TREC Incident Streams – 2018 Planning Workshop

RICHARD MCCREADIE[1] AND IAN SOBOROFF[2]

[1]UNIVERSITY OF GLASGOW, [2]NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Background

Internationally, civil protection, police forces and emergency response agencies are under increasing pressure to more quickly and effectively respond to emergency situations.

- 50,000 people per-year on average die during natural disasters internationally

Situational Awareness

The corner-stone that forms the basis of successful response actions is ***situational awareness***

- Situational awareness is derived from accurate knowledge of what is occurring at the current moment (the operational picture)

Command and Control centre staff need an accurate and complete operational picture to:

- Choose effective actions to remedy the situation
- Take preventative steps to avoid further loss of life/damage.

To build an operational picture, command and control centre operators receive updates from local responders and members of the affected public, as well as other services (e.g. weather)

Social Media

The mass adoption of mobile internet-enabled devices paired with wide-spread use of social media platforms for communication and coordination has created new ways for the public on-the-ground to contact response services.

FEMA

"Social media is an important part of the whole community approach because it helps facilitate the vital two-way communication between emergency management agencies and the public."
Craig Fugate (FEMA Administrator, 2009-2017)

USF UNIVERSITY OF SAN FRANCISCO

Moreover, a recent study reported that 63% of people now expect responders to answer calls for help on social media

Who is Monitoring Social Media?

In many regions, social media is not monitored by response services

- Lack of manpower
- Considered 'too risky' to support due to lack of effective tools

In some regions, social media is monitored by volunteers groups (Europe) or by public information officers (U.S.)

- But have to manually find information using their personal accounts

Incident Streams Task

AIMS AND DEFINITIONS

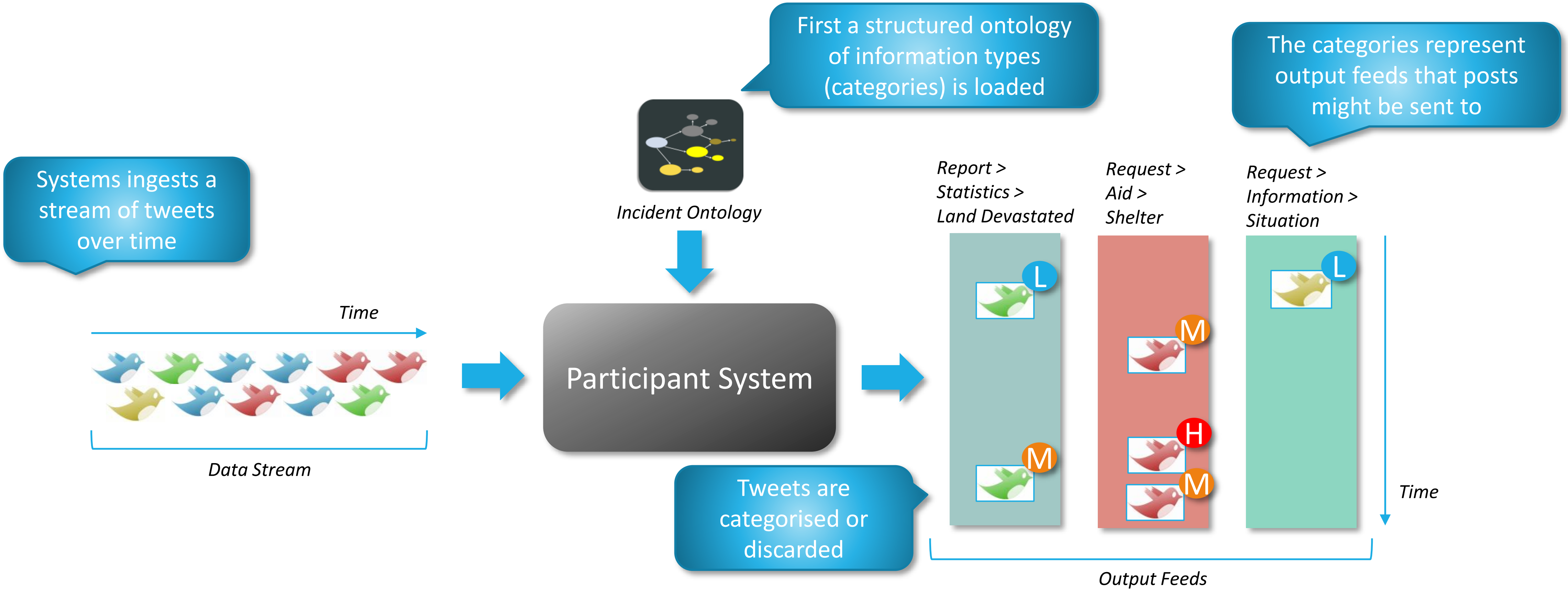
Track Aims

Promote state-of-the-art research into filtering algorithms to better support response services to harness social media during emergencies.

