KCE_DAlab @ EventXtract-IL-FIRE2017: Event Extraction using Support Vector Machines

SharmilaDevi V, Kannimuthu S
Department of Information Technology,
Karpagam College of Engineering,
Coimbatore, India

Safeeq G
Department of Information Technology,
Sri Ramakrishna Institute of Technology,
Coimbatore, India

Anand Kumar M
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering,
Coimbatore, Amrita Vishwa Vidyapeetham, India

ABSTRACT

Nowadays, Social media has become a major part to transfer the message that must be shared with the people ideas and express the information globally in our day-to-day life. Through social media can able to connect the people together, they are vulnerable to crimes like Identity thefts, false information, and identity masking etc. Identifying the event from the social media messages and news headlines are the important area of research in the current era. This paper illustrates work done on Event Extraction for Indian language shared task which is conducted in Forum for Information Retrieval Evaluation (FIRE) 2017. For this Event extraction task, organizers release the dataset with three languages Tamil, Hindi, and Malayalam. Each language dataset consists of two files Original Tweet file and Annotation files. We only participated in the Tamil event extraction task. In this task, we converted the original tweet file into Bio-format to apply the machine learning directly. Then analyzing each chunk of the word is an event is said to [B] beginner and the other events will be given as Intermediate and the others are assigned as O tag. Each word or chunk should be trained whether it is the event or not an event with the help of rich features and SVM classifier. Here we also find out the Cross- Validation accuracy using Natural language techniques.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Language resources**; **Feature selection**;

KEYWORDS

Indian Language, Event extraction, Social Media, Text Classification

1 INTRODUCTION

Natural Language Processing (NLP) is a field that covers computer understanding and manipulation of human languages. It focuses on the interaction between human language and computer is called Natural language processing. Event Extraction is an important stream of information extracted it has greatly gained in popularity due to the advent of big data and the developments in the related fields of text mining in Natural Language Processing. One common application of text mining is event extraction which encompasses deducing specific knowledge concerning incidents referred to in texts. Most of the data is initially unstructured. Using NLP techniques, information is extracted from texts from various sources such as new messages and blogs that must be stored in a structured way eg. Databases. The event can be useful in some applications like risk analysis, monitoring systems and decision making supporting tools. The event must be used in three methods that is data to driven knowledge, extract knowledge through representation and exploitation of expert knowledge and hybrid event extraction.

With the enormous content of data and the impact of digital data sources are easily extracted. Most of the data is in an unstructured format that is human can easily understand the language.The data that are given here is to be converted to machine understandable language. The application that is mainly used in Information retrieval and Information Extraction Methods. Information Extraction is the method of automatically extracting structured information from the unstructured or semi-structured machine-readable documents. In related work, the open dataset for event extraction for the English language is explored in [6]. Here the corpus raises two main issues. It was annotated with templates describing all events with the same set of slots. The methodology used in this type is Annotation and ASTRE corpus. In this paper, the ASTRE corpus, a new corpus dedicated to the evaluation of event schema induction. Template-based Information Extraction without template is discussed in [2]. The template defines a specific type of event with a set of semantic roles for the typical entities involved in such an event. The methodology in this paper is learning templates from raw text and clustering on event distance.

Distant supervision approach to template-based event extraction, focusing on the extraction of passenger counts, aircraft types, and other facts concerning airplane crash events is explored in [8]. They also presented a publicly available dataset and event extraction task in the plane crash domain based on Wikipedia infoboxes and newswire text.

Event extraction is treated as an dependency parsing in [5]. Here, authors proposed a simple approach for the extraction of such structures by taking the tree of event-argument relation. This gives the better performance in the extraction of a biomedical event.

In [9], Event Extraction from unstructured text data was explained. Authors extended the bootstrapping method that was initially developed for extracting relations from web pages to the problem of content extraction from short unstructured text. The event extraction method proposed in this paper attained less accuracy for the Twitter dataset as compared to the enterprise dataset.

An overview of event extraction from text was described in [3]. This literature survey discussed the text mining techniques that are employed for various event extraction purposes. Here knowledge driven event extraction and Hybrid driven even extraction methods are discussed elaborately.

Table 1: EventXtract-IL Tamil Dataset

Files	Training	Testing
Annotations	1109	-
News Headlines	3843	5304
Unique Authors	1799	2509

The rest of the paper is organized as follows. Section 2 presents the overview of the shared task and the details regarding the dataset. Section 3 describes the proposed system developed for the event extraction task while Section 4 shows the evaluation results of three submissions for Tamil event extraction shared task. Finally, Section 5 concludes the paper.

2 DATASET DETAILS

The task contains two files such as Original tweet file and Annotation files. The first two column must contain Tweet ID and user ID. The third column must represent the event phrase of the ID. The Fourth column will mention the index where this phrase starts in the tweet string and the fifth column is the string length of the event phrase. The events are given as Natural disasters, Man-made disasters, political events and cultural /social events.

