# Development of an Event Extraction System from Newswires and Social media texts for Malayalam- An Experiment

Manju K
College of Engineering,Cherthala
Cherthala, Kerala, INDIA
manju@cectl.ac.in

Sumam Mary Idicula
Department of Computer Science,CUSAT
Kochi, INDIA
sumam@cusat.ac.in

David Peter S
Department of Computer Science,CUSAT
Kochi, INDIA
davidpeter123@gmail.com

## ABSTRACT

In recent years with the advancement in internet technologies and increase of smartphone adoption among youngsters has made information easily accessible in different forms such as text, images, audio and videos. Today communication has become very fast, it is very much possible that an event that happens in any part of the world gets communicated in few seconds/minutes to the rest of the world. This demands for a system that can automatically identify events such as bomb blasts, floods, cyclone, fires, political events etc., reported in various Newswires and Social Media text. In this paper we explore the development of an event extraction system for Malayalam Language. This work was done as part of the shared task on Event Extraction from Newswires and Social Media Text in Indian Languages (EventXtract-IL) in Forum for Information Retrieval and Evaluation(FIRE 2017). The experiments were done on the standard data set provided and the results showed that the system was able to give performance comparable to the methods employing more sophisticated procedures.

## CCS CONCEPTS

• **Natural Language Processing**; • **Information Retrieval** → *Event Extraction*; • **Text Summarization**;

## KEYWORDS

Event extraction, preprocessing, structured information

## 1 INTRODUCTION

Information on a web is increasing at infinitum. There exists a plethora of information in electronic and digital form. This information load demands for some automatic help. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine readable documents[1]. Extracted structured information can be used for variety of enterprise or personal level task of varying complexity. Event extraction is a subproblem of Information Extraction which aims to extract meaningful information called events from from different sources of information.

The objective of this task is to encourage in development of systems for the identification of Events in the various types of text such as Newswire, Facebook, and Twitter etc., for Indian languages. This extraction system tries to identify crisis events such as bomb blasts, floods, cyclone, fires, political events etc., reported in various Newswires, Social Media text,etc.. The events are categorised into groups such as natural disasters(floods, earthquakes etc), man made disasters (accidents, crime etc), political events (inaugurations by political leaders, political rallies etc), cultural/social events (Seminars, Conferences, light music events etc)..

## 2 TASK DESCRIPTION

In this initiative the organizers suggested three Indian languages Hindi, Malayalam and Tamil and we have chosen Malayalam Language. For training the model we were provided with two files, one which is the raw file and another file containing the annotations. The raw file consists of the twitter obtained from the Twitter API. The second file, the annotation file, consists of annotation of tweets which are events. This is a five column file First two columns are the TweetID and UserID as found in the Original Tweet File. The third column is the Event phrase of the tweet, Fourth column is the index where this phrase starts in the tweet string. Fifth column is the string length of the event phrase. In some instances full tweet string is considered as Event Phrase.

The test data, provided was tweets file similar to the original Tweets files provided during the training. We were asked to submit the annotations file similar to the annotations file provided during training.

### 2.1 Methodology

When a typical event extraction system processes documents in a large collection, it primarily uses prior knowledge in the form of extraction patterns, classifiers trained on annotated corpora, ontologies, and so on [2].

Events extracted in our proposed framework are represented as a 5-tuple $\langle U, T, S, I, L\rangle$ where U is the userid, T is the twitterid, S the event string, I the index value and L the length of the event string. Our proposed framework consists of three main steps, pre-processing, event extraction and writing the annotations to an outputfile. The details of our proposed framework are described below.

*2.1.1 Preprocessing.* The preprocessing consists of the application of some Natural Language Processing (NLP) tools to the raw text in order to segment it into sentences and remove the unwanted characters. This is followed by extracting the userid, twitterid, event phrase and url into respective variables.

*2.1.2 Event Extraction.* From the url using some web scraping tool such as Beautiful Soup, the tweet string can be extracted. The Tweet Id can be used to determine the crisis type of the event. This can me done by maintaining a hash table. Locate the event phrase in the Tweet string and extract the sentence as the event string.With the Event string and Tweet string the start index of the event string in tweet string as well as the length of the event phrase can be determined.

*2.1.3 Annotation File.* As a last step the value of userid, twitterid, event phrase, index and length are written into a file.

# 3 EXPERIMENTATION AND RESULTS

We have carried out the evaluation of the proposed system using the data provided by the shared task.The system was developed based on the training data provided and the evaluation was done based on the test data. For analysisng the performance of the system we have determined their precision, recall and F-measure. Let $E$ be the set of automatically annotated events in the test corpus and let $\acute{E}$ be the set of events annotated by a human expert. We call the latter set the ground truth. The recall is the ratio of correctly detected events and all events in the ground truth, i.e. recall $=\mid \frac{E \cap \acute{E}}{\acute{E}} \mid$ . We define precision as the fraction of events, which were correctly annotated by the framework as crisis events, i.e. precision $=\mid \frac{E \cap \acute{E}}{E} \mid$. F-measure is defined as the mean of precision and recall, i.e. Fmeasure $= \frac{(2 * precision * recall)}{(precision + recall)}$ presents the top results obtained for precision, recall and F-measure separately. Table 1 shows the result obtained interms of precision, recall and accuracy.

**Table 1: Results**

| Language | Submission 1 | | |
|---|---|---|---|
| | Precision | Recall | F-measure |
| Malayalam | 21.43% | 67.17% | 32.40% |

# 4 CONCLUSIONS

In this paper, we have proposed an event extraction system where in the preprocessing phase we have used NLP Tools to prepare the data followed by event extraction and annotation. The framework was evaluated based on the data set provided by the shared task for event extraction.

# REFERENCES


[1] J. Graf, V. Koroteyev, E. Mikhaylov, E. Bricker, B. Levy, and A. Wong. 2006. Extracting data from semi-structured text documents. (Oct. 26 2006). https://www.google.com/patents/US20060242180 US Patent App. 10/565,611.

[2] Peiquan Jin, Lin Mu, Lizhou Zheng, Jie Zhao, and Lihua Yue. 2017. News Feature Extraction for Events on Social Network Platforms. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 69–78.