Foundations of a Multilayer Annotation Framework for Twitter Communications During Crisis Events

William J. Corvey* **Sudha Verma***** **Sarah Vieweg**** **Martha Palmer*** **James H. Martin*****

*Department of Linguistics
**ATLAS Institute
***Department of Computer Science
University of Colorado, Boulder CO
{william.corvey,sudha.verma,sarah.vieweg,martha.palmer,james.martin}@colorado.edu

Abstract

In times of mass emergency, vast amounts of data are generated via computer-mediated communication (CMC) that are difficult to manually collect and organize into a coherent picture. Yet valuable information is broadcast, and can provide useful insight into time- and safety-critical situations if captured and analyzed efficiently and effectively. We describe a natural language processing component of the EPIC (Empowering the Public with Information in Crisis) Project infrastructure, designed to extract linguistic and behavioral information from tweet text to aid in the task of information integration. The system incorporates linguistic annotation, in the form of Named Entity Tagging, as well as behavioral annotations to capture tweets contributing to situational awareness and analyze the information type of the tweet content. We show classification results and describe future integration of these classifiers in the larger EPIC infrastructure.

Keywords: Crisis Informatics, Annotation, Computer-Mediated Communication

1. Motivation

During crisis events, the popular microblogging service Twitter serves as an outlet to offer and receive useful information; it provides a way for those experiencing a mass emergency to obtain information beyond what is available via traditional methods of dissemination, such as mainstream media broadcasts. In these time-sensitive situations, specific requirements arise; people may need food, shelter, and medical care, among other essentials. In addition, affected populations require information. The pervasiveness of information and communication technology (ICT), including social media sites and microblogging services, has greatly increased the volume of information disseminated in times of crisis. So much information is now broadcast during mass emergencies that it is not possible for people to locate, organize and understand it, much less make meaningful decisions. Our aim is to help those faced with mass emergencies to find relevant, tactical information, which involves the implementation of computational methods to aid human efforts in information gathering.

We know that useful, timely information is broadcast via Twitter during mass emergencies (Starbird et al., 2010; Vieweg et al., 2010). Yet the challenge is in locating the right information. In addition to broadcasting valuable, actionable information, many also send general information that is void of helpful details, or communicate supportive messages that lack tactical information.

The EPIC (Empowering the Public with Information In Crisis) Project (Palen et al., 2010) extends previous research on computer-mediated communication during mass emergency situations (Palen et al., 2009; Qu et al., 2009; Starbird et al., 2010; Vieweg et al., 2010) with the goal harnessing behavioral and linguistic insights to analyze and distill information passed over Twitter during mass emergency. We annotate a subset of the total data collected; the data collection methodology is described below.

2. Data Collection

The data used in this paper were collected during five disaster events: Hurricane Gustav (2008), the 2009 Oklahoma Fires, the 2009 and 2010 Red River Floods, and the 2010 Haiti Earthquake.[1] When we learned of each disaster situation, software engineering colleagues began to query the Twitter Streaming API to collect tweets that contain one or more of the terms listed in Table 1 below; see (Anderson and Schram, 2011) for a full description of data collection infrastructure.

Dataset	Search Terms
Gustav	*gustav, hurricane*
Haiti	*haiti, earthquake, quake, shaking, tsunami, ouest, port-au-prince, temblement, tremblement de terre*
Oklahoma Fires	*okfire, oklahoma, grass fire, grassfire*
Red River 2009	*red river, redriver*
Red River 2010	*fmflood, flood10, red river, redriver, ccflood, fargoflood*

Table 1: Dataset Search Terms

Once all the data are collected, various sampling methods are employed to create datasets that are manageable for

[1]The EPIC data are not currently public, in part due to Terms of Use agreements with the service provider, but we hope to release the data in the future.

manual annotation. For a detailed description of the sampling method for the Oklahoma Fires and 2009 Red River Flood data, see (Starbird et al., 2010; Vieweg et al., 2010).

