GeoCorpora: Building a Corpus to Test and Train Microblog Geoparsers

Jan Oliver Wallgrün[a*], Morteza Karimzadeh[b*], Alan M. MacEachren[b*], and Scott Pezanowski[b*]

[a] *Independent Researcher Affiliated with GeoVISTA Center and ChoroPhronesis (Pennsylvania State University), Forsthof Hagen 4, 22926 Ahrensburg, Germany;* [b] *Department of Geography, GeoVISTA Center, Pennsylvania State University, USA*

(*Received 00 Month 200x; final version received 00 Month 200x*)

In this article, we present the GeoCorpora corpus building framework and software tools as well as a geo-annotated Twitter corpus built with these tools to foster research and development in the areas of microblog/Twitter geoparsing and geographic information retrieval. The developed framework employs crowdsourcing and geovisual analytics to support the construction of large corpora of text in which the mentioned location entities are identified and geolocated to toponyms in existing geographical gazetteers. We describe how the approach has been applied to build a corpus of geo-annotated tweets that will be made freely available to the research community alongside this article to support the evaluation, comparison, and training of geoparsers. Additionally, we report lessons learned related to corpus construction for geoparsing as well as insights about the notions of place and natural spatial language that we derive from application of the framework to building this corpus.

Keywords: Geoparsing, Corpus Building, Microblogs, Twitter, Geo-Annotation, Named Entity Recognition

1. Introduction

Location matters, and not just for real estate. The rapid proliferation of GPS-enabled devices and other positioning sensors has dramatically increased our ability to understand the world and make decision in all contexts from business, to leisure, to crisis management (Crooks *et al.* 2012, Tsou *et al.* 2013, Chae *et al.* 2014). But, there is still probably much more geographic information available in text, in the form of place references, than

[*] Corresponding authors. Email: {wallgrun,karimzadeh,maceachren,scottpez}@psu.edu

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

from sensing devices that use GPS and other means to generate location data directly. Leveraging that information resource is an ongoing research challenge that has been addressed over the past decade or more by research and development in Geographic Information Retrieval (GIR) (Jones *et al.* 2002).

One key aspect of GIR is the recognition of references to places in unstructured text and subsequent geolocation of these references. This task, often termed geoparsing, typically involves the application of advanced Natural Language Processing (NLP) techniques. Identification of place entities is often based on available tools for Named Entity Recognition (NER) (Hu and Ge 2007, Woodward *et al.* 2010). Once entities are recognized, various disambiguation (toponym resolution) strategies are employed to determine the referenced locations (Gelernter and Balaji 2013, Hu and Ge 2007, Speriosu and Baldridge 2013), e.g. through various strategies to select the most likely candidate toponym from a geographic gazetteer of georeferenced entities.

Deciding whether a phrase refers to a location entity or not is a non-trivial problem, even for humans[1]. For example, Alonso *et al.* (2013) report on research directed to the challenge of regular polysemy (also referred to as metonymy; figure-of-speech usage in which a term is used to refer to something that the term is closely associated with) in relation to determining the semantic type of a word/phrase (that words can switch between semantic types depending on context of use). They focus on location, noting that location names can often behave as a member of a different semantic type (e.g., an organization or a nationality). One example that they provide is "England was being kept busy with other concerns." In this case, the sense is (probably) that of an organization (a political entity) rather than a place (location). Using a comparison between laypeople and experts, they demonstrate that there are situations in which the sense of a place word can be underspecified and not possible to determine reliably; experts are more adept at making this judgement while laypeople are more likely to pick one sense or the other.

Geoparsing has traditionally been applied to the content of web pages, news stories, or other types of unstructured longer textual documents. However, the drastically increased importance of microblog data in areas such as (social) sciences, policy, and humanitarian relief (Ghosh and Guha 2013, Blanford *et al.* 2014, Starbird *et al.* 2014, Tapia *et al.* 2011) has produced efforts to extend geoparsing methods to microblog and other more cryptic social media posts (Gelernter and Balaji 2013). Since Twitter added the capability to add a location to their tweets, substantial research attention has been given to leveraging that information to support a variety of applications including situational awareness for crisis events (Crooks *et al.* 2012) and mapping of social activities related to elections and other events (Tsou *et al.* 2013). However, being able to geoparse and interpret the tweet text itself (and location-related information users encode in their profile) opens up a potentially much larger source of place-related information: While only about 1-2% of tweets include geolocation (Leetaru *et al.* 2013), 10% or more include references to places in the text (MacEachren *et al.* 2011), and our own analysis has found that approximately 53% of Tweeters include some identifiable place related information in the location field of their profile(MacEachren *et al.* 2013). The focus on geolocated tweets, thus, misses a much larger source of place-relevant information.

However, microblog text comes with its own set of particular challenges for geoparsing tools due to the lack of direct context, the presence of abbreviations, special language and notations, etc. (see Gelernter and Balaji 2013). A number of geoparsers and related software tools and services exist for identifying and geolocating place names in different

[1] The work reported on in this article has only confirmed this assessment (see discussions in Section 3).

types of unstructured text (Ahlers and Boll (2008), Lieberman and Samet (2012), Katz and Schill (2013), Yosef *et al.* (2011), see also D'Ignazio *et al.* (2014)), and in microblog or Twitter messages in particular (Gelernter and Balaji 2013, Karimzadeh *et al.* 2013, Dredze *et al.* 2013). However, their performance has not yet reached a satisfactory level and assessment and comparison of different tools is a major challenge. For example, in research taking a machine learning approach, Speriosu and Baldridge (2013) starting with recognized entities and focusing only on the toponym disambiguation component of the problem achieved 85% accuracy with their best combination of methods. In complementary research, Moncla *et al.* (2014) focus on the whole process of both recognizing and geolocating toponyms, but limit the text in their experiments to hiking descriptions. They report results of only 55% of toponyms being located correctly using rule-based methods without a spatial inference component and an increase to 84% with spatial inference methods included. While this improvement is impressive, 84% is far from acceptable in many applications, and the focus documents (while having the challenge of many place names that are not found in standard gazetteers) are in many ways easier than text in general since, as the authors report for these hiking documents, "almost all named entities are spatial entities." One study, by Gelernter and Balaji (2013) reports 96% accuracy for coding locations in tweets. While the results are impressive and the appendices of the article provide valuable clues to details of their approach, the authors constrain their problem to tweets about events in two narrowly defined locations (the 2011 earthquake in Christchurch, NZ and the 2011 wildfire in Austin,TX). Thus, the task they judge their approach against is much simpler than that of leveraging location from the full Twitter firehose. Additionally, the corpus is not public, thus there is no opportunity for results to be replicated. Also, there are some issues with the "gold standard" Gelernter et al. developed to test their algorithms against; the issues relate to some of our corpus building findings reported below. Based on the description of the small sample of annotations from the corpus that is provided in the article (Gelernter and Balaji (2013), section 4.2), the "locations" that they test their algorithms against are not all actually locations; they include a variety of attributive nouns (a noun that modifies another noun). Of the 16 examples they provide from their corpus, 25% are attributive nouns where the entity is either a person, object, or organization (e.g., New Zealand News Service). Anotherl entity is a metonym (" 'Christchurch' welcomes you"); they acknowledge this metonym, but not the attributive nouns. One other feature that is common to all published results of geoparsing tools we have seen so far is a failure to discuss the existence of ambiguous place names that overlap one another (e.g., "Gaza" which is used to indicate both Gaza Strip and the city of Gaza, with the one the author intended sometimes being impossible to determine due to insufficient context).

We contend that a key obstacle for progress on geoparsing and GIR in general is the current lack of adequate publicly available ground-truth corpora of geoparsed text to facilitate the evaluation, comparison, and training of geoparsing methods and systems. Similar efforts in other areas of (spatial) language processing, such as the GeoCLEF activities (Mandl *et al.* 2009), have led to major advances in text processing tools. While the testing of toponym resolution algorithms based on corpora of news stories has been discussed in the literature (Leidner 2006, 2007, Katz and Schill 2013) and a few corpora aiming at the NER part only do exist (such as the corpus of 2400 manually annotated tweets used by Ritter *et al.* (2011) to demonstrate that their tool outperforms Stanfords CRF NER on tweets) as well as other general corpora related to human spatial language (see, for instance, Stock *et al.* 2013, Fort *et al.* 2009), the work we present in this article is to the best of our knowledge the first attempt to address this gap with a focus on microblogs and the goal of providing a publicly available benchmarking corpus and an analysis of the insights about place references in language gained in the process of corpus construction.

In this article, we present a research initiative we call GeoCorpora in which we develop strategies, a framework, and tools for geospatial corpus construction. We describe the GeoCorpora tool set and how it has been employed to build a ground-truth *geo-microblog corpus* of Twitter posts with manually marked and geolocated location entities that will be made available to the research community along with the publication of this article. The GeoCorpora tools and corpus have been redesigned and rebuilt based on lessons learned from previous corpus building work (Wallgrün *et al.* 2014) and the adoption of a geovisual analytics approach to building a geo-microblog corpus. Our approach leverages crowdsourcing by employing an Amazon Mechanical Turk visual interface to annotate location entities in event-based microblog posts, and expert-guided, visual-computational methods to facilitate geographic disambiguation (toponym resolution) and geolocation of place names (geo-annotation). The corpus consists of 6711 tweets. Approximately ∼30% of these contain at least one place name. Mentions of location entities are annotated with a unique ID and geographic coordinates stemming from the GeoNames geographical gazetteer[1]. To deal with the incompleteness of GeoNames, we employ an iterative process for dealing with places that are not currently in GeoNames: Initially, we locate the place to a 'representative' nearby location, adding a special 'Not-in-Geonames' tag to these locations in our corpus to distinguish them from those with 'true' locations. Then, assuming the feature is robust enough to justify its addition, we determine a location and add the place to GeoNames making it available to the wider community.

