EMTerms 1.0: A Terminological Resource for Crisis Tweets

Irina Temnikova
Qatar Computing Research Institute
itemnikova@qf.org.qa

Carlos Castillo
Qatar Computing Research Institute
chato@acm.org

Sarah Vieweg
Qatar Computing Research Institute
svieweg@qf.org.qa

ABSTRACT

We present the first release of EMTerms (Emergency Management Terms), the largest crisis-related terminological resource to date, containing over 7,000 terms used in Twitter to describe various crises. This resource can be used by practitioners to search for relevant messages in Twitter during crises, and by computer scientists to develop new automatic methods for crises in Twitter.

The terms have been collected from a seed set of terms manually annotated by a linguist and an emergency manager from tweets broadcast during 4 crisis events. A Conditional Random Fields (CRF) method was then applied to tweets from 35 crisis events, in order to expand the set of terms while overcoming the difficulty of getting more emergency managers' annotations.

The terms are classified into 23 information-specific categories, by using a combination of expert annotations and crowdsourcing. This article presents the detailed terminology extraction methodology, as well as final results.

Keywords

Terminological resource, crises, Twitter

INTRODUCTION

Social media are being used more and more for communicating, detecting, tracking, and extracting information about currently occurring or recently passed crises (see Blanchard et al., 2012; Cobb et al., 2014; Denef et al., 2013; Denis et al., 2012; Fraustino et al., 2012; Hughes and Palen, 2012; Hughes et al., 2014; Imran et al. 2014a; Olteanu et al., 2015; Sarcevic et al. 2012; Starbird et al. 2010; Starbird, 2013; St. Denis et al., 2014; Vieweg et al. 2010). Due to the opportunity to share a message with a potentially large audience, members of the public often use social media to communicate about crises (Blanchard et al., 2012; Imran et al., 2014c; Vieweg, 2012).

Social media are also used to break stories; recent high-profile cases include the Asiana flight crash in July 2013, a picture of which was posted on Twitter 30 seconds after the crash[1], and the first *tweet* (Twitter message) about the Westgate Mall attack[2] on September 2013, which was posted the first minute after the event occurred (Imran et al., 2014b).

[1] http://www.slideshare.net/shanxz/asiana-flight-214-crash-in-sfo-crises-management-case-study-and-analysis.

[2] http://community.ihub.co.ke/blogs/16012/how-useful-is-a-tweet-a-review-of-the-first-tweets-of-the-westgate-attack.

However, when it comes to retrieving information posted on social media sites, the task faced by emergency managers is daunting. Large crises can generate millions of social media messages (e.g. 3.5 million messages in one day, during the Hurricane Sandy in 2012[3]). Manually checking each message and filtering those which are relevant is a cumbersome - if not impossible - task. Several automatic systems have been built to help emergency managers identify and filter useful information posted to social media sites (Cataldi, 2010; Hampton, A., 2014; Imran et al., 2013; Imran et al., 2014a; Mizuno and Inui, 2013; Pohl et al., 2012; Temnikova et al., 2013; Varga et al., 2013).

Among these methods, focusing on linguistic expressions used to describe crises in social media is a relatively recent development (Hampton, 2014; Imran et al., 2013; Olteanu et al., 2014; Roy Chowdhury et al., 2013; Temnikova et al., 2014; Vieweg, 2012). Some of these methods have produced linguistic resources, including the 380-term CrisisLex vocabulary (Olteanu et al., 2014). However, no extensive linguistic resource exists that covers the language used in social media to communicate information about crises. To address this gap, we present the first release of an extensive terminological resource for crisis management in English, which reflects *the real, observed linguistic expressions used in Twitter to describe a wide-range variety of crises*, and, at the same time, *focuses on the information needs of emergency managers*. To ensure this last criterion, the seed set of terms used for developing the resource were manually selected by an emergency manager, and the terms were divided into 23 information-specific categories, useful to emergency managers. The resource contains over 7,000 manually annotated terms, collected from tweets sent during 35 crisis events occurring between 2012-2014. The events include 22 natural hazards and 13 human-induced crises, ranging from earthquakes and floods, to terrorist attacks and building collapses.

The resource can be used by crisis managers, linguists, and computer scientists developing crisis computer applications for Twitter (for example for information categorization and extraction, ontology population, and text summarization). Crisis managers can use EMTerms 1.0 to: 1) Get an overview of the linguistic expressions used by citizens in Twitter to refer to crises; 2) Retrieve tweets relevant to a certain category via Twitter or via a crisis management social media platform (like AIDR[4] or Sahana Eden[5]); 3) Build a detailed informative picture of a crisis, as "handling highly dynamic scenarios … requires lots of information about the situation" (Wucholt et al., 2011); 4) To serve as a basis for translation of crisis terms into languages other than English.

The next sections will present the related work, provide details about the terminology collection methods, as well as the final results and statistics.

