of MSIR. Transliteration of any term can be extracted from parallel or comparable corpora in extraction approach whereas in generation, transliteration is generated depending on phoneme, grapheme or syllable-based rules.

4. APPROACHES

Our approaches for the solution of Subtask-1 and Subtask-2 have been described in subsections below.

4.1 Query Word Labeling

We have considered *word labeling* as classification issue for the tags annotation to the given terms w_i. Terms can be classified either manually or using any classifier. Manual classification and tagging is not feasible on the large dataset. MaxEnt a supervised classifier is used for classification and labeling of words from utterances. The Stanford's MaxEnt implementation is used for this purpose [4].

For the classification, model was *trained* on development data and then terms from utterances of test dataset were classified based on extracted features during training.

4.1.1 Training

For the training purpose input terms and annotations are tokenized and made align with proper tags.

Features used.

Features value with default parameter were used some of which are listed below:

- *useNGrams* accept boolean value true or false to make features from letter n-grams where true is assigned here.
- *usePrefixSuffixNGrams* makes features from prefix and suffix substrings of the string and accept boolean value where we have assigned true.
- *maxNGramLeng* takes integer value and size beyond the assigned number will not be used in the model. Maximum length 4-grams was used.
- *minNGramLeng* also takes integer number and n-grams below this size will not be used in the model. It must be a positive integer and we have set it 1.
- *sigma* is a parameter to several of the smoothing methods,usually gives a degree of smoothing as standard deviation. Here this number is 3.0.
- *useQN* accepts boolean value where true indicates Quasi-Newton optimization will be used if it is set to true.
- *tolerance* is convergence tolerance in parameter optimization and set 1e-4.

Classification model was train on above parameter values and 23 classes were identified during the training.

4.1.2 Classification

Given terms from utterances of test dataset were tokenized and parsed on trained model. Tokens of test data are classified and annotated with different tags such as for Hindi terms *hi*, English terms *en*, proper names (name of the person *NE_P*, location *NE_L*).

4.2 Mixed Script Information Retrieval

Subtask-2 has queries for Hindi song lyrics, astrological data and movies reviews related documents retrieval. Proposed approach consist three modules: documents indexing, query formulation and documents retrieval.

4.2.1 Document Indexing

Simple bags-of-words approach may retrieve noisy documents for lyrics retrieval. Because in lyrics consecutive terms are important as change in position changes the context of a song. Hence, we have chosen *block indexing* with block-size 2 *words* in addition simple indexing. Two approaches simple indexing (bags-of-words) and block indexing (phrase retrieval) were used to index the collection with block size one word and two words respectively.

4.2.2 *Query Formulation (expansion)*

As documents in the corpus are in mixed script, seed value only can't give good result for retrieval. Hence, the query must be reformulated to enhance the performance of the system. In query formulation, script of the query is identified and then transliteration is extracted using Google transliteration API. There are many terms for which API gives more than one transliteration for such term first one is chosen. For the submission of run ISMD2 and ISMD4 we have used formulated mixed script query as shown in Table 1.

Table 1: Query formulation table

Query Type	Queries
Original Query	tujo nahi lyrics
Transliterated Query	तुजो नहीं लिरिक्स
Formulated Query	tujo nahi lyrics तुजो नहीं लिरिक्स
Original Query	सूर्य रेखा कर्क राशि
Transliterated Query	suyra rekha kark rashi
Formulated Query	सूर्य रेखा कर्क राशि suyra rekha kark rashi

4.2.3 *Document Retrieval*

Poisson model with Laplace after-effect and normalization 2 of Divergence From Randomness (DFR) framework has been used to measure the similarity score between documents d and query Q [5]. For the implementation we have used *terrier 4.0*.

$$Score\,(d,Q) = \sum_{n\in Q} \left(qtfn \cdot w\,(t,d)\right) \qquad (1)$$

$$qtfn = \frac{qtf}{qtf_{max}} \qquad (2)$$

where $w(t,d)$ is the weight of the document d for a query term t and $qtfn$ is the normalized frequency of term t in the query. And qtf is the original frequency of term t in the query, and qtf_{max} is the maximum qtf of all the composing terms of the query for details see[5].

5. RESULTS AND ANALYSIS

Our approaches have been evaluated on the provided test data for query word labeling and MSIR. In both the subtasks our approaches performed moderate.

5.1 Subtask-1

MaxEnt based classifier worked moderate as depicted in table 2. In some of measure our approach performed well with scores 0.46 strict f-measure NE and 0.24 in strict f-measure NE_P. For some metrics we performed moderate and in others poor as well. Some terms are misclassified e.g. Input utterance:

```
<utterance id="186">
ei path jodi na shesh hoy lyrics
</utterance>
```

Annotated utterance:

```
<utterance id="186">
bn hi bn bn bn bn en
</utterance>
```

The token 'path' in input utterance should have a Bengali term and has same meaning in Hindi and English also but misclassified in Hindi due to ambiguity as same term exist in Hindi. But 'path' seems to be 'poth' in Bengali due to regional accent.

