An Approach to Selecting Keywords to Track on Twitter During a Disaster

Kenneth Joseph
Carnegie Mellon University
kjoseph@cs.cmu.edu

Peter M. Landwehr
Carnegie Mellon University
plandweh@cs.cmu.edu

Kathleen M. Carley
Carnegie Mellon University
kathleen.carley@cs.cmu.edu

ABSTRACT

Several studies on Twitter usage during disasters analyze tweets collected using ad-hoc keywords pre-defined by researchers. While recent efforts have worked to improve this methodology, open questions remain about which keywords can be used to uncover tweets contributing to situational awareness (SA) and the quality of tweets returned using different terms. Herein we consider a novel methodology for uncovering relevant keywords one can use to search for tweets containing situational awareness. We provide a description of the methodology and initial results which suggest our approach may lead to better keywords to use for keyword searching during disasters.

Keywords

Social Media and Disasters, Twitter, Sampling from Twitter, Data mining

INTRODUCTION

Researchers and practitioners alike have used ad-hoc keyword searches of Twitter to try and find tweets containing *situational awareness (SA)*, defined as "all knowledge that is accessible and can be integrated into a coherent picture, when required, to assess and cope with a situation" (Vieweg et al., 2010, citing Sarter and Woods, 1991). The corpora of tweets resulting from these ad-hoc searches are constrained to the keywords selected by human intuition and do not represent all SA tweets made during a given disaster. Recent work both within the disaster response community and outside of it has recognized this problem and suggested approaches for automatically selecting the keywords used to search Twitter for tweets relevant to a particular topic. For example, Abel et al. (2012) designed a system that takes named entities (including locations) from both official reports on disasters and previously captured tweets. They show that doing so provides a more relevant sample than simply using pre-defined keywords on a variety of datasets. Li et al. (2013) developed a more general system that works with any classifier to iteratively find tweets most relevant to a particular topic as time progresses.

While such work has presented important new methods for automatically adapting keyword searches, the extent to which any specific keyword or class of keywords captures a relevant number of disaster-related tweets is largely unknown. Indeed, researchers have only recently begun to understand the lexicon that differentiates a tweet which provides SA from one that does not (Verma et al., 2011; Vieweg, 2012). In the present work, we use a novel approach for uncovering and ranking keywords that, when searched for on Twitter, may have increased SA surrounding the 2010 Haitian earthquake. Our approach first pulls these terms from another crowd-sourced dataset available at the time: detailed reports submitted by volunteers running an instance of the crowd-sourced crisis mapping tool Ushahidi[1]. We then create a novel metric for ranking keywords by their expected usefulness in finding SA tweets.

[1] http://www.ushahidi.com

After explaining our methodology, the present work provides a brief overview of the usefulness of our methodology. Specifically, we show how well our method is able to uncover keywords that were relevant to the disaster, how these keywords related to ad-hoc terms utilized by previous researchers, and some examples of SA tweets our model uncovers from a dataset of 81M tweets from the week after the disaster. We end with a discussion of the implications of our work and of future efforts in line with those shown here.

METHODOLOGY

In the present study, we use two sources of data. The first is a corpus of approximately 90M mostly English tweets pulled from the gardenhose from January 7th through January 20th of 2010; 81M tweets were sent after the earthquake[2]. At the time the tweets were drawn, the gardenhose returned about 15% of all public tweets. The second data source are 1105 reports from the crowd-sourced mapping site Ushahidi. Within a few hours of the 2010 Haiti earthquake, volunteer crisis workers created a deployment of the crowd-sourced mapping tool Ushahidi and began combing Twitter, media sources and organization websites for information on the situation (Morrow et al., 2011; Munro, 2013). Starting days after the earthquake, volunteers also began receiving SMS messages from Mission 4636, a collaboration with local phone services to collect all text messages sent by Haitians to the number "4636". Ushahidi volunteers used the information to file reports on actionable information items which were then logged electronically (Munro, 2011). It is these reports, made publicly available for study, which we use here. Each report consists of seven fields, four of which are used in the present work. They are: a volunteer-specified report title, a volunteer-specified location, the content of the volunteer's report, and the data and time of the report. Of particular interest is the content section, from which other fields were often derived.