The constructed corpus enables researchers and developers working on (microblog) geoparsing to (a) train their geoparsing systems (both NER and toponym disambiguation), (b) evaluate the performance of their algorithms, and (c) quantitatively compare the performance of geoparsing approaches based on a benchmarking dataset. Beyond these applications, the corpus is also a valuable source for linguistic and cognitive studies, for instance, for analyzing and gaining an understanding of how people conceptualize place and place names, or for comparing the results and performance of experts to that of crowd-sourced laypersons. We adopt Baker's argument that "... corpora allow researchers to find evidence of rare or unusual cases of language, as well as shedding light on very frequent phenomena." (Baker 2010). Given the lack of existing knowledge about the kinds and relative frequencies of geoparsing challenges and errors, it was essential for us to avoid making assumptions on the kinds of challenges and errors, and instead, use the corpus building process to generate that knowledge as a primary product of our research. In our analysis of the crowdsourcing results (reported on below), we identified interesting causes of disagreement and gained novel insights regarding the use of place names as attributive nouns (noun adjuncts), metonymy, and occurrences of ambiguous references with spatially overlapping toponym candidates. We expect these results to have a significant impact on future corpus building work as well as automated geoparser algorithm design.

The remainder of this article is organized as follows: In Section 2, we describe the underlying corpus building framework and the design decisions made in its development. In Section 3, we present an analysis of the results from the crowdsourcing-based place identification and geo-annotation components of our framework, look at reasons for dis-

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

agreement between AMT workers and insights gained from the discussions between the experts, and discuss how we addressed difficult cases in our current corpus. In Section 4, we close the article with concluding remarks and an outlook on future work.

2. GeoCorpora Corpus Building Framework

The aim of the framework and GeoCorpora software tools presented in this section is to build corpora of tweets (and other text sources) in which all mentions of location entities in the textual content have been marked, and each such place name is resolved to one (or more) toponym(s) from GeoNames. Using GeoNames for identifying toponyms allows for a unique ID for each toponym as well as compatibility with different geoparsing tools (e.g., Katz and Schill 2013, Gelernter and Zhang 2013, Karimzadeh *et al.* 2013), as well as with the linked-data paradigm since all GeoNames toponyms have a unique URL with a corresponding RDF web service[1]. An additional advantage of using GeoNames is that it contains vernacular, historic, and other forms of alternate names for its toponyms.

In the concrete corpus of tweets discussed in this article, place names are annotated with (a) a unique GeoNames ID to identify it, (b) its coordinates, and (c) descriptive tags to indicate special kinds of place names and place name candidates as discussed further in Sections 2.4 and 3.3. Coordinates are given as WGS 1984 latitudes and longitudes. These coordinates also stem from GeoNames and typically represent the centroids or the seat of a unit, such as the capital of a country / administrative division of the respective place. For instance, in the tweet

> *“Where are the best #steakhouses in New York City? Find out: http://t.co/KJLKJL #nyc”*

that refers to New York City twice, once as ”New York City” and once as ”nyc”, each occurrence would be geo-annotated with the following information: (1) toponym: New York City; (2) GeoNames ID: 5128581; (3) latitude: 40.71427, longitude: -74.00597; (4) no special tags. It is worth noting that GeoNames keeps getting improved and updated by official sources and volunteers, and some toponyms such as countries already have polygon representations as well that can be accessed via the GeoNames ID.

The GeoCorpora corpus building framework and software tool set employ two visual interfaces designed to support two main steps: (a) crowdsourcing-based place name identification and (b) expert-based geocoding of those place names (referred to as *geo-annotation* in this article to set it apart from other meanings of the term geocoding) by resolving them to GeoNames toponyms. For the first, since determining whether any word or phrase refers to a place is a subjective judgement by speakers of the language about meaning of statements in context, we use Amazon Mechanical Turk to generate a majority vote among multiple English speakers about whether a word or phrase is intended as a reference to a place. For locating the place entities identified, the individual generating the reference is assumed to be referring to a single specific location; thus the task is an objective one of locating the intended place. For this, we utilize geographers with expertise in using maps and toponym databases along with various components of context (that include profile locations of tweeters, other places mentioned in their tweets or linked documents, and more) to determine the intended location. To minimize errors, we use a collaborative system that requires pairs of geography experts to come to a consensus on which specific location was intended. The architecture of our framework is

[1] http://www.geonames.org/ontology/documentation.html

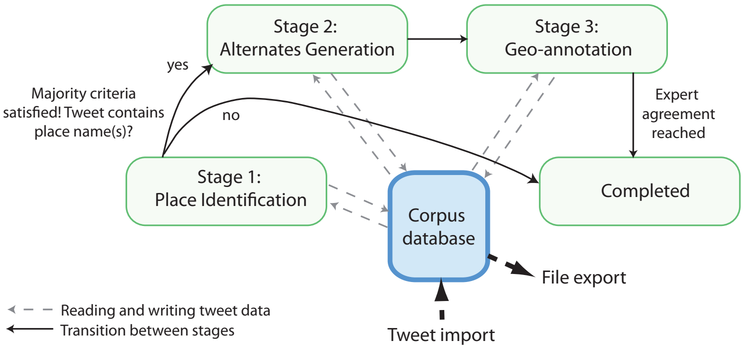


Figure 1. Architecture: In Stage 1, place names in the text are marked by AMT workers. For each identified name, a list of potential toponyms is generated in Stage 2. In Stage 3, an interactive visual interface is used to geo-annotate the place names by picking the correct toponym.

shown in Figure 1: A central database stores the tweets (or other text documents) making up the corpus with all required additional information. Each tweet passes through different stages (unprocessed → place identification → generation of GeoNames toponym candidates (which we call alternates) that the place name may refer to → geo-annotation → processing completed). The main components including the three main phases (stages 1-3) of corpus generation are further explained in the following.

2.1. *Tweet Collection and Sampling*

For the current corpus of tweets, a temporally stratified random sample of 6000 tweets was drawn from approximately 1 billion tweets we collected in 2014 and 2015 by querying for particular keywords: 700 tweets were drawn for each of 12 terms, *earthquake, ebola, fire, flood, flu, malaria, measles, protest, rebels, riot, tornado, violence*, chosen from three kinds of events representing (a) natural disasters, (b) infectious disease, and (c) human threats/violence. Our goal is not to create a representative sample of tweets but rather a sample that has a higher number of place names than a completely random sample because the corpus is primarily meant for training and evaluating geoparsers. While it is guaranteed that a selected tweet contains at least one of the 12 keywords, the keyword may in some cases occur in a different meaning than the intended event-related one. However, identifying such cases would be difficult and potentially even subjective, and more importantly, is not required for our goal of achieving a higher number of place names within the sample. Temporal stratification is used to reduce the chance that multiple tweets in the sample are about the same event and, thus, mention the same places.

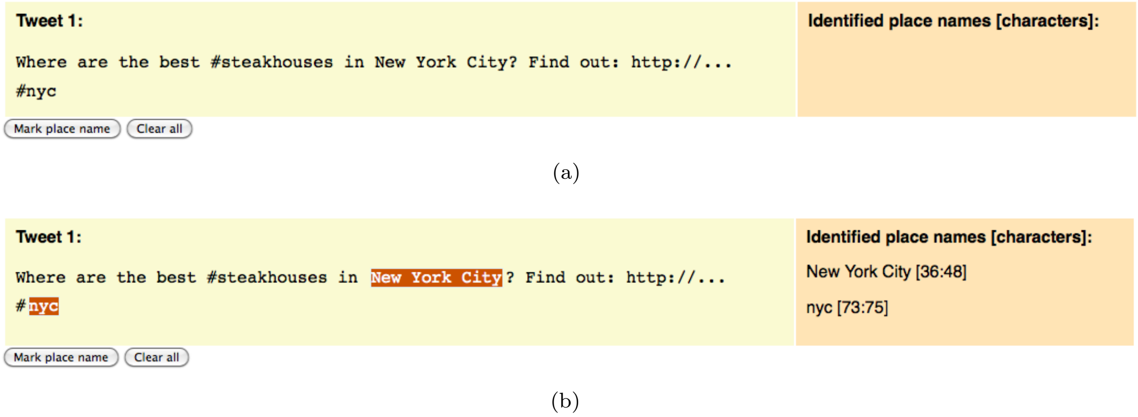
Of the 700 tweets per keyword, we then created an initial set of 500 tweets per keyword, making sure that only tweets that were primarily in English were selected and that no two selected tweets were the same or very similar (e.g., one is a re-tweet of another) using a minimal normalized string edit distance of 0.45 as a criterion. We imported the resulting 6000 tweets into our central corpus database, making them available for Stage 1, the place identification step. After gathering first experiences with a predecessor of our current corpus building framework (Wallgrün *et al.* 2014), we realized that 711 tweets of our 6000 had in the meantime been deleted on Twitter by their respective authors. We added replacements for these tweets to our corpus database by drawing additional ones from the 700 tweets per keyword selected as described above. As a result, the corpus database at the time of this writing contains 6711 tweets. To be in agreement with Twitter's privacy policies, the published version of the corpus will only contain tweet IDs and our annotations, but not the tweet text itself. We will provide software tools to easily acquire the tweet text for a given tweet ID. The original text of tweets that have been deleted in the meantime will, however, not be accessible anymore; as a result, the number of tweets usable for evaluation and training in the corpus will be at or below 6000 and may decrease over time. Furthermore, there are, of course, no guarantees about the longevity of any technology company. Thus, we cannot guarantee that our corpus will be available permanently. However, as far as we know, it will be the only one that is publicly available at this point of publication and, hence, it should prove to be a valuable resource for at least near term research on microblog geoparsing.