RELATED WORK

This section introduces the related work most relevant to us, namely: 1) terminological resources in other domains; 2) terminological and lexical resources in crisis computing for Twitter; and 3) previous research in tweet classification by information category.

Terminological Resources in Other Domains

In order to present previous work, we need to define the notions of *term*s, *terminological resources, and automatic terminology recognition*.

- *Terms* can be defined as single words and multi-words expressions, which are "highly domain-specific" (Hanks, 2010), and which correspond to important concepts in a specialized domain (Ahmad and Collingham, 1996). They are usually more frequent in this specific semantic domain, and less frequent in other domains (Drouin, 2003; Korkontzelos and Ananiadou, 2014).
- *Terminological resources* are *collections of terms* narrowly characterizing a closed semantic domain, and corresponding to key concepts. More complex terminological resources (Thompson et al., 2011) provide additional information about the terms, including the relationships between different concepts. Terminological resources have been developed, among others, for areas like the medical and bio-medical domains (UMLS, Lindberg et al., 1993; BioLexicon, Thompson et al., 2011), environment and ecology (EcoLexicon, Faber et al., 2014) the legal domain, aeronautics, and computer science (Godman, 1984).

[3] http://www.cbsnews.com/news/social-media-a-news-source-and-tool-during-superstorm-sandy/.

[4] http://aidr.qcri.org.

[5] http://sahanafoundation.org/products/eden/.

- *Automatic Terminology Recognition (ATR)* (or *terminology extraction*) from domain-specific texts is a subfield within Natural Language Processing, which tackles the recognition and extraction of terms from specialized *corpora,* or text collections (Bourigault et al., 2001; Jacquemin, 2001; Korkontzelos and Ananiadou, 2014; L'Homme, 2004). ATR tools collect terms to be stored in terminological resources, or to be used by other NLP applications, such as information retrieval, information extraction, and ontology population.
  A variety of ATR approaches exist. Most methods recognize the terms by applying a mixture of linguistic and statistical approaches: TF-IDF (Manning et al., 2008); C-Value (Frantzi et al., 2000); Kyoto Scoring (Bosma & Vossen, 2010); simplemaths (Kilgarriff, 2012). These methods, however, work on large text documents and are not applicable to tweets. Methods borrowed from information retrieval (Olteanu et al., 2014) apply statistical tests such as chi-square, or Point-wise Mutual Information (PMI) to obtain the terms that statistically discriminate crisis tweets. As we want to ensure that the extracted terms are approved by human experts, their method is not applicable here. For this reason, we start with manually annotated terms, and apply several stages of human reviewing and cleaning.

Terminological and Lexical Resources in Crisis Computing for Twitter

To the best of our knowledge, there are only 3 lexical and terminological resources for crisis tweets, which are made available and freely distributed.

CrisisLex (Olteanu et al., 2014), available along with an extensive tweets collection on http://crisislex.org/ (Last accessed on November 20th, 2014), represents a collection of 380 single-word terms, common across crises, and found to be statistically frequent and discriminative for crisis tweets. The terms have been collected via manual annotation of those tweets, which talk about a crisis. An automatic extraction of those terms, which are considered to be the most frequent and statistically discriminative for crisis tweets, is performed after that. A final manual curation is done by crowdsource workers who validate the automatically extracted terms.

There are two fundamental differences between EMTerms and CrisisLex: (i) EMTerms is almost 20 times larger; and (ii) EMTerms are separated into information-specific categories. Additionally, CrisisLex terms have been selected by their discriminative power.

The second resource is a previous work of the authors (Temnikova et al., 2014), consisting of a preliminary investigation of the linguistic nature of terms used in Twitter in 2 types of crises. The terms have only been annotated manually and various statistics about them described. The resource makes available over 1000 terms, manually annotated by two annotators in 500 tweets for each of four events.

The third resource (Vieweg, 2012) is a list of 39 verbs, grouped in 9 VerbNet classes, observed in tweets coming from four different crisis events. The identified VerbNet classes highly correlate with tweets that mention the social, built, or physical environment, defined to correspond to situational awareness tweets. The author states that adding this knowledge as features to a Machine Learning (ML) classifier would enhance tweet categorization. Similar to this resource, EMTerms 1.0 provides terms, divided into information-type categories. This advantage allows for the ability to search for different categories of tweets that include information about different aspects of a crisis situation.

Tweet Classification per Categories

As the immediate application of our resource would be to assist in classifying crisis tweets into information-specific categories, this section introduces the relevant related work.

One previous approach to the classification of crisis tweets into various combinations and number of categories shows that tweets can mostly be split into three major categories: *informativeness, information type*, and *source* (Olteanu et al., 2015). From these categories, almost all of EMTerms 1.0's categories are pulled from the *information type* category. EMTerms 1.0 also defines the category "witness," which can be classified into the major category *source* in Olteanu et al. (2015).