3. Annotation

This paper describes two genres of annotation, linguistic and behavioral, that form the basis of a classification system to support information extraction. Linguistic annotation focuses on named entity tagging; behavioral annotations consist of labels describing the information content of a tweet drawn from a taxonomy of disaster-specific information types, and a binary annotation indicating a tweet's relevance to situational awareness (see Section 3.2.1). These two genres of annotation are not currently related in a single machine-learning task, however integration is in progress.

3.1. Linguistic Annotation

A description of an on-going mass emergency event must include details about the hazard, the people and places affected, and the existence and allocation of aid and other resources. Named entity (or nominal entity) tagging (Bikel and Weischdel, 1999) targets extraction of this linguistic information. We turn to the Automatic Content Extraction (ACE) guidelines (LDC 2004) to provide typical labeled entities, which include: Person, Location, Organization, and Facility as four maximal entity classes.

Our preliminary annotation task consisted of identifying the syntactic span and entity class for these four types of entities in a pilot set of Twitter data (200 tweets from the 2009 Oklahoma grassfires dataset) (Corvey et al., 2010). Through iterative development of the annotation guidelines, we have expanded this initial ontology; Artifacts (such as supplies and vehicles) are now marked and a variety of events are annotated, along with their participants and syntactic extent. Annotations are performed using Knowtator (Ogren, 2006), a tool built within the Protégé framework (http://protege.stanford.edu/).

Recent annotation work has applied this annotation scheme to a wide range of data. Named entity tagging has been completed for datasets from tweets related to the 2008 Hurricane Gustav, the 2009 Oklahoma Fires, the 2009 and 2010 Red River Floods and the 2010 Haiti Earthquake. Table 2 reports inter-tagger agreement for entities annotated in each dataset. We define precision as the agreement rate for dual-annotated instances; inter-tagger agreement (ITA) is the raw agreement between annotators, which counts single-annotated instances as disagreements.

Dataset	Precision	ITA
Hurricane Gustav	.9100	.6430
Oklahoma Fires	.8921	.8020
Red River 2009	.7968	.5234
Red River 2010	.7143	.3181
Haiti	.8285	.5550

Table 2: Named Entity Agreement Rates

Relatively high precisions indicate that when two annotators label a span of text, they agree on the correct label a majority of the time. This suggests that the labels are appropriate to the data and that the annotators understand how to use the labels. However, because of the large number of entities, annotators can miss instances, leading to disagreement and a lower (ITA). We believe that because of the nature of the data, named entity tagging in this domain is a cognitively difficult task. Unlike in other domains, in Twitter entities may not correspond to grammatical roles, and there may be many more entities per sentence than annotators would otherwise be expected to identify. We continue to work to reduce the number of missed instances.

3.2. Behavioral Annotation

Broadly, behavioral annotations describe how community members use tweets during the hazard period recorded in our data. For information extraction in the crisis informatics domain, disaster-related tweet content is of particular interest. Therefore, tweets are coded for the type of information that they convey, based on a multi-layered annotation scheme described below.

3.2.1. Annotation for Situational Awareness

Prior to annotating features of tweet information content, tweets are annotated with respect to whether situational awareness is demonstrated in the text (this is an annotation task in its own right; see (Verma et al., 2011) and Section 4.2 below for discussion of annotation and classification specifics). Simply stated, situational awareness is an understanding of a situation as a whole; obtaining situational awareness is a complex process that requires the perception and comprehension of what is happening in one's environment (Endsley, 1995; Endsley and Garland, 2000). Those who find themselves in circumstances that require situational awareness are constrained by time, and are often in potentially dangerous environments. During times of mass emergency, attaining situational awareness involves knowledge of elements in the environment and an understanding the significance of those elements; affected populations must grasp the meaning of the information they are receiving (Harrald and Jefferson, 2007).

During a mass emergency situation, information broadcast via Twitter may include the location of evacuation centers, the state of the hazard agent, where building and infrastructure damage has taken place, and the number and location of injured people and/or animals. Such knowledge provides decision-makers with information that contributes to an overall understanding of emergency situations, and can help them choose what actions to take.

