Pinpointing the Eye of the Hurricane — Creating A Gold-Standard Corpus for Situative Geo-Coding of Crisis Tweets Based on Linked Open Data

Andrea Salfinger*, Caroline Salfinger*, Birgit Pröll†, Werner Retschitzegger*, Wieland Schwinger*

*Dept. of Cooperative Information Systems, †Inst. for Application Oriented Knowledge Processing,
Johannes Kepler University Linz,
Altenbergerstr. 69,
4040 Linz,
Austria
*{andrea.salfinger, caroline.salfinger, werner.retschitzegger, wieland.schwinger}@cis.jku.at, †bproell@faw.jku.at

Abstract

Crisis management systems would benefit from exploiting human observations of disaster sites shared in near-real time via microblogs, however, utterly require *location information* in order to make use of these. Whereas the popularity of microblogging services, such as Twitter, is on the rise, the percentage of GPS-stamped Twitter microblog articles (i.e., tweets) is stagnating. Geo-coding techniques, which extract location information from text, represent a promising means to overcome this limitation. However, whereas geo-coding of news articles represents a well-studied area, the brevity, informal nature and lack of context encountered in tweets introduces novel challenges on their geo-coding. Few efforts so far have been devoted to analyzing the different types of geographical information users mention in tweets, and the challenges of geo-coding these in the light of omitted context by exploiting situative information. To overcome this limitation, we propose a *gold-standard* corpus building approach for evaluating such situative geo-coding, and contribute a human-curated, geo-referenced tweet corpus covering a real-world crisis event, suited for benchmarking of geo-coding tools. We demonstrate how incorporating a semantically rich Linked Open Data resource facilitates the analysis of types and prevalence of geo-spatial information encountered in crisis-related tweets, thereby highlighting directions for further research.

Keywords: Corpus Building, Geo-parsing, Geo-coding, Toponym Resolution, Social Media

1. Introduction

Social Media for Crisis Management. Nowadays, timely situational update information on crisis events can frequently be retrieved from social media, such as eyewitness reports of disaster sites shared via microblogging[1] (Olteanu and others, 2015). Whereas initial prototypes have already demonstrated the potential of incorporating social media data, such as Twitter microblog articles (i.e., tweets), in emergency management systems (cf. surveys in (Imran and others, 2015; Salfinger and others, 2015a)), these utterly depend on *location information*, as emergency managers and first responders ultimately need to know *where* their assistance is required. Therefore, the actual sparsity of GPS-tagged tweets (Gelernter and Mushegian, 2011; Imran and others, 2015; Schulz and others, 2013) represents a major bottleneck for exploiting social media data for crisis management, requiring additional means to utilizing GPS tags to obtain essential location information. Apart from locations actually associated with the tweet's author (the user's current location and location profile (Ikawa and others, 2013)), location information can also be extracted from its *textual content*.

Extracting Location Information. However, the detection of place names in such free-form text (i.e., *toponym recognition* or *geo-parsing*), and the mapping of a place name to its corresponding geographic location by assigning appropriate coordinates (i.e., *disambiguation*, *toponym resolution*, *geo-coding* or *(spatial) grounding*) (Leidner, 2007; Martins and others, 2005), represents a non-trivial task due to both, *geo-/non-geo-ambiguity*[2], as well as *geo-/geo-ambiguity*[3] (Amitay and others, 2004). Whereas traditional news articles frequently allow to resolve these ambiguities, since these provide their reader with contextual information required to understand the situation described therein, the ultimate brevity and real-time nature of tweets introduce unprecedented challenges on their geo-coding: Tweets are typically written in informal, localized language and expose specific characteristics — for instance, location information may also be obscured in multi-word hashtags (e. g., “#Hawaiihurricane”, “#bigislandoutage”). To meet the imposed length limitations[4], tweets frequently lack discourse context, as humans tend to omit contextual information that is shared between the correspondents[5]. Thus, this *context deletion* represents a severe obstacle for an automated extraction of (otherwise valuable) situational update information from social media. Consequently, georeferencing of tweets needs to stretch beyond conventional topoynm recognition and resolution, as developed for news prose (e.g., (C. D'Ignazio and others, 2014; Leidner and Lieberman, 2011; Lieberman, 2012; Quercini and others, 2010; Samet and others, 2014)) and longer web documents

[1] Examples of microblogging platforms include Twitter (www.twitter.com), Tumblr (www.tumblr.com) and Sina Weibo (www.weibo.com).

[2] For example, “Jordan” may refer to a basketball player or a country.

[3] For example, “Sydney” may refer to a city in Australia or Canada.

[4] Up to 140 characters per tweet.

[5] For instance, Vieweg et al. could identify several tweets where people simply referred to “the river” when actually meaning the “Red river” in their studies of tweets on the Colorado flooding events in 2009 (Vieweg and others, 2010).

(such as Wikipedia or online news, e.g., (Amitay and others, 2004; Woodward and others, 2010)).

Disambiguation Context. Although geo-coding approaches fine-tuned towards the characteristics of tweets have been developed (Flatow and others, 2015; Gelernter and Balaji, 2013; Ikawa and others, 2013; Karimzadeh and others, 2013; Schulz and others, 2013), current approaches provide limited account for this *context deletion*: During the development of our social media-sensing Situation Awareness system for crisis management (Pröll and others, 2013; Salfinger and others, 2015b; Salfinger et al., 2016a; Salfinger et al., 2016b), we encountered many tweets that were not appropriately resolved by presently available geo-coding tools. From our empirical observations, we noted that such toponym resolution errors frequently could be attributed to the common error of not incorporating sufficient *context* for toponym disambiguation, which can be classified into the following two context classes: (i) "in-tweet-context", i.e., unambiguous toponym disambiguation *within* a single tweet is possible based on the joint context of all location mentions occurring in this tweet, and, (ii) "between-tweets-context" or "situative context", which refers to the event-level context of the monitored scenario - i.e., the associated event-context would allow to derive valuable disambiguation cues guiding toponym resolution, as proposed in (Salfinger et al., 2016b).