From the title, location, and content fields of the Ushahidi reports, we extracted three different types of terms that we believed might be useful in uncovering SA tweets: entities, actions and locations. These word types were selected based loosely on Munro's (2011) definition of an "actionable" request, which requires a location, a specific need (here, action to be performed) and implicitly, a requesting entity. Report text was processed using the *nltk* toolkit for Python and a number of special purpose heuristics[3].

To find actions, we used nltk's part of speech (POS) tagger to process only the content fields; verbs and verb phrases were considered to be de facto actions. To find entities and locations, we utilized the content field, the title, and the location field. We first used nltk's Named Entity Recognizer (NER) to pull entities and locations from the content field. While the NER had the capability to distinguish between entities and locations, it did a poor job for our particular use case and so we classified all extracted terms as either an entity or a location by hand. We assumed all words and phrases extracted by the NER that we did not know to be locations were entities.

Unlike the content fields, which held unstructured text, the location and title fields both held highly structured data. The location field specified places in a hierarchical manner, split by commas (e.g. "Texaco Station, Port-au-Prince, Haiti"). We split the location field by commas and used each term as a unique location. The title field held a concise summary of the content of the report. To find both locations and entities in the title, we extracted all capitalized words and phrases from the text. Any of the extracted words or phrases that followed the words "at", "in", "on", or "by" was assumed to be a location; the rest were assumed to be entities.

After collecting locations, entities and actions, we manually filtered for keywords that do not represent these types of terms. For actions, this meant removing terms mis-categorized by the POS tagger and removing Creole terms which did not translate to verbs according to Google Translate. For entities, we removed nonsense terms and terms that could never refer to a specific collection of entities (e.g. "everyone"). For locations, we removed those that obviously referred to places outside of Haiti, as well as nonsense terms. After cleaning, we were left with 708 actions, 1651 entities and 842 locations.

After creating the cleaned list of locations, actions and entities, we needed to distinguish the importance of each remaining term in searching for disaster-related tweets, particularly tweets that provided SA. After studying the keywords by hand, we determined five general *categories* that keywords fell into (regardless of type). First, there were terms like "Jacmel" (a city in Haiti), which were expected to provide a large number of relevant tweets and very few tweets unrelated to the disaster. Second were terms like "trapped", which were likely to

[2] Thank you to Brendan O'Connor for providing the data.

[3] Code for all methods presented in this article is available at https://github.com/kennyjoseph/iscram_14

Work	Keywords	Sample Size
(Munro, 2011)	#haiti	40,000
(Verma et al., 2011; Vieweg, 2012)	haiti, earthquake, quake, shaking, tsunami, ouest, port-au-prince, tremblement, tremblement de terre	4M (230,000)
(Sarcevic et al., 2012)	earthquake, port-au-prince, ouest, tsunami, haiti, tremblement	3.28M (16,000)

Table 1. Known articles on the Haitian Earthquake, the search terms used, and the size of the resulting dataset. Sizes in parentheses represent sizes of final corpora after additional sampling and subsampling techniques were applied

produce tweets relevant to the disaster but which were also likely to produce some noise because of the variety of contexts they could be used in. Third were terms such as "food", which were expected to product a significant amount of noise but that may also have been relevant in a select set of SA tweets. Fourth were highly specific terms found in only one of the Ushahidi reports, like street addresses, which would likely prove important if found but were unlikely to be observed in any tweets. Finally were terms found in only a single Ushahidi report, such as (possibly mis-spelled) Haitian first names, which even if found were unlikely to produce information relevant to SA.