In terms of methods of tweets classification, many approaches apply a manual reading and content analysis of tweets, and then manual assignment into categories (e.g. Denef et al., 2013; Hughes et al., 2014; Qu et al., 2011; Sreenivasan et al., 2011). The shortcoming of using a purely manual classification approach is that it limits the amount of tweets to be analyzed, and of categories identified. Several approaches use manual annotation and supervised ML techniques (Diakopoulos et al., 2012; Imran et al., 2013; Starbird et al., 2012). From these, only Diakopoulos et al. (2012) used a vocabulary of 700 words (not publicly released).

OVERVIEW OF THE APPROACH

The creation of a terminological resource is a resource-intensive process in which multiple objectives need to be balanced. One of the key objectives is to be *comprehensive*, i.e. to include as many terms as possible. Another key objective is to be *specific*, i.e. to avoid including terms which are not related to the given domain. A third objective in our case is to be *accurate*, i.e. to assign the correct category label to each of the terms. These objectives to some extent compete with each other. If we keep human effort constant, given a finite budget of human annotations, increasing the number of terms means risking errors in terms of specificity and accuracy.

The method we adopt includes some degree of automatic discovery of new terms, in order to make the resource more comprehensive. However, methods for automatic discovery of terms are not 100% accurate, and thus necessarily involve a post-filtering step in which terms that are not related to the domain need to be removed.

At a high-level, our method consists of four steps:

1. Manual annotation of a set of seed terms.
2. Automatic expansion of the set of seed terms, to discover new candidate terms.
3. Manual filtering of the automatically-discovered candidate terms.
4. Crowdsourced verification of all the terms.

The first step into building EMTerms 1.0, the **manual annotation of a set of seed terms**, was based on previous work of the authors, which consisted of automatically collecting 500 tweets for each of 4 different crisis events (two floods and two protests) by searching for specific hashtags. The terms were manually annotated by an emergency manager and a linguist. Insights into the linguistic nature of the collected terms were drawn. In the course of *this work*, these terms were further categorized into 23 different classes, described in the following sections. The validation of the seed terms by one crisis manager only is sufficient at this preliminary terms collection stage. When the list of terms will be sufficiently extended, we envisage validating the completed EMTerms with a large number of crisis managers. This is the first terminological resource which relies on the annotations and validation of crisis managers.

The second step, **automatic expansion of the set of seed terms, to discover new candidate terms**, was performed using a well-known automatic information extraction method: Conditional Random Fields (CRF). This is a statistical ML method which takes as input a series of annotated *training* examples, which in our case are the set of seed terms, and can be applied over new, unseen data, to uncover new terms similar to the ones given as training examples. In our case, the new data corresponded to tweets from 35 different crisis events. This resulted in 7,841 candidate terms. Other existing automatic terminology recognition methods were not applied due to the reasons described in the related work section.

During the third step, **manual filtering of the automatically-discovered candidate terms**, each of the 7,841 candidates was reviewed by a linguist who approved or rejected their assignment to each category, or moved or copied them into a different category.

Finally, a **crowdsourced verification of all the terms** was performed asking 3 different crowdsourcing workers to review each of the terms and indicate if s/he believed the term was related to disaster response and corresponded to the assigned category. Whenever there was agreement among them, we accepted the terms.

The final resource contains 7,241 terms, divided into 23 categories. In the next sections we present in detail each of the steps we have outlined.

INITIAL SEED SET ANNOTATION

The initial data consisted in 500 tweets from each of the following four events:

- 2013 Russia-China floods (Event1),
- 2013 Pakistan-Afghanistan floods (Event2),
- 2013 Bohol earthquake (Event3),
- 2012 Colorado wildfires (Event4).

The tweets discussing Event1 and Event2 came from previous work (Temnikova et al., 2014) and were manually annotated by an emergency manager and a linguist using the GATE software (Maynard et al., 2008). Tweets from Event3 and Event4 were annotated by a linguist.

As a second annotation step, a linguist split the extracted terms from Event1 and Event2 into 23 pre-defined categories, listed in Table 1. A spreadsheet program and a series of simple scripts were used for this purpose.

The categories originate from 3 sources. First, categories T01-T11 correspond to information type classifiers used by default in AIDR classifiers (Imran et al., 2014a), which in turn are based on existing classes described in the social media for emergency management literature. Second, categories C01-C08 form the United Nations Humanitarian Cluster System[6], which is the standard way in which humanitarian efforts are organized by UN OCHA and other organizations. Finally, Categories O01-O04 ("Other") complement the above, and are based on consultations with specialists and a study of the current literature on the topic.

In total, 1,892 terms were annotated during this phase, distributed according to the last column of Table 1.

