TREC Incident Streams Track Proposal

Richard McCreadie (University of Glasgow) and Ian Soboroff (NIST)
richard.mccreadie@glasgow.ac.uk, ian.soboroff@nist.gov

Background

Internationally, civil protection, police forces and emergency response agencies are under increasing pressure to more quickly and effectively respond to emergency situations. Moreover, such emergencies are common and recurring. For example, 50,000 people per-year on average die during natural disasters internationally, meanwhile in the U.S. 80 deaths per-year are directly attributed to tornadoes alone. Moreover, even when lives are not at stake, slow or ineffective emergency response can result in increased property/livelihood damage.

The corner-stone that forms the basis of successful response actions is *situational awareness (SA)*. SA is derived from accurate knowledge of what is occurring at the current moment (the operational picture), which can be used to take effective action to remedy the situation, as well as take preventative steps to avoid further loss of life/damage. Response services rely on direct contact with local first-responders and emergency services to form an operational picture. This is then augmented with information from the public, passed to the incident commander by contact points such as emergency call-centre operators. However, depending on the disaster severity, travel distance to the affected area, the state of the local roads and communication infrastructure, it can take hours before accurate information becomes available

The mass adoption of mobile internet-enabled devices paired with wide-spread use of social media platforms for communication and coordination has created ways for the public on-the-ground to contact response services. Moreover, a recent study reported that 63% of people expect responders to answer calls for help on social media.

Proposal

In the U.S. under the National Incident Management System (NIMS) contact with the public is handled by Public Information Officers (PIOs) for each region, With the rise of social media PIO's are now expected to monitor those channels to answer questions from the public, as well as report requests for aid to the Incident Commander. However, PIOs do not have adequate tools or manpower to effectively monitor social media, due to the large volume of information posted on these platforms and the need to categorise, cross-reference and verify that information.

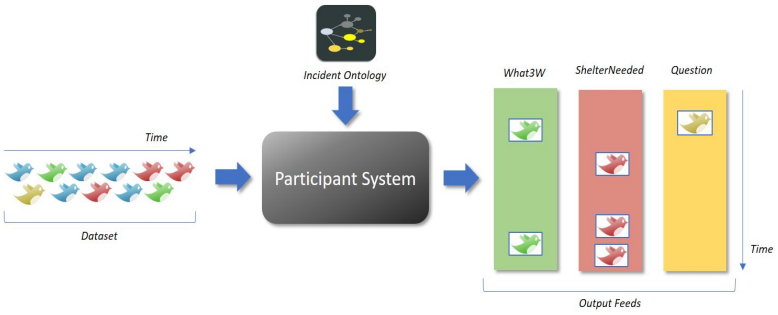
The TREC Incident Streams track (TREC-IS) is a new initiative designed to promote state-of-the-art research into tooling to better support response services harness social media during emergencies. In particular, it will develop a test collection and evaluation methodology for automatic and semi-automatic filtering approaches that aim to identify and categorize information and aid requests made on social media during crisis situations. This will support the advancement the technology readiness level (TRL) of current social media crisis monitoring solutions and better support social media monitoring by PIOs and other stakeholders in the future.

The aim of the TREC-IS task is to produce a series of curated feeds containing social media posts, where each feed corresponds to a particular type of information request, aid request, or report containing a particular type of information. These 'types' are defined based on existing hierarchical incident management information ontologies, such as MOAC (Management of a Crisis), For instance, for a flash flooding event, feeds might include, 'requests for food/water', 'reports of road blockages', and 'evacuation requests'. In this way, during an emergency, individual PIO's and other stakeholders can register to access to the subset of feeds within their

domain of responsibility providing access to relevant social media content without the large costs of monitoring all social media content produced during an emergency.

Task Definition

Methodology: The task will follow a classical TREC evaluation methodology whereby a dataset will be provided beforehand to the participants containing a stream of social media posts from a past event. Where possible the post stream will be provided with all the necessary content included, however participants may need to perform additional work to resolve content/metadata for individual posts depending on the terms and conditions of the data source. An ontology will also be provided, which represents the different information needs of the user. Each participant will develop a system that will process the post stream in time order, as if the event was occurring in real-time. As a system processes the stream, it will emit individual posts over time it identifies as matching one or more information types in the ontology, discarding the rest. Due to the time-critical nature of the task, decisions for each post must be made immediately, i.e. a system much chose to emit or discard a post immediately as it is processed. Emitted posts will be written to a 'run' file. A participating group will be allowed to submit a number of run files to TREC for evaluation by expert assessors. At the TREC conference, the performance of individual runs will be released for comparison across groups. Post-TREC, the assessments will be released to the community for reproducibility and for use by future researchers.



Input: Participating systems will ingest a stream of social media posts from a major platform in time order. It is anticipated that Twitter will be the primary social media data source for the first year of the track. However, access to other social media feeds for emergency management organisations, e.g. Facebook pages will also be considered (based on data availability). Each post will contain the following fields:

- ID: A unique string identifier for the post
- Content: This is a text string that contains the post message
- PostTimestamp: This is a long value that corresponds to the UNIX time the post was made.

Posts may also contain additional metadata about the post, author or source platform, depending on data source.

Additionally, participants will be provided with a structured ontology file containing the information needs that the participant system is to find related posts for. Each entry (information type) in the ontology will contain:

- ID: A unique string identifier for the information type
- Description: A free text description of what type of information would match this entry.
- Parent: The ID of the parent entry if one exists.

