Event Detection from News in Indian Languages Using Similarity Based Pattern Finding Approach

Shubham Basak[a]

[a]*Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700 108, WB, India*

Abstract

In this work, we propose a rule based method to identify the event type and create a frame for that event. With the help of Natural Language Toolkit (NLTK) and preloaded SpaCy models, we have tried to define certain methods to identify the event (for task 1) and create the event frame from the given article (for task 2).

Keywords

Natural language processing, rule based system, bag-of-words approach, natural language toolkit

1. Introduction

The use of language processing modules such as parser, chunkers, stemmers has been growing steadily in recent years. The event extraction is a significant and necessary aspect of the natural language processing (NLP) & computational linguistics.

The identification[1][2] or detection of event stimuli is an important and vital process of extracting events, since the same occurrence can exist in different trigger words and language may represent multiple event forms in different contexts.

Event detection is the mechanism by which event types are analysed so as to identify collections of events relating to event patterns in the sense of an event. The patterns of the occurrence and context describe forms of events. If during the study a series of events following the sequence of an event type is detected, then the event type perpetrators should be alerted. Typically, the analysis involves filtering and integrating the events.

Event frame creation[3] is a work to create a suitable collection of frames to identify any event correctly. We need an event type & subtype to identify what kinds of event happened. In order to know, where and when the event took place, we need to find a place and a time argument to denote the key places and date & time respectively. Lastly the main tasks e.g. why the disaster happened & for this disaster, what is the number of casualties, need to be recovered from the given data.

Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad

shubhambsk@gmail.com (S. Basak)

0000-0003-4243-7983 (S. Basak)

2. Methodology

In order to detect a disaster, whether manmade or natural and their subtypes, we store certain terms and categories from the train datasets that define the type of disaster. Then using the bag-of-words approach, we find the similar words in test datasets. For each similar word in a file, we increase counter for the type & subtype that the word is representing. We denote those that has maximum occurrence with the word.

SpaCy is an open source software library available for advanced natural language processing. It has pre-built neural network models to perform many operations such as parts-of-speech tagging, named entity recognition tagging, dependency parsing etc. such works on mainly European languages. We use preloaded SpaCy models to perform named entity recognition[4] tagging operation and identify the places and time variables correctly.

To find the reason data, we search for all possible substrings that has a common keyword from disaster data, then we continue until any specified stopword occurs. If we get the reason substring that contains only very few number of words i.e. not be understandable by human being, we discard those and move forward until we find any such specific sentence.

For casualty statement, we find such substrings that contains a keyword such as injure or kill or dead along with a cardinal to represent the number of casualties occurred during that disaster. If we can find only some cardinal data, but no such specified keywords can be found within that, we generally discard that substring and looking forward for another substring with similar kind of data.

3. Experimental Result

We ran our program for both event detection (task 1) & event frame creation (task 2) for English Dataset and only event detection task for Bengali Dataset. The precision, recall & f1_score for our results are given by EDNIL 2020 organizers[5]. We got the result listed in these tables below-

Dataset (Task)	Precision	Recall	F1_score
English Dataset (Task 1)	0.3109475621	0.3400402414	0.3248438251
English Dataset (Task 2)	0.1128436602	0.1093439364	0.1110662359
Bengali Dataset (Task 1)	0.09563994374	0.1073401736	0.1011528449

Table 1: Precision, Recall & F1_score of our experimental result

4. Limitations of Our Implementation

Despite of our method being rule based, we faced some constraints during the implementation. And as our results are not a state-of-art being such simple technique. Here are some of our restrictions-

- As we use SpaCy models, it is very difficult for this model to identify arguments in regional Indian languages.
- As we use bag-of-words approach, so if there exists any other type of disaster in the given data, our method may not capture that.
- For generating the reason argument string, we have to define certain stopwords, otherwise it will be impossible for the program to select those strings.

References

[1] X. Feng, L. Huang, D. Tang, H. Ji, B. Qin, T. Liu, A language-independent neural network for event detection, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 66–71. URL: https://www.aclweb.org/anthology/P16-2011. doi:10.18653/v1/P16-2011.

[2] H.-H. Chen, L.-W. Ku, An NLP & IR Approach to Topic Detection, Springer US, Boston, MA, 2002, pp. 243–264. URL: https://doi.org/10.1007/978-1-4615-0933-2_12. doi:10.1007/978-1-4615-0933-2_12.

[3] Y. Wei, L. Singh, B. Gallagher, D. Buttler, Overlapping target event and story line detection of online newspaper articles, in: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016, pp. 222–232. doi:10.1109/DSAA.2016.30.

[4] G. Kumaran, J. Allan, Text classification and named entities for new event detection, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 297–304. URL: https://doi.org/10.1145/1008992.1009044. doi:10.1145/1008992.1009044.

[5] B. Dave, S. Gangopadhyay, P. Majumder, P. Bhattacharya, S. Sarkar, S. L. Devi, Overview of the FIRE 2020 EDNIL track: Event Detection from News in Indian Languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, CEUR Workshop Proceedings, CEUR-WS.org, 2020.