Code	Name	Description	Seed terms
T01	Caution and advice	Warnings issues or lifted, guidance, and tips.	659
T02	Injured people	Casualties (injured) due to the crisis.	20
T03	Dead people	Casualties (deceased) due to the crisis.	131
T04	Infrastructure damage	Buildings or roads damaged or operational; utilities/services interrupted or restored.	283
T05	Money	Money requested, donated, or spent.	32
T06	Supplies needed or offered	Needs or donations of supplies such as food, water, clothing, medical supplies or blood.	52
T07	Services needed or offered	Services needed or offered by volunteers or professionals.	37
T08	Missing, found, or trapped people	Questions and/or reports about missing or found people.	19
T09	Displaced and evacuated people	People who have relocated due to the crisis, even for a short time (includes evacuations).	78
T10	Animal management	Pets and other non-human animals, living, missing, displaced, or injured/dead.	20
T11	Personal updates, sympathy	Status updates about individuals or loved ones; emotional support, thoughts and prayers.	369
C01	Children and education	Children's well-being and education.	27
C02	Food and nutrition	Nutritional well-being. Needs food, or able to provide food.	12
C03	Health	Mental, physical, emotional well-being and health.	16
C04	Logistics and transportation	Delivery and storage of goods and supplies.	21
C05	Camp and shelter	Shelter required or offered; condition and location of shelters and camps.	16
C06	Water, sanitation, and hygiene	Availability of clean water, waste and sewage disposal, access to hygienic facilities.	12
C07	Safety and security	Protection of people/property against harm such as violence or theft.	21
C08	Telecommunications	Mobile and landline networks, Internet.	15
O01	Weather	Updates about the weather.	75
O02	Response agencies in place	Formal response agencies present (and acting) at the crisis location.	164
O03	Witnesses' accounts	Direct accounts by eyewitnesses of the crisis.	2
O04	Impact of the crisis	Negative consequences of the crisis.	2

Table 1: EMTerms 1.0 Categories.

AUTOMATIC EXTRACTION OF CANDIDATE TERMS

[6] http://www.unocha.org/what-we-do/coordination-tools/cluster-coordination.

After the manual annotation of the seed terms, we performed an automatic expansion phase to find new candidates. This expansion was done by training an automatic information extractor to recognize similar terms in a larger collection.

Automatic information extraction is usually done using machine learning (ML), specifically structured and supervised ML. *Structured* because the input is not a vector but a more elaborate data structure, in this case a list of words forming a sentence. *Supervised* because the method needs to be "trained" using a set of elements for which the labels are known.

The method we chose to perform this extraction is Conditional Random Fields (CRF), a probabilistic method for learning on sequences which is the state-of-the-art for many Natural Language Processing (NLP) operations. CRF takes a structured input, in this case, a sequence of features representing each word, and produces a structured output, which is a sequence of labels. The specific implementation we used is the one in ArkNLP,[7] which is a Twitter-specific system in which features and parameters have been optimized for collections of tweets. The features used to represent each word include generic and Twitter-specific features: word length, capitalization, presence of an "@"-sign, etc.

The data used to "train" this automatic information extractor were all the 1,892 terms, which we refer to as *the seed set*, each one accompanied by an example tweet in which this term was used. The input was reformatted as a sequence of <word, marker> tuples in which the marker was 1 if the word did not belong to the extracted term, and 0 otherwise. For instance, for category C1 ("Caution and advice"), this was one of the input items:

- Term: "major earthquake"
- Tweet: "Death toll from major earthquake in central Philippines ..."
- Formatted training item: < <Death,0>, <toll,0>, <from,0>, <major,1>, <earthquake,1>, <in,0>, <central,0>, <Philippines,0>, … >

Each category-specific trained model was then applied over a dataset consisting of data from 35 different crises: (i) the entire set of tweets from the CrisisLexT26 (Olteanu et al., 2015) collection, consisting of 26 events; (ii) plus data from 7 events provided by the lead author of that study and collected in a similar way; (iii) plus the 2 initial flood events from which our initial sample was obtained. The complete list of crisis events is provided in Table 2.

Crisis	**Type**	**Sub-Type**
2012 Italy earthquakes	Natural hazard	Geophysical
2012 Costa Concordia ship accident	Human-induced	Accidental
2012 Colorado wildfires	Natural hazard	Climatological
2012 Philippines floods	Natural hazard	Hydrological
2012 Venezuela refinery explosion	Human-induced	Accidental
2012 Costa Rica earthquake	Natural hazard	Geophysical
2012 Guatemala earthquake	Natural hazard	Geophysical
2012 Typhoon Pablo	Natural hazard	Meteorological
2013 Brazil nightclub fire	Human-induced	Accidental
2013 Queensland floods	Natural hazard	Hydrological
2013 Russian meteor	Natural hazard	Others
2013 Boston bombings	Human-induced	Intentional
2013 Savar building collapse	Human-induced	Accidental
2013 West Texas explosion	Human-induced	Accidental
2013 Alberta floods	Natural hazard	Hydrological
2013 Singapore haze	Mixed	Others
2013 Lac-Mégantic train crash	Human-induced	Accidental
2013 Spain train crash	Human-induced	Accidental
2013 Manila floods	Natural hazard	Hydrological

[7] http://www.ark.cs.cmu.edu/TweetNLP/.