- ➢ Enable volunteer groups and public information officers quickly see new requests for help and other event relevant information that can then be escalated to an appropriate response officer

Participants will develop systems that process a real-time stream of social media posts during an event, filtering and categorizing that content based on an pre-defined ontology of user information needs.

Task Visualisation



Systems ingests a stream of tweets over time

First a structured ontology of information types (categories) is loaded

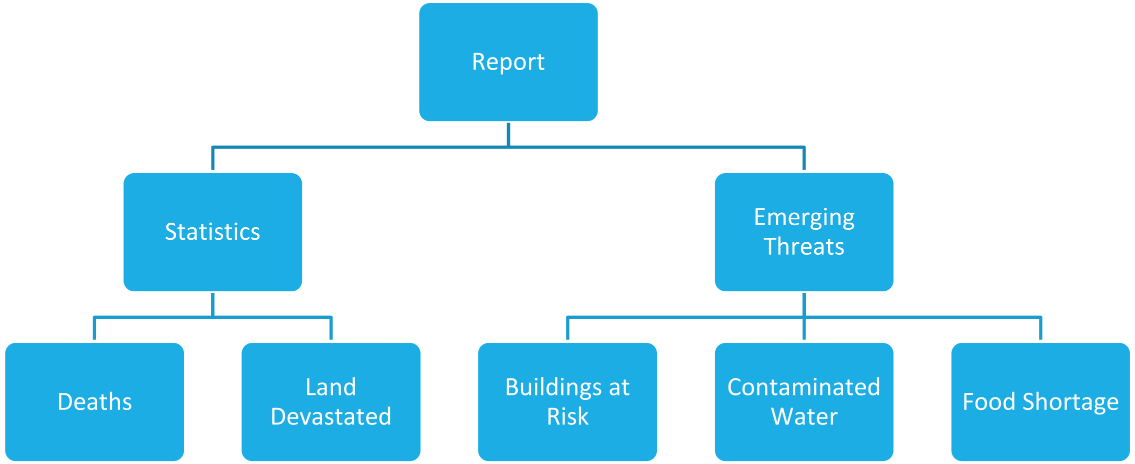
The categories represent output feeds that posts might be sent to

Tweets are categorised or discarded

System Inputs

INCIDENT ONTOLOGY

A hierarchy of information types defined by emergency management organisations



SOCIAL MEDIA POST STREAM

A stream of tweets captured for an event (~topic) in JSON format

```
{
        “EventID”: “TREC-IS-E001”,
        “ID”: “561324141413345”,
        “Text”: “#nswfires NE of Mount Coramba, Type: Bush fire, Status:
                out of control, Size 1ha”
        “PostTimestamp”: 1459468800,
        "user": {},
        "entities": {}
}
```

System Outputs

For each tweet, the system needs to decide

- Is this relevant to an emergency response operator (does it match any of the intents in the ontology)?
  - If so, which information needs does it match? A tweet may match multiple information needs.
  - How confident is the system that it matches those information needs
  - How critical is this information?
- Else, discard that tweet

```
{
        "EventID": "TREC-IS-E001",
        "ProcessingType": "Automatic",
        "ID": "561324141413345",
        "Content": "#nswfires NE of Mount Coramba, Type: Bush fire, Status:
                out of control, Size 1ha"
        "PostTimestamp": 1459468800,
        "InformationTypes: [
                {
                        "OntologyID": "What3W",
                        "Confidence": 1.0
                },
                {
                        "OntologyID": "Where3W",
                        "Confidence": 0.7
                }
        ],
        "Priority": 0.6
}
```