5.2 Subtask-2

Submitted four Runs for subtask-2, with combinations of simple indexing and original query, simple indexing and formulated query, block (size=2 words) indexing and original query and block (with size=2 words) indexing and formulated query. From the score in Table 3 we have observe that Run with block indexing and formulated queries better and the order in higher to lower performance on NDCG@10 is $ISM4 > ISM2 > ISM3 > ISM1$.

Overall the retrieval approaches performed moderate compare to other teams. Some challenges remains un-addressed in approaches: spelling variation in transliterated (Roman) text, combined term (such as 'kabhi-kabhi' could be 'kabhi', 'kabhi', 'tujo' could be 'tu', 'jo') and translation (some document consist information in another language such as सूर्य रेखा कर्क राशि could be translated into Line of Sun for Cancer) of query text. One more challenging issue is partial transliteration and translation. For example query number 69, query is "shani dashaa today for a 20 year old" in that first two tokens are Hindi terms. Hence either Hindi terms will be translated to English or other terms need to be translated into Hindi and then transliterate into Roman text.

6. CONCLUSIONS

Our work comprises two subtasks annotation and retrieval. We have used learning based classifier for word labeling. Label accuracy was moderate for submitted runs. We identified some terms were incorrectly labeled by the classifier. Perhaps this happened due an important reason i.e. term ambiguity where same term exist in more then one classes. For the MSIR, simple and block indexing both used separately during document indexing. In the query formulation transliterations are extracted using Google API. To measure the similarity score a DFR framework is used which performed moderate. Some query expansion approach can address MSIR retrieval issues. In future we are looking to address the unresolved issues mentioned above.

7. REFERENCES

[1] Choudhury, M., Chittaranjan, G., Gupta, P., and Das, A. Overview and datasets of fire 2014 track on transliterated search. In *Pre-proceedings 6th workshop FIRE-2014* (2014), Forum for Information Retrieval Evaluation (FIRE).

[2] Gupta, P., Bali, K., Banchs, R. E., Choudhury, M., and Rosso, P. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (2014), ACM, pp. 677–686.

[3] King, B., and Abney, S. P. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *HLT-NAACL* (2013), pp. 1110–1119.

[4] Klein, D. The stanford classifier. http://http://nlp.stanford.edu/software/classifier.shtml, 2003.

Table 2: Query word labeling score

Metric	ISM_Score	Aggregate_Mean	Aggregate_Median	Max_Score
MIXesAccuracy	12.5	5.0595	0	25
NEsAccuracy	13.253	36.0103	35.9459	63.964
NEsCorrect	22	199.8571	199.5	355
strict f-measure NE	0.461728395	0.371410272	0.07536114	0.461728395
strict f-measure NE_L	0	0.0426	0	0.2114
strict f-measure NE_P	0.2486	0.1086	0.1133	0.2486
strict f-measure X	0.9612	0.8989	0.9379	0.9668
strict f-measure bn	0.7113	0.7073	0.7549	0.8537
strict f-measure en	0.9052	0.8067	0.8356	0.9114
strict f-measure gu	0.1383	0.1338	0.1331	0.3484
strict f-measure hi	0.6618	0.6168	0.6413	0.8131
strict f-measure kn	0.6373	0.5752	0.6062	0.8709
strict f-measure ml	0.4871	0.4762	0.4757	0.7446
strict f-measure mr	0.5636	0.5994	0.6469	0.8308
strict f-measure ta	0.718	0.7261	0.749	0.8911
strict f-measure te	0.5439	0.4654	0.4817	0.7763
TokensAccuracy	**77.0648**	**71.1137**	**75.5563**	**82.7152**
UtterancesAccuracy	**17.298**	**14.6645**	**17.1086**	**26.3889**
Average F-measure	**0.613402366**	**0.539559189**	**0.113420527**	**0.69174727**
Weighted F-Measure	**0.767831108**	**0.698989963**	**0.095876627**	**0.829929229**

Table 3: Subtask-2 scores

Team	Block_Size	Query_Formulation	NDCG@1	NDCG@5	NDCG@10	MAP	MRR	RECALL
ISMD1	1 word	X	0.4133	0.4268	0.4335	0.0928	0.244	0.1361
ISMD2	1 word	✓	0.4933	0.5277	0.5328	0.1444	0.318	0.2051
ISMD3	2 words	X	0.3867	0.4422	0.4489	0.0954	0.2207	0.1418
ISMD4	2 words	✓	0.4967	0.5375	0.5369	0.1507	0.3397	0.2438

Online; accessed 19-02-2014.

[5] Plachouras, V., He, B., and Ounis, I. University of glasgow at trec 2004: Experiments in web, robust, and terabyte tracks with terrier. In *TREC* (2004).

[6] Prabhakar, D. K., and Pal, S. Ism@fire2013 shared task on transliterated search. In *FIRE '13 Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation* (2013), ACM New York, p. 6.

[7] Roy, R. S., Choudhury, M., Majumder, P., and Agarwal, K. Overview and datasets of fire 2013 track on transliterated search. In *Pre-proceedings 5th workshop FIRE-2013* (2013), Forum for Information Retrieval Evaluation (FIRE).