2013 Colorado floods	Natural hazard	Hydrological
2013 Australia wildfires	Natural hazard	Climatological
2013 Bohol earthquake	Natural hazard	Geophysical
2013 Glasgow helicopter crash	Human-induced	Accidental
2013 LA Airport shootings	Human-induced	Intentional
2013 NYC train crash	Human-induced	Accidental
2013 Sardinia floods	Natural hazard	Hydrological
2013 Typhoon Yolanda	Natural hazard	Meteorological
2013 Jakarta floods	Natural hazard	Hydrological
2013 Nairobi airport fire	Human-induced	Accidental
2013 North India floods	Natural hazard	Hydrological
2013 Pakistan earthquake	Natural hazard	Geophysical
2013 Solomon Islands earthquake	Natural hazard	Geophysical
2013 Toronto floods	Natural hazard	Hydrological
2013 Pakistan-Afghanistan floods	Natural hazard	Hydrological
2013 Russia-China floods	Natural hazard	Hydrological

Table 2: Crisis Events Statistics.

As shown on Table 2, the events cover a variety of countries and include 12 human-induced crises, 22 natural hazards crises, and 1 crisis of mixed nature[8].

The ArkNLP CRF received as input a file that is equivalent to this example from the category C1 ("Caution and advice"):

- Input: "Massive earthquake in southwest Pakistan ..."
- Output: <<Massive,1>, <earthquake,1>, <in,0>, <southwest,0>, <Pakistan,0>, … >

Sequences of consecutive words for which the output label was 1 were concatenated to create new candidate terms. Candidate terms where then sorted by frequency in descending order, starting from the terms that appeared in most tweets. Finally, terms were automatically labeled according to whether they existed in the seed set or not. A total of 7,564 candidate terms were produced. An example output for the category C06 ("Water, sanitation, and hygiene") includes[9]:

1. "water", Existing term, frequency 47
2. "river", New term, frequency 25
3. "floods", Existing term, frequency 14
4. "flood water", Existing term, frequency 12
5. "sewage water", New term, frequency 7
6. "relief work", New term, frequency 6
7. …

As we can see from this example, the most frequent terms contain a combination of (i) terms existing in the input data, such as "water" in the example (ii) new terms discovered by the CRF method, and specific to the relevant class, such as "sewage water" in the example, and (iii) new terms discovered by the CRF method, but not specific to the relevant class, such as "relief work" from the example above. Out of the 7,564 terms recognized in the 35 events crisis tweets, 353 (i.e. 4.6%) were terms, existing in the input data, and 7,209 (i.e. 95.3%) were new, unseen terms.

[8] The Singapore haze is caused by a combination of climatological factors and intentional fires to clear land.

[9] Examples for candidate terms omitted for brevity.

Qualitatively, and across different categories, we observe that the high-frequency extracted terms are usually relevant for their intended category; however we also observe relatively low-frequency terms, which are also relevant. For this reason, we examine even candidates terms that appear only once.

ANNOTATION OF CANDIDATE TERMS

Due to the fact that both high-frequency and low-frequency terms (including those which appear only once) were worthy of examination, all of the new candidate terms identified by the previous step were verified manually by one of the linguist authors of this paper. In this manual verification step, the candidate term and the context in which it was found were examined. At this step, the following outcomes were possible for each of the candidates:

- Accept the candidate term in this category and with this example tweet
    - Optionally, add it additionally to other categories.
- Move the candidate term to a different category, with the same example tweet.
- Reject the candidate term.

After this annotation, out of the 7,209 new candidate terms produced by the CRF, 4,721 terms (65.5%) were accepted (moved or kept in the same category), while 2,488 terms (i.e. 34.5%) were rejected. This indicates that the CRF-based automatic generation of candidates cannot be used in isolation but requires a manually cleaning step.

CROWDSOURCED QUALITY ASSURANCE

Finally, we used crowdsourcing to locate cases in which the annotation was done mistakenly, or cases in which the interpretation of the term as belonging to the category would not be understandable or shared by the public.

Crowdsourcing was done with CrowdFlower[10]. We passed all terms (both the initial ones, and the ones obtained via the CRF candidate generation and annotation) through crowdsource workers. The total number of terms given to CrowdFlower workers to annotate, were 7,841. The question posed was the one in Figure 1:

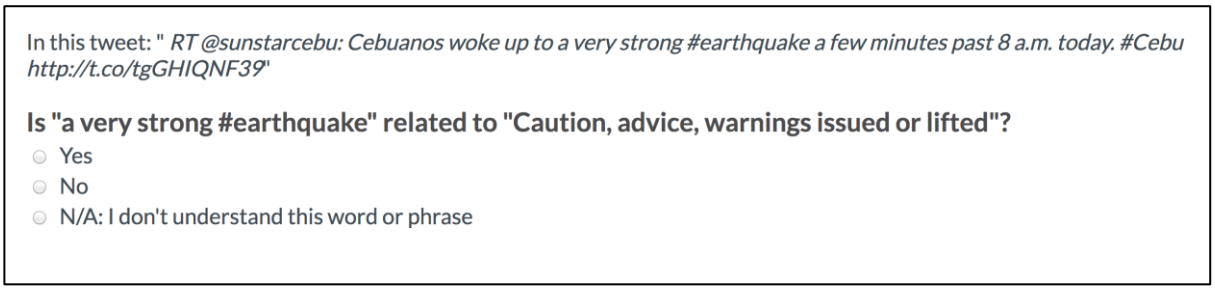

Figure 1: Crowdsourcing annotation task.