3 EVENT EXTRACTION FOR TAMIL LANGUAGE

Normally, for Text mining and Information Extraction, preprocessing is the mandatory step and it is necessary for the Twitter dataset. The methodology which is followed in the entity extraction is followed in the event extraction too [7] [1] . The preprocessing step encompass Normalization and Tokenisation methods. In Tokenisation, based on the white spaces, sentences are partitioned into tokens. These tokens are further normalized where superficial variations are extracted. However, normalization of Twitter messages is desired to prevailing the non-standard words, spelling digression, lengthen the unconstrained abbreviations (eg., tmrw for tomorrow), and prevailing the phonetic alternation. For English language, case folding is a relevant one where case variations must be obtained but it does not feel necessary for Indian language where no such variation exists.The methodology of the proposed system is illustrated in Figure 1. The training dataset consists of two files such as raw tweets and extracted type annotated entities. The tweet file will be expressed by "Tweet ID","User ID" and tweets. The entity file must be expressed of "Tweet ID","User ID", Entity type, entity, starting index and length. We have merged these files and converted into conventional BIO formatted text in which B-XXX tag refers the Beginning word of the entity type and I-XXX is needed for the following chunks of an entity. The tag other than the event is represented as O. In tokenization the tweets are further partitioned into small chunks called as tokens. Training and testing tweets must be tokenized properly in one token-per line format. Annotated events and tokenized training tweets are combined to create the BIO format. Features are extracted in Tamil and train the system with support vector machine-based classifier, SVMLight [4]. Finally, the BIO format tokens are converted into the given annotation format and the event is extracted.

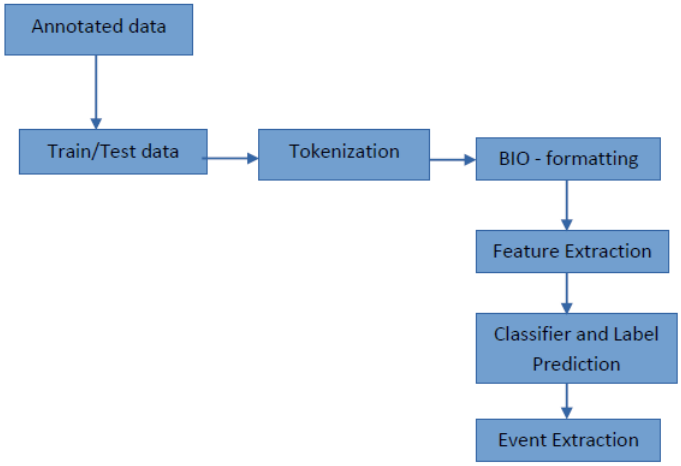

Figure 1: Methodology

Table 2: EventXtract-IL Results

	Tamil		
KCE_DAlab	Prec %	Rec %	F-m%
Submission1	39.1	62.28	48.04
Submission2	38.05	51.81	43.88
Submission3	38.44	61.14	47.2

3.1 Features for Event extraction

In this work, feature extraction is essential as this decides the accuracy of the machine learning based system. The traditional features like words, prefixes, and suffixes of the word, binary feature, shape features are used in the feature extraction step. Binary feature and shape feature is binary features where if it is present in the tweet then it is marked as '1' or else '0'. For prefix and suffix feature maximum up to five characters before and after the current character are taken as features. The punctuation mark such as question mark, exclamatory marks, comma, and full stop are also used as features.

4 RESULTS

This section explains the submission details and the results obtained. The results are shown in Table.2, Submission-2, is the baseline system and submission-1 undergone the C-parameter tuning of SVM. In, submission-3 the parameters are fixed based on 10-fold cross-validation.

5 CONCLUSION AND FUTURE SCOPE

The work is submitted as a part of Shared Task on Event Extraction for the Tamil Language in FIRE 2017. The task organizer provided the twitter file and annotation file. Three submissions were submitted for the task using the traditional features. The system was trained and tested using SVM classifier. In future, POS tagging and the NER features along with word embedding can be added to improve the performance of the event extraction system.

ACKNOWLEDGEMENT

We would like to thank organizers of Forum for Information Retrieval Evaluation 2017 for providing the shared task platform to the researchers. We would also like to thank the organizers of the EventXtract-IL task.

REFERENCES

[1] M. Anand Kumar, S. Se, and K. Soman. Amrita-cen@fire 2015: Extracting entities for social media texts in indian languages. volume 1587, pages 85–88, 2015.

[2] N. Chambers and D. Jurafsky. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 976–986. Association for Computational Linguistics, 2011.

[3] F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong. An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011)*, volume 779, pages 48–57, 2011.

[4] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[5] D. McClosky, M. Surdeanu, and C. D. Manning. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1626–1635. Association for Computational Linguistics, 2011.

[6] K.-H. Nguyen, X. Tannier, O. Ferret, and R. BesanÄğon. A dataset for open event extraction in english. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1939–1943, 2016.

[7] G. Remmiya Devi, P. Veena, M. Anand Kumar, and K. Soman. Amrita-cen@fire 2016: Code-mix entity extraction for hindi-english and tamil-english tweets. volume 1737, pages 304–308, 2016.

[8] K. Reschke, M. Jankowiak, M. Surdeanu, C. D. Manning, and D. Jurafsky. Event extraction using distant supervision.

[9] C. Shang, A. Panangadan, and V. K. Prasanna. Event extraction from unstructured text data. In *International Conference on Database and Expert Systems Applications*, pages 543–557. Springer, 2015.