3.2.2. Information Type

The information type annotation effort involves assigning qualitative codes to tweets based on the information they contain; this process aims to identify what information tweets convey at the behavioral level. Annotators first assign one of three mutually exclusive codes to a sample of tweets collected during a given disaster. One code indicates the tweet is off-topic, meaning it contains no information about the disaster event; another code indicates that the tweet is on-topic, meaning it does include information about the event, but lacks any information relevant to situational awareness; and a third code indicates that the tweet

is both on-topic, and includes information that contributes to situational awareness. After the first round of coding is complete, annotators perform a second pass of coding, and focus only on those tweets that are on-topic, and include situational awareness information.
The second pass of coding involves assigning one of three non-mutually exclusive codes, meaning each tweet may be coded with one, two or all three codes. These codes indicate whether tweets include information about the social environment, the built environment, or about the physical environment, including the hazard agent and hazard conditions. After the second pass of coding, tweets go through a third pass of coding. This process involves assigning specific information types to each tweet, based on the second pass codes. Second and third-pass codes for on-topic tweets are shown in Table 3 below.

Second Pass Code	Third Pass Code
Social Environment	Advice - Information Space Animal Management Caution Crime Death Evacuation General Population Info. Injury Missing Offer of Help Preparation Recovery Request for Help Request for Information Rescue Response - Community Response - Formal Response - Miscellaneous Respose - Personal Sheltering Status - Community/Population Status - Personal
Built Environment	Damage Status - Infrastructure Status - Personal Property Status - Personal
Physical Environment	General Area Information General Hazard Information Historical Information Predictions Status - Hazard Weather

Table 3: Macro- and Micro-level Information Categories

The third pass coding categories were iteratively developed over several years and reflect the empirical analysis of datasets from flood, fire and earthquake events. As researchers analyzed Twitter data, and came to understand what Twitter users communicate in disaster situations, categories of information emerged to describe the tweet content. Of the third pass categories listed here, the majority are represented in the datasets examined. Examples of tweets coded with various second and third pass categories are shown below.

(1): "@User Its so dry hre that grass fires can start anywhere. We R on the edge of the city hre, but 20-30 miles from the fire

- Second pass annotations: social environment, hazard agent and conditions
- Third pass annotations: general area information, general hazard information, status - hazard, status - personal

(2): "Carter County, officials reported at least 15 fires in progress and & several homes destroyed

- Second pass annotations: built environment, hazard agent and conditions
- Third pass annotations: damage, status - hazard, status - personal property

These examples provide a glimpse of the amount of information contained in tweets. From these, we can see how those faced with a mass emergency situation may use this information to make informed decisions.
Table 4 reports inter-tagger agreement (expressed as a Kappa (Cohen, 1960)) for all coding passes for datasets currently coded with the behavioral annotation categories.

Dataset	Kappa Calculation		
	Pass 1	Pass 2	Pass 3
Oklahoma Fires	.900	.822	.891
Red River 2009	.906	.862	.826
Red River 2010	.940	.906	.898

Table 4: Information Type Agreement Rates

4. Classification

To determine the utility of annotations in sifting through the large and diverse stream of tweets during crises, we implemented a set of classifiers to identify situational awareness in tweets and to detect location and events using named-entity tagging. The predictions these classifiers generate form important features in a system used to classify information content, currently under development.

4.1. Classification of Situational Awareness

As a preliminary to future behavioral classifications for tweet information content, we implemented a classifier to categorize tweets that contribute to situational awareness (SA tweets) (Verma et al., 2011). We envision this classifier as the first in a pipeline of classifiers used to extract information during crisis, as contribution to situational awareness is a key characteristic of tweets containing useful

information for processing and dissemination. Qualitative analysis of the tweets suggested a correlation between SA tweets and certain linguistic characteristics, such as objectivity and register. SA tweets were found to be written in a formal and impersonal style and tended to be objective in nature. Following these intuitions gained from empirical analysis, we implemented a coding scheme to incorporate these linguistic characteristics as features in a machine learning system to identify SA tweets. As such, each tweet was independently annotated with the following information:

1. whether a tweet was SA or not
2. whether a tweet showed formal or informal register
3. whether a tweet was objective or subjective in nature
4. whether a tweet was written in impersonal or personal tone.