We selected annotators living in English-speaking countries, asking for 3 independent labels per each <term, category, example> element. The results from Crowdflower include a confidence score for each annotation (a combination of inter-annotator agreement and worker trust, a measure of the extent to which annotators agreed with a set of gold standard elements provided by us). The gold standard was composed of 10 clearly positive and 10 clearly negative cases that were interleaved with the tasks containing elements to be labeled, following standard practices in this crowdsourcing platform.

Crowdsource annotators rejected about 9.1% of the terms with confidence greater than 80%. Qualitatively, we observe two classes of rejections: mistakes done by the author-annotator, which naturally occur given the size of this annotation task, and different interpretations on how narrow a category should be.

[10] http://crowdflower.com/.

As a final step to have a uniform interpretation, all the authors of this paper jointly reviewed each rejection case, applying a more "strict" approach, and in this way keeping a small fraction (16.2%) of the rejected cases. The total number of final terms which stayed in the resource are, in this way, **7,241**.

Figure 2 shows some of the rules the authors followed in this final clean-up.

- Remove the term, if, according to the context, the term is not strictly related to the category it was assigned to, or not related at all:
  - Term: "bears evacuated by"
  - Category: *Displaced and evacuated people*
  - Tweet: "Amazing rescue: Bears evacuated by helicopter in flood-hit Russia's Far ..."
- Remove the term, if it is event-specific, or if it is mentioning a geographic location:
  - Term: "Colorado's wildfires"
  - Category: *Caution and advice*
  - Tweet: "@TIME: WATCH: Four harrowing videos of Colorado's wildfires | http://t.co/Bb2p7Kj"
- Remove the term if it could rarely be interpreted as a crisis:
  - Term: "island emerges"
  - Category: *Caution and advice*
  - Tweet: "@DazMSmith MRT @RichardJConway: New island emerges from sea following Pakistan earthquake: http://t.co/GCNafwW9AF via @TIME @TIMEPicture"

Figure 2: Rules followed by authors for final clean-up.

AVAILABILITY AND EXAMPLES

EMTerms 1.0 is available for download at *http://crisislex.org/crisis-lexicon.html*. The terms are listed in a .csv file, with one record per line of comma-separated fields. Each line contains the following fields:

- Term
- Category code
- Category name
- Category description
- Example tweet text, containing the term as belonging to the associated category

Table 3 includes some examples from the resource. For matters of space, the category descriptions (same as in Table 1) are omitted. For clarity, in Table 3 we have also underlined the term inside the example.

Term	Category code	Category name	Example tweet text
{Number} killed	T03	Dead people	[#WNewsIreland]: At least **232 killed** in Brazil nightclub blaze 'after musician set fireworks of... http://.../ #Ireland #Dublin
volunteers needed	T07	Services needed or offered	RT @COEmergency: RT @weldgov: Food and **volunteers needed** at the Weld Food Bank...http://.../ #COFlood
impassable roads	C04	Logistics and transportation	**Impassable roads** hamper Philippines quake rescue efforts http://t.co/xY5JfML0sk http://.../
rescue workers	O02	Response agencies in place	Philippines struggles to help victims of killer quake: **Rescue workers** were forced Thursday onto boats and heli... http://t.co/e48dMePWMU
injured are children	C01	Children's well being & education	RT @cnnbrk: Injuries in #Boston Marathon terror attack now at 183; 23 critical; at least 9 of the **injured are children**. http://t.co/9YZV ...
shelter material	C05	Sheltering	RT @theOFDA: The @USAID heavy-duty sheeting will provide temporary **shelter material** for approx 20,000 families in need #PabloPH #ReliefPH
extreme rainfall	O01	Weather Conditions	Flooding from **extreme rainfall** is a hazard for ON too. Emergency in #Alberta should be a call to get a 72 hour kit. Better to be prepared.

Table 3. Examples from EMTerms 1.0.

As it can be seen, in EMTerms 1.0, some examples do not follow the usual terms "noun-phrase" (NP) structure (e.g. "injured are children"). This is because such were many of the terms annotated by the crisis manager (Temnikova et al., 2014).

CONCLUSIONS

We have described the methodology used to create a terminological resource for analyzing crisis-related messages broadcast on Twitter. The methodology consists of a manual annotation of a set of seed terms, an automatic expansion by using an information extraction algorithm, a second round of manual annotation, crowdsourced verification, and a final round of manual inspection of the cases disputed by crowdsource workers.