Ground-truth Data Sets. In order to examine and systematically study the challenges of toponym disambiguation, however, a *ground-truth data set* would be required, which reflects the way a human monitoring the crisis scenario would resolve encountered location descriptions by incorporating contextual reasoning. Although valuable work on corpus-building of geo-parsed and/or geo-referenced tweet corpora have been undertaken (which we will review in Sec. 2.), these mainly focus on general toponym recognition aspects, such as identification of proper place names (Wallgrün and others, 2014), or detection of locative expressions without considering the mapping of these to real-world locations (Liu and others, 2014). Little focus so far has been on studying toponym disambiguation problems, especially from the social media-specific *context deletion* perspective. Therefore, we set to systematize and share our experiences by creating a human-curated *ground-truth* data set suited to study such geo-coding challenges from a crisis management perspective.

Linked Open Data. However, the creation of such a shareable evaluation data set for toponym resolution tasks is complicated by the inter-dependency between the employed geographical reference frame, i.e., the topographical information used to determine the mapping from textual entities to geographical space, and the resulting geo-referenced corpus. Thus, corpora created with different geographical reference frames may not be directly comparable to each other (e.g., due to different toponym resolution granularity) (Leidner, 2006). Recently, however, the growth in Linked Open Data (LOD) initiatives provides a remedy towards this problem: Geographical ontologies, such as GeoNames[6], represent a semantically rich, comprehensive and global-coverage source of geographical knowledge, providing an extensive basis for geographical reference, and tend to become the de-facto standard for geographical reference sources utilized in geo-parsing tools (Wallgrün and others, 2014).

Contributions. Therefore, we introduce a gold-standard corpus building methodology involving publicly available annotation tools and LOD to create shareable language resources (LRs) for studying situative toponym resolution, and report on the resulting corpus building initiative. We propose an event-driven corpus sampling strategy to allow for incorporating situative context, an annotation schema involving a LOD resource which also comprises annotation types for assessing implicitly specified geographical information, describe the developed annotation process, and contribute the resulting human-curated, geo-referenced gold standard tweet corpus on a specific crisis event for benchmarking and training of geo-coding techniques. We further outline how the semantic richness of the employed LOD resource benefits the analysis of the resulting corpus, by examining the types and prevalence of geo-spatial information encountered in this corpus from a crisis management perspective. We specifically also assess *implicit* and qualifying geo-spatial information to outline which potentially valuable spatial cues could be exploited for crisis management applications, but which remain unused by presently available geo-coding tools, thereby indicating directions for further research. This is further underpinned by a comparative evaluation of state-of-the-art geo-parsing tools on this data set, which highlights current performance limitations. We hope that our proposed methodology encourages similar initiatives in creating sharable LRs supporting the analysis of situative geo-coding.

Structure of the Paper. In the next section, we compare our approach to related endeavors on gold standard corpus building for geo-coding purposes. In Sec. 3., we describe the set-up of our collaborative annotation project, before analyzing the resulting *gold standard corpus* in Sec. 4., and concluding our lessons learned in Sec. 5.

2. Related Work

In this section, we explain how our gold standard corpus creation extends valuable findings reported in other work. We first assess related tweet corpora, before discussing more widely related work on news corpora.

Social Media. Gelernter and Mushegian describe the building of a geo-annotated tweet corpus on the 2011 earthquake in Christchurch, New Zealand (Gelernter and Mushegian, 2011), thus, focusing on a specific crisis event, as in our study. Whereas they defined a location upon a diverse set of types (such as countries, buildings, street addresses) and also incorporated hashtags and abbreviations, as well as generic places, i.e., non-proper place names (e. g., "city", "house", "home"), they did not devise a specific annotation scheme for discriminating these types in order to study the distribution of encountered types, as in our approach. They neither did include place names being part of multi-word tokens, which we included to examine the frequency of place names encountered in multi-word hashtags.

Wallgrün et al. employed a crowd-sourcing approach to

[6] http://www.geonames.org

Table 1: Comparison of highly related approaches. Abbreviations: ? = not stated, K = 1000

	Annotations				Corpus Characteristics		
Approach	Toponym Recognition	Locative Expressions	Toponym Resolution	Employed Geographical Gazetteer	Event-specific	Annotators/Message	Volume
(Gelernter and Mushegian, 2011)	✓	?	?	—	✓	3	1.4K
(Liu and others, 2014)	✗	✓	✗	—	✗	3	1K
(Wallgrün and others, 2014)	✓	✗	planned	GeoNames	✗	5	6K
this work	✓	✓	✓	GeoNames	✓	3	4K

create a geo-annotated tweet corpus, and provided an extensive discussion of encountered annotator errors (Wallgrün and others, 2014). 6K tweets have been annotated for identified place names in a crowd-sourcing project on the Amazon Mechanical Turk platform, which, as opposed to our approach, were not confined to a specific event, but sampled according to different criteria. In the present work, we base upon their findings by incorporating their characterization of annotator errors into the definition of our annotation schema. Furthermore, Wallgrün et al. proposed to employ the GeoNames ontology for toponym resolution, which they planned to address in future work. Following their suggestion, our annotation schema thus encodes manually resolved toponyms by their Geonames identifiers (IDs). However, whereas Wallgrün et al. solely focused on proper place names, our annotation schema also involves annotation types for resolving implicitly stated geo-spatial information, since we also aimed at detecting implicit or vague spatial information in order to quantify the proportions and types of implicit information encountered in crisis-related tweets.

Liu et al. focused on the annotation of Locative Expressions (LEs) on corpora of different web document types (e.g., Twitter, Blogs, Youtube comments) (Liu and others, 2014), i.e., any expressions referring to a location (such as “in my cozy room”, “at home” or “around the city”). Their manually annotated corpora provided the basis for comparing Precision, Recall and F-score of six different geoparsers. Their focus, however, has been on entity recognition, i.e., identifying the text chunks comprising LEs, not on their actual geo-coding (i.e., mapping to geographic coordinates).

