Twitter in Mass Emergency: What NLP Techniques Can Contribute

William J. Corvey[1], Sarah Vieweg[2], Travis Rood[1] & Martha Palmer[1]
[1]Department of Linguistics, [2]ATLAS Institute
University of Colorado
Boulder, CO 80309
William.Corvey, Sarah.Vieweg, Travis.Rood, Martha.Palmer@colorado.edu

Abstract

We detail methods for entity span identification and entity class annotation of Twitter communications that take place during times of mass emergency. We present our motivation, method and preliminary results.

1 Introduction

During times of mass emergency, many turn to Twitter to gather and disperse relevant, timely information (Starbird et al. 2010; Vieweg et al. 2010). However, the sheer amount of information now communicated via Twitter during these time- and safety-critical situations can make it difficult for individuals to locate personally meaningful and actionable information. In this paper, we discuss natural language processing (NLP) techniques designed for Twitter data that will lead to the location and extraction of specific information during times of mass emergency.

2 Twitter Use in Mass Emergency

Twitter communications are comprised of 140-character messages called "tweets." During times of mass emergency, Twitter users send detailed information that may help those affected to better make critical decisions.

Our goal is to develop techniques to automatically identify crucial pieces of information in these tweets. This process will lead to the automatic extraction of information that helps people understand the situation "on the ground" during mass emergencies. Relevant information would include such things as warnings, road closures, and evacuations among other timely information.

3 The Annotation Process

A foundational level of linguistic annotation for many natural language processing tasks is Named Entity (or nominal entity) tagging (Bikel 1999). Typical labeled entities that were included in the Automatic Content Extraction (ACE) guidelines (LDC 2004) are: Person, Location, Organization, and Facility, the four maximal entity classes. Our preliminary annotation task consists of identifying the syntactic span and entity class for these four types of entities in a pilot set of Twitter data (200 tweets from a data set generated during the 2009 Oklahoma grassfires). In future annotation, the ontology will be expanded to include event and relation annotations, as well as additional subclasses of the entities now examined. Annotations are done using Knowtator (Ogren 2006), a tool built within the Protégé framework (http://protege.stanford.edu/). The ontology development is data-driven; as such it is likely that certain ACE annotations will never emerge and other annotations (such as disaster-relevant materials) will be necessary additions.

Three annotators undertook pilot annotation as part of the construction of preliminary annotation guidelines; the top pairwise ITA score is reported below. Twitter data makes reference to numerous entity spans that are of specific interest to this annotation task, such as road intersections and multi-word named entities. The example below, from the pilot annotation set, shows a relatively simple span delineation.

[PERSON Velma area residents]: [PERSON Officials] say to take [FACILTY Old Hwy 7] to [FACILTY Speedy G] to safely evacuate. [LOCATION Stephens Co Fairgrounds] in [LOCATION Duncan] for shelter

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 23–24,
Los Angeles, California, June 2010.

Because of the varying length of entities, annotators cannot be given simple rules for deciding the spans for annotations. This difficulty is reflected in markedly lower rates for span identification inter-annotator agreement (IAA) rates than for simple class assignment.

4 Preliminary Results

IAA calculations were performed using the Knowtator IAA functionality. When annotations are required to be both the same span and class, the pilot annotation yielded an F-score of 56.27 (An additional 4% have exact span matches but different classes). However, when annotations are required to have the same class assignment but only overlapping spans, this F-score rises to 72.85. While Facility and Location are the most commonly confused classes, span-matching remains a difficult issue for all entity classes.

5 Discussion

While these ITA rates are significantly lower than published results from previous ACE annotation efforts (LDC 2004), we believe that the crisis communications domain, particularly with regard to Twitter analysis, provides challenges not encountered in newswire, broadcast transcripts, or newspaper data. First, determining the maximal span of interest for a given class assignment is non-trivial. The constraint of 140 characters necessarily results in very limited syntactic and semantic contexts, making spans and entity class assignments much harder to determine.

A large source of disagreement was on the treatment of coordinated or listed noun phrases. In certain contexts, each entity (*cities* below) requires its own span (e.g. "Firestorms in Oklahoma. [*Midwest City*], [*Lake Draper*]. Some houses lost"), whereas in other contexts we find *multiple* entities per span (e.g. "Midwest City to evacuate between SE 15th and Rena and Anderson and Hiwassee also [*Turtlewood, Wingsong, and Oakwood additions*]"). Equally, class assignment cannot be a mechanistic process or accomplished by reference to lists, as it is important to distinguish between cases where terms have been elided due to limited space and cases where no elision has taken place. For instance, the entity "Attorney General" (as opposed to "Attorney General's Office") might be annotated 'Person' or 'Organization' depending on context, or simply ambiguous, i.e. lacking sufficient context. It is primarily these unclear cases of class assignment that will require careful discussion in the annotation guidelines and in future mappings to an ontology.

In summary, this pilot study represents a new application of ACE annotation practices to a uniquely challenging domain. We outline issues that place special demands on annotators and future directions for ongoing research. We are confident that as we refine our guidelines and provide more cues and examples for the annotators that the determination of spans and entity classes will improve.

Acknowledgments

This work is supported by the US National Science Foundation IIS-0546315 and IIS-0910586 but does not represent the views of the NSF. This work was conducted using the Protégé resource, supported by grant LM007885 from the US NLM.

References

Daniel M. Bikel, Richard Schwartz and Ralph M. Weischedel. 1999. *An Algorithm that Learns What's in a Name*. In: the Machine Learning Journal Special Issue on Natural Language Learning.

George Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. (2004). *The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation.* In: Proceedings of Conference on Language Resources and Evaluation (LREC 2004).

Kate Starbird, Leysia Palen, Amanda L. Hughes and Sarah Vieweg. 2010. *Chatter on The Red: What Hazards Threat Reveals About the Social Life of Microblogged Information.* In: Proc. CSCW 2010. ACM Press.

LDC, 2004, Automatic Content Extraction [www.ldc.upenn.edu/Projects/ACE/]

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

Sarah Vieweg, Amanda L. Hughes, Kate Starbird and Leysia Palen. 2010. *Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness.* In: Proc. CHI 2010. ACM Press.