The result is a large and unique linguistic resource including over 7,000 terms in different categories of relevance for emergency managers. The resource can be accessed at *http://crisislex.org/crisis-lexicon.html.*

The construction of most large linguistic resources is iterative in nature; accordingly, this is the 1.0 release of EMTerms. Future work will include the expansion of the resource through the usage of other linguistic resources (for instance, looking for synonyms, paraphrases, or hyponyms), refinements of the categories used and/or addition of new information typologies (e.g. detailed linguistic information), plus the addition of new crisis-related datasets to increase the coverage, variety, and precision of the included terms. In order to address the two known problems in managing crisis terminology, namely that 1) the terminology used by citizens differs from the terminology of emergency professionals (Reuter et al., 2012; Temnikova et al., 2014), and 2) emergency professionals from different organizations (e.g. fire department, police, etc.) use different terminology (Reuter et al., 2012; Wucholt et al., 2011), we plan to organize public events involving a large number of emergency professionals to evaluate the usefulness of the resource and to disambiguate and come to a common meaning of its terms.

ACKNOWLEDGMENTS

We would like to thank all the reviewers for their valuable comments.

REFERENCES

1. Ahmad, K., & Collingham, S. (1996) POINTER Final Report. University of Surrey.
2. Blanchard, H., Carvin, A., Whitaker, M. E., Fitzgerald, M., Harman, W., and Humphrey, B. (2012) The case for integrating crisis response with social media.
3. Bosma, W., & Vossen, P. (2010) Bootstrapping Language Neutral Term Extraction. *In 7th Language Resources and Evaluation Conference (LREC)*.
4. Bourigault, D., Jacquemin, C., & L'Homme, M. C. (Eds.) (2001) *Recent advances in computational terminology* (Vol. 2). John Benjamins Publishing.
5. Cataldi, M., Di Caro, L., & Schifanella, C. (2010) Emerging topic detection on Twitter based on temporal and social terms evaluation. *Tenth International Workshop on Multimedia Data Mining* (p. 4). ACM.
6. Cobb, C., McCarthy, T., Perkins, A., Bharadwaj, A., Comis, J., Do,B., and Starbird, K. (2014) Designing for the Deluge: Understanding & Supporting the Distributed, Collaborative Work of Crisis Volunteers. *Proceedings of CSCW*.
7. Denef, S., Bayerl, P. S., & Kaptein, N. A. (2013) Social media and the police: Tweeting practices of British police forces during the August 2011 riots. *SIGCHI Conference on Human Factors in Computing Systems* (pp. 3471-3480). ACM.
8. Denis, L. A. S., Hughes, A. L., & Palen, L. (2012) Trial by fire: The deployment of trusted digital volunteers in the 2011 Shadow Lake Fire. *9th International ISCRAM Conference*.
9. Faber, P., Pilar L. A., and Reimerink, A. (2014) Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era*, Educational Linguistics 19, ed. E. Bárcena, T. Read and J. Arhus. Berlin, Heidelberg: Springer.
10. Fellbaum, C. (1998). *WordNet*. Blackwell Publishing Ltd.
11. Frantzi, K., Ananiadou, S., & Mima, H. (2000) Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, *3*(2), 115-130.
12. Fraustino, J. D., Liu, B., & Jin, Y. (2012) Social media use during disasters: A review of the knowledge base and gaps. National Consortium for the Study of Terrorism and Responses to Terrorism.
13. Godman, A. (1984) The Cambridge Illustrated Thesaurus of Computing Terms. Cambridge University Press, Cambridge.
14. Hampton, A. (2014) Computational Filtering and Management of Social Media Traffic to Assist Crisis Response Coordination.
15. Hanks, P. (2010) Lexicography, Terminology, and Phraseology. *Proceedings of Euralex 2010*. Leeuwarden.
16. Hughes, A. L., & Palen, L. (2012) The evolving role of the public information officer: An examination of social media in emergency management. *Journal of Homeland Security and Emergency Management*, *9*(1).
17. Hughes, A. L., St. Denis, L. A., Palen, L., and Anderson, K. (2014) Online public communications by police and fire services during the 2012 hurricane Sandy. *Proceedings of CHI* (2014).
18. Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014a) AIDR: Artificial intelligence for disaster response. Proceedings of *The companion publication of the 23rd international conference on World Wide Web companion* (pp. 159-162).
19. Imran, M., Castillo, C., Lucas, J., Meier, P., & Rogstadius, J. (2014b) Coordinating human and machine intelligence to classify microblog communications in crises. *Proceedings of ISCRAM*.
20. Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2014c) Processing Social Media Messages *Proceedings of* Mass Emergency: A Survey. *arXiv preprint arXiv:1407.7071*.
21. Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., & Meier, P. (2013) Extracting information nuggets from disaster-related messages in social media. *Proceedings of ISCRAM, Baden-Baden, Germany*.
22. Jacquemin, C. (2001) Spotting and discovering terms through natural language processing. MIT press.
23. Kilgarriff, A. (2012) Getting to know your corpus. In *Text, Speech and Dialogue* (pp. 3-15). Springer Berlin Heidelberg.
24. Korkontzelos, I. and Ananiadou, S. (2014) Term Extraction. In: Oxford Handbook of Computational Linguistics (2nd Ed.)

