AmritaCEN_NLP @ FIRE 2015 Language Identification for Indian Languages in Social Media Text

Rahul Venkatesh Kumar RM
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India
rahulvks@gmail.com

Anand Kumar M
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India
m_anandkumar@cb.amrita.edu

Soman KP
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India
kp_soman@amrita.edu

ABSTRACT

The progression of social media contents, similar like Twitter and Facebook messages and blog post, has created, many new opportunities for language technology. The user generated contents such as tweets and blogs in most of the languages are written using Roman script due to distinct social culture and technology. Some of them using own language script and mixed script. The primary challenges in process the short message is identifying languages. Therefore, the language identification is not restricted to a language but also to multiple languages. The task is to label the words with the following categories L1, L2, Named Entities, Mixed, Punctuation and Others This paper presents the AmritaCen_NLP team participation in FIRE2015-Shared Task on Mixed Script Information Retrieval Subtask 1: Query Word Labeling on language identification of each word in text, Named Entities, Mixed, Punctuation and Others which uses sequence level query labelling with Support Vector Machine.

CCS Concepts

• Theory of computation~Support vector machines
• Computing methodologies~Natural language Processing
• Information systems~Information extraction • Human-centered computing~Social tagging systems

Keywords

Language Identification, Support Vector Machine (SVM), Information retrieval, Mixed Script, Short Message.

1.INTRODUCTION

This paper describes our system for FIRE 2015 Shared Task on Query Word Labeling on Mixed Script Information Retrieval. The faster growth of internet in current period the Webpages are not limited to English, social media content in other languages increasing rapidly [1]. Now a day's webpages can be found in every popular non English language which includes Indian languages too. In social media users are generally using their native languages in Romanized form to express their thoughts [2][3]. To handle this Multi-lingual text processing problem, we need to label the token into corresponding languages. The idea of Multi-Script IR was first introduced by P Gupta, Kalika Bali, R E Banchs, M Choudhury, P Rosso in 2013 SIGIR conference [3]. This task addresses problem of language identification in code mixed queries. Task focuses on sentence level language identification in code mixed queries in English and any 8 Indian Languages (L) Hindi, Bengali, Tamil, Guajarati, Marathi, Kannada, Telugu and Malayalam. Our language identification system uses Support Vector Machine for word level classification.

2.RELATED WORKS

The problem of language identification is researched for half century (Gold,1967) and code switching for several decades. But there has been less work on automatic language identification for mixed script analysis in social media websites and forums. Research showed that the predominant language used in Twitter and Face book in their earlier days was English [4][5]. With the worldwide growth social media, people started to write in their own language with the help of roman script. Number of people who using mixed script in social media commutation has increased tremendously. According to the report 45% of users using mixed script in facebook,40% of people using English for communicating and 15% people used their native language [6][7]. Identification of the language in social media content and their analysis is essential for extracting information which can be further used in aiding search engines and monitoring online behavior so as to ensure security [8][9]. Few years back documents were written only in a single language. With the emergence of social media these day's documents were written in mixed script [10].

3.DATA SET DESCRIPTION

In training data set, input query is constructed and annotated with their label. The query is written in roman script. Input query and annotated set are given as a part of the Subtask. The training data contains annotation and input file each have 2908 sentences (Tokens 54,088). The test data contains 792 sentences (Tokens 11,999). Tokens are person name, location, organization and abbreviation comes under NER label.

Table 1: Tag set and Total Count.

Language Token Count	41,515
Named Entities	2,391
Mixed (Mix)	70
Punctuations (X)	7710
Others (O)	11
Total Tag Count	54,088

Table 2: Language Data Training Set count.

Language	Token Count
Tamil	3169
English	18017
Hindi	4615
Bengali	3556
Guajarati	890
Marathi	1960
Kannada	1674
Telugu	6474
Malayalam	1160

Table 3: Named Entities Training Set count.

Named Entities	Token Count
NE	2028
NE_P	257
NE_L	29
NE_O	22
NE_PA	7
NE_LA	1
NE_X	38
NE_XA	5
NE_OA	24

Tokens which constructed two parts, each coming from a different language are labelled as MIX, Emoticons, hash and punctuation are labelled as MIX. Foreign languages are labelled as O. There is no extra data set is used in this task. Input query many contain mixture of 1 or 2 languages, named entities, mixed, punctuation and others. Table 1 contains the counts for mixed, punctuation and others with overall token count. The languages token count is mentioned in Table 2. Named Entities have nine different tag set and total count of the NER tokens are mentioned in the Table 3.

4.METHODOLOGY AND FEATURE DESCRIPTION

We participated in the Query Word Labeling task which is described very briefly as follows: Suppose that q: w_1 w_2 w_3 is a query which is written in Roman script. The words, w_1 w_2 etc., could be standard English words or transliterated from another language L = {Bengali (Bn), Gujarati (Gu), Hindi (Hi), Kannada (Ka), Malayalam (Ml), Marathi (Mr), Tamil (Ta), Telugu (Te)}. The task is labeling the words as En or L. In Query word labeling we used Support Vector Machine classifier to predict language of a particular word which belong to either Indian language or English. As the training corpus is very huge, the words from the corpus are taken as features. As a method of preprocessing, the input raw data taken as token per sequence is annotated with corresponding tag set. This annotated data set is assigned as input for the machine from which the features are extracted. Various features are taken for better labelling of language. The three prefixes and suffixes of the current word, length of the present token, position of current word are taken as features. Punctuation, comma, colon/Semi Colon, dot and word starting with '@' and '#' are taken as binary features. This set of feature has been mainly used to identify Indian languages. They constitute checks on token endings in terms of presence of certain characters. Along with the features machine also learns from the training data set which is already labelled. When it comes to test data, same preprocessing step is carried out. Annotated test data is given as input for Support Vector Machine Classifier and classified output is taken. Sample output is given in Table 4.