In order to capture these different categories of tweets quantitatively, we score terms based both on their relevance to the disaster, as determined by their frequency of use in the Ushahidi reports, and on their relevance outside the context of the disaster, as determined by their frequency of use in tweets we know were not related. The frequency of use in Ushahidi reports was determined by counting the number of reports in which each term appeared one or more times. This produced the set U, where u_k refers to the score for a specific keyword k. To determine the frequency of use in tweets not relevant to the disaster, we calculated the average number of tweets each term appeared in across nine independent sets of 1M tweets that were sent *before the earthquake occurred*. This produced the set T, where t_k refers to the score for a specific keyword k. After obtaining U and T, we take the logarithm of all values. For terms where $t_k = 0$ (and thus the logarithm was undefined), we set $t_k = \min(\{t_x | \ x \ is \ found \ at \ least \ once\}) - 1$, thus scoring them an order of magnitude lower than any other term. Following this log-scaling, we then further scale each set to have a standard deviation of 1 and a mean of $\min(U) + 1$ and $\min(T) + 1$, respectively (to ensure there are no values less than 1). The expected benefit score of each keyword is then calculated as $benefit\ score(k) = \frac{u_k}{t_k}$. Note that for terms that never appeared in the pre-disaster tweets and appeared in only one Ushahidi report, the benefit score equaled one. Such terms, which fell into the fourth and fifth categories discussed above, were thus given a "medium" weight to represent our uncertainty as to whether they were noise or not.

As a basis for comparison, in addition to selecting keywords from Ushahidi we also use a set of ad-hoc keywords from four studies of the Haitian disaster and Twitter. Table 1 shows these articles, the search terms used in them and the resulting size of each study's dataset. Note that in some cases, these "keywords" really consisted of multiple words. In keeping with previous work on Twitter, we will assume that keywords represent phrases (e.g. "tremblement de terre") that must match in their entirety to be considered "found" in a particular tweet.

RESULTS

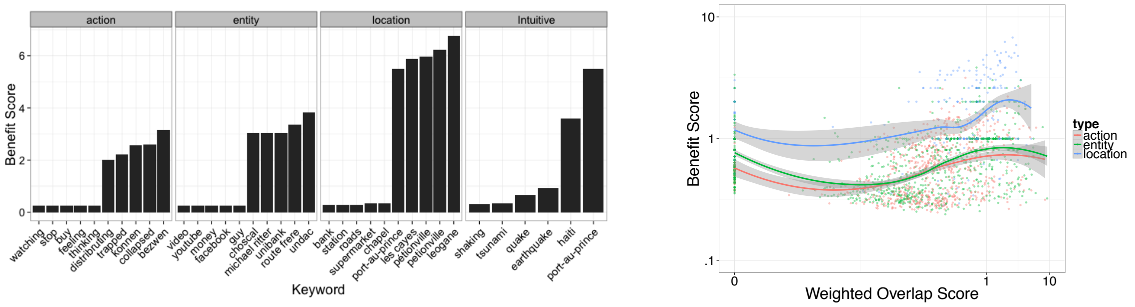

Figure 1. a) Left, the top five scoring and bottom five scoring terms for actions, entities and locations. The five terms in the left of each sub-plot are the lowest scoring terms; the five on the right of each sub-plot are the highest scoring terms. We also show scores for any ad-hoc keyword that appeared in the Ushahidi reports in the subplot furthest to the right. b) Right, comparing the benefit score (y-axis) to the weighted overlap score (x-axis, described below) for all unintuitive keywords. Each keyword type is represented by a different color. Each keyword type also has a non-linear regression line computed using a LOESS fit, where grey represents the 95% confidence interval of the regression.

"RT @***: RT @*** Foodmax located in Petion-Ville is now open. Pass the news along to people who live in PAP & need grocer ..."
"Plz HELP orphange La Maison Des Anges 21 Rur Clairsine Tabarre Glasys Maximillien plz call *** NEED WATER #haitiquake #rescuehaiti"
"RT @***: This is a call for help for *** and the 50 children of the Nursery Nid d'Amour. URGENT NEED SUPPLIES Tabarre ..."
@*** Riviere Froide SW of PAP after Leogane, Les Petites Soeurs de Ste Therese,house collapsed,need aid+food+h2o call ***"
"Was the hospital that collapsed in Petion-Ville l'hopital des Petits-Freres? I am looking desperately for a nurse there...***"

Table 2. Five tweets hand-drawn from the top 100 relevant tweets as suggested by our model. Twitter handle names and person names have been replaced with ***

Figure 1a shows the top five and bottom five scoring terms for the locations, actions and entities found in the Ushahidi reports and the benefit scores for several of the ad-hoc keywords from previous work. We see that top locations scored higher than top entities, which in turn faired slightly better than actions. This matches the widely held notion that locations are a requisite piece of information for SA, as suggested by the fact that, for example, Ushahidi only accepts location-tagged information. We also see that more specific terms tended to be ranked higher than general terms. For example, we see that for actions, where almost by definition all terms can be used outside the context of the earthquake, those more specific to the disaster (*bezwen* is Creole for "need", *konnen* for "know") score highly while those centering on less relevant verbs like "feeling" score low.

Figure 1a also shows that the scoring metric provides a conceptually appealing ordering to the search terms used in previous work that were also in the Ushahidi dataset. Haiti, while used frequently in the Ushahidi reports, did not score as highly on our metric as "Port-au-Prince" because it was used relatively often in tweets prior to the earthquake. The metric suggests that Haiti was used in more general, global conversation of the earthquake than the term Port-au-Prince. This claim is supported by prior work which suggests people involved in disasters often use localized terminology (Vieweg et al., 2010). Additionally, Figure 1a shows that "Petionville" and "Pétionville", though referring to the same location, both had high benefit scores. While it would have been trivial to disambiguate these two and combine them into one term, one would then only be searching Twitter for the single term used in the disambiguation. This presents a subtle but useful benefit of drawing keywords from the crowd-sourced Ushahidi data. As opposed to official reports or intuition, which promote correctly spelled terms, pulling from crowd-sourced data also allows us to automatically uncover common misspellings that might exist in tweets providing SA.

Figure 1b shows which keywords frequently occurred in tweets that also contained ad-hoc keywords utilized in previous work. The figure plots the benefit score of each term versus its *weighted overlap*, defined as $\log(|I_k| + 1) * \frac{|I_k|}{|Tw_k|}$, where $|I_k|$ is the number of times the keyword was found in tweets that had intuitive keywords and $|Tw_k|$ was the number of times the keyword was found in any tweet. Also shown in Figure 1b is a regression curve fitted using a localized nonlinear regression model. The *weighted overlap* and the related curve show the extent to which terms having different benefit scores were found both frequently in general and frequently in tweets having intuitive keywords.

Figure 1b shows that keywords that were never found in the overlap (those having a weighted overlap score of 0) tended to have slightly higher benefit scores than terms found rarely in the overlap. Ignoring terms never found in the overlap, however, we see a positive correlation between benefit score and weighted overlap. We also observe that locations were significantly more likely to be in the intersection than either entities or actions.

These observations give both pros and cons to approaches which begin with a small set of initial intuitive keywords and expand to a larger, more refined set by pulling new terms from tweets captured via this initial set (Li et al., 2013; Lin et al., 2011). On the one hand, we find that for terms found in one or more tweets holding intuitive keywords, higher weighted overlap scores suggested higher benefit scores. On the other had, there were several terms with high benefit scores that were never found in tweets that also contained intuitive keywords. This provides additional evidence that drawing search terms from outside of Twitter in addition to capturing new search terms from Twitter itself may improve the ability of these types of approaches (Abel et al., 2012).

This observation also suggests that there were tweets uncovered by keyword searches with terms our model found relevant that would not have been uncovered by ad-hoc keyword searches. Table 2 shows five tweets which contain few terms anyone without specific domain knowledge could reasonably have thought to search for, and none of the keywords used in prior work on Haiti. These keywords all contain locations mentioned in the Ushahidi reports, and so could have been uncovered using the approach provided here.

CONCLUSION

In the present work, we develop a novel approach to obtaining and ranking keywords that might be useful in finding tweets containing SA. We are able to automatically uncover terms that might not have been intuitive to search for, alternative spellings of important keywords and the relative importance of a large number of distinct entities, locations and actions. Our preliminary results also show we can uncover SA tweets using our methodology which would not have been discovered via ad-hoc keywords used in previous approaches. Results thus provide some evidence that our method, which is based on a novel combination of two crowd-sourced datasets, may be a viable means to uncover useful keywords to search Twitter for tweets containing SA during disasters.

While we have developed initial means to deploy our methodology in real-time (Joseph et al., 2012), the most important drawback of the present work is the assumption that one would have access to Ushahidi data as a disaster unfolds. However, while the work here relies on Ushahidi, the methodology presented requires only that one have a collection of terms that are expected to be somewhat relevant to the disaster. With only slight modifications to the (open-source) approach provided here, one could utilize, for example, Wikipedia entries about the location of a disaster to rapidly obtain keywords that might be useful in uncovering SA tweets.

As we hope future work will show, the scored unigram model we have built here is also amenable to rapid classification of tweets as containing SA or not. It is also amenable to active learning, which will, we hope, eventually allow us to rely on only a small number of hand-labeled data points. Furthermore, by scoring each keyword independently, we can also manage API limitations by thresholding limits on the required score for a given term to be included in the search. Efforts here thus serve as the basis for the further development of a new approach similar to the efforts of Abel et al. (2012), Li et al., (2013) and others that couple sampling from the Streaming API with classification of SA tweets.

REFERENCES

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., Tao, K., 2012. Semantics+ filtering+ search= twitcident. exploring information in social web streams, in: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. pp. 285–294.

Joseph, K., Landwehr, P., Carley, K.M., 2012. Using the Crowd to Mine the Crowd: Combining Ushahidi and Twitter Data from the Haitian Earthquake of 2010, in: InfoSocial '12. Chicago, IL.

Li, R., Wang, S., Chen-Chuan, K., 2013. Towards Social Data Platform: Automatic Topic-focused Monitor for Twitter Stream. Proc. VLDB Endow. 6.

Lin, J., Snow, R., Morgan, W., 2011. Smoothing techniques for adaptive online language models: topic tracking in tweet streams, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11. ACM, New York, NY, USA, pp. 422–429.

Morrow, N., Mock, N., Papendieck, A., Kocmich, N., 2011. Independent Evaluation of the Ushahidi Haiti Project (Harvard Humanitarian Initiative Report). Active Learning Network for Accountability and Performance in Humanitarian Action.

Munro, R., 2011. Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 68–77.

Munro, R., 2013. Crowdsourcing and the crisis-affected community. Inf. Retr. 16, 210–266.

Sarcevic, A., Palen, L., White, J., Starbird, K., Bagdouri, M., Anderson, K., 2012. "Beacons of hope" in decentralized coordination: learning from on-the-ground medical twitterers during the 2010 Haiti earthquake, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. ACM, New York, NY, USA, pp. 47–56.

Sarter, N.B., Woods, D.D., 1991. Situation awareness: A critical but ill-defined phenomenon. Int. J. Aviat. Psychol. 1, 45–57.

Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A., Anderson, K.M., 2011. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. Proc ICWSM.

Vieweg, S., 2012. Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications. University of Colorado at Boulder.

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness, in: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI '10. ACM, New York, NY, USA, pp. 1079–1088.