25. L'Homme, M. C. (2004) *La terminologie: Principes et techniques*. Pum.
26. Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993) The Unified Medical Language System. *Methods of information in medicine*, *32*(4), 281-291.
27. Manning, C. D., Raghavan, P., & Schütze, H. (2008) *Introduction to information retrieval* (Vol. 1, p. 496). Cambridge: Cambridge university press.
28. Maynard, D., Li, Y., and Peters, W. (2008) NLP Techniques for Term Extraction and Ontology Population.
29. Mizuno, J., & Inui, K. (2013) NICT Disaster Information Analysis System. In *Sixth International Joint Conference on Natural Language Processing* (p. 29).
30. Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014) CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. Proceedings of ICWSM.
31. Olteanu, A., Vieweg, S., & Castillo, C. (2015) What to Expect When the Unexpected Happens: Social Media Communications Across Crises. *Proceedings of CSCW 2015 (forthcoming)*.
32. Pohl, D., Bouchachia, A., & Hellwagner, H. (2012) Automatic identification of crisis-related sub-events using clustering. In Proceedings of *Machine Learning and Applications (ICMLA),* vol. 2, pp. 333-338. IEEE.
33. Qu, Y., Huang, C., Zhang, P., & Zhang, J. (2011) Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. Proceedings of *the ACM 2011 conference on Computer supported cooperative work* (pp. 25-34). ACM.
34. Reuter, C., Pipek, V., Wiedenhoefer, T., & Ley, B. (2012) Dealing with terminologies in collaborative systems for crisis management. In L. Rothkrantz, J. Ristvey, & Z. Franco (Eds.). *Proceedings of ISCRAM* (pp. 1–5). Vancouver, Canada.
35. Roy Chowdhury, S., Imran, M., Asghar, M. R., Amer-Yahia, S., & Castillo, C. (2013) Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In *10th International ISCRAM Conference*.
36. Sarcevic, A., Palen, L., White, J., Starbird, K., Bagdouri, M., and Anderson, K. (2012) Beacons of hope in decentralized coordination: Learning from on-the-ground medical twitterers during the 2010 Haiti earthquake. *Proceedings of CSCW*. ACM, 47–56.
37. Schuler, K. K. (2005) VerbNet: A broad-coverage, comprehensive verb lexicon.
38. Sreenivasan, N. D., Lee, C. S., & Goh, D. H. L. (2011) Tweet me home: exploring information use on twitter in crisis situations. In *Online Communities and Social Computing* (pp. 120-129). Springer Berlin Heidelberg.
39. Starbird, K., Muzny, G., & Palen, L. (2012) Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions. *Proceedings of ISCRAM 2012*.
40. Starbird, K., Palen, L., Hughes, A. L., and Vieweg, S. (2010) Chatter on the red: what hazards threat reveals about the social life of microblogged information. *Proceedings of CSCW. ACM*, 241–250.
41. Starbird, K. (2012) Crowdwork, Crisis and Convergence: How the Connected Crowd Organizes Information during Mass Disruption Events (Doctoral dissertation, University of Colorado).
42. Starbird, K., & Palen, L. (2011) Voluntweeters: Self-organizing by digital volunteers in times of crisis. Proceedings of *the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1071-1080). ACM.
43. Starbird, K. (2013) Delivering patients to Sacré Coeur: Collective intelligence in digital volunteer communities. *Proceedings of CHI*. ACM, 801–810.
44. St. Denis, L., Palen, L and Anderson, K. (2014) Mastering Social Media: An Analysis of Jefferson County's Communications during the 2013 Colorado Floods. *Proceedings of ISCRAM* 2014.
45. Temnikova, I., Biyikli, D., and Boon, F. (2013) First steps towards implementing a Sahana Eden Social media dashboard. *Proceedings of SMERST*.
46. Temnikova, I., Varga, A., & Biyikli, D. (2014) Building a Crisis Management Term Resource for Social Media: The Case of Floods and Protests. *Proceedings of LREC 2014*. Reykjavik, Iceland, May 26-31, 2014.
47. Thompson, P. et al. (2011) The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, *12*(1), 397.
48. Varga, I., Sano, M., Torisawa, K., Hashimoto, C., Ohtake, K., Kawai, T., & De Saeger, S. (2013) Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster. In *ACL (1)* (pp. 1619-1629).
49. Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010) Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. *Proceedings of CHI*.

50. Vieweg, S. E. (2012) Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications. Doctoral dissertation, University of Colorado.
51. Wucholt, F., Yildirim-Krannig, Y., Mähler, M., Krüger, U., & Beckstein, C. (2011) Cultural Analysis and Formal Standardised Language – a Mass Casualty Incident Perspective. *Proceedings of ISCRAM*.