2.2. *Stage 1: Crowdsourcing-based Place Identification*

To be able to build a large corpus, our framework employs a crowdsourcing approach for the place identification stage, that uses Amazon's Mechanical Turk (AMT) platform[1]. AMT enables the creation and distribution of small tasks, referred to as Human Intelligence Tasks (HITs), over a large base of workers, handling recruitment and payment and allowing for collecting a large number of results quickly and efficiently. Over the past 5-7 years, crowdsourcing has been used increasingly to annotate corpora and for corpus building in general. Sabou *et al.* (2014) review 13 of these studies and use the analysis of the existing work to propose best practice guidelines for crowdsouring of text annotation designed to create large-scale, high quality linguistic resources. We model our approach on a strategy reported in Potthast (2010) in which a 2/3 agreement among annotators was required to accept the result. Otherwise, additional annotators were added. In contrast to the geo-annotation stage, the place identification stage is particularly well-suited for crowdsourcing because (a) the number of tweets that need to be processed is much higher than can be coped with by a small number of experts (those ∼70% that do not contain at least one place name will not enter the geo-annotation stage), (b) less geographic knowledge and interaction between the annotators is required than for the geo-annotation task, and (c) the question of what constitutes a place reference and what does not is a difficult one for which no strict rules can be established; hence, having the place identification stage based on many laypeoples' opinions allows for identifying and taking into account prototypical and boundary cases of place references (MacEachren 2014).

To adopt the AMT platform for our purposes, we designed a web interface allowing an AMT worker to tag a tweet by marking the place names in it. After accepting one of our HITs, the worker is presented with a web page with detailed instructions and two tweets to tag. The main part of the web interface in which a tweet has to be tagged is illustrated in Figure 2, before and after marking the place names. Workers select the character sequence they want to mark as a place name and then press a button, at which point the selection highlighting will become fixed. Identified place names are also listed in the box to the right of the tweet. Our GUI employs a direct manipulation style (Shneiderman 1982) utilizing selection and highlighting rather than using radio buttons or checkboxes, which constitute the default UI elements available in AMT. It minimizes worker errors and increases their efficiency. For each tweet individually as well as for the entire HIT, workers are also able to leave comments in corresponding text fields.

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



(a)

Tweet 1:

Where are the best #steakhouses in New York City? Find out: http://...
#nyc

Identified place names [characters]:

(b)

Tweet 1:

Where are the best #steakhouses in New York City? Find out: http://...
#nyc

Identified place names [characters]:

New York City [36:48]

nyc [73:75]

Figure 2. Place identification web interface in AMT. (a): The tweet before tagging. (b): The tweet after occurring place names have been marked by highlighted them.

The web page includes an animated image to illustrate the highlighting process, instructions to highlight multi-word names (e.g., New York) together and names separated by a comma as in “Boston, MA” separately, and to not include spaces or punctuation before or after a name. Appendix A of the Supplementary Online Materials lists these and additional instructions exactly as they appear on the web page presented to the AMT workers. The final place identification results in the corpus are based on the AMT output plus adjustments developed collaboratively by the experts during geo-annotation. This will be further discussed in Section 3.3.

We paid AMT workers 4 cents per assignment consisting of two tweets which resulted in the overall cost of approximately $700 to collect at least 5 AMT results for each tweet (roughly 17000 AMT assignments). We utilized AMT’s capabilities to only allow for workers with an acceptance rate of $\geq 95\%$ from previous HITs at the outset, later increased to 97%, with the goal of achieving a high quality of the collected results.

Worker results were compared by a submodule of our framework to see whether the following two agreement criteria were satisfied: (1) At least 5 results from different workers exist for the tweet; (2) for each character in the tweet, there is at least a 70% majority agreement on whether that character is part of a place name or not, and if so, which characters the entire place name consists of. If the criteria were satisfied, a ground-truth tagging reflecting the majority vote over all results was created and stored. This majority vote was taken for each place name occurring in the tweets textual content individually. In the case that the ground-truth place identification result for a tweet did not contain any place names, processing of this tweet was concluded. Otherwise, the tweet entered Stage 2, the alternate generation.

Our corpus collection software automatically generated additional HITs for tweets when the criterion of at least 70% agreement was not satisfied after collecting at least 5 results. However, creation of new HITs was stopped if the agreement did not increase substantially with the additional results. Remaining disagreements were then inspected and discussed by the authors and ultimately either resolved manually or submitted to the GeoAnnotator tool (described below) for analysis by pairs of experts (experts included three of the authors plus two senior Geography majors) as part of the geo-annotation work from the 3rd stage. This will be further described in Section 2.4.

2.3. *Stage 2: Alternates Generation*

Since in the geo-annotation stage (see Section 2.4) coders were supposed to choose the correct toponym for each place name in a tweet on a map interface, we produced an initial list of alternative toponyms with GeoNames ID and WGS 1984 coordinates for each identified place name. The software module for this purpose used the relevance ranking and geocoding component from our own geoparser, GeoTxt (Karimzadeh *et al.* 2013), but bypassed the NER part of that tool. Essentially, the module queried a local Apache Solr index originally created from a GeoNames data export and produced up to 100 most likely toponyms (in decreasing order of likelihood) for a given place name as ranked by our algorithm based on name similarity, population, geographic prominence and custom GeoNames feature code ranking. The lists were stored along with that tweet in the corpus database and the tweet was moved on to the geo-annotation stage.

2.4. *Stage 3: Geo-Annotation*

As discussed, we refer to the process of manually resolving the previously identified place names to the corresponding toponyms with geographic coordinates as *geo-annotation*. Our software package for this stage of the corpus building process is called GeoAnnotator.

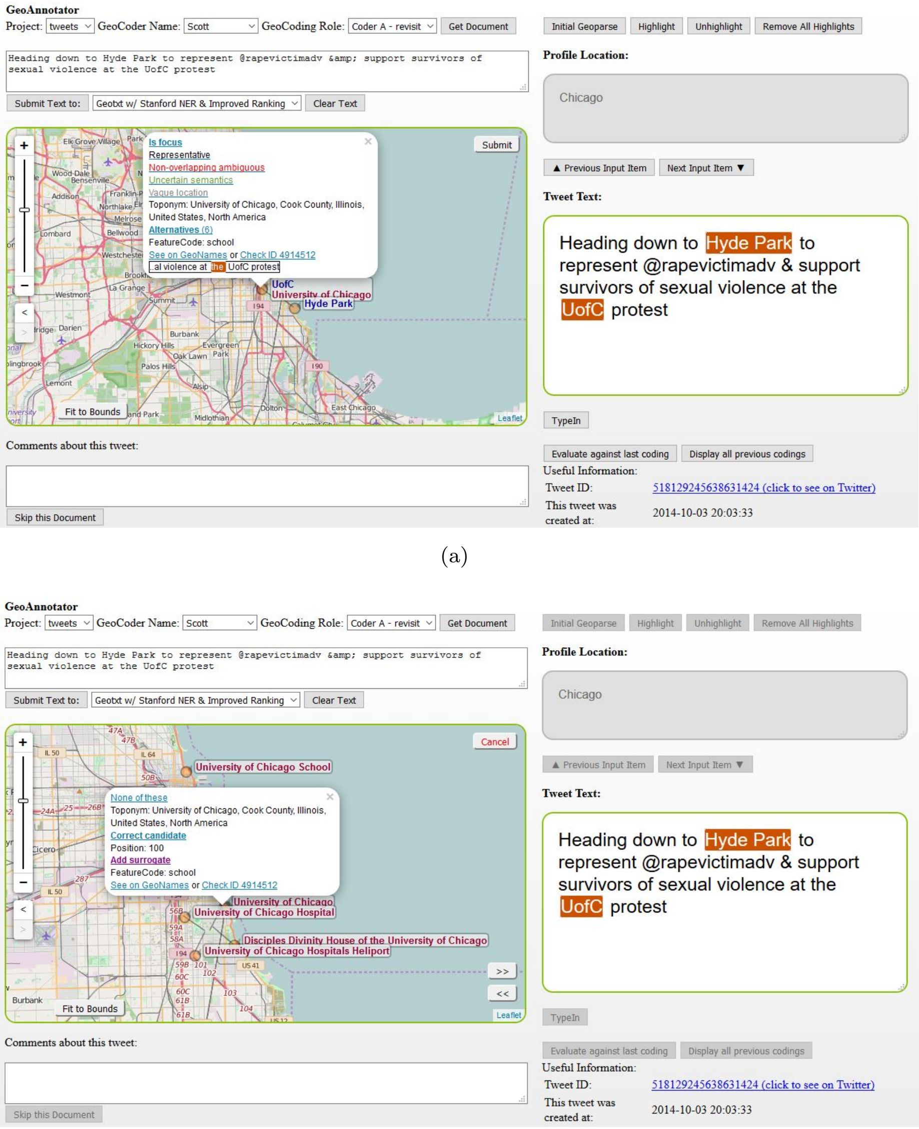
We had four reasons for using experts for this stage: (1) Deciding whether a token is a place name requires a higher number of votes because the concept of a place name can be considered to be a radial category with easily identified exemplars but also many boundary cases (MacEachren 2014), as also witnessed by the disagreement categories in Section 3.2; whereas in geo-annotation, there is almost always a place that the speaker (or writer) meant when naming it. (2) Only tweets that have a place name in them or for which there is unresolved disagreement between the AMT workers (32% of the initial sample according to the AMT results) would reach this stage, making this approach using individuals with expertise in geography feasible. (3) Geo-annotation is a harder task compared to recognizing place names in that it requires either local knowledge of the place (Leidner 2006) or general geographical knowledge and the willingness to spend more time researching the existence and location of such a place as well as its most likely toponym candidate, typically using different web sources and putting together fragments of information. (4) In case of disagreement in geo-annotation between annotators, we wanted them to be able to communicate, collectively make decisions, and establish guidelines for similar cases, instead of always having the case assessed by a new person without establishing a fundamental understanding of the reasons leading to the disagreement.

