EventXtract-IL: Event Extraction from Social Media Text in Indian Languages @ FIRE 2017 – An Overview

Pattabhi RK Rao and Sobha Lalitha Devi
AU-KBC Research Centre
MIT Campus of Anna University, Chrompet,
Chennai, India
+91 44 22232711
{pattabhi, sobha}@au-kbc.org

ABSTRACT

Today through social media platforms the communication has become exceptionally fast that people across the world get to know any event happening at the nook and corner of the world in a fraction of a second. The penetration of smart phones, tabs etc has significantly changed the way people communicate. Facebook and Twitter are two most popular social media platforms, where people post about events, their personal daily activities and plans. And also post their thoughts, responses or reactions for any public cause or issue. In the recent times we have seen how the facebook posts and twitter tweets have helped in mobilizing people in states such as Tamil Nadu (TN) and Jammu & Kashmir (J&K) in India. The mass public protests for the "Jallikattu" event in TN and stone pelting protests in J&K are prominent examples of how social media has impacted the common man. The information about events or happenings in real time is very valuable to the administration for disaster management, crowd control, public alerting. These information which is used in the development of recommender systems adds value for the growth of business enterprises. Thus there is great need to develop automatic systems for automatic event extraction. This paper presents the overview of the task "Event extraction in Indian languages", a track in FIRE 2017. The task of this track is to extract events from the social media text, The Twitter. Some of the main issues in handling of such social media texts are i) Spelling errors ii) Abbreviated new language vocabulary such as "gr8" for great iii) use of symbols such as emoticons/emojis iv) use of meta tags and hash tags and v) Code mixing, though in this track, we have not considered code mixing. Though event extraction from Indian language texts is gaining attention among Indian research community, however there is no benchmark data available for testing the systems. Hence we have organized the Event Extraction in social media text track for Indian languages (EventXtract-IL) in the Forum for Information Retrieval Evaluation (FIRE). The paper describes the corpus created for three languages, viz., Hindi, Malayalam and Tamil and present the overview of the approaches used by the participants.

CCS Concepts

• **Computing methodologies ~ Artificial intelligence**
• **Computing methodologies ~ Natural language processing**
• **Information systems ~ Information extraction**

Keywords

Event Extraction; Social Media Text; Twitter; Indian Languages; Tamil; Hindi; Malayalam; Event Annotated Corpora for Indian Language data.

1. INTRODUCTION

Over the past decade, Indian language content on various media types such as websites, blogs, email, chats has increased significantly and it is observed that with the advent of smart phones more people are using social media such as twitter, facebook to comment on people, products, services, organizations, governments, etc. Thus it is seen that content growth is driven by people from non-metros and small cities who generally are comfortable with their own mother tongue rather than English. The growth of Indian language content is expected to increase by more than 70% every year. Hence there is a great need to process these data automatically. This requires natural language processing software systems which extracts events, entities or the associations of them. Thus an automatic Event extraction system is required.

The objectives of the evaluation are:

- Creation of benchmark data for Event Extraction in Indian language Social Media text.
- To encourage development of Event extraction systems for Indian language Social Media text.

Event extraction has been actively researched for over last decade. Most of the research has, however, been focused on resource rich languages, such as English, French and Spanish. The scope of this work covers the task of event recognition and extraction in social media text (twitter data) for Indian languages. In the past there were events such as Workshop on NER for South and South East Asian Languages (NER-SSEA, 2008), Workshop on South and South East Asian Natural Language Processing (SANLP, 2010&2011) conducted to bring various research works on NER being done on a single platform. NER-IL tracks at FIRE (Forum for Information Retrieval and Evaluation) in 2013, 2014, and 2015; Code Mix Entity Extraction (CMEE-IL) in 2016 have contributed to the development of benchmark data and boosted the research towards NER for Indian languages. But it is observed that there are very little works in Indian language event extraction. The user generated texts such as twitter and facebook texts are diverse and noisy. These texts contain non-standard spellings and abbreviations, unreliable punctuation styles. Apart from these writing style and language challenges, another challenge is concept drift (Dredze etal., 2010; Fromreide et al., 2014); the distribution of language and topics on Twitter and Facebook is constantly shifting, thus leading to performance degradation of NLP tools over time.

Some of the main issues in handling of such texts are i) Spelling errors ii) Abbreviated new language vocabulary such as "gr8" for

great iii) use of symbols such as emoticons/emojis iv) use of meta tags and hash tags v) Code mixing.
For example:

"Muje kabi bhoolen gy to nhi na? :(Want ur sweet feedback about my FC ? mai dilli jaa rahi hoon".

