Experimentation Standards for Crisis Informatics

Fernando Diaz
Microsoft Research
New York, NY
fdiaz@microsoft.com

1. INTRODUCTION

Easy access to online media has led to an escalation of researchers and companies developing information analysis systems for use in crisis response. These systems use state of the art text mining algorithms to extract information from online media to support crisis responders and other users.

Published research in this area has primarily focused on systems leveraging public social media feeds such as Twitter. These results often are supported by samples of social media manually analyzed using a combination of hired assessors and data mining algorithms.

Unfortunately, despite leveraging open data like Twitter, the current experimental environment for crisis informatics does not rely on open test collections or evaluation methodologies. Social media samples are often collected by individual research sites and not shared. Assessment is often conducted by crowd-workers or in-house editors, at times with little attention paid to inter-site consistency or to real use cases. Consequently, crisis informatics results are difficult to compare across research publications and to apply to real situations.

In this position paper, I will make the argument that *information access algorithms that support crisis response require standard evaluation metrics and experimental corpora.* There are two objectives for such a initiative: (a) encourage open and reproducible experimentation on crisis informatics, and (b) develop standard evaluation metrics for use by decision-makers.

2. INFORMATION ACCESS AND CRISIS RESPONSE

A variety of information access problems exist immediately after the onset of an unanticipated crisis. Those immediately affected by the event may be interested in information about shelter, food, or infrastructure damage [17]. Those socially connected to these individuals may be interested in communications[1] or providing donations [13]. Crisis responders may be interested in situational awareness information such as displaced populations, food needs, and infrastructure damage [8].

Online information provides a rich, realtime stream of data potentially useful for these different information needs. Social media in particular has received a great deal of attention because of the broad coverage and its ability to, at times, precede official or standard media communications [20]. This has resulted in interest from the crisis response community in mining social media for actionable information [3].

Research in this area has primarily applied existing algorithms from the text mining community to the crisis domain. Like any domain-specific instance of a problem, several nuances emerge in the crisis space. These include domain-specific side-information such as gazetteers, seismic and weather data, and official crisis response channels. This certainly involves some research overhead but that is common to any domain-specific application.

However, to dismiss these tasks as 'applied research' ignores some fundamental research questions involved. Take estimating the number of fatalities from a tornado. This sounds like a standard information extraction problem and there is a huge body of research suggesting appropriate methods to apply. All the same, very little of this literature focuses on the streaming data, where reliability of data may be dynamic. Another common theme across crisis tasks is the very low latency requirements. Users are interested in information about the world as soon as it happens. Knowing about the collapse of a bridge a week after it occurs is not as valuable as knowing about the collapse a few seconds after it happens.

3. A PROPOSAL

Information access systems for crisis response require very careful development. They need to maintain a high level of reliability in the face of low latency demands. Certainly similar situations exist in, for example, finding information about celebrity scandals, sports events, or political sentiment. However, crisis response involves a more direct humanitarian mission.

In order to address the problems with current crisis informatics systems, I propose the development of a standardized experimental methodology for information access supporting crisis response.

[1] `https://google.org/personfinder/global/home.html`

Standards for rigorous experimental frameworks exist for evaluating other technologies, including emergency response equipment (e.g. fire helmets) and medical equipment (e.g. pharmaceuticals). When a computer technology emerges in an area traditionally subject to testing, such as with computer aided medical diagnosis, new standards are developed.

The development of a standardized experimental methodology for information access supporting crisis response requires three efforts: (a) definition of shared tasks, (b) development of a shared data, and (c) development of shared metrics.

3.1 Background

Before outlining this proposal, it may be worth reflecting on a similar exercise conducted in the information retrieval community.

Text mining and information retrieval researchers study the development of systems for, amongst other things, keyword-based search over large text corpora. In the 1980s, information retrieval researchers worked with small, often-proprietary corpora and site-specific queries. This made comparison of algorithms very difficult. In other words, the situation looked a lot like what crisis informatics research looks like today.

Partially to address this lack of experimental standards in the information retrieval community, the National Institute of Standards and Technology (NIST) organized the first annual Text Retrieval Conference (TREC) in 1992 [18].[2] It has run every year since then. TREC provides researchers access to a shared data, task definition, and evaluation methodology. Information retrieval researchers participate in various TREC tracks, each representing a different information access problem. These have included, *ad hoc* retrieval, information filtering, question answering, and cross-lingual retrieval.[3]

The importance of TREC to the information retrieval community should not be underestimated. In 2008, NIST contracted a third party to evaluate the impact of roughly twenty years of TREC [15]. The report found that TREC had had significant impact on information retrieval research and commercial information retrieval systems. We believe that applying similar experimental methodology to crisis informatics will result in similar innovation.