The last point above proved to be particularly important in relation to some of the results about references to place that are reported below. Specifically, the communication among expert geo-annotators provided many insights on the challenges of geographic disambiguation, including insights on what kinds of ambiguity are typical in microblog (or other kinds of text) references to place (see Section 3.3). Also, we used this conversational mechanism to improve GeoAnnotator's capabilities over time. Annotators had the option to "skip" a tweet that they found philosophically challenging to annotate or problematic because of technical shortcomings. We used lists of these skipped tweets along with group discussions to improve GeoAnnotator and fed skipped tweets to GeoAnnotator again. The collaborative, expert-based geo-annotation process had two important outcomes that can benefit future place corpus building. First, as toponyms were disambiguated, rules for place name identification were refined; these rules should lead to more robust crowd-sourced place entity recognition in the future. Second, the iterative refinement of GeoAnnotator to deal with various kinds of ambiguity in place names and with lack of some place names in GeoNames has resulted in a more robust visual analytics tool for this part of the process. While we still contend that the geo-annotation part of place name corpus building cannot be done reliably by generic crowdsourcing through AMT, it may be practical to develop a citizen science process for geo-annotation in which committed volunteers are willing to invest the time needed to solve the cases that result from underspecification in context and/or from abbreviations, alternative names, and other issues we encountered that are explained in Section 3.3.

The GeoAnnotator interface enabled the annotators to see the results from the place identification stage, to confirm or edit them as needed (editing was needed when AMT workers had insufficient agreement, when their agreement was counter to the rules provided to them, or when updating rules or guidelines during the expert geo-annotation step; see Sections 2.2, 3.2 and 3.3), and to locate the corresponding places on an interactive map. The interface also allows for adding tags to highlighted place references to indicate special cases that we further explain in Section 3.3. GeoAnnotator's interface also displays some useful metadata for each tweet that may help annotators develop a better understanding of the tweet context. These metadata include links to the original tweet on Twitter, timestamp for when the tweet was generated, link to the tweeter's profile, as well as the description that the tweeter had listed on their profile.

The interactive map component of GeoAnnotator's user interface renders all the places mentioned and identified in the tweet. For each, the automatically determined toponym with highest relevance (according to GeoTxt as described in Section 2.3) is provided as the default choice for each place name pinpointed on the map (Figure 3(a)). In the case that an automatically selected toponym is not the correct one, the annotator is able to click on it to see the alternate candidates for that place name and then choose the correct one, if it exists in GeoNames. Annotators are also able to query for additional alternative candidates if the correct candidate is not already listed (Figure 3(b)).

Each tweet was assigned to two experts. If there was disagreement between their annotations, the tweet was sent back to both annotators one by one in order, giving them a chance to reconsider their annotation and to leave some explanatory comments on the annotation describing their reasoning. When a tweet with geo-annotation disagreement is returned to one of the original annotators, the annotator can see their own as well as the other annotator's annotations in the form of highlights in text and toponyms on maps in addition to the comments they have left on each submission (Figure B1 in Appendix B of the Supplementary Online Materials shows an example of this). This enables the conversation between the experts about the reasoning for their choices. This conversational mechanism helped us establish consistent guidelines on geo-annotating tweets. The special tags explained in Section 3.3 were established during face-to-face meetings in which the experts discussed remaining disagreements. Using GeoAnnotator, 2185 tweets were successfully geo-annotated in more than 250 person hours of work. Sixty-three of these tweets turned out to not include a place name in the tweet text, but were passed to the geo-annotation stage because of AMT worker disagreement. Thus, the outcome of coding was 2122 tweets out of 6711 containing one or more annotated places.



GeoAnnotator

Project: tweets GeoCoder Name: Scott GeoCoding Role: Coder A - revisit

Heading down to Hyde Park to represent @rapevictimadv &amp; support survivors of sexual violence at the UofC protest

Submit Text to: Geotxt w/ Stanford NER & Improved Ranking

Is focus
Representative
Non-overlapping ambiguous
Uncertain semantics
Vague location
Toponym: University of Chicago, Cook County, Illinois, United States, North America
Alternatives (6)
FeatureCode: school
See on GeoNames or Check ID 4914512
al violence at the UofC protest

Profile Location:

Chicago

Tweet Text:

Heading down to Hyde Park to represent @rapevictimadv & support survivors of sexual violence at the UofC protest

Comments about this tweet:

Useful Information:

Tweet ID: 518129245638631424 (click to see on Twitter)

This tweet was created at: 2014-10-03 20:03:33

(a)

GeoAnnotator

Project: tweets GeoCoder Name: Scott GeoCoding Role: Coder A - revisit

Heading down to Hyde Park to represent @rapevictimadv &amp; support survivors of sexual violence at the UofC protest

Submit Text to: Geotxt w/ Stanford NER & Improved Ranking

None of these
Toponym: University of Chicago, Cook County, Illinois, United States, North America
Correct candidate
Position: 100
Add surrogate
FeatureCode: school
See on GeoNames or Check ID 4914512

Profile Location:

Chicago

Tweet Text:

Heading down to Hyde Park to represent @rapevictimadv & support survivors of sexual violence at the UofC protest

Comments about this tweet:

Useful Information:

Tweet ID: 518129245638631424 (click to see on Twitter)

This tweet was created at: 2014-10-03 20:03:33

(b)

Figure 3. Interactive geocoding interface. (a) Right panel: places highlighted in the tweet profile location and tweet text. Left panel: Initial situation with automatically assigned toponyms. (b) Alternative toponyms rendered allowing the annotator to select the correct one.

3. Analysis of the Corpus and Corpus Building Work

In this section, we provide an analysis of the properties of the corpus we built and discuss results and insights gained from the corpus building work. As expected, we encountered numerous problematic cases of potential place names in the tweets that lead to disagreement between the AMT workers at the place identification stage and between the experts at the geocoding stage, often leading to long discussions. In some cases, it had to be accepted that no clear decision could be made on whether a phrase is a place name or to which particular place a phrase refers, given the unclear language and very limited amount of context information we experience with tweets. We provide classification schemes for common causes of disagreement and explain how we have dealt with them in our corpus. We also discuss what can be learned from them for future corpus building activities.

We start this section with some statistical properties of the constructed corpus, potentially relevant to future users of the corpus for evaluating or training geoparsing algorithms.

3.1. *Corpus Properties*

The statistics in this section (as well as Appendix C of the Supplementary Online Materials) are based on the cases we consider "unproblematic" in the sense that an agreement has been reached on entities being place names and the respective place names have been resolved to single entities in GeoNames. Put differently, a phrase that constitutes a place name candidate is "unproblematic" if there is no uncertainty about whether it is meant to refer to a place, no ambiguity about which location entity it refers to and that entity has clearly defined boundaries, and a toponym for this entity exists in GeoNames. 81.6% of the 2966 place name candidates found in the corpus were unproblematic in this sense, and 80.9% of the 2122 tweets with at least one place candidate only contained unproblematic place names[1].

3.1.1. Place Name Frequency

As already discussed, we intentionally set out to create a corpus with a higher frequency of place name than one would encounter in a completely random sample. Our analysis of the frequencies of identified place names shows that the event-based sample of tweets used results in 69.77% of the tweets containing no place names, 21.79% containing one place name, 6.32% containing two place names, and 2.13% containing more than two place names, with 7 being the maximum number of place names in a tweet in the corpus. These percentages are significantly higher than the 10% of tweets containing place names we encountered in previous work querying Twitter for about 150 terms with greater diversity (MacEachren *et al.* 2013). The mean number of place names per tweet is 0.36. Only considering tweets with at least one place name, the mean is 1.20 and the median is 1.