The research in analyzing the social media data is attempted in English through various shared tasks. Language identification in tweets (tweetLID) shared task held at SEPLN 2014 had the task of identifying the tweets from six different languages. SemEval 2013, 2014 and 2015 held as shared task track where sentiment analysis in tweets were focused. They conducted two sub-tasks namely, contextual polarity disambiguation and message polarity classification. In Indian languages, Amitav et al (2015) had organized a shared task titled 'Sentiment Analysis in Indian languages' as a part of MIKE 2015, where sentiment analysis in tweets is done for tweets in Hindi, Bengali and Tamil language.

Named Entity recognition was explored in twitter through shared task organized by Microsoft as part of 2015 ACL-IJCNLP, a shared task on noisy user-generated text, where they had two sub-tasks namely, twitter text normalization and named entity recognition for English. The ESM-IL track at FIRE 2015 was the came up with the entity annotated benchmark data for the social media text, where the data was in only one language. where users use only one language. But there are no such shared task for event identification and Extraction. Thus there is a need to develop systems that focus on social media texts for event extraction.

The paper is organized as follows: section 2 describes the challenges in event extraction on Indian languages. Section 3 describes the corpus annotation, the tag set and corpus statistics. In section 4 the overview of the approaches used by the participants are described and section 5 concludes the paper.

2. GENERAL CHALLENGES IN INDIAN LANGUAGE EVENT EXTRACTION

The challenges in the development of event extraction systems for Indian languages from social media text arise due to several factors. One of the main factors being there is no annotated data available for any of the Indian languages, Apart from the lack of annotated data, the other factors which differentiate Indian languages from other European languages are the following:

a) **Ambiguity** – Ambiguity between common and proper nouns. Eg: common words such as "Roja" meaning Rose flower is a name of a person.
b) **Spell variations** – One of the major challenges is that different people spell the same entity differently. For example: In Tamil person name -Roja is spelt as "rosa", "roja".
c) **Less Resources** – Most of the Indian languages are less resource languages. There are no automated tools available to perform preprocessing tasks required for NER such as part-of-speech tagging, chunking which can handle social media text.

Apart from these challenges we also find that development of automatic event recognition systems is difficult due to following reasons:

i) Tweets contain a huge range of distinct event types. Almost all these types are relatively infrequent, so even a large sample of manually annotated tweets will contain very few training examples.

ii) In comparison with English, Indian Languages have more dialectal variations. These dialects are mainly influenced by different regions and communities.

iii) Indian Language tweets are multilingual in nature and predominantly contain English words.

The following examples illustrate the usage of English words and spoken, dialectal forms in the tweets.

Example 1 (Tamil):

Ta: Stamp veliyittu ivaga ativaangi .....

En: stamp released these_people get_beaten ….

Ta: othavaangi .... kadasiya <loc>kovai</loc>

En: get_slapped … at_end kovai

Ta: pooyi pallakaatti kuththu vaangiyaachchu.

En: gone show_tooth punch got

("They released stamp, got slapping and beating … at the end reached Kovai and got punched on the face")

This example is a Tamil tweet where it is written in a particular dialect and also has usage of English words.

Similarly in Hindi we find lot of spell variations. Such as for the words "mumbai", "gaandhi", "sambandh", "thanda" there are atleast three different spelling variations.

3. CORPUS DESCRIPTION

The corpus was collected using the twitter API in two different time periods. The training partition of the corpus was collected during June 2017. And the test partition of the corpus was collected during Aug 2017. As explained in the above sections, in the twitter data we observe concept drift. Thus to evaluate how the systems handle concept drift we had collected data in two different time periods. In this present initiative the corpus is available for three Indian languages Hindi, Malayalam and Tamil. The Tables and figures show different aspects of corpus statistics.

ANNOTATION TAGSET

The corpus for each language was annotated manually by trained experts. Event Extraction task requires to identify event trigger keyword and the full event predicate and represent it with a tag. In this work, the data is tagged with one single tag "Event" where a single phrase consisting of Event trigger and the event predicate. For example "Governor for Tamil Nadu appointed". We find that in most of the works in Event extraction in English, Automatic Content Extraction (ACE) Event tag set has been used. In the present work for this track we have only focused on just the extraction one event phrase, which consists of the Even trigger and the whole event predicate which gives the information of where and when the event has happened and who all participants involved in the event. As there is no much work in this area in Indian languages, and to keep the task definition simple, in this edition we have not taken identification of event types, where and who of the events individually.