The data sets for evaluating the geo-coding techniques proposed in (Flatow and others, 2015; Schulz and others, 2013) use the GPS locations of the user’s device as geo-reference. However, the user’s current location may be disparate from the *focused location* (Ikawa and others, 2013) of the tweet, i.e., the location the user writes about, which is actually the location of interest in our crisis management application domain.

News Articles. Extensive studies on toponym resolution have been conducted by J.L. Leidner, however, with a focus on news prose (Leidner, 2006; Leidner, 2007). The two human-curated gold standard datasets created in the course of this work (Leidner, 2006) therefore consist of news articles obtained from the REUTERS Corpus Volume I.

Since toponym recognition and resolution also represent core algorithmic tasks for Geographical Information Retrieval Systems (Geo-IR), the need for standardized evaluation procedures and appropriate benchmarking data sets also led to corpus building efforts in this research domain (cf. (Martins and others, 2005) for an overview), however, with a focus on newswire texts (e. g., Geo-IR evaluation tracks GeoCLEF[7] 2005, 2006, 2007 and 2008).

3. Methodology

In the present section, we describe our gold-standard corpus building methodology, for which we followed the best practice guidelines on collaborative annotation projects suggested in (Sabou and others, 2014).

Scenario. Due to our application domain of crisis management, we pursued an event-driven approach for corpus sampling, by assembling a corpus characterizing a specific real-world crisis event. Our initial tweet corpus has been retrieved with the aim of monitoring the effects of hurricanes Iselle and Julio on the Hawai’ian islands, in August 2014[8]. We recorded tweets matching keywords associated with that crisis[9] from the public Twitter Stream[10], yielding roughly 212 600 tweets collected between August the 9th to 21st, 2014. This event-driven approach allows us to study the challenges of geo-locating tweets within a real-world crisis context, as opposed to open-domain geo-coded corpora, such as created in (Wallgrün and others, 2014). We can thus specifically examine whether the studied tweets also contain *context-sensitive* geo-spatial information, i.e., information which cannot be interpreted if the general event context is lacking, thus making it impossible even for human annotators to understand. The selected data set conforms to a highly-localized event, involving small-scale locations on the Hawai’ian islands. Hawai’i furthermore proves to be challenging with respect to (w.r.t.) toponym resolution, since it comprises many ambiguous locations (i. e., multiple locations with the same name exist on different islands, e. g., Wailea, Wailua) which need to be properly resolved to aid crisis management tasks.

Corpus Sampling. In order to optimize the allocation of human work force, we designed a dedicated data preprocessing protocol to narrow down the data set to a manageable yet representative proportion of the collected tweets, and eliminate near-duplicate tweets. We restricted our data set to English-language tweets (provided by the tweet’s language tag) in the time range between Aug., 9th - 16th, 2014, resulting into 137K tweets, 83K of those were actually textually distinct. In the first step, *background knowledge* regarding the monitored events was employed in or-

[7] http://ir.shef.ac.uk/geoclef

[8] www.latimes.com/nation/nationnow/la-na-nn-hawaii-storm-iselle-juliio-20140808-story.html

[9] tracked by a filter query leaving language and location deliberately unspecified and the following keywords: Hurricane, #HurricaneIselle, #HurricanePrep, #hiwx, #HIGov, Iselle, #updatehurricaneiselle, #Genevieve, #Iselle, #Julio

[10] Twitter Streaming API: https://dev.twitter.com/streaming/public

der to reduce the data set to presumably disaster-relevant tweets. This is due to the fact that keyword-based queries frequently return tweets not related to the disaster, i.e., in which the corresponding term is used in a different context (e. g., "#Mystery, #Romance #Humor a #Hurricane' what more could you want! http://t.co/13el5JKptR @rpdahlke #Bargain 99"). We also included several *location* terms, a-priori known to be crisis-relevant in the chosen scenario, in this initial filtering[11] to guarantee a high number of geo-referenceable tweets. Furthermore, we wanted to investigate user-generated content (i. e., tweets ideally written by human on-site observers), as opposed to the plethora of tweets containing news headlines, which refer to news articles and external web sources, since our major goal was studying the characteristics of social media and not - unintentionally - examining news prose. According to our empirical analysis, tweets referring to and advertising external content (e.g., consisting of news headlines and a URL to the corresponding news agency) tend to correlate with specific Twitter clients[12] (e. g., IFTTT and Hootsuite, a social media marketing tool for enterprises). We therefore filtered the tweets on Twitter clients that, according to our experience, more likely contain original content[13]. By focusing on content sent from mobile devices, we thus seek to increase the proportion of content published by end-users with a non-commercial focus. We further noted that even after filtering on textual distinctness, in a semantic sense, many duplicates remain. This is attributable to the different URLs generated from URL shorteners, which are commonly employed on Twitter to meet the strict length limitations. Thus, we receive many duplicates in terms of slightly modified and "broadcast" message content, such as "Pound of prevention' pays off for Hilo Medical Center during Iselle http://t.co/HSl9pX9JEI #hawaii" and "Pound of prevention' pays off for Hilo Medical Center during Iselle: Hilo Medical Center had to switch to gen... http://t.co/HHXodZT0O0". We therefore aimed at eliminating the effect of shortened URLs by replacing them with a specific token. Upon this URL-coding, we could discard tweets which have a too low string distance to other tweets, by using the `stringr` and `stringdist` R packages to filter out textually highly similar tweets (M.P.J. van der Loo, 2014).

