SAVITR: A Real-time Location Inference System during Emergencies

Ritam Dutt, Kaustubh Hiware, Avijit Ghosh, Rameshwar Bhaskaran

Indian Institute of Technology, Kharagpur

Abstract

We present SAVITR, an automated, real-time system which utilises the information posted on Twitter to observe and analyse emergency situations. Our system employs natural language processing techniques to extract locations in an unsupervised fashion and projects the results on a map-based platform. SAVITR is designed specifically for efficient performance, achieving a decent F1-score of 0.81 and is approximately two orders of magnitude faster than currently available systems. A prototype of the system is deployed live at http://savitr.herokuapp.com.

Salient Features

- Savitr infers the location from the tweet **automatically**, unlike Ushahidi [3] and MapBox which needs explicit labelling.
- The proposed method achieves the **highest F-score** amongst all baseline methods.
- Our system is several orders of magnitude **faster** and thus suitable for real-time deployment.

Flowchart of our proposed methodology

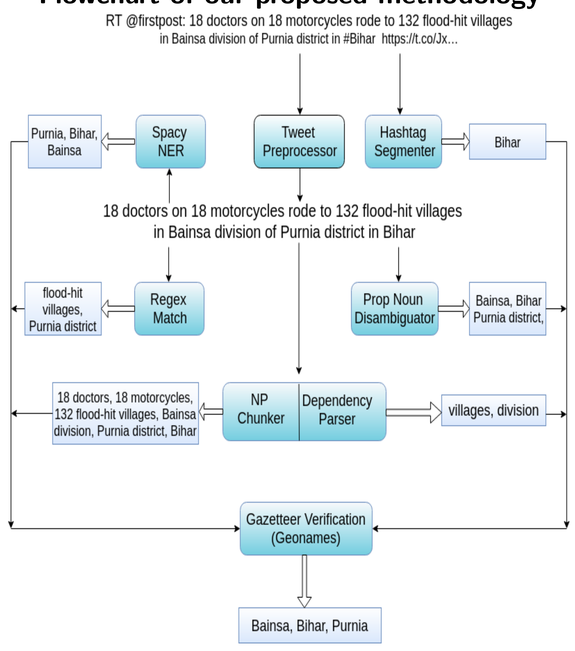

Figure 1:

Proposed Methodology

- Statistically segmented hashtag (#Nepalquake) into distinct words [2], but also retained the original one (Bengaluru yields Bengal and Uru)
- Preprocessed the tweet text to remove URLs, mentions, ellipses, emojis and stray characters. Abstained from case folding and stemming.
- Disambiguated Proper Nouns using a POS Tagger and considered those which were preceeded by viable location-based prepositions (in, from) or appended by common suffixes of locations.
- Compiled a list of common location-based suffixes and direction-based prefixes. Applied regex similarity to identify the words surrounding these suffix and prefix elements as viable locations.
- Performed Dependency Parsing on the processed tweet text to obtain nouns which are at short dependency distance from the emergency word. We also account for noun phrase like [2] and the words identified as locations by a publicly available NER like coreNLP, as used by [2, 1].
- Gazetteer Verification of these potential phrases and words to retain only those that correspond to real-world locations. The gazetteer choice depends on the granularity and precision of our location and also on performance speed.

Type	Common Examples
Landforms	lake, steam,island, valley, mountain
Roads	street, boulevard, rd, lane, bridge
Buildings	hospital, school, villa, mosque
Towns	city, district, village,town, nagar
Directions	north, eastern, NE, north east
Diseases	dengue, cholera, zika, malaria
Disasters	earthquake, floods, landslide, tsunami

Table 1: Examples of suffixes and emergency-related words

Dataset

Twitter Streaming API was used to collect tweets pertaining to 'dengue' or 'flood'. The 317,567 tweets were preprocessed to remove duplicates and non-English tweets resulting in 239,276 distinct tweets.

Gazetteer Employed

We employ a gazetteer to verify if a place is located within India's geographical boundaries and obtain its geo-spatial coordinates. The gazetteers used are

- The data publicly scraped from Geonames has 449, 973 locations within India, but lacks finer-grained information such as roads, buildings.
- The Open Street Map (OSM) gazetteer has a comprehensive list of all addresses in India, but API calls take considerable time than GeoNames.

We consider two variants of our proposed algorithm.

- GeoLoc - Our algorithm using Geonames.
- OSMLoc - Our algorithm using OSM

Baseline Methodology

- UniLoc- Take the unigrams in the processed tweet text and infer if they are viable locations.
- BiLoc- Similar to UniLoc, except we consider both unigrams and bigrams in the tweet text.
- StanfordNER - Employ the NER of coreNLP
- TwitterNLP - Employ the Twitter NLP NER
- GoogleCloud - Use the Google Cloud Natural Language Platform to infer locations.
- SpaCyNER - Use the SpaCy NER tagger.

Observation

Method	Precision	Recall	F-score	Time(in s)
UNILoc	0.3848	0.7852	0.5165	0.0553
BILoc	0.4025	0.8590	0.5482	0.0624
StanfordNER	0.8103	0.6322	0.6988	175.0124
TwitterNLP	0.6356	0.5474	0.5882	28.0001
GoogleCloud	0.6321	0.5339	0.5789	NA
SpaCyNER	**0.9883**	0.5555	0.7113	1.0891
GeoLoc	0.7987	0.8300	**0.8141**	1.1901
OSMLoc	0.3383	**0.8888**	0.4901	711.5817

Table 2: Evaluation performance of the different methods.

We have considered precision, recall and F1-score, along with the average performance time over a dataset of 100 tweets as the evaluation metrics.

Savitr System

The Savitr system is deployed live at http://savitr.herokupapp.com. It supports multiple search queries, and displays both tagged tweets and untagged tweets along with the daily statistics.

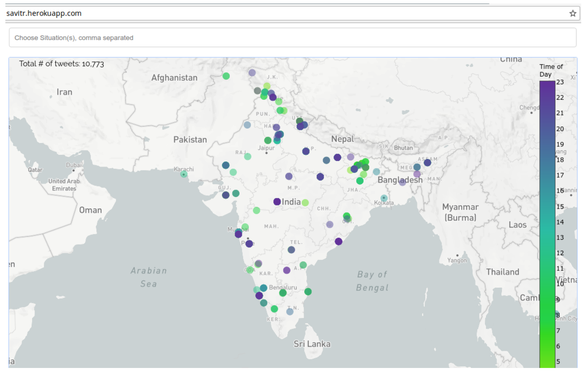

Figure 2: Snapshot of the SAVITR system:

Discussions

We intend to modify the algorithm to include tweets posted in languages other than English as well as extend the reach of the system to tweets posted outside India. The massive information influx would necessitate implementing automated summarization algorithms to capture and display summaries on the system.

References

[1] Judith Gelernter and Wei Zhang. "Cross-lingual geo-parsing for non-structured data". In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*. ACM. 2013, pp. 64–71.

[2] Shervin Malmasi and Mark" Dras. "Location Mention Detection in Tweets and Microblogs". In: *Computational Linguistics*. Ed. by Kôiti Hasida and Ayu Purwarianti. Singapore: Springer Singapore, 2016, pp. 123–134. ISBN: 978-981-10-0515-2.

[3] *Ushahidi*. `https://www.ushahidi.com/`. Accessed: 2018-01-22. 2008.