DATA FORMAT

The participants were provided the data with annotation markup in a separate file called annotation file. The raw tweets were to be separately downloaded using the twitter API. The annotation file

is a column format file, where each column was tab space separated. It consisted of the following columns:

i) Tweet_ID
ii) User_Id
iii) Event string
iv) Event Start_Index
v) EventString_Length

For example:

Tweet_ID:890123456782341
User_Id:987654321
EventString: TN Governor appointed
Index:43
Length:21

Index column is the starting character position of the Event string calculated for each tweet and the count starts from '0'. The participants were also instructed to provide the test file annotations in the same format as given for the training data.
The dataset statistics is as follows:

Table 1. Corpus Statistics

Language	**No. of Tweets**	**No. of Events**
Hindi	5476	1533
Malayalam	7391	1733
Tamil	9147	2074

The data has events from different types such as cyclones, floods, accidents, disease outbreak and political events. And the majority of the types were the disasters and political events such inaugurations/opening ceremonies by political leaders. Also the data had events on movie or audio release functions.

4. SUBMISSION OVERVIEWS

The evaluation metrics used for this task is Precision, Recall and F-measure, which is the widely used metric for this task. A total of 16 teams registered for participating in the track. The final submission was done by 4 teams among the 16 teams. They submitted their test runs for evaluation with multiple runs. A total of 11 test runs were submitted for evaluation. Only 1 team had participated for all the three languages. One teams each participated for Hindi, Tamil and Malayalam.

We had developed a base system without using any pre-processing and lexical resources. The base line system was developed using a CRF classifier which will mark if a word is part of an event phrase or not. The base line system was developed so that it would help in making a better comparative study. The system performance is: precision of 23.87% and recall of 29.67%. It is observed that all the teams outperformed the base system. In the following paragraphs we briefly describe the approaches used by each team. The results of the teams are given in Table 3.

a) Alapan team had used Neural Networks, to develop the system. They had used CNN algorithm in combination with LSTM. They first remove the URLs, emoticons etc from tweets. There is no NLP pre-processing such as POS and Chunking done to the tweets. This team participated in all languages and had submitted 2 runs each for each language.

b) Sharmila team used SVMs for developing the system. The data was preprocessed for tokenization and no cleaning is performed. The task is modeled as simple binary classification task. The team submitted participated for Tamil and submitted three runs.

c) Nageshbhattu team used CRFs for the task. This team pre-processed the data for Part-of-Speech (POS) tagging. They have used POS tags and words in the Window of 5 as features for the CRFs learning. One interesting aspect is that the POS tagger for the general texts has been used for the Tweet data. It will be interesting to know how well a general Newswire POS engine performs on Tweet data. This team participated for Hindi and submitted one run.

d) Manju team used an open source tool called BeautifulSoap to identify the events. This tool is used for website scrapping but here they have used for event classification. The choice of the tool is not appropriate for this task. Infact this method can be said as a "blind mrthod", where almost all the input tweets are marked as events, and by default 1/5th of it has come out correct. This team participated in Malayalam and submitted one run.

The different methodologies used by the teams are summarized in Table 2.

Evaluation

Evaluation metrics used are precision, recall and f-measure. All the systems have been evaluated automatically by comparing with the gold data. The results obtained for each participant is shown in table 3.
One main condition in the Event phrase identification is related to the event span. The span or extent of the Event phrase is to be optimally minimum, it should include Event trigger and the Predicate. Consider the example below

Hi: bahut dinom se kahi jA rahi rAjyapAl ki niyukti, tamilnadu me naye rAjyapAl ki niyukti huA.

Here the event phrase is "tamilnadu me naye rAjyapAl ki niyukti". It can not be just "rAjyapAl ki niyukti".Here the event trigger is "niyukti". The event predicate is " tamilnadu me naye rAjyapAl", from which we get the information where and what.

So the participating system need to identify this exact event phrase. Any system output which has tagged anything more than this extent is considered as wrong.

Thus we define:
Precision,P=(No. Correctly identified Events by the system)/(Total No. of Events identified by the system)

Recall, R=(No. Correctly identified Events by the system)/(Total No.of Events identified in the Gold)

F-measure= (2*P*R)/(P+R)

5. CONCLUSION

The main objective of creating benchmark data representing a few of the popular Indian languages has been achieved. And this data has been made available to research community for free for research purposes. The data is user generated data and is not any genre specific. Efforts are still going on to standardize this data and make it perfect data set for future researchers. We observe that the results obtained are almost similar for all the languages. We hope to see more publications in this area in the coming days from these different research groups who could not submit their results. Also we expect more groups would start using this data for their research work.

This EventXtract-IL track is one of the first efforts towards creation of Event annotated user generated data for Indian languages. The data being generic, this could be used for developing generic systems upon which a domain specific system could be built after customization. In the next edition of this track we plan to add more data and also include identification and extraction of event types, event cause-effects and event participants.