3.1.2. Geographic Distribution of Identified Place Names

We derived the geographical distribution of place names identified in the corpus (not counting problematic cases) on a per-continent basis and on a per-country basis. Both show a bias towards English-speaking areas but also towards countries particularly struck by natural disasters, diseases, or violence in the queried time span. Given the focus on English language tweets and the general geographic distribution of Twitter activity (Leetaru *et al.* 2013), the much higher number of places located in North America (in particular the US) is not really surprising. The details can be looked up in Tables C1 and C2 of Appendix C.1 in the Supplementary Online Materials. Additionally, we have created an interactive web map[2] that shows all identified locations hierarchically.

3.1.3. Distribution of Entity Types of Identified Place Names

We also analyzed the distribution of the identified unproblematic place names in the corpus over the GeoNames feature types to get an idea of how many of the place names

[1] This means overall, 94% of the 6711 tweets in the corpus contain only unproblematic place names.
[2] https://www.geovista.psu.edu/GeoCorpora/viz/

refer to continents, countries, cities, natural physical features, etc. A small proportion of these place names lack feature codes in GeoNames, thus these are excluded from our statistics. Our results show a clear focus on country names, state names, and names of populated places like cities and towns. The lack of natural physical features we found (except for streams and islands) can be partially explained by a focus of GeoNames on administrative and populated places and the fact that names of natural physical features such as mountains and forests are typically structured more complexly which leads to a higher number of problematic cases (see Sections 3.2 and 3.3). Table C3 of Appendix C.2 in the Supplementary Online Materials lists the 20 most frequent feature types.

3.2. *Disagreement and Conclusions from the Place Identification*

The disagreement between AMT workers can have a multitude of reasons and as argued by Wong and Lee (2013) is potentially due to ambiguity in natural language. Thus, analyzing cases of disagreement can provide important insights on that ambiguity. Approaches for dealing with disagreement according to Wong and Lee include: providing specific annotation guidelines, using expert adjudication (by a subject matter expert), enabling discussion among annotators, removing entities for which no agreement can be reached, relaxing criteria allowing slightly differing judgments to be counted as the same, and employing crowd wisdom by adding annotators until a certain level of agreement is reached. Wong and Lee further argue that the approach taken in many corpus building efforts to remove all disagreement and keep only entities that sufficient agreement has been reached on is flawed because it fails to consider the "legitimate" disagreement that is rooted in natural language ambiguity. In agreement with this conclusion, our corpus building approach combines several of the measures suggested by Wong and Lee (further explained in this and the next subsections) and still includes tweets with problematic place references in the corpus, clearly marked and distinguished by special tags to give the corpus user full control of how these are employed in an evaluation or training scenario.

For 9.45% of the 6711 tweets in our corpus, no agreement was reached in the AMT-based place identification stage, meaning that they contain at least one place reference candidate which did not achieve the threshold of 70% agreement between AMT workers. In the following, we provide an overview on the reasons for disagreement in place identification, employing a novel classification scheme we established for this purpose. We grouped the specific categories for causes of disagreement into three main classes: *Tweet Deficiencies*, *Worker Errors*, and the most interesting category of *Disagreement based on Differing Definitions* of place name. In a few cases, a clear categorization was not possible as there were several plausible explanations for the disagreement. We here typically only provide a single tweet example for each category but more examples can be found in Table D1 of Appendix D of the Supplementary Online Materials.

3.2.1. Tweet Deficiencies (TweetDef)

This category comprises cases in which the disagreement is due to spelling or language deficiencies such as typos, cut off words at the end of the tweet, etc. occurring in the original tweet and leading to AMT worker uncertainty about whether something is a place name that should be marked or not. In the following example, the misspelling of Ferguson caused several workers to not recognize it as a place reference:

Example 1: *"#furgeson I didn't know it was, "wear a hoodie to the riot" day. Damn pullover!"* (Tweet ID 537064161977434113)

3.2.2. AMT Worker Errors

In this category we collect cases in which workers made mistakes that clearly violate the instructions we gave to them. The category is divided into the following subcategories:

Grouping errors (GroupingErr): In our instructions, we told the AMT workers to mark the words making up a single place name (e.g."New York City") together and to mark consecutive names separately, even in cases with a hierarchical relationship (e.g. pairs of city-state as in "Hartford, CT"). One reason for this was that there are often multiple entities in the world that match such place combinations, e.g., "Springfield, VA" exists in 7 counties. Tagging each separately provides the possibility of locating the state correctly even when the city is not, which in many applications will still be useful. Despite our instructions and the animated example demonstrating how to highlight the words making up a single place name together, we found cases where individual worker submissions clearly violated these rules. In the GroupingErr category, we count all errors related to either (a) not marking words forming a single place name together and (b) incorrectly marking consecutive place names as a single place reference.

Example 2: *"RT @paulturkey: 2,000 people turned out in the wind and rain today in Arklow Co Wicklow to protest against the water charges well done all"*
(Tweet ID 529227404452327424)

This rather complex case of a grouping issue resulted in three different grouping variants by the AMT workers: "Arklow Co Wicklow","Arklow" and "Co Wicklow", and "Arklow Co" and "Wicklow". "Arklow" is a city in Ireland located in "County Wicklow". We believe that unfamiliarity of many workers with the involved places, a tweet deficiency in the form of a missing comma separating the two involved entities, the use of the abbreviation Co for County and potentially unfamiliarity with the British and Irish way of putting "County" in front of a county name rather than behind it (as practiced in the U.S.) have strongly contributed to the varying AMT results we got.

Missed place errors (MissedPlaceErr): Under this category, we subsume all cases where the disagreement is due to some workers erroneously not marking a place name at all that they clearly should have marked according to the instructions, and where this is not due to some tweet deficiency (see category TweetDef above) or reasonable disagreement on word sense or definition of place (see Section 3.2.3 below). The rules related to this are that workers were told to mark all occurrences of a place name in a tweet, including named buildings and areas, street names, abbreviations and nicknames, and place names occurring as part of a hashtag. The reasons for a missed place cannot clearly be identified from an outside perspective, but presumably range from unfamiliarity with a place name or abbreviation/nickname of a place, through not reading the instructions carefully enough to understand that entities such as named buildings or streets should be included, or interpreting categories such as 'buildings' or 'street names' too narrowly, to simply overlooking a place mention in the tweet.

Example 3: *"RT @EuromaidanPR: Comparing the uniforms of "rebels" in the #Donetsk region and of #Russian SOF in #Grozny @Fbeyeee —EMPR http://t.co/8NJpM"*
(Tweet ID 540582148907220992)

In this example, several workers did not mark the mention of the city of "Grozny" in the tweet, presumably because of not being familiar with the name. Another explanation here could be that these workers did not take into account the rule to include place names that are part of a hashtag, but at least some of them marked "Donetsk" which also is part of a hashtag, so unfamiliarity seems like the most likely explanation.

Not-a-place errors (NotAPlaceErr): This category subsumes all cases where workers identified words or part of words (e.g., in the case of hashtags) as a place name when the tweeter was not intending to refer to a place and, in particular, if the marked phrase falls into one of the categories that workers were explicitly told not to mark, such as place names that are part of business or organization names ("Baltimore Ravens") or used as descriptors ("U.S. dollar"), see Section 2.2. Please note that cases where it is debatable whether something is a reference to a place or not, will be discussed in Section 3.2.3.

Example 4: *"#News: At least 1,119 Iraqis died in violence in September, toll excludes slayings by IS group: The U.N. http://t.co/QwphtAUm2B #TU"* (Tweet ID 517254978889457665)

In the results we got from the AMT workers, we frequently found cases in which a worker marked the name of a country that appears as a prefix of the term for the people from that country. While we did not provide an explicit rule for this case, our view is that this is a reference to a particular group of people rather than a direct reference to a place and, hence, should not be tagged. However, the question where exactly to draw the boundary is a tricky one when one considers the spectrum of phrases from "Iraqi" to "Iraqi people", to "people *of* Iraq", to "people *in* Iraq" (see further discussion of this topic in Section 3.3). Another example from this category was the term "West Nile" in a context that made it clear that it was a reference to the disease, not a place reference.

3.2.3. Disagreement Based on Differing Definitions

This category refers to cases where even under careful consideration, laypeople and even geography experts may disagree about whether words in a tweet refer to a place or not. They are of particular interest for both theoretical and practical reasons. Theoretically, examining them can potentially provide us with insight into how people recognize and refer to places, and help us make progress on the overall question of "what is a place?". Practically, these cases constitute the most common source of disagreement. Hence, we need to understand them to maximise our yield of identifications for the corpus.