3.2 Shared Tasks

As mentioned in Section 2, information access tasks have, in general, reflected themes currently established in the text mining community. These include (a) new event detection, (b) event tracking, (c) classification, (d) information extraction, and (e) summarization. In the rest of this section, we will touch on these problems in a crisis response domain.

New event detection refers to detection of news events in a stream of documents [1]. In the context of crisis informatics, events of interest are high level crisis types such as earthquakes [16, 14]. Although earthquake detection specifically may be unnecessary (since precise seismic sensors exist for that purpose), other events, such as landslides or civil disturbances (where accurate sensors are lacking), benefit more from social media.

Event tracking refers to the detection of news articles relevant to a given event [1]. In the context of crisis informatics, post-crisis social media tends to follow certain phases of change, admitting novel algorithms to the problem [9, 4].

Classification refers to the detection of documents in pre-defined categories. In the context of crisis informatics, these categories often align themselves with the IASC Cluster System, a method for organizing the set of crisis response responsibilities [8]. Unlike other classification tasks, although the categories may be pre-defined, each crisis might use unique vocabulary for documents in each category. Geographically-specific named entities are a basic example of words unlikely to be consistently relevant from one crisis to the next. Others include event-specific reactions unlikely to occur from one event to the next.

Information extraction refers to the detection of structured information from an individual or group of documents. In the context of crisis informatics, extraction and aggregation occurs over a stream of documents of varying reliability. The streaming context results in a lack of redundancy, a feature necessary for many information extraction algorithms. As a special case, *geolocation* refers to the extraction of the location where the document was written and is the motivation for many 'crisis mapping' systems.[4]

Summarizaton refers to the extraction of unstructured information representing a core theme of an individual or group of documents. In the content of crisis informatics, summarization occurs over a stream of documents, again, of varying quality. Like extraction, summarization methods rely upon redundancy in documents and, as a result of streaming, many of these algorithms fail [7].

3.3 Shared Data

In most information access tasks, shared data consists of three parts: the corpus (a sample of the text stream), the topics (the events during which you want to evaluate performance), and the labeled data (an example of desired system behavior). In streaming experiments such as information filtering or topic detection and tracking, researchers simulate system behavior by processing documents in time order.

3.3.1 Corpus

Although many possible corpora of timestamped documents exist, social media in particular has received more attention because of its sensitivity to changes in the underlying social and physical environment. The traditional choice was news collections, which offer rich text and timestamps; whereas, at a loss of text cleanliness, social media provides a much more topically and temporally granular textual reflection of the world.

Use of social media carries technical, commercial, and ethical issues. Social media, even for a few months, can be several orders of magnitude larger than other text mining corpora. This makes both processing and–more importantly–distribution non-trivial. Secondly, social media is extremely valuable to social media service providers. As such, releasing social media, even to academics, represents a significant risk to social media service providers. Finally, since social media has been created by individuals with varying levels of understanding about possible experiments, there may be ethical issues to address in using these posts.

[2] `http://trec.nist.gov/`

[3] Machine learning researchers may be familiar with the impact of other NIST initiatives. In particular, the NIST OCR Conference on digit recognition [19] resulted in the development of the MNIST digit recognition dataset.

[4] `http://crisismappers.net/`

Fortunately, some of these issues have been addressed by TREC. The Microblog track has organized search and filtering tasks using Twitter as a corpus. To address technical and commercial issues, the organizers built a service for participants to interact with during experimentation [10]. The Knowledge Base Acceleration (KBA) track has organized information filtering specifically for entities.[5] Participants use a very large corpus of news, weblog, and many other timestamped documents. Because of the size of the collection, participants need to work with the data through a cloud service.

3.3.2 Topics

Topics, in the TREC sense of the word, refer to points or conditions under which to measure performance. For keyword search, each topic would be an individual information need expressed as a query. In our situation, we are interested in the performance of a system during a crisis event; this will be measured by simulating the conditions of the corpus as the event developed.

Deciding on a shared set of events involves several considerations. First, events should be well-represented in the corpus. Clearly picking an event that occurred outside of the timespan of the corpus would be meaningless. But so would picking an event that may have occurred during the timespan of the corpus but was not well-represented by social media. This occurs especially with localized events. Second, events should involve categories of interest to the eventual users. Categories include dimensions such as impact (e.g. size of effected population), preparedness (e.g. degree of anticipation), cause (e.g. seismic, weather), and duration (e.g. number of days to run the system).