6. ACKNOWLEDGMENTS

We thank the FIRE 2017 organizers for giving us the opportunity to conduct the evaluation exercise. We also thank the Language Editors in CLRG, AU-KBC Research Centre.

7. REFERENCES

[1] Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel Campos, Iñaki Alegría Loinaz, Nora Aranberri, Aitzol Ezeiza, Víctor Fresno. 2014 TweetLID@SEPLN 2014, Girona, Spain, September 16th, 2014. CEUR Workshop Proceedings 1228, CEUR-WS.org 2014

[2] Mark Dredze, Tim Oates, and Christine Piatko. 2010. "We're not in kansas anymore: detecting domainchanges in streams". In *Proceedings of the 2010 Conferenceon Empirical Methods in Natural LanguageProcessing*, pages 585–595. Association for ComputationalLinguistics.

[3] Hege Fromreide, Dirk Hovy, and Anders Søgaard.2014. "Crowdsourcing and annotating ner for twitter#drift". *European language resources distributionagency.*

[4] H.T. Ng, C.Y., Lim, S.K., Foo. 1999. "A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation". In *Proceedings of the {ACL} {SIGLEX} Workshop on Standardizing Lexical Resources {(SIGLEX99)}*. Maryland. pp. 9-13.

[5] Preslav Nakov and Torsten Zesch and Daniel Cer and David Jurgens. 2015. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).*

[6] Nakov, Preslav and Rosenthal, Sara and Kozareva, Zornitsa and Stoyanov, Veselin and Ritter, Alan and Wilson, Theresa. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*

[7] Rajeev Sangal and M. G. Abbas Malik. 2011. *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (SANLP)*

[8] Aravind K. Joshi and M. G. Abbas Malik. 2010. *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing* (SANLP). (http://www.aclweb.org/anthology/W10-36)

[9] Rajeev Sangal, Dipti Misra Sharma and Anil Kumar Singh. 2008. *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. (http://www.aclweb.org/anthology/I/I08/I08-03)

[10] Pattabhi RK Rao, CS Malarkodi, Vijay Sundar R and Sobha Lalitha Devi. 2014. Proceedings of Named-Entity Recognition Indian Languages track at FIRE 2014. http://au-kbc.org/nlp/NER-FIRE2014/

Table 2. Participant Team Overview - Summary

Team	Languages & System Submissions	Approaches (ML method) Used	Pre-Processing Step	Lexical Resources Used	Open Source NLP Tools Used	Variation Between Runs
Alapan – IIT-Kgp	i)Hindi: 2 runs ii) Malayalam: 2 runs iii) Tamil: 2 runs	Run1: Neural Networks – CNN architecture with LSTM , pipelined process flow Run2: Neural Networks – CNN architecture with LSTM , non- pipelined process flow	Tweet Preprocessor alone used to eliminate http links, emoticons	NIL	CNN – ML tool	Pipelined Process Flow and Non-pipeline process flow
Sharmila – Karpagam Eng. College (KEC)	i) Tamil: 3 runs	SVMs – words, prefixes, suffixes and shape features used	Tweet cleaning and Tokenization	NIL	SVM Tool kit	Run 1: C-parametre of SVM is tuned Run 2: Without any parametre tuning of the SVM tool kit Run 3: Tuning of all other parameters of the SVM tool kit
Nageshbhattu - IDRBT	i) Hindi: 1 run	CRFs –	NLP pre-processing – Uses general Text POS tagger	NIL	POS tagger and CRFs tool kit	N/A
Manju – CEC, Chertala	i) Malayalam:1 run	NIL	Tweet cleaning	NIL	BeautifulSoap – a web scrapping tool	N/A

Table 3. Evaluation Results of Participating Systems

Team	**Language**	**Submission 1**			**Submission 2**			**Submission 3**		
		Prec %	Rec %	F-m%	Prec %	Rec %	F-m%	Prec %	Rec %	F-m%
IIT Kgp	**Hindi**	**36.58**	**79.02**	**50.01**	31.42	56.37	40.35	NA	NA	NA
	Malayalam	**32.98**	**90.20**	**48.29**	39.98	57.50	47.17	NA	NA	NA
	Tamil	**43.16**	**64.77**	**51.80**	39.73	49.33	44.01	NA	NA	NA
IDRBT Hyderabad	Hindi	31.56	71.39	43.77	NA	NA	NA	NA	NA	NA
KEC, Coimbatore	Tamil	39.10	62.28	48.04	38.05	51.81	43.88	38.44	61.14	47.20
CEC Cherthala	Malayalam	21.43	67.17	32.40	NA	NA	NA	NA	NA	NA