EventXtract-IL: Event Extraction from Newswires and Social Media Text in Indian Languages @ FIRE 2018 – An Overview

Pattabhi RK Rao, Sobha Lalitha Devi
AU-KBC Research Centre, MIT, Anna University, Chromepet, Chennai, India
pattabhi@au-kbc.org, sobha@au-kbc.org

Abstract. Today communication has become very fast and is happening in real time. An event that happens in any part of the world gets communicated in few seconds/minutes to the rest of the world. For example the recent twin bomb blasts in Damascus, Syria was known to the world within few minutes. This event was broadcasted in various media channels. . The penetration of smart phones, tabs etc., has significantly changed the way people communicate. 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 a great need to develop systems which can automatically identify various events such as bomb blasts, floods, cyclone, fires, political events etc., reported in various Newswires, Social Media text. This is the 2nd edition of the track. The first edition of this track was conducted last year at FIRE 2017. In that edition the task was to identify only the event and event span given in the data. Thus further going ahead in this track, along with the identification of event and its span, it is necessary to identify the cause and effects of a given event. The actual real time applications will be benefited only if the full information related to the event is identified. For example for a bomb blast, it will be required to know where it has occurred, when it has occurred, who and what all got effected, what are the causalities etc. In this edition of the track we propose to provide data annotated with the cause and effect details of an event and participants are required to identify these details along with event identification. And as in the last year, the focus is on Indian languages text. This paper presents the overview of the task "Event extraction in Indian languages", a track in FIRE 2018. The task of this track is to extract events and all other associated arguments or information such as locations, cause, and its effects from the text. 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 this track in the Forum for Information Retrieval Evaluation (FIRE). The paper describes the corpus created for two Indian languages, viz., Hindi, and Tamil and present the overview of the approaches used by the participants.

Keywords: Event Extraction, Social Media Text, Indian Languages, Tamil, Hindi, 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 newswire, social media text such as facebook for Indian languages. In the past there were events such as Workshop on NER for South and South East Asian Languages [9], Workshop on South and South East Asian Natural Language Processing [7][8] 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 [10]; 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 [3][4] the distribution of language and topics on Twitter and Facebook is constantly shifting, thus leading to performance degradation of NLP tools over time.

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 [2] 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 [1] 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 came up with the named entity annotated benchmark data for the social media text. And CMEE-IL Track of 2016 came for named entity annotation detection for code-mixed data. The task of Event identification in Indian languages is at nascent stage. EventXtraction track at FIRE 2017 is the first step towards creating benchmark data and boosting the research in Indian language event extraction. This edition of the EventXtraction track at FIRE 2018 focusses on complete extraction of events and their associated arguments.

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 text arise due to several factors. One of the main factors being there is no annotated data available for any of the Indian languages. In general the following well known linguistic characteristics in Indian languages also make the task more challenging.

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

ii) Indian Language text are multilingual in nature and predominantly contain English words.

iii) Event triggers are ambiguous and require context, just occurrence of a trigger key term need not necessarily indicate an occurrence of event.

iv) Identifying all the linking arguments associated with an event is more difficult problem. For example identifying the cause of event is relatively easy in English but more difficult in Indian languages. This is due to the fact that discourse markers in English have explicit words whereas it is not so in Indian languages, it is expressed as inflection markers.

3 Corpus Annotation

The corpus was collected in two different time periods. The training partition of the corpus was collected during June 2018. And the test partition of the corpus was collected during Aug 2018. In this present initiative the corpus is available for two Indian languages Hindi and Tamil along with English.

3.1 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. An event can be an occurrence happening in certain place during a particular interval of time with or without the participation of human agents. It may be a part of chain of occurrences or an outcome or effect of preceding occurrence or a cause of succeeding occurrences. An event can occur naturally or it can be because of human actions. In this work, we have focused on disaster and entertainment events. And also we need to identify the arguments of the event such place of happening, people involved, cause and effects of the event to get a complete information. In this track the data is annotated with the complete information. The participants have to identify and extract the event and all the relevant arguments of the event. One of the well-known event annotation tagset used most of the works in English is Automatic Content Extraction (ACE) Event tag set. In this work we have developed our own annotation guidelines which is inspired by the ACE guidelines. We differ in defining what to be tagged as event and also the type of arguments for an event type. The event tag phrase consists of Event Trigger and the event predicate. For example in the sentence “The central government had appointed a full time Governor for Tamil Nadu.”. In this we tag the phrase “appointed a full time Governor for Tamil Nadu” as the event. The arguments for this event are “who did the appointment”, “who is the appointee” and place. The generic tagset for different event types is as follows:

i) EVENT TAG
ii) ARGUMENTS:
  a. EventType: Types such as Eg:Natural/Manmade/Meeting]
  b. Location: Place of occurrence
  c. Agents involved:
    i. Cause: Living
    ii. Cause: Thing
    iii. Effect: Living
    iv. Effect: Thing
  d. Temporal
    i. Date

ii. Time

e. Miscellaneous

3.2 Data Format

Here we have followed HTML style of annotation in this work. The general syntax for the event tagging is as given below.