Annotation Schema. We provided our annotators with a dedicated annotation schema for marking explicit and implicit spatio-temporal information encountered in tweets, and geo-referencing this information based on a semantically-rich LOD resource, the GeoNames ontology[14]. By linking to the corresponding ontology instances, we are not only able to unambiguously refer to a specific toponym and retrieve its geographic coordinates, but can also examine additionally provided geographic meta-data, such as a toponym's administrative division (allowing to discriminate coarse-grained — such as country-level — from fine-grained information, such as districts and villages). A screenshot showing the resulting annotation editor dialog is shown in Fig. 1. We incorporated the findings presented in (Wallgrün and others, 2014) into the definition of this annotation schema, which comprises the following annotation types:

◇ *Proper Place Name (PPN)* for marking named location entities, such as the names of populated places (i.e., countries, cities, etc.) or other geographical features (i. e., mountains, islands, etc.). This annotation type also includes several additional annotation features that should be specified, such as a free text field titled GeonamesID. By looking up recognized place names on the Geonames search interface, annotators should manually perform toponym resolution by identifying the appropriate location candidate from the resulting Geonames toponyms list, and enter its corresponding Geonames ID, which uniquely identifies this location. Since Geonames also allows a map-based inspection of its retrieved toponyms, annotators were encouraged to carefully analyze and disambiguate results. Furthermore, annotators were required to specify whether a location name is part of a single-word or multi-word hashtag (annotation feature "hashtag complete", e. g., "#hawaii", or "hashtag partial", e. g., "#bigislandoutage", respectively), whether it is used attributively (e. g., "One week later: This is how Hawaii Island residents have to live after #Iselle."), its name is specified informally (e. g., "#Iselle about to make landfall on Big Hawaiian Island."), or abbreviated (e. g., "Many still wo power in Puna District on Big Is."). Misspelled place names should be annotated (e. g., "Hawai"), but following (Gelernter and Mushegian, 2011; Wallgrün and others, 2014), we explicitly excluded place names part of an organization name (e. g., "Hawaiian Airlines") or a Twitter user handle (i.e., "mention", e. g., "@akeakamaihawaii"). Annotators should mark each occurrence of a PPN, even if specified multiple times in the same tweet.

◇ *Point of Interest (POI)* corresponds to distinctive locations that cannot be found on Geonames, but are known to a greater audience (e. g., "Iselle Relief: Plate lunches Available at Nanawale Community Center Today"), thus, mostly denote specific buildings or well-known spots.

◇ *Place Qualifier (PQ)* corresponds to a locative expression which further spatially restricts a given location (e. g., south California or upper Manhattan). Since, for crisis management applications, we are interested in the most fine-grained locative description possible, we are thus interested in examining such spatial restrictions, which would require spatial reasoning capabilities to be appropriately geo-coded in an automated fashion.

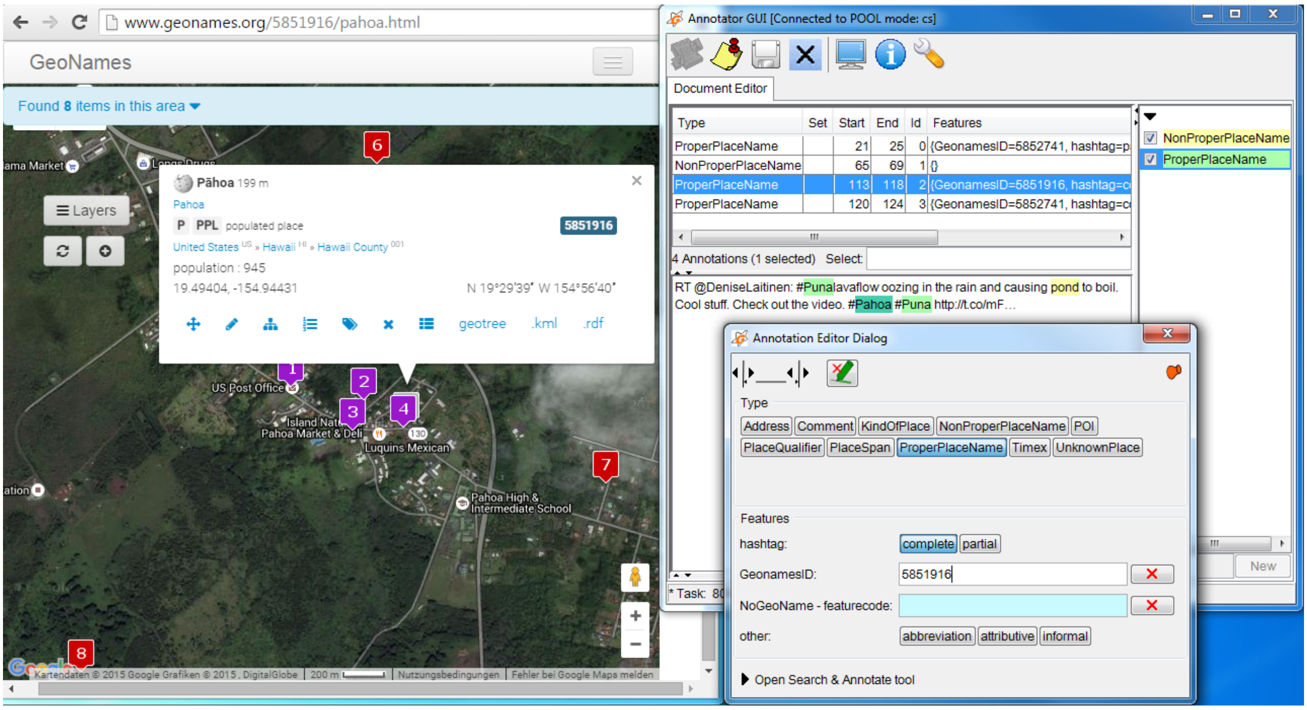
◇ *Non-proper Place Name (NPPN)* denotes general spatial

[11] Notably, filtering on the following terms: "Hawaii", "Pahoa", "Puna", "Kona", "Hilo", "iselle", "honolulu", "oahu", "maui", "kauai", "BigIslandOutage", "big island", "#HIwx", "HELCO"

[12] The Twitter client the tweet has been sent with can be retrieved from the tweet's meta-data.