Tweet user ID-related (TweetUserID): Place names are commonly used as part of Twitter usernames as in the following tweet:

Example 5: *"RT @DFID_UK: Summary of pledges from countries, charities & companies made at the Defeating Ebola conference now online: http://t.co/VWTTKd"*
(Tweet ID 517737521196064768)

While we provided instructions that places in hashtags should be marked, we did not provide any instructions for Twitter user IDs such as "@DFID_UK" in this example. A place occurring in a Twitter name can give some hint about where the tweet author is located (or perhaps where they were from) and some of these were tagged by AMT workers while others were not. After discussing these cases, we decided that Twitter IDs, first of all, are references to persons or organizations and not direct references to places, even if they include place names. Therefore, we established the additional rule that place references in Twitter IDs should never be considered place mentions and marked.

Address related (AddressRelated): While we included the rule to mark names of streets, etc., we did not say anything about addresses occurring in tweets:

Example 6: *"BACK TO THE FUTURE ARTIST SHOWCASE — AUG. 24th @ FIRE & ICE! 312 MARKET ST. 5pm - 10pm! Tickets* 10&*15 at the door! ?? BE HERE!!"*
(Tweet ID 496371470722150400)

Some AMT workers did not mark "312 MARKET ST." at all, others marked only "MARKET ST.", or others marked the entire phrase. It is our view that an address constitutes a named place and, hence, should be marked and that the house number and street name should be considered as a single place reference to be marked together.

Descriptive and adjectival usage (DescAdjUsage): This category refers to cases where place names are used as a descriptor or adjective (including noun adjuncts, also called attributive nouns, and possessive forms) for a different object that is not itself a place. While we had a rule in the instructions that such cases should not be tagged, it became clear when looking at the AMT results that this is problematic and many of these cases are boundary cases where it is difficult, if not impossible, to decide whether a place name is used as a place reference or not. Let us look at a representative example:

Example 7: *"RT @Watcherone: South Sudan rebels have killed several Uganda soldiers in the Upper Nile in renewed fighting in the country."* (Tweet ID 521571268362252288)

Both "South Sudan" and "Uganda" are noun adjuncts used as descriptors for groups of people in this tweet. Such noun adjuncts and other forms of adjectival use can refer to the place, the people, the language, or to the government of a place. In the example here, the purpose most likely is to describe the origin of the respective group of people.

Other examples falling into this category include the phrase "Hanshin-Awaji Earthquake", an example of a historic natural disaster that can be considered a reference to a place as much as a reference to the event, and "Kansas State University campus", an example in which a specific location or geographic entity is referred to with a qualifier that contains a proper name (potentially a proper place name), the name of "Kansas State University" (which can be seen as both an organization name as well as a reference to a place). We assume some workers considered the proper names in these as descriptors (noun adjuncts in this case), while others considered the entire expression to be a place reference. This resulted in the realization that these situations can indeed be difficult to distinguish, and that it can also be hard to differentiate when a proper name is a qualifier from when it is part of a proper name itself. This ultimately was one of the reasons for us to introduce a special tag called "Uncertain semantics" for annotating such place reference candidates (see Section 3.3). All tweets with place references from the *DescAdjUsage* category were submitted to the geography experts in the geo-annotation stage for discussion and guideline establishment.

Organization related (OrganizationRelated): As discussed, AMT workers were told not to mark names or part of names of organizations and businesses, even if their names contained a place name. The logic behind these instructions is that these instances are first and foremost references to organizations instead of direct references to places, even though an association between the organization and a place may exist. While certain AMT results were clear violations of this rule, most cases turned out to fall into the gray area where it was difficult to decide whether a phrase involving a place name actually refers to an organization or something else, e.g. a building or geographic area.

Example 8: *"RT @we_support_PTI: All Pakistani's In USA - COME OUT to Protest #GoNawazGo, as the #FakePrimeMinister visits United Nations. #ImranKhan ht"*
(Tweet ID 515036649768890368)

In this example, "United Nations" can refer to the United Nations organization but also be seen as a reference to the United Nations Building. "United Nations," as the object to "visits," hints at this interpretation. But, we found even less ambiguous examples in the corpus where the name appears with the spatial preposition "at". In conclusion, there can be very small changes in wording that will sway one's interpretation of a phrase between organization, place, or also other things such as government of a country, event, etc. (see also Alonso *et al.* 2013, Markert and Nissim 2002). Similar to the *DescAdjUsage* category, we decided to annotate boundary cases with the "Uncertain semantics" tag.

Kind of place qualifiers (KindOfPlaceQual): We often find the name of a place in combination with a noun describing its kind, such as in "City of London", "Island of Saint Louis", "South Georgia Island", "Ganges River", etc. It can be difficult to decide when such place qualifiers are actually part of the proper name and when they are added for other reasons, e.g. for disambiguation of the generic object (e.g., river, island, etc.).

Example 9: *"RT @specterm: Will smbdy have to die before ppl get the message? RT @APHL Measles outbreaks hit 18-year high in Washington state http://t."*
(Tweet ID 487688329643565056)

In this example "state" is not part of the proper place name but most likely used to distinguish it from Washington D.C., often also only referred to as Washington. Some AMT workers marked just "Washington" and some marked "Washington state". Another example was the expression "at the Battle of Oulart Hill, County Wexford" containing two kinds of place qualifiers: "Hill" in "Oulart Hill" and "County" in "County Wexford". The official name of the county is "County Wexford" which distinguishes it from its county town just called "Wexford". While AMT results for "County Wexford" were split half-and-half between marking just "Wexford" and marking "County Wexford", there was sufficient agreement (according to our 70% criterion) that "Oulart Hill" should be marked together. For many physical features such as lakes, mountains, etc., the kind of place noun is typically considered part of the proper name. From a geoparsing perspective, this is very relevant because otherwise, it would not be possible to distinguish the physical feature from a similarly named populated place as in Lake Michigan vs. the state of Michigan. For other entities, for example, cities or states, it can be very difficult to decide even for experts and with thorough background research.

Vaguely qualified place name (VaguelyQualified): This category deals with expressions in which a proper place name has a spatial qualifier that narrows down or extends its spatial footprint, but at the same time introduces some vagueness. The most common case of vague qualifiers are cardinal directions as in "west Africa" and "Eastern Indonesia", but we also find examples in which it is another spatial term such as "upper" or "central", or where the vagueness is introduced by a trailing "region" or "area" as in:

Example 10: *"RT @EuromaidanPR: Comparing the uniforms of "rebels" in the #Donetsk region and of #Russian SOF in #Grozny @Fbeyeee —EMPR http://t.co/8NJpM."*
(Tweet ID 540582148907220992)

While the proper place name occurring as part of a VaguelyQualified entity can typically be geocoded to a toponym in GeoNames, this is usually not the case for the region described by the entire expression. In many cases, it can also be extremely difficult to determine whether the expression refers to a clearly defined region with fixed boundaries or is meant in a vague way. While our original instructions to the AMT workers were to not mark the qualifier part of vaguely qualified place names, considering these disagreements, we revised that decision and instead marked the entire expression and introduced the special tag "Vague" to annotate such place references in our corpus (Section 3.3). Due to this change, marked place references in tweets with AMT agreement had to be revisited manually and checked for vaguely qualified place names. The small number of cases found were then manually corrected to follow our new rule.

3.2.4. Resolving the Disagreements

Many of the tweets with disagreement at the AMT stage are particularly interesting cases for evaluating and training geoparsing software because they often contain several place names, sometimes in a particular spatial relationship, e.g., a hierarchical relationship as in some examples from the *GroupingErr* category. Hence, resolving the disagreement and keeping these tweets in the corpus is important to not introduce a bias into the final output. After discussing different options for dealing with annotator disagreement (see again Wong and Lee 2013), our approach was the following:

(1) Clear cases for the different categories were resolved manually and then moved to the next stage as if they were tweets with AMT agreement. Almost all cases from the *TweetDef*, *GroupingErr*, *MissedPlaceErr*, and *NotAPlaceErr* categories were clear cases. The categories *TweetUserID* and *AddressRelated* could also almost completely be resolved manually. In future corpus building work, updated instructions should help reduce cases of disagreement from these categories.

(2) All other cases of place references with disagreement received a special tag in our corpus database and were presented to the experts performing the geo-annotation stage. That means the experts did not only perform the geo-annotation step of the marked place references, but also performed the place identification for these tweets again. The experts then used special tags to annotate the problematic place references where they also were not able to make a clear decision. This was the case for most of the candidates falling into the *DescAdjUsage*, *OrganizationRelated*, *KindOfPlaceQual*, and *VaguelyQualified* categories.

3.3. *Disagreement and Conclusions from the Geo-Annotation Stage*

As described in Section 2.4, in the geo-annotation stage, geography experts used GeoAnnotator to look at the place names identified by AMT workers and geocode them to toponym from the GeoNames gazetteer. In addition, the experts had to decide on the remaining unresolved cases from the place identification stage where AMT workers disagreed and no easy manual resolution was possible. In the ideal case, the experts would agree on four aspects of a place reference candidate in a tweet: (1) that a particular considered section of a tweet is indeed a reference to a place, (2) which geographic entity (toponym) the mentioned place refers to (no ambiguity), (3) that the place is a concrete place, not a vague place (more on this below), and (4) that this place exists as a toponym in GeoNames (and which GeoNames toponym it corresponds to). We will continue to refer to such a place name candidate as an unproblematic case. A tweet is considered “unproblematic” if all place candidates in it are unproblematic. However, as discussed, the concept of place is a difficult one and tweets and natural language in general are full of difficult and boundary cases with regard to these four aspects. As a result, we had many cases of place reference candidates where: (a) experts did not agree initially (but came to consensus after discussion) or (b) experts agreed that one or more of the four points above did not hold, and thus tagged the case as uncertain (thus did not meet 1), ambiguous (thus did not meet 2), vague (thus did not meet 3), and/or that it was missing from GeoNames (thus did not meet 4). For this purpose, we introduced a special tagging scheme to label place reference candidates in the tweets. These tags are included in the published version of the corpus and enable users of the corpus to isolate these particular tweets when training or evaluating geoparsing software (e.g., to exclude the problematic cases or to focus on one or more aspects of those cases).