<EVENT ID=”number” TYPE=“abc” SUBTYPE_1=“xyz” SUBTYPE_2=“def”>Event Trigger</EVENT>

Here, this event tag has attributes:

i) ID -- This is a number which will be unique for each event in a given document.

ii) TYPE – This is the type of the event such as “manmade disaster”

iii) SUBTYPEs – These are the subtype category names of the particular event.

Event arguments such as participants, time of occurrence, and location of occurrence are also annotated using HTML style. For example the “time of occurrence” attribute of an event will be annotated as follows:

<TIME-ARG REF-ID=”eventID”> abc </TIME>

Each argument tag of an event will have the attribute “REF-ID”, which is a number that refers to the event ID of the event to which the argument belongs.

The different types of event argument tags are as follows:

a) <TIME-ARG>

b) <CAUSE-ARG>

c) <CAUSUALITIES-ARG>

d) <PLACE-ARG>

e) <EFFECTS-ARG>

Example:

RAW TEXT:

On 29 December 2017 a massive fire broke in Kamala Mills, Mumbai the capital of Maharastra, killed at least 14 people and injured several.

ANNOTATED TEXT:

On *<TIME-ARG REF-ID=”1”>*19 JULY 2018*</TIME-ARG>*, a massive *<EVENT TYPE= “Manmade Disaster” ID=”1” SUBTYPE_1= “Accident” SUBTYPE_1.1= “Fire Accident”>fire broke</EVENT> in <PLACE-ARG REF-ID=”1”>Kamala Mills, Mumbai the capital of Maharastra</PLACE-ARG>, killed <CAUSALITIES-ARG REF-ID=”1”>at least 14 people and injured several</ CAUSALITIES-ARG>.*

The participants were provided the data with the above explained annotation markup in a separate file called annotation file. 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	Number of Docs		No. of Events	
	Training	**Testing**	**Training**	**Testing**
English	100	803	934	1040
Hindi	107	311	943	380
Tamil	64	1438	1274	1652

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

A total of 10 teams registered for participating in the track. The final submissions were done by 2 teams among the 10 teams. They submitted their test runs for evaluation. A total of 5 test runs were submitted for evaluation. Only 1 team had participated for all the three languages. Another team had participated for English and Hindi.

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 phrase is an event phrase or not. The baseline system performed with a Precision of 0.4521 and Recall of 0.6522 for event identification. The different methodologies used by the teams are summarized in Table 2.

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**
Alapan et al – IIT-Kgp	Hindi: 1 run Tamil: 1 run English: 1 run	Neural Networks – CNN architecture with Bi-LSTM	Preprocessor alone used to eliminate http links, emoticons	NIL	fastText Toolkit

Anita et al – IIT BHU	Hindi: 1 run English: 1 run	Rule based – based on patterns and POS, NER features used	cleaning and Tokenization	NIL	NLTK Tool kit

4.1 Evaluation

Evaluation metrics used are the well measures precision, recall and f-measure. All the systems have been evaluated automatically by comparing with the gold data. 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)

The methodology for calculating the Precision and Recall will be field based average score. For example, for an Event E1, if there are 6 fields such as Event Type, Event Location, Event Date, Event Actors/Participants, Causes, Effects. Then for that event E1, if all these fields are identified correctly then the system gets full score of 7/7 else according to the identified fields the score will be modified. And finally micro and macro-average of the Precision and Recall will be calculated and final score is arrived at. The results obtained for the system runs is presented in Table 3.

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 online Newswire data. 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 encouraging but scores are low and need lots of improvement for real time use. We aim to provide a more standard and corrected data for these 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. In this edition of the track we have provided data which can be used to develop a complete Event extraction engine, so that real time systems can be developed in near future. We plan to add few more other languages data.

Table 3. Evaluation Results of Participating Systems

Team	***Language***	***Submissions***		
		Precision %	Recall %	F-measure%
Alapan - IIT Kgp	Hindi	62.85	29.02	**39.71**
	Tamil	59.98	27.20	**37.42**
	English	65.16	28.77	**39.91**
Anitha – IIT Bhu	Hindi	29.65	61.39	**39.98**
	English	34.54	64.87	**45.07**

6 Acknowledgements

We thank the FIRE 2018 organizers for giving us the opportunity to conduct the evaluation exercise. We also thank the Language Editors in our CLRG lab, AU-KBC Research Centre for annotating the data.

7 References

1. Amitava Das, Dipankar Das, Manish Shrivastava, Rajendra Prasath. 2015. Shared Task on Sentiment Analysis in Indian Languages Tweets in MIKE 2015 (SAIL 2015).
2. 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
3. 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 Computational-Linguistics.
4. Hege Fromreide, Dirk Hovy, and Anders Søgaard.2014. "Crowdsourcing and annotating ner for twitter#drift". *European language resources distributionagency.*
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/