[13] Twitter for iPad (http://twitter.com/#!/download/ipad), Twitter for iPhone (http://twitter.com/download/iphone), OS X (http://www.apple.com/), Twitter for Windows Phone (http://www.twitter.com), Twitter Web Client (http://twitter.com), Facebook (http://www.facebook.com/twitter), TweetDeck (https://about.twitter.com/products/tweetdeck), Twitter for Android (http://twitter.com/download/android), Twitter for Android Tablets (https://twitter.com/download/android), Instagram (http://instagram.com), Instagram (http://instagram.com), Mobile Web (M2) (https://mobile.twitter.com)

[14] www.geonames.org



www.geonames.org/5851916/pahoa.html

GeoNames

Found 8 items in this area

Pāhoa 199 m

Pahoa

P PPL populated place 5851916

United States US » Hawaii HI » Hawaii County 001

population : 945

19.49404, -154.94431 N 19°29'39" W 154°56'40"

geotree
.kml
.rdf
Layers
6
1
2
3
4
7
8
ama Market
Longs Drugs
US Post Office
Island Nat
Pahoa Market & Deli
130
Luquins Mexican
ation
Pahoa High & Intermediate School
Kartendaten © 2015 Google Grafiken © 2015 , DigitalGlobe | 200 m | Nutzungsbedingungen | Fehler bei Google Maps melden

Annotator GUI [Connected to POOL mode: cs]

Document Editor

Type	Set	Start	End	Id	Features
ProperPlaceName		21	25	0	{GeonamesID=5852741, hashtag=p
NonProperPlaceName		65	69	1	{}
ProperPlaceName		113	118	2	{GeonamesID=5851916, hashtag=c
ProperPlaceName		120	124	3	{GeonamesID=5852741, hashtag=c

☑ NonProperPlaceName
☑ ProperPlaceName

4 Annotations (1 selected) Select:

RT @DeniseLaitinen: #Punalavaflow oozing in the rain and causing pond to boil. Cool stuff. Check out the video. #Pahoa #Puna http://t.co/mF…

Annotation Editor Dialog

Type: Address | Comment | KindOfPlace | NonProperPlaceName | POI | PlaceQualifier | PlaceSpan | [✓ ProperPlaceName] | Timex | UnknownPlace

Features

hashtag: [✓ complete] partial

GeonamesID: 5851916

NoGeoName - featurecode:

other: abbreviation | attributive | informal

Open Search & Annotate tool

Figure 1: Screenshot showing the annotation task view.

descriptors (e. g., “south shore”, “islands”), which lack an explicit identifier. However, these frequently correspond to the omission of discourse context encountered in tweets, and actually refer to specific locations known by the communicating people (cf. “the river” denoting the “Red river” example in Sec. 1.). Therefore, we demanded the annotation of such NPPNs, in order to examine whether the referred-to location could be inferred given the external *situative context* of the tweet (i.e., the general event context). Thus, if our annotators could infer the actual location due to their human intelligence and provided event context, the feature “Reference to Geoname ID” allowed to add the corresponding GeoNames ID. Thus, this annotation scheme provides a means to study geo-coding techniques for NPPNs which operate on techniques exploiting the situative context, as proposed in (Salfinger et al., 2016b), which sets our approach apart to other approaches discussed in Sec. 2..

◇ *Kind of Place (KOP)* represents a redundant description to a given PPN (e.g., “city of London”). As discussed in (Wallgrün and others, 2014), these are frequent causes for annotator disagreement, as it is difficult to discriminate in which cases a KOP expression is actually a part of the PPN itself. As Wallgrün et al. discussed, this is more frequently the case for physical features, such as mountains and lakes (e. g., “Lake Michigan”, for which “lake” is a part of the PPN, in order to discriminate it from the state of Michigan), than for populated places (e. g., regarding “city of London”, city is redundant information). In our guidelines, we were following the notion of (Wallgrün and others, 2014), i.e., KOP only corresponds to a separate annotation if it is clearly redundant, otherwise (e. g., if part of the PPN given on Geonames), the text should be included in the PPN annotation. Although we presented examples in the annotation guidelines to clarify on this, we still received many annotator disagreements, who often misconceived this annotation with others (such as NPPN and PQ).

◇ *Address (ADR)* comprises separate annotation features for marking street names, postcodes and house numbers.

◇ *Place Span (PSP)* represents a meta-annotation for marking a location span (e. g., “from Hilo to Pahoa”), in order to assess their frequencies.

◇ *Timex (TMX)* for annotating temporal expressions, which may be of sub-type “date”, “time”, or “other”, if the previous two do not apply.

◇ *Comment (COM)* Annotators also were given the possibility of attaching their own, free-form remarks.

Project Execution. Regarding the technical setup, we employed the collaborative annotation platform GATE Teamware (Bontcheva and others, 2013), which provides a client-server-architecture for managing the set-up and distribution of collaborative annotation tasks. Annotators used a web interface to retrieve their assigned annotation tasks, which were executed using a GATE-based annotation editing interface (cf. Fig. 1) in conjunction with the GeoNames web interface. Regarding the assignment of annotation tasks, the sampled tweet corpus (in total 4 117 tweets) has been partitioned across twelve human annotators, who were conducting these tasks in the course of a summer internship at our institution (age range 15 - 19, two females, ten males). Our annotators had an educational background at high school level and were non-native English speakers, but have been learning English for at least five years, therefore, had a solid language level[15]. In an initial preparation meeting, our annotators have been presented with the required

[15] Place name identification has been recognized as relatively easy annotation task also for non-local users (Gelernter and Mushegian, 2011).

background information regarding the events covered in the data sets, i.e., have been given a summary on the key events, in order to understand the scope and situative context of the assigned tweets. We feel the option of realizing such introductory meetings presents an advantage of such lab-study based annotation projects over the use of online crowd-sourcing platforms, as it allows the introduction of more complex annotation tasks by first providing the annotators with essential background knowledge. We also sought to address the frequently reported unfamiliarity problem (Gelernter and Mushegian, 2011; Wallgrün and others, 2014) (i. e., non-local annotators may overlook place names since they do not know the corresponding text fragment represents a location name, due to their unfamiliarity with the corresponding location), by pointing the annotators towards relevant geographic characteristics, locations and their popular abbreviations of the event site Hawai'i, thereby increasing their geographic awareness. Furthermore, annotators were given a set of guidelines regarding the devised annotation schema, involving detailed screenshots and examples. The annotation tasks were introduced by means of a small *pilot study*, in which annotators could get acquainted with the annotation interface (shown in Fig. 1) by experimenting with example tasks, and were encouraged to ask questions, before we assigned the actual annotations tasks. We demanded at least three annotators per tweet, thus corresponding to a total annotation effort of 12351 tweets. To avoid introducing any group bias, we split the entire data set into several batches, and permuted group composition of the three annotators allocated per batch across the different batches.