Previous research has curated specialized sets of events by either manually selecting candidate events or through a semi-automatic process [7].

3.3.3 Labeled Data

Labeled data refers to a definition of expected optimal system performance for each topic. For keyword search, this involves deciding on the optimal ranking for each query. In practice, this decision is decomposed to simpler per-document relevance judgements which can be used to measure rank performance.

In our situation, the nature of the labeled data depends on the task (Section 3.2). However, one important labeling consideration is the use of retrospective information. Take information extraction of 'number injured' during a hurricane as an example. The question needs to be raised whether the target value is a statement about the attribute at simulation time (e.g. 'number injured one hour after landfall') or about the projected value of the attribute at assessment time (e.g. 'total number injured after dissipation'). A separate question is whether the target value uses only information available at simulation time (e.g. 'number injured given data only about a subset of individuals available at landfall') or all information available at assessment time (e.g. 'number injured given data about all individuals retrospectively'). The answers to these questions again depend on the use case (e.g. projection vs. awareness). Another consideration–one stressed by crisis responders–is that of veracity. Unreliable information exists throughout social media, and conditioning on a crisis does not address this fact [11]. Assessors should make sure the retrospective labels are sensitive to the desired treatment of rumors, misinformation, or unsubstantiated claims.

[5] http://trec-kba.org/

3.4 Shared Metrics

Currently, the development of performance metrics consists of translating the potential user's expectations of system behavior into a function of the system performance based on a given combination of topic, corpus, and labeled data. Usually this metric involves establishing a model of user behavior and calculating expected user satisfaction given simulations of both the user and the system.

Because many of the tasks described in Section 3.2 are based on existing problems, information access tasks will likely be evaluated using similar metrics with some modification.

One theme underlying all of these tasks is a focus on low latency. New events must be detected as close to when they really occurred as possible. Bursts of on-event documents should be detected early. Classification, if used for situational awareness, should be up to date. Information extraction systems should modify aggregate estimates using current data. And summarization systems should present updates as they occur, with minimal latency. We should express this latency penalty using models based on real world user tolerance to latency. For example, using models based on cognitive psychology could provide grounding for these preferences [12]. Some attempts have been made to incorporate temporal sensitivity in information access problems, usually involving some manner of relevance demotion [2, 6]. Other approaches allow for arbitrary temporal preferences, not just a focus on speed [5].

Accurately modeling the user task is necessary for any metric to provide reliable estimates of real world performance. Therefore, a dialog needs to occur between those individuals in the crisis response community and metric designers.

4. AN EXPERIMENT

The TREC Temporal Summarization Track[6] was started as an exercise in developing a standard experimental approach to text summarization during a crisis event [2, 7]. Started in 2013, the track evaluates systems designed to provide online updates on a crisis event as it unfolds. Participants processed documents in time order for each event, flagging sentences likely to be relevant, comprehensive, and novel, all with low latency. The track uses the KBA corpus of timestamped documents; natural disasters and human tragedies extracted from Wikipedia; and gold standard timestamped sub-event incidents defined retrospectively by human assessors. For evaluation metrics, time-sensitive versions of precision, recall, and novelty were derived.[7] A sample summary is presented in Table 1.

The first year saw seven sites participating, mostly from the information retrieval community. Systems demonstrated a range of performance as participants became familiar with task, corpus, and metrics. The second year is currently underway, bringing participation from those in the text summarization community.

The Temporal Summarization Track definitely has areas

[6] http://trec-ts.org

[7] http://www.trec-ts.org/trec2014-ts-metrics.pdf

Table 1: Example summary from TREC 2013 Temporal Summarization Track. Updates reflect new or updated information as it is reported.

time	update
Nov 21 10:52:29 2012	Tel Aviv bus bombing; 13 injuries; reported on a bus line 142; occured on Shaul Hamelech street; No claims of responsibility.; 3 badly hurt; occured in the heart of Tel Aviv near military hdqtrs
Nov 22 20:49:57 2012	occurred in an area with many office buildings; occured in area with heavy pedestrian traffic; first notable bombing in Tel Aviv since 2006; At least 28 people were wounded; Hamas' television featured people praising the attack; Khaled Mashal, leader of Hamas, categorically rejected any connection of the bombing to his group; UN Secretary-General Ban Ki-moon deplored the attack; The White House called the bombing a terrorist attack against innocent Israeli civilians; The Russian foreign ministry termed the bombing a "criminal terrorist act"; 21 wounded in terror attack on Tel Aviv bus
Nov 26 04:33:27 2012	an Israeli Arab man was arrested on charges of planting the explosive device on the bus; Suspect was reportedly connected to Hamas; Suspect was reportedly connected to the Islamic Jihad
Nov 26 14:49:35 2012	The Romanian Foreign Minister condemned the bombing,
Nov 29 04:55:26 2012	govt rep refers to attack as terrorist attack
Nov 30 05:22:02 2012	Fears about Bus Bomb Before the Cease-Fire: Could derail peace talks
Nov 30 06:47:40 2012	The suspect remotely detonated the explosive device; suspect hid device in advance on the bus; The explosive device contained a large quantity of metal shrapnel designed to cause maximum casualties; The suspect later on confessed to carrying out the bus attack; suspect prepared the explosive device; suspect chose the target of the attack; suspect purchased a mobile phone; suspect received an Israeli citizenship as part of a family unification process