An example information type is shown below:

```
{
        "ID": "ShelterNeeded",
        "Description": "Records information regarding additional requests for shelter made by a member of the public",
        "Parent": "Needs3W"
}
```

Output: Systems will output a subset of the social media posts that they have identified as matching one or more of the information types in JSON format. Each post should include the following fields:

- EventID: A unique identifier for the event being processed.
- ProcessingType: Either 'Automatic', 'Semi-Automatic' or 'Manual'. Automatic indicates that the post was selected based on a fully automatic process that only used data from before the event started to make the decision. 'Semi-Automatic' indicates that the post was selected based on a fusion of human effort and automatic classification, e.g. if the system used active learning and requested a human label for the post. 'Manual' should be used if a human decided to emit this post without input from an automatic system.
- ID, Content and PostTimestamp: Fields from the input post.
- InformationTypes: The information type(s) identified by the platform that the post matches. This should be a list of JSON objects containing ontology ids, e.g. 'ShelterNeeded' and confidence values on a 0-1 scale, where 1 is the highest confidence of a match and 0 is the lowest confidence of a match.

Evaluation

The aim of TREC_IS evaluation is to test the performance of a system in terms of its ability to identify relevant content for each of the ontology entries for the PIOs, as well as estimate how useful that information is. To achieve this, the evaluation is split into two main components, namely the *filtering component* and the *grouping component*, described below:

Filtering Component

The first component is focused on evaluating how effectively a system can identify and categorize emergency-related content into each of the selected ontology entries. In effect, it is designed to measure how relevant the content returned is and to what extent the content returned was allocated to the correct category. To achieve this, we first need a ground truth. To create the ground truth, the updates returned by each participant system will be pooled and assessed as matching one or more of the ontology entries (information types) by an expert. The task is inherently recall-focused, as missing information can have a very high cost (particularly in the case of aid requests). However, assessment resources will be limited. As such, after run submission, a subset of the ontology entries will be selected by the organisers. The pools of posts emitted by systems for these ontology entries will be completely assessed.

Performance will then be measured using classical filtering metrics per selected ontology entry. More precisely, the target metrics are:

- Precision: The proportion of posts returned for the ontology entry are relevant for that entry.
- Recall: The proportion of all posts identified for the current ontology entry that were returned.
- Accuracy: The proportion of correctly classified posts (both true positives and true negatives) among the total number of posts pooled.

Overall filtering performance on the task is the average accuracy over all selected ontology entries. In this way, the core evaluation metrics represent how effectively a participant system can identify and categorise emergency-related posts for the PIO.

Grouping Component

As the TREC-IS task focuses on streaming scenario, where a post may not be unique/independant, merely measuring filtering effectiveness will not give a true estimation of system performance. In particular, as redundancy exists within the post stream, for some ontology entries (particularly high level entries, such as those relating to 'where' or 'when') multiple posts containing the same information will be common. This raises two main issues. First, a post returned by a participant system containing a piece of information X, may not be the earliest

instance of X within the stream. As such, information might be returned too late to be useful. Hence, it is critical that systems that return information in a timely fashion can be distinguished from those that tend to return later/out-of-date instances of that information. Second, as redundant posts exist, a system may return multiple of those posts to the PIO, thereby (potentially) wasting their time. Hence, it would also be desirable to capture when a system does this.

To evaluate this, a second evaluation assessment phase will take place. In particular, from each system, a fixed-size sample of the posts returned by each participant system will be selected (e.g. based on system confidence scores). These posts will be provided to the expert assessors along with a search system containing all posts for the event. The assessors will then be provided a fixed time period per post to find and tag all other posts that contain the same information. In this way, a cluster is formed for each post.

A participant system will then be evaluated based on the clusters produced for its sampled posts. The following two metrics will be reported:

- Latency: For a selected post and its cluster, the time difference between the earliest post in the cluster and the post's PostTimestamp will be calculated. This will be averaged across all selected posts to generate the final latency score. Latency will be reported in minutes, lower is better.
- Redundancy: Proportion of posts that were identified as redundant. For a selected post and its cluster, the number of other posts from the cluster that were returned by the system will be counted. This count will be summed over all selected posts and then divided by the number of posts returned. Lower is better.

Potential Participants

Information management on social media during emergencies is a growing research area with a range of international research groups who might be interested in participating. Indeed, a range of workshops in this area have been well attended over the last couple of years such as Social Web for Disaster Management (SWDM'16) co-located with CIKM and Exploitation of Social Media for Emergency Relief and Preparedness (SMERP'17) co-located with ECIR. Indeed, SMERP ran a small data challenge pilot on a related topic which attracted participation from six international research groups. In addition, there are three other communities that might be interested in participating. First, there are a number of European Commision-funded consortia working in the emergency management space that target more effective social media use during emergencies, such as SLÁNDÁIL (http://slandail.eu/), SUPER (http://super-fp7.eu/) and DRIVER (http://driver-project.eu). Second, there are a number of international research hubs that focus on developing solutions to support emergency management, such as the QATAR computing Institute and their AIDR social media tagging platform. Third, the track may also be of interest to groups working on real-time filtering and summarization technologies for use on social media, such as past participants to the past TREC Temporal Summarization track. It also synergizes well with the current TREC Real-time Summarization track, as TREC-IS is also concerned with the real-time processing of social media data streams. Although the focus and challenges tackled by the two tracks are different, participants that want to take part in both should be able to re-use some components from their current solutions to bootstrap them for the new track.

Assessment Resources

As discussed in the evaluation section above, TREC-IS will require human assessors for both components of assessment. In the first component will involve tagging of social media posts with respect to an emergency management information ontology. This could be performed by traditional NIST assessors, or if available regional PIOs with prior experience. The second stage will involve human assessors searching for matching content using a search engine provided. This can be performed by traditional NIST assessors. For both components, the pooling and sampling sizes can be adjusted to fit the available assessment resources available.