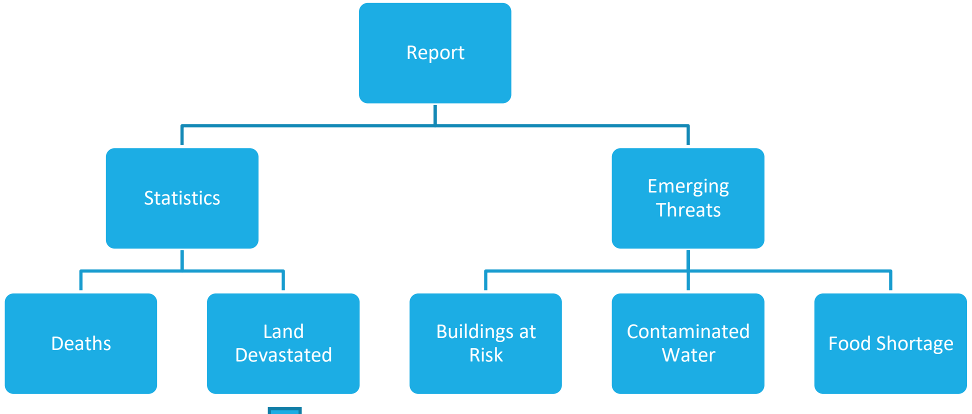
Evaluation

DATASETS AND METRICS

Ontology and Information Types

We are currently constructing the full ontology of information types to track

- There is only one ontology, which will cover information types from a range of events
- Some information types will be generic to many events, while some will be specific to particular events
- One (complex) information need → many events



```
{
        "OntologyID": "LandDevastated",
        "Parents": ["Report", "Statistics"],
        "Content": "Represents posts that quantify how much land was
                    damaged or destroyed due to the disaster. This is
                    common in wild fires and hurricane/tornado events.",
        "Keywords": ["acres", "sq.", "mi", "miles"]
}
```

Each information type will have:

- OntologyID: Unique identifier
- Parents: Its parent information types (if it has them)
- Description: A natural language description of what the information type is about
- Keywords: A small sample of terms that may help identify posts about that information need

Datasets

TRAINING

We aim to provide participants with a training dataset comprised of past events with example information type labels

- We will have human annotators categorise a sample of tweets from a range of events based on the ontology
- Example coverage will vary across the information types

CrisisLexT26 is the dataset we will be drawing example events from for training

- http://crisislex.org/data-collections.html#CrisisLexT26

TESTING

We will be following a classical 'submitted run' evaluation scenario.

Around June next year we will release a testing dataset containing tweet streams from a series of events

- Participants run their systems over each stream, writing out a filtered stream with category annotations to a 'run' file.
- Participants upload their 'run' files to NIST

Assessing Runs

FILTERING AND CATEGORIZATION

Runs will first be evaluated in terms of their ability to identify and categorize emergency-related content into each of the selected ontology entries

Performance is calculated per information type:

- 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.

LATENCY AND PRIORITY

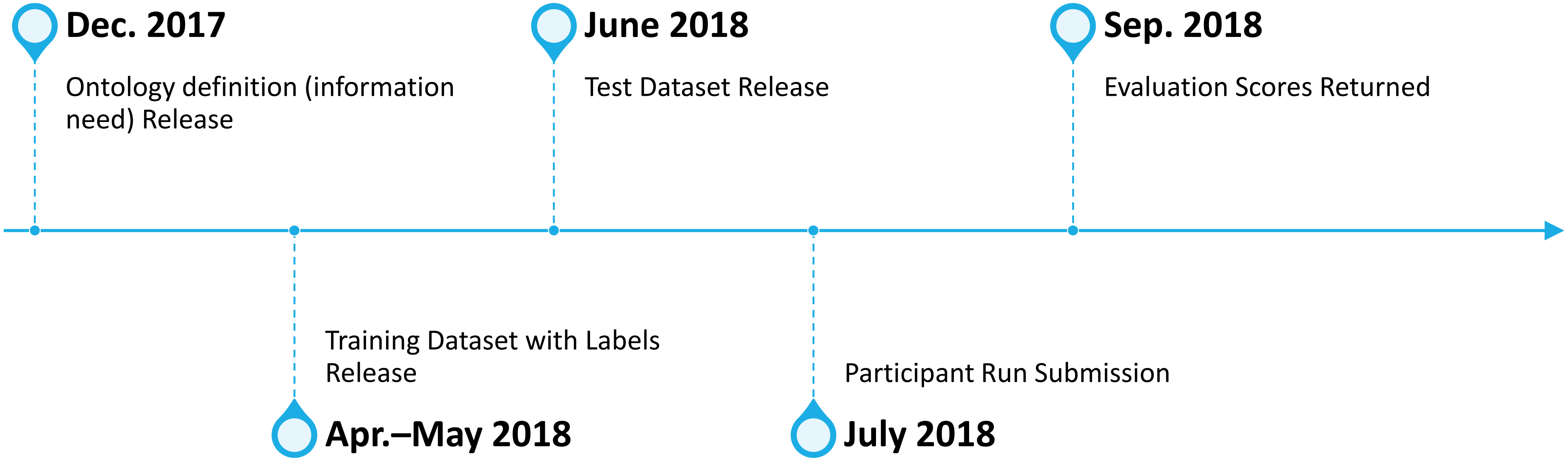
The task also has information prioritisation and timeliness aspects we want to capture. We are considering evaluating:

- Latency: For a selected post, the time difference between the earliest post containing the same information and the post's PostTimestamp will be calculated.
- Priority: The degree of prediction error between a post's priority and its labelled cluster priority

Other Details

COMPARISONS AND DISCUSSION POINTS

Initial Timeline



Dec. 2017

Ontology definition (information need) Release

Apr.–May 2018

Training Dataset with Labels Release

June 2018

Test Dataset Release

July 2018

Participant Run Submission

Sep. 2018

Evaluation Scores Returned

How is this different from RTS?

REAL TIME SUMMARIZATION

Use-case: Generating a News timeline

Information Need: User Interest Profiles

Input: Random Tweet Sample

Focus: Precision, minimal redundancy

Evaluation: Living Lab

Metrics: Expected Gain, Gain Minus Pain, Latency

INCIDENT STREAMS

Use-case: Filtering and categorizing emergency content

Information Need: Structured information types

Input: Tweets matching event-related terms

Focus: Recall, Categorization, Information Priority

Evaluation: Cranfield-style

Metrics: Precision, Recall, Latency, Priority

Why is this task difficult?

Classical Filtering/Summarization Challenges

- Semantic Relevance/Vocabulary Missmatch: While keywords are provided for each information type (to help bootstrap participant systems), it is not expected that these will be sufficient to identify all content relevant to an information type.
- Timeliness/Effectiveness Trade-off: Participating systems might want to use repetition of information to gain confidence that a piece of information is important. However, some information is time critical. Systems may need type-dependant solutions to this (where can we afford to wait?).

Scenario-Specific Challenges

- Type Hierarchies: The nature of our information need defines relationships between our information types. Systems can choose to try to leverage this.
- Priority Estimation: This is a new research area – how to we estimate priority? For instance, is it a feature of an information type or a feature of a post’s content?
- Recall is Critical: Unlike in summarization, redundancy is not a major concern. Missing critical information is a much more important risk!
- Categorization: We anticipate that categorizing tweets to different information types will have wide-ranging difficulty – some types will be very rare (e.g. requests for aid), but are very important!

Open Questions

WHAT ARE WE CONSIDERING?

Do we allow systems to return tweets late?

In real-time summarization we allowed systems to ‘delay’ returning tweets to allow them to collect more evidence – enabled systems to capture a popularity signal

- Tweets were clustered and latency was measured against the earliest tweet in a cluster

Does this make sense for real-time filtering?
(for much of the information we are interested in popularity will not be a strong signal)

The alternative is to mandate that systems make a final decision for a tweet before moving on to the next

- This significantly simplifies evaluation, but systems will not be able to ‘batch’ process tweets

Event Types

We need to define the set of events types that the track will cover in its first year

To support an event type we need

- Datasets containing past tweets of that event type (in English!)
- The event type needs to be common enough that we will see at least one example of this event type in the next 6 months or so (so we can add to our test data)
- We need engaged end users for that event type that we can get feedback from

Do you have event types you want to see that match these criteria?

Participant Collaboration

A continual problem with TREC tracks is that each group builds their own system, runs it on the track data… then that code disappears (storage dies, developer moves on, etc.)

- This makes it really difficult for groups to make significant advances from year to year
- Are we actually building better systems year-on-year?

One potential way around this is to develop a modular framework into which participants can plug in their own custom modules with their secret sauce

- At the end of each year after the results are out, the participants would then upload their modules for people to build on in the following years

If we could build such a framework, is this something people would be interested in perusing?

What's Next?

Join the Google Group!

https://groups.google.com/forum/#!forum/trec-is/

Notes

Look for prior works in tct4d

Ontology guy – via Doug Ord – Soergel Dagobert

Dependencies between info priority – add checkbox in assessment ui