in which it could improve. First, the metrics are based heavily on existing information retrieval metrics and have not been rigorously vetted against real users. Second, the corpus, while exhibiting the size and granularity of social media, does not directly contain documents from major social networks (e.g. Facebook, Twitter) for legal reasons. Nevertheless, the track demonstrates an preliminary example of standardizing crisis informatics experimentation.

5. CONCLUSION

In this position paper, I have advocated an initiative to standardize experimentation for information access supporting crisis response. This argument is motivated by the current lack of coordination of research and successful NIST evaluations such as TREC. Crisis response, because of its interest to government and non-profit organizations, seems well-suited for a coordinated evaluation program.

6. REFERENCES

[1] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*, volume 12 of *The Information Retrieval Series*. Springer, New York, NY, USA, 2002.

[2] J. A. Aslam, F. Diaz, M. Ekstrand-Abueg, V. Pavlu, and T. Sakai. Trec 2013 temporal summarization. In *Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013)*, 2013.

[3] H. Blanchard, A. Carvin, M. E. Whitaker, M. Fitzgerald, W. Harman, B. Humphrey, P. P. Meier, C. Starbird, J. Solomon, and R. Zeiger. The case for integrating crisis response with social media. Technical report, American Red Cross, August 2010.

[4] S. R. Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo. Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In *Proc. of ISCRAM*, 2013.

[5] L. Dietz, J. Dalton, and K. Balog. Time-aware evaluation of cumulative citation recommendation systems. In *Proceedings of SIGIR 2013 Workshop on Time-aware Information Access, TAIA, 2013*, 2013.

[6] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 11–20, New York, NY, USA, 2010. ACM.

[7] Q. Guo, F. Diaz, and E. Yom-Tov. Updating users about time critical events. In P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, editors, *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 483–494. Springer Berlin Heidelberg, 2013.

[8] IASC Sub-Working Group on the Cluster Approach. Reference module for cluster coordination at the country level. Technical report, Inter-Agency Standing Committee, 2012.

[9] A. Iyengar, T. Finin, and A. Joshi. Content-based prediction of temporal boundaries for events in twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 186–191. IEEE, 2011.

[10] J. Lin and M. Efron. Evaluation as a service for information retrieval. *SIGIR Forum*, 47(2):8–14, December 2013.

[11] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In *Proceeding of the KDD 2010 Workshop on Social Media Analytics*, 2010.

[12] M.-H. Peetz and M. de Rijke. Cognitive temporal document priors. In P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, editors, *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 318–330. Springer Berlin Heidelberg, 2013.

[13] H. Purohit, C. Castillo, F. Diaz, A. Sheth, and P. Meier. Emergency-relief coordination on social

media: Automatically matching resource requests and offers. *First Monday*, 19(1), 2014.

[14] B. Robinson, R. Power, and M. Cameron. A sensitive twitter earthquake detector. In *Proc. of WWW (companion)*, pages 999–1002. International World Wide Web Conferences Steering Committee, 2013.

[15] B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic impact assessment of nist's text retrieval conference (trec) program. Technical Report Project Number 0211875, RTI International, Research Triangle Park, NC, July 2010.

[16] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 851–860, New York, NY, USA, 2010. ACM.

[17] S. Vieweg. *Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications*. PhD thesis, University of Colorado at Boulder, 2012.

[18] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[19] R. A. Wilkinson, J. C. Geist, S. Janet, P. J. Grother, C. J. Burges, R. Creecy, B. Hammond, J. J. Hull, N. W. Larsen, T. P. Vogl, and C. L. Wilson. The first census optical character recognition systems conference. NIST Interagency/Internal Report 4912, National Institute of Standards and Technology, January 1992.

[20] E. Yom-Tov and F. Diaz. Location and timeliness of information sources during news events. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1105–1106, New York, NY, USA, 2011. ACM.