Tag	%
no tag (= unproblematic)	81.76
uncertain semantics	6.00
vague	4.28
non-overlapping ambiguous	0.67
overlapping ambiguous	3.07
not-in-geonames	8.60

Table 1. Percentages of tags applied over the 2966 place name candidates (multiple tags possible).

In the unproblematic case described above, the place mention will not be tagged with any special tags. The geocoding result consists of the name and GeoNames ID of the chosen GeoNames toponym as well as the WGS84 coordinates provided for it by GeoNames. This information allows for training NER and geocoding components and geoparsing tools as well as evaluating the performance (e.g., precision and recall) of existing approaches either based on the ID (in systems that are themselves based on GeoNames or another geographic knowledge base that is linked to GeoNames) or based on the geographic distance derived from the coordinates. The following example shows the information provided for a tweet that only contains unproblematic cases of place mentions.

Example 11: *"#USGS #Breaking? M 2.9, 8km NW of The Geysers, California http://t.co/FfdfHlGpmQ #PastHour #286 #earthquake #tsunami #prayfromjapan"*
(Tweet ID 546169066524651520)

Place references:

"The Geysers" (tweet characters 35 to 45)
Toponym name: The Geysers
GeoNames ID: 5352058
WGS84: 38.77491, -122.75527

"California" (tweet characters 48 to 57)
Toponym name: California
GeoNames ID: 5332921
WGS84: 37.25022, -119.75126

"japan" (tweet characters 127 to 131)
Toponym name: Japan
GeoNames ID: 1861060
WGS84: 35.68536, 139.75309

For problematic cases, we use the five tags "Uncertain semantics", "Vague", "Non-overlapping ambiguous" / "Overlapping ambiguous", and "Not-in-GeoNames" to categorize the reasons why a place candidate is problematic. The information we provide as the result of the geocoding stage varies depending on which of these categories the place candidate falls into (further explained below). Additional examples can be found in Table E1 of Appendix E of the Supplementary Online Materials. Table 1 shows the percentages of how often the different tags were applied over the 2966 place name candidates.

3.3.1. "Uncertain Semantics"

The tag "Uncertain semantics" is used for boundary cases where expert annotators found it difficult to decide whether a phrase should be considered a reference to a named place or not (or decided that it was clearly a boundary case with mixed word sense). Most cases from the *DescAdjUsage* category discussed in Section 3.2.3 led to discussions between the experts, and many of these ended up tagged as "Uncertain semantics". The experts established a guideline that DescAdjUsage forms should be marked but not receive the "Uncertain semantics" label if the noun they qualify is a physical event like a natural disaster (as in "Luzon earthquake") or a non-natural event taking place at a particular location like "riot" in the example below:

Example 12: *"#XclusiveOldPost Full Official Statement Of The Whole Caleb University Riot http://t.co/61ouBuXd4B"* (Tweet ID 530978755377127424)

For place names used as attributive nouns that modify a generic place noun, it was agreed to highlight and code the composite entity as a single place (similarly to the approach taken with street number and name).

Example 13: *"@themoonandyou that guy beheading that poor woman, or the fire at the Chicago airport, etc.). I don't mind that they have an article on them"*
(Tweet ID 516000748098699266)

Regarding possessive forms, it was agreed to mark the place name since the focus is on something that belongs to the place rather than the place as a descriptor for the noun:

Example 14: *"RT @ezralevant: Watch the riot videos. Listen to the victims of this violence. And then help me hire Calgary's best lawyer to sue the offen"* (Tweet ID 491432559159414786)

In all other forms of DescAdjUsage cases, the experts decided to apply the "Uncertain semantics" tag to indicate that these are borderline cases. In addition, the experts agreed to tag cases from the *OrganizationRelated* category (Section 3.2.3) as "Uncertain semantics" unless it is absolutely clear that the reference is to an organization which has the place name as part of the organization name. In particular, metonymical uses of place names in this context were tagged as "Uncertain semantics" and AMT workers marking the place name was taken into account as an indication that multiple interpretations are possible. Below is one such example that was tagged as "Uncertain semantics".

Example 15: *"Presbyterian Hospital Dallas says they admitted a patient into strict isolation based on the patients symptoms and recent travel. #Ebola"* (Tweet ID 517056733022130176)

In this example, "Presbyterian Hospital Dallas" is the place where the patient was admitted but the statement is from the hospital administration as an organization. Hence, it is an example of metonymical use of a place name.

Despite not being able to make a clear decision, phrases tagged with "Uncertain semantics" were still further processed in the same way as clear mentions of places, meaning the experts attempted to geocode them to a toponym from GeoNames. Hence, users of the corpus can still look at the geocoding results of these cases if they decide to do so. It also means that these phrases are subject to be tagged with the additional labels "Vague", "Non-overlapping ambiguous", "Overlapping ambiguous", and "Not-in-GeoNames" if they also satisfy the criteria for any of these categories (described in more detail in the following) and that the geocoding result provided for such a place candidate depends on whether additional tags are applied or not.

3.3.2. "Vague", "Non-overlapping ambiguous", "Overlapping ambiguous", "Not-in-GeoNames"

This group of tags is used to signal that a place reference is problematic with regard to the aspects 2-4 (vagueness, ambiguity, and not being listed in GeoNames) under the assumption that the phrase in question should indeed be considered a place reference. We further explain these tags in the following and provide examples.

"Vague": If a place name refers to an area, neighborhood, or region whose boundaries are not clearly defined or agreed upon, we consider it an instance of the vague category and tag it accordingly. Whether or not an entity is considered vague is itself an interpretation. In practice, the experts relied on geographical knowledge and/or secondary sources to make many of these decisions. If, for example, a region is used by governments or NGOs to define a territory for administrative purposes, the region was not tagged as vague (e.g., the "West Midlands" in the UK, which is an official region for statistical purposes). On the other hand, if a source such as Wikipedia or a research publication indicates that there are differences in opinions on the bounds of a region, that region was tagged as "Vague". For intance, Southeast Texas shows in Wikipedia as having a core that everyone agrees upon and a periphery where there is no agreement[1]. For the Middle East, the CIA Factbook and Encyclopedia Britanica have 15 countries that match and 7 that do not. In addition, in cases where there are no standard sources that detail where a region is (whether cultural or physical), the experts used capitalization as a criterion. With capitalization, the directional qualifier is included as part of the name and tagged as vague. With no capitalization and no sources that imply a recognized region, we followed instructions given to AMT workers and did not include the directional qualifier and did not tag the place itself as vague. The following tweet is an example in which the experts agreed to tag "Bay Area" as "Vague" because multiple definitions of the San Francisco Bay Area exist[2]:

Example 16: *"Waking up to news of the Bay Area earthquake. Hope all my friends there are ok!"* (Tweet ID 503558049697513472)

Vague regions are in some cases listed in GeoNames. Therefore, the result of the geocoding stage for place references tagged as "Vague" can be one of two things: either a GeoNames toponym (name, ID, WGS84 coordinates) as in the case of an unproblematic place reference, or a representative point (WGS84 coordinates) freely chosen by the experts, in which case the place reference will also be tagged as "Not-in-GeoNames". "North India" appearing in another tweet is an example for this. In contrast, the "Bay Area" from Example 16 above is listed under the toponym "San Francisco Bay Area" (ID 10630414) in GeoNames and has been annotated accordingly.

"Non-overlapping ambiguous" / "Overlapping ambiguous": The aspect of ambiguity is divided into two categories with the respective tags "Non-overlapping ambiguous" and "Overlapping ambiguous", which turned out to be mutually exclusive in our corpus although one could think of cases where both types of ambiguity occur simultaneously. The tag "Overlapping ambiguous" is applied to place references for which it is not possible to confidently determine which of multiple candidate toponyms that are spatially overlapping is the correct one. For example, "Gaza" in the following tweet could refer to "Gaza Strip" (GeoNames ID 281132), "Gaza City" (GeoNames ID 281133), and in principle also "Gaza", a second-order administrative division of Israel (GeoNames ID 7870269), but the latter is an unlikely interpretation considering the tweet content:

Example 17: *"RT @saidshouib: #Gaza_Under_Attack — This is not the effect of an #earthquake, also it's not a #meteor. It's an Israeli Missile. http://t.c"* (Tweet ID 488353532714967041)

Often, but not always, the alternative candidates appear at different levels in the hierarchy of administrative regions, and it is just not possible to determine which level is being referred to (e.g. city vs. state or economic region). A very typical example is a reference to "New York" where it is not clear whether this is a reference to New York City or the state of New York. The geocoding result for "Overlapping ambiguous" cases is a list of GeoNames toponyms (name, ID, WGS84 coordinates) instead of a single toponym.