Data Evaluation. We assessed the annotators' agreement using GATE's Inter-Annotator Agreement plugin, measuring Precision, Recall and F1 in a strict sense. Whereas we received good Inter-Annotator Agreement for PPNs (F1: mean: 0.74, standard deviation: 0.07, measured strictly and incorporating equality of the specified Geonames ID), and acceptable results regarding the TMX (F1: mean 0.46, standard deviation: 0.05)), the agreement regarding the other annotation types was insufficient[16]. Thus, we can confirm the findings presented in (Wallgrün and others, 2014), who classified these annotation types into the most common cause of annotator errors, and furthermore, can show that even when provided with an annotation schema and guidelines addressing these error sources (as suggested by Wallgrün et al.), humans face difficulties in reaching an agreement w.r.t. the corresponding type. It also appears that human annotators have mainly focused on the detection of PPNs, as all other type have been frequently overlooked, which may explain the fact that other work solely focusing on the annotation of temporal expressions reported higher F1 scores.

Corpus Delivery. For ultimately aggregating the different annotators' mark-ups to the final gold standard data set, the following steps were performed: First, a majority voting component copied these annotations to the *consensus set*, if a majority of annotators agreed strictly (for which we required that the annotation span was equal, i.e., not overlapping, and *all* annotation features were equal). Second, one of the authors performed manual adjudication of the remaining annotations, using the GATE Developer Tool: By comparing the annotators' opinions using the annotation stack tool, the adjudication manager resolved conflicting cases by copying the correct annotation to the consensus set, annotating overlooked entities or merging differing annotations.

4. Discussion

In the present section, we outline how the semantic richness of the employed LOD resource enables a fine-grained analysis of the resulting corpus, by providing additional metadata allowing for a faceted analysis.

4.1. Characteristics of the Resulting Corpus

The resulting geo-referenced tweet corpus is publicly available for research purposes[17], in the widely used GATE document XML serialization format[18]. We furthermore also provide lists of the encountered annotated texts and their frequencies of identified PQs (mostly corresponding to orientation relations, such as cardinal directions), NPPNs and POIs, as well as the proposed annotation schema.

Finding the Needle in the Haystack. For 99% of PPNs, a corresponding GeoNames toponym could be identified, yielding in total 244 unique identified GeoNames references. However, whereas this may, at first sight, seem to benefit applications such as crisis management, an inspection of the most frequent toponyms in Tab. 2 also highlights disguised challenges: The - by far most frequent - toponym refers to the entire state of Hawai'i, which clearly is expected. For crisis management applications, however, this information is of limited use, as a more detailed localization of affected areas - such as severely hit cities and villages - would be required, which we indeed encounter on rank 3, 6 and 10 (Pahoa and its surrounding Puna District have been damaged the by hurricane). Thus, the *granularity* of provided spatial information (in terms of their corresponding administrative division — e.g., state-level information versus city-level information) should ideally be attributed with corresponding weights, rewarding highly localized information (e.g., Pahoa) with higher priority fur further processing than area-/country-level information (e.g., State of Hawai'i). However, this also induces the challenges on how to track such information on Twitter, which in times of such crisis is flooded by corresponding news headlines from all over the world, which, however, mostly contain coarse-grained information (e.g., that Hawai'i is threatened by a hurricane), but provide limited value for actual crisis management tasks.

Need for Hashtag Decomposition. 13% of PPNs are obscured in multi-word hashtags, thereby requiring geo-parsers capable of extracting the toponym chunks from these.

[16] F1, mean: NPPN: 0.10, PQ: 0.12, PSP: 0.44, POI: 0.15, KOP: 0.06

[17] https://weizenbaum.tk.jku.at/owncloud/public.php?service=files&t=6076c0c9b7f3e03fc6204b1607a8b0e1

[18] Including the final, adjudicated gold-standard annotations, the annotations of each individual annotator, and the results obtained with analyzed tools, for reasons of reproducibility.

Table 2: Most frequent Toponyms.

Rank	Place Name	Geo Names ID	Freq.
1	State of Hawai'i	5855797	1996
2	Island of Hawai'i	5855799	482
3	Puna District	5852741	412
4	Maui County	5850871	152
5	O'ahu	5851609	134
6	Hilo	5855927	99
7	Kaui County	5848514	74
8	Honolulu	5856195	67
9	Hawaiian Islands	5855811	66
10	Pahoa	5851916	57

Table 3: Annotation Type Distribution.

Anno. Type	Total	Freq.	Feature	Total	Freq.
PPN	4177	67%	GeoN. ID	4155	99%
			Hashtag	1300	31%
			- complete	771	18%
			- partial	529	13%
			other	959	23%
TMX	1113	18%			
NNPN	500	8%	Ref. to Geo. ID	62	12%
PQ	202	3%			
POI	165	3%			
ADR	25	0%			
KOP	23	0%			
PSP	9	0%			

Low Frequencies of Other Spatial Information. Regarding annotation types other than PPN and NNPN, we observe low frequencies in this dataset. The rare occurrences of KOP annotations may be attributable to the length restrictions imposed by Twitter, as the limit of 140 characters per tweet probably forces users to eliminate redundant information such as KOP. However, the scarcity of qualifying spatial information (PQ), fine-grained spatial information such as POIs and ADR (which will most likely be provided by local users familiar with the geographical situation), and specification of place spans, demands further investigation. Intuitively, one would expect these types of spatial information to correlate with the provision of fine-grained situational update information (e.g., which areas may be severely affected, at which addresses shelters would be provided etc.). Therefore, further studies involving different crisis datasets would be required to analyze whether the observed low frequencies are attributable to the sampling strategy employed in the generation of the current corpus, or these annotation types are indeed generally rarely observed in crisis-specific data sets.

Need for Situative Context-Aware Toponym Resolution. Discriminating the most frequent toponyms, i.e., rank 1 and 2 in Tab. 2, represents a major challenge, since both are typically referred to by the text "Hawai'i" in the tweet, but correspond to different toponyms and spatial granularity: Rank 1 comprises the entire group of islands, whereas rank 2 solely denotes the largest Hawaiian island, making a key difference for crisis management purposes. Therefore, toponym resolution techniques capable of reasoning on the current situative context to extract the adequate toponym are required.

4.2. Benchmarking of Geoparsers

Ultimately, the most interesting question is how well existing geo-referencing tools perform on this ground-truth data set. We thus examined the performance of advanced state-of-the-art systems (C. D'Ignazio and others, 2014), notably CLAVIN-NERD[19], and GeoTxt[20] (Karimzadeh and others, 2013), cf. Tab. 4, both capable of resolving toponyms based on the GeoNames ontology. Both tools are built for recognizing and resolving PPN annotations only, therefore, the following experiments solely evaluate their performance on detecting and resolving PPN annotations. Since CLAVIN-NERD does not provide support for parsing hashtags, we thus preprocessed the tweet texts by replacing "#" tokens with blanks, which should — at least — enable it to resolve single-word hashtags accordingly. Following (Martins and others, 2005), we provide a separate evaluation of *toponym recognition* and *toponym resolution*, to pinpoint performance lacks to the corresponding phase. Whereas these tools yield high Precision, Recall is below 50%, thus, the majority of geo-spatial information actually contained in tweets (from a human's perspective) remains unused.

Toponym Resolution Errors. We furthermore analyzed the most frequently incorrectly disambiguated toponyms, cf. Tab. 6. As assumed in Sec. 4.1., the resolution of small-scale locations tends to be problematic, which, however, is crucial for application domains such as crisis management. To examine this assumption, we conducted another experiment to separately evaluate the tools' performance on such small-scale locations, by excluding annotations corresponding to a populated place of an administrative division 1 and 2, as provided by the Geonames ontology, which indeed yields a severe drop in Recall (cf. Tab. 4). A closer analysis of the mapped locations suggests that incorporating a geo-spatial reasoning aware of the situative context could potentially improve toponym resolution, as several of the toponyms selected by these tools are located highly disparate from the actual event location (notably, are located even at different continents).

5. Conclusions and Lessons Learned

In the present work, we contributed a geo-referenced, manually curated tweet corpus, described the employed corpus building methodology, and provided an analysis of the resulting corpus. We examined the availability and prevalence of geospatial information in tweets from the requirements perspective of a crisis management application, thereby identifying several research challenges for future work. Our evaluation of state-of-the-art geo-parsing

[19] https://clavin.bericotechnologies.com,
https://github.com/Berico-Technologies/CLAVIN-NERD

[20] http://www.geotxt.org

Table 4: Corpus statistics, **A** = gold standard data set, **B** = results obtained with the geo-referencing tool listed in the left-most column.

Tool (**B**)	Match (Correct)	Only **A** (Missing)	Only **B** (Spurious)	Overlap (Partial)	Prec. **B/A**	Rec. **B/A**	F1	Match (Correct)	Only **A** (Missing)	Only **B** (Spurious)	Overlap (Partial)	Prec. **B/A**	Rec. **B/A**	F1
	All PPN annotations.							**Small-scale PPN annotations.**						
	Toponym Recognition — F1.0-score strict on PPN annotations, without considering Geonames ID.													
CLAVIN-NERD	1963	2052	221	162	**0.84**	0.47	0.60	618	1235	175	58	**0.73**	0.32	**0.45**
GeoTxt Stanford h	2234	1700	441	243	0.77	**0.53**	**0.63**	640	1162	397	109	0.56	**0.33**	0.42
GeoTxt Gate h	no results retrieved													
	Toponym Resolution — F1.0-score strict on PPN annotations, incorporating Geonames ID.													
CLAVIN-NERD	1727	2431	600	19	**0.74**	0.41	0.53	408	1499	439	4	**0.48**	0.21	**0.30**
GeoTxt Stanford h	2023	2136	877	18	0.69	**0.48**	**0.57**	453	1452	687	6	0.40	**0.24**	**0.30**

Table 5: Toponym Resolution errors, GS = gold standard data set, F. = Frequency.

GS	Clavin-Nerd	F.
Island of Hawai'i (5855799), HI, US	Big Island (4747418), Virginia, US	90
Island of Hawai'i (5855799), HI, US	Republic of Estonia (453733)	27
Puna District (5852741), HI, US	Pune, India (1259229)	20
Kailua-Kona (5847504), HI, US	Cona, Italy (3178217)	11
Island of Hawai'i (5855799), HI, US	Hawaii, FL, US (6463769)	7

Table 6: Toponym Resolution errors, GS = gold standard data set, F. = (Total) Frequency.

GS	GeoTxt Stanford h	F.
Puna District (5852741), HI, US	Pune, India (1259229)	64
Pacific Ocean (2363254), HI, US	Pacific, MO, US (4402300)	28
Island of Hawai'i (5855799), HI, US	Big Island (4747418), Virginia, US	18
Kailua-Kona (5847504), HI, US	Cona, Italy (3178217)	13
State of Hawai'i (5855797), HI, US	Hawaii, FL, US (6463769)	6

tools' performance on our gold standard corpus revealed that further research on tackling the specifics of tweets is utterly needed, as current tools provide unsatisfactory Recall, especially regarding small-scale locations. Thus, only a fraction of geo-spatial information can be used at the moment, hindering valuable use cases for Twitter data, such as benefiting crisis management. Since Recall can only be measured given a comprehensive ground truth data set, we therefore hope that the contribution of our gold standard corpus may aid in the development of effective location entity recognition and geo-coding techniques for tweets.

Naturally, our current gold standard corpus is limited in terms of generalizability, since only a single crisis event is covered and we only incorporated English-language tweets. For future work, we thus seek to extend our corpus building endeavor towards other crisis events and languages, allowing to further examine potential country- or language-specific characteristics in social media usage. By devising and describing a corpus building methodology involving publicly available annotation tools and LOD resources, we hope to encourage other research groups to join these efforts in creating shareable, inter-operable LRs for studying situative geo-coding, similarly to related efforts for collaboratively created ground truth data sets for examining social media characteristics across crises, such as the extensive CrisisLex26 data set for informativeness classification of crisis-related tweets (Olteanu and others, 2015).

6. Acknowledgements

This work has been funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant FFG BRIDGE 838526, FFG TALENTE 850859, 852071, 850855 and 850594, and WTZ AR10/2015. We further thank Paula Maria Fischer, Theresa Gierlinger, Josef Haider, Philipp Hofer, Paul Huber, Leo Jungmann, Felix Kastner, Erik Reichl, Lorenz Reichl, Lukas Riegler, Maximilian Schartner and Martin Stoiber for performing the annotation tasks.

7. Bibliographical References

Amitay, E. et al. (2004). Web-a-where: Geotagging Web Content. In *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 273–280. ACM.

Bontcheva, K. et al. (2013). GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

C. D'Ignazio et al. (2014). CLIFF-CLAVIN: Determining geographic focus for news. In *NewsKDD: Data Science for News Publishing*.

Flatow, D. et al. (2015). On the Accuracy of Hyper-local Geotagging of Social Media Content. In *Proc. of the Eighth ACM Int. Conf. on Web Search and Data Mining*, WSDM '15, pages 127–136. ACM.

Gelernter, J. and Balaji, S. (2013). An Algorithm for Local Geoparsing of Microtext. *Geoinformatica*, 17(4):635–667.

Gelernter, J. and Mushegian, N. (2011). Geo-parsing Messages from Microtext. *Transactions in GIS*, 15(6):753–773.

Ikawa, Y. et al. (2013). Location-based Insights from the Social Web. In *Proc. of the 22Nd Int. Conf. on World Wide Web Companion*, WWW '13 Companion, pages

1013–1016. Int. World Wide Web Conferences Steering Committee.

Imran, M. et al. (2015). Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys*, 47(4).

Karimzadeh, M. et al. (2013). GeoTxt: A Web API to Leverage Place References in Text. In *Proc. of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pages 72–73. ACM.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, 3(2):5–11.

Leidner, J. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417.

Leidner, J. L. (2007). *Toponym resolution in text*. Ph.D. thesis.

Lieberman, M. D. (2012). *Multifaceted Geotagging for Streaming News*. Ph.D. thesis, College Park, MD, USA.

Liu, F. et al. (2014). Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis. In *Proc. of the 4th Int. Workshop on Location and the Web*, pages 9–16. ACM.

Martins, B. et al. (2005). Challenges and Resources for Evaluating Geographical IR. In *Proc. of the 2005 Workshop on Geographic Information Retrieval*, GIR '05, pages 65–69. ACM.

M.P.J. van der Loo. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1):111–122.

Olteanu, A. et al. (2015). What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In *Proc. of the 18th ACM Conf. on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 994–1009. ACM.

Pröll, B. et al. (2013). crowdSA - Crowdsourced Situation Awareness for Crisis Management. In *Proc. of Social Media and Semantic Technologies in Emergency Response (SMERST)*.

Quercini, G. et al. (2010). Determining the Spatial Reader Scopes of News Sources Using Local Lexicons. In *Proc. of the 18th SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, GIS '10, pages 43–52. ACM.

Sabou, M. et al. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proc. of the Ninth Int. Conf. on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 859–866.

Salfinger, A. et al. (2015a). Crowd-Sensing Meets Situation Awareness - A Research Roadmap for Crisis Management. In *Proc. of the 48th Annual Hawaii Intl. Conf. on System Sciences (HICSS-48)*.

Salfinger, A. et al. (2015b). crowdSA - Towards Adaptive and Situation-Driven Crowd-Sensing for Disaster Situation Awareness. In *Proc. of IEEE Int. Multi-Disciplinary Conf. on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA 2015)*.

Salfinger, A., Retschitzegger, W., Schwinger, W., and Pröll, B. (2016a). Towards a Crowd-Sensing Enhanced Situation Awareness System for Crisis Management. In Galina L. Rogova et al., editors, *Fusion Methodologies in Crisis Management*, pages 177–211. Springer International Publishing.

Salfinger, A., Schwinger, W., Retschitzegger, W., and Pröll, B. (2016b). Mining the Disaster Hotspots - Situation-Adaptive Crowd Knowledge Extraction for Crisis Management. In *Proceedings of the 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 219–225. IEEE.

Samet, H. et al. (2014). Reading News with Maps by Exploiting Spatial Synonyms. *Commun. ACM*, 57(10):64–77.

Schulz, A. et al. (2013). A Multi-Indicator Approach for Geolocalization of Tweets. In *Int. AAAI Conf. on Weblogs and Social Media*.

Vieweg, S. et al. (2010). Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '10, pages 1079–1088. ACM.

Wallgrün, J. O. et al. (2014). Construction and First Analysis of a Corpus for the Evaluation and Training of Microblog/Twitter Geoparsers. In *Proc. of the 8th Workshop on Geographic Information Retrieval*, GIR '14, pages 4:1–4:8. ACM.

Woodward, D. et al. (2010). A Comparison of Approaches for Geospatial Entity Extraction from Wikipedia. In *Semantic Computing (ICSC), 2010 IEEE Fourth Int. Conf. on*, pages 402–407.