Table 4: Input query with desired output.

Input Query	Output
And ibruna meet maadid kushinu aythu !!!	And\en ibruna\kn meet\en madid\kn kushinu\kn aythu\kn !\X
Dhoni risk edutha gumbala risk edukanam !	Dhoni\NE Risk\en edutha\ta gumbala\ta risk\en edukanam\ta !\x

5.PROPOSED SYSTEM

The query and corresponding tag set is given as a training data of the shared task and these are annotated as preprocessing procedure. Flow of the proposed system is illustrated in Fig 1. From the annotated data the features are extracted. Along with the extracted feature sequence of lines are given as an input for the Support vector machine classifier in which in creates a module file. The test data and module file is given to the classifier and output is extracted. Further the output is processed in which the utterance id is properly paired with test data.

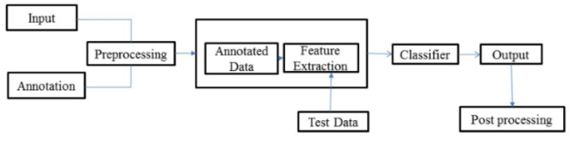

Fig 1: Proposed system flow diagram.

6.RESULT AND CONCLUSION

In this paper, we described our system for Subtask 1 in FIRE 2015 - Query Word Labelling on Mixed Script Information Retrieval. The query word labelling is very useful in search engines. We used SVM classifier to identify languages, Punctuations, NEs, Mixed and Other. SVM uses set of features guaranteeing reasonable accuracy for mixed languages query and other tags. In proposed language identification system, the word sequences are divided into tokens which are trained using SVM classifier and the system is evaluated against the given test data. System is elevated separately for each tag in language pair, Mixed and Named Entities using Recall, Precision and F1-Score.Concentration is required more on Mixed Script and NEs. As a future work words from the language dictionary and word as distributed vector can also be included as feature which will improve the accuracy of the system. Overall scores for tags set is mentioned in Table 5.

Table 5: Summary of Scores.

Mixes Accuracy	8.3333
NEs Accuracy	36.3964
Token Accuracy	76.6231
Utterance Accuracy	16.9182
Average F -Measure	0.682876
Weighted F-Measure	0.766462

7.REFERENCE

[1] Irshad Ahmad Bhat(IIT-H), Vandan Mujadia(IIT-H), Aniruddha Tammewar(IIT-H). IIT-H System Submission for *FIRE2014 Shared Task on Transliterated Search.*

[2] P Gupta, Kalika Bali, R E Banchs, M Choudhury, P Rosso.Query Expansion for Mixed-Script Information Retrieval*. In Processing's of the 37th international ACM SIGIR conference on Research & development in information retrieval2014.*

[3] Dinesh Kumar Prabhakar, Sukomal Pal (Indian School Of Mines) ISM@FIRE-2014: *Shared task on Transliterated Search FIRE 2014.*

[4] Channa Bankapur, Adithya Abraham Philip, Saimadhav A Heblikar (PES University). Query Word Labeling using Supervised Machine Learning: *Shared task report by PESIT team 2014.*

[5] Utsab Barman, Amitava Das, Code Mixing: A Challenge for Language Identification in the Language of Social Media. Joachim Waanger and Jennifer Foster CNGL Center for Global Intelligent Content National Center for Language Identification 2014.

.

[6] Induja, Indu M, P.C Reghu Raj. Text Based Language Identification System for Indian Languages Following Devanagari. International Journal of Engineering Research & Technology (2014) *(IJERT) IJERTIJERT ISSN: 2278-0181.*

[7] Abinaya.N, Neethu John, Dr.M. Anand Kumar and Dr.K.P. P Soman - Amrita University.AMRITA@FIRE-2014: Named Entity Recognition for Indian Languages *FIRE 2014.*

[8] Kalika Bali, Yogarshi Vyas, Monojit Choudhury– Microsoft India and University of Maryland.POS Tagging of English-Hindi Code-Mixed Social Media Content.*Proceedings of the 2014 EMNLP pages 974–979, October 25-29 (2014).*

[9] Supriya Anand, Bangalore. India. FIRE-2015 Language identification for transliterated forms of Indian Languages queries.

[10] Anupam Jamatia,Amitava Das.Part-of-Speech Tagging System for Indian Social Media Text on Twitter. *Proceedings Workshop on Language Technologies For Indian Social Media(SOCIAL-INDIA), Pages 21-28).*

[11] Yogarshi Vysas, Spandana Gella,Jatin sharma,Kalika Bali, Monojit Choudary.POS Tagging of English-Hindi Code-Mixed Social Media Content. *(EMNLP) Conference on Empirical Methods in Natural Language Processing-2014, Pages 974-979.*