For "Non-overlapping ambiguous" place references it is also not possible to confidently pick the correct toponym from a set of alternative candidates. However, these are not

[1] https://en.wikipedia.org/wiki/Southeast_Texas
[2] see https://en.wikipedia.org/wiki/San_Francisco_Bay_Area

geographically overlapping, but rather disjoint alternatives such as "Washington DC" vs "Washington State" for "Washington", or all the different cities and towns called "Bloomington" that exist in 18 U.S. states and in Canada for the tweet example below.

Example 18: *"It's funny cuz the news doesn't say anything about a tornado in bloomington my my room tells a different story"* (Tweet ID 531929908952707072)

The geocoding result can be a list of alternatives similar to "Overlapping ambiguous". For instance, for one tweet we found two possible candidate locations for "Saint Joe Regional Medical Center", namely Saint Joseph Regional Medical Center Mishawaka Campus (GeoNames ID 4925876) and Saint Joseph Regional Medical Center Plymouth Campus (GeoNames ID 8103007). Both are listed as alternatives in the result. However, there were cases in which the number of alternatives was too large to identify and list all of them (e.g., "Bloomington" in Example 23 above). In some of these, context provided enough clues to allow 2-4 to be listed as the probable options. In other cases, there were many possible places and no clues in the context, thus an arbitrary one of the many was selected and tagged as non-overlapping ambiguous (this is easily recognized in the corpus where only one candidate is listed). This means if corpus users decide to employ a tweet with a non-overlapping ambiguous place reference, they should take into account that there is a high chance that this toponym is actually not the one referred to in the tweet.

"Not-in-GeoNames": Lastly, if a place (or candidate place in the case of ambiguous place references) does not exist in GeoNames, so that it currently is not possible to provide a GeoNames ID for it, it will receive the tag "Not-in-GeoNames". This could be (a) places that are in general not listed in GeoNames (for instance street names and addresses) for which a different geographic knowledge base is needed, (b) places whose boundaries are so vaguely defined that adding them to GeoNames is not justified, or (c) named places which simply have not been added to GeoNames yet. In the latter case, we have started adding these places to GeoNames, and our progress with this task will be reflected in future releases of the corpus. One example for a place reference tagged as "Not-in-GeoNames" in the current corpus appears in the following tweet:

Example 19: *"An Earthquake in the Muslim World Imam Calling for Jewish Prayer on the Temple Mount! via The Middle East... http://t.co/HrGTJYIVFn"*
(Tweet ID 548978575136022528)

No matching entity for the natural geographical feature of the Temple Mount could be found in GeoNames at the time of this writing, although we believe it should be added. The geocoding result for all such place references that are not currently in GeoNames is a representative point picked by the experts. The representative points were selected without an attempt to provide a precise location with the mentioned plan of replacing these in future releases once a corresponding toponym has been added to GeoNames. We typically either used the coordinates of entities at about the same geographical level (e.g., a building or park that was in GeoNames to represent a building that was not) or the coordinates of a surrounding political entity (e.g., a city) to provide a less precise but 'in the ballpark' location.

3.4. *Concluding Discussion*

The corpus produced includes a particularly rich perspective on the challenges of recognizing and locating place references in text. Our goal was to create a set of annotations that provided the most accurate reflection we could achieve short of contacting each tweet author to ask what they meant. Thus, we and the additional geo-annotators made use of authors' Twitter profiles, news stories about events discussed, and additional web searches to remove uncertainty about whether a reference was to a place, and if so, which one. Thus, the corpus provides a target that will be difficult for computational geoparsing methods to match unless those methods also leverage context that is not part of the tweet text itself. An interesting research question is to which extent places in this corpus can be recognized and located without going beyond the tweets and what is gained through integration of different sources of context.

4. Conclusions and Outlook

We presented our GeoCorpora initiative for developing a corpus building framework plus tool set and constructing a geo-microblog corpus of geo-annotated tweets to facilitate the evaluation, comparison, and training of geoparsing software and foster future GIR research. The article will be accompanied by the release of an initial version of the corpus on Github[1] to make it freely accessible to the research community together with a software tool for accessing the corpus instances. The tool will provide an easy way for the user to select which corpus instances will be used based on our tagging scheme. In particular, it will allow the user to restrict the application of the corpus to unproblematic instances, excluding place references tagged as "Uncertain semantics" from the calculation of precision and recall statistics, which we expect to be the common way of using the corpus for evaluation purposes.

The corpus building framework we described adopts geovisual analytics and crowdsourcing components, and we expect this framework to play a key role in future corpus building work, for instance in creating corpora for other kinds of text, or to produce more domain specific or localized corpora. Refined and improved guidelines as well as extensions of the GeoAnnotator interface resulting from the corpus building and analysis work will greatly benefit these future endeavors. The analysis of place identification results from AMT workers and the discussions between the experts performing the geo-annotation task resulted in important insights gained on the usage of place references in text and microblog text in particular. Among the most important insights are: (a) that use of place names as attributive nouns (noun adjuncts) represents a challenge for automated geoparsing, since not all such uses should be considered equivalent, (b) metonymy is fairly common, and in some cases, place names have more than one sense simultaneously, and (c) while ambiguity has been considered in prior research, few if any studies have addressed the challenge of overlapping ambiguous references. This is potentially due to the location of the candidate entities being so similar, but disambiguation among overlapping ambiguous references still matters in situations where feature type matters or where linking semantic web sources using unique IDs is required. More generally, it is important to recognize that no place reference corpus can be perfect due to the nature of language; different individuals are likely to interpret the same text differently, and in some cases, the individual who generated the text may not even agree with themselves on its meaning at some later point in time. While we are convinced that employing experts for the geo-annotation step at this point of our research was the right approach, we are planning to explore different ways of using AMT to improve future corpus building and create larger corpora in the presence of a trade-off between quantity and quality: It is

[1] https://github.com/geovista/GeoCorpora

clear that one could create a larger corpus quickly by reducing the qualification requirements for AMT workers and relying more on AMT workers and less on experts even for the geo-annotation step, but this would probably come at the cost of lower quality of the resulting corpus. Gaining a better understanding of this trade-off in light of the general task of improving geoparsing approaches and developing solutions that offer a better compromise are important goals for future experimental work. As suggested by one of the reviewers of this article, tweets could also be distributed to AMT workers based on an estimate of the locations the tweets are about and the location of the AMT worker, something that is possible in AMT at a country and state level. The understanding gained by analyzing causes of disagreement could now also be used to create specialized corpora to evaluate geoparser performance for a particular problem class.

The tagging scheme we introduced allows for dealing with the challenges discussed above on a coarse level and we plan to further refine it with future releases of the corpus, potentially based on feedback from the research community. Future releases of the corpus may also be based on additional geographical knowledge bases to deal with feature types such as street addresses not represented in GeoNames. As a first step, we have started to integrate OpenStreetMap data into our local GeoNames database. One of our next planned steps is the application of the corpus to quantitatively evaluate and improve the performance of our own geoparser GeoTxt (Karimzadeh *et al.* 2013). As part of this process, we intend to employ the results from the AMT place identification stage to further train the NER components in our system. We expect that these activities will result in new insights that will in turn lead to further improvements of the corpus and corpus building framework. One important future research question in this context is to study the relationships between success in geoparser training (using a range of geoparsing methods), corpus size, and other corpus characteristics (e.g., whether constructed using a stratified sample by event type mentioned in the text or through a completely random sample of text). While it is obvious that larger corpora are needed to support deep learning methods than to support more standard machine learning approaches, the required increase in corpus size is unknown. Additionally, research is needed to determine the impact of corpus size and sampling method on the relative success of rule-based parsers (e.g., Gate) and machine learning based parsers (e.g., Stanford).

Acknowledgements

This work was partially funded by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0003. We thank Frank Hardisty and Yiting Ju for many fruitful discussions and their contributions to earlier versions of our corpus building framework, Nathan Gould and Connor Henderson for their help with geo-annotating the tweets, and the anonymous reviewers for invaluable feedback and suggestions.

References

Ahlers, D. and Boll, S., 2008. Retrieving address-based locations from the web. *In*: *Proceeding of the 2nd int. workshop on Geographic information retrieval*, 27–34.

Alonso, H.M., Pedersen, B.S., and Bel, N., 2013. Annotation of regular polysemy and underspecification.. *In*: *ACL (2)*, 725–730.

Baker, P., 2010. Corpus methods in linguistics. *In*: L. Litosseliti, ed. *Research methods in linguistics*. London: Continuum International Publishing Group, 93–116.

Blanford, J., *et al.*, 2014. Tweeting and Tornadoes. *In*: S. Hiltz, M.S.P., L. Plotnick and A. Robinson, eds. *Proceedings of the 11th Inter. ISCRAM Conference, University Park, PA: ISCRAM*, 319–323.

Chae, J., *et al.*, 2014. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *In*: *Computers & Graphics*, Vol. 38, 51–60.

Crooks, A., *et al.*, 2012. #Earthquake: Twitter as a Distributed Sensor System.. *Transactions in