Training data consisted of roughly 2,000 tagged tweets from four emergency events: 2009 and 2010 Red River Floods, 2009 Oklahoma grassfires and the 2010 Haiti Earthquake. In addition, we also created a uniform dataset consisting of 500 SA and 500 non SA tweets across all events. We used the Mallet (McCallum, 2002) implementation of a Maximum Entropy to implement classifiers to predict subjectivity, register and tone of the tweet. The predicted tags from these classifiers were then used as an input feature for the SA classifier. Performance of these classifiers was evaluated by taking the mean accuracy over 10 fold cross-validation.
We normalized the tweet by replacing URLs and Twitter-specific symbols, such as, "RT", "@username" and the hash symbol (#) with unique symbols. We then tokenized the tweets and used basic lexical features such as, words and their frequency (coded "W" in Table 5 below). We also used the part-of-speech tags for the tweet obtained from the Stanford part-of-speech (POS) tagger (coded "P" in Table 5 below). The SA classifier used these features along with the predicted subjectivity (S), register (R) and tone (T) of the tweet obtained from the pipeline of classifiers. Results are shown in Table 5 below. Classification results are shown for all datasets individually and for the uniform dataset (coded "U" in Table 5 below).

Features	RR09	RR10	Haiti	OKFire	U
W,P	79.1	87.8	83.9	82.6	83.3
W,P,S	**84.8**	88.4	85.3	84.7	82.3
W,P,R	82.1	87.4	84.7	84.4	81.3
W,P,T	83.9	87.1	83.7	85.1	81.5
All	84.1	**88.6**	**88.8**	**87.1**	**84.5**

Table 5: Average 10-fold cross validation accuracies of SA classifier

We found that the basic bag of words model with part of speech tags gave good results, showing an accuracy of 83.3%. From this we interpret that during emergency events, users employ a specific vocabulary to convey tactical information on Twitter. For example, when flooding occurs, tweets containing the words "water", "level" and "evacuate" are generally tactical in nature.
The additional derived linguistic features, such as objectivity, register and impersonal or personal tone of the tweet helped improve accuracy of classifying SA tweets for most datasets. Using all features on the uniformly distributed dataset reduced error rate by 11%. The high accuracy confirms our belief that annotating and classifying for situational awareness of a tweet can be an important step to locate useful information during crisis.

4.2. Classification of Entity Types

A set of classifiers currently under development utilizes the ACE annotations to categorize for location and event categories. We pilot our entity classification task by looking specifically at location annotations; location is an important attribute during crisis since it provides important context for an event. For example, information on road closures, earthquake epicenters, and the overall impact of an event often reference location. As an example, following tweet collected during the Oklahoma fires, gives a specific location that's being evacuated - "Midwest City to evacuate between SE 15th and Rena and Anderson and Hiwassee also Turtlewood, Wingsong, and Oakwood additions #okfire".
We use Mallet's implementation of Conditional Random Field (CRF) for labeling Locations and Events. CRF has been known to give good results for this task, named entity recognition. We used lexical/syntactic input features, such as, token, previous token, capitalization, and POS to classify Location. We evaluated the performance for both partial matches and exact-span matches on tweets from North American events annotated with ACE categories. Exact-span match is incremented when the whole phrase is detected by the classifier. A partial match is true if any of the words in the annotation are classified correctly. Because exact-span matches are a subset of partial matches, evaluation on partial matches performs better. We used 60% of the annotated data for training and 40% held-out data for testing. Results are shown in Table 6 below.

Match Criteria	Precision	Recall	F1
Exact Match	69.5	63.4	66.3
Partial Match	86.7	79.1	82.8

Table 6: Entity Classifier Performance for Location Spans

5. Discussion

Our early classification results are encouraging. We have good preliminary results and will explore features that help reduce false positive and negative rates. We will also explore domain adaptation and active learning techniques so small amounts of annotated data can be effective over varying types of crisis events. At present, Named Entity annotations are not connected to our work on Situational Awareness. We are incorporating entity information as a feature in future behavioral classifiers; a system is in progress to predict information types within SA tweets using entity information as a feature.

Future work will also include an integration of PropBank style semantic role labeling (Palmer et al., 2005) as a more verb-specific annotation of tweet elements. Preliminary examination suggests that gold-standard semantic role labeling could be generated through hand-correction of the output of an in-house automated semantic role labeling system (Choi and Palmer, 2011a; Choi and Palmer, 2011b).
Finally, NLP system described here will be embedded within a larger system that analyzes features found outside of the tweet text, in the metadata provided by Twitter.

6. Acknowledgements

We gratefully acknowledge sponsorship from the US National Science Foundation Grant IIS-0910586. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. References

K. Anderson and A. Schram. 2011. Design and implementation of a data analytics infrastructure in support of crisis informatics research. *Proceedings of the 33rd International Conference on Software Engineering (ICSC 2011)*.

R. Bikel, D. M. Schwartz and R.M. Weischdel. 1999. An algorithm that learns what's in a name. *The Machine Learning Journal Special Issue on Natural Language Learning*.

J.D. Choi and M. Palmer. 2011a. Getting the most out of transition-based dependency parsing. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL: HLT '11)*, pages 687–692.

J.D. Choi and M. Palmer. 2011b. Transition-based semantic role labeling using predicate argument clustering. *Proceedings of ACL Workshop on Relational Models of Semantics (RELMS'11)*, pages 37–45.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

W.J. Corvey, S. Vieweg, T. Rood, and M. Palmer. 2010. Twitter in mass emergency: What nlp techniques can contribute. *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 23–24.

M.R. Endsley and D.J. Garland, editors, 2000. *Situation Awareness Analysis and Measurement*, chapter Theoretical Underpinnings of Situation Awareness: A Critical Review. CRC Press.

M.R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37:32–64.

J. Harrald and T. Jefferson. 2007. Shared situational awareness in emergency management mitigation and response. *Proceedings of the Hawaii International International Conference on Systems Science (HIICS 2007)*.

Linguistic Data Consortium (LDC). 2004. Automatic content extraction. `www.ldc.upenn.edu/Projects/ACE/`.

A.K. McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

P. Ogren. 2006. Knowtator: A protégé plug-in for annotated corpus construction. *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

L. Palen, S. Vieweg, S.B. Liu, and A.L. Hughes. 2009. Crisis in a networked world: Features of computer-mediated communication in the april 16, 2007 virginia tech event. *Social Science Computer Review*, 27(4):1–14.

L. Palen, K.M. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald. 2010. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. *Proceedings of the Association of Computing Machinery and British Computing Society's 2010 Conference on Visions of Computer Science*.

M. Palmer, D. Gildea, , and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Y. Qu, P.F. Wu, and X. Wang. 2009. Online community response to major disaster: A study of tianya forum in the 2008 sichuan earthquake. *Proceedings of the Hawaii International International Conference on Systems Science (HIICS 2009)*.

K. Starbird, L. Palen, A.L. Hughes, and S. Vieweg. 2010. Chatter on the red: What hazards threat reveals about the social life of microblogged information. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)*.

S. Verma, S. Vieweg, W. Corvey, L. Palen, J.H. Martin, M. Palmer, A. Schram, and K.M. Anderson. 2011. Natural language processing to the rescue?: Extracting "situational awareness" tweets during mass emergency. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*.

S. Vieweg, A.L. Hughes, K. Starbird, and L. Palen. 2010. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. *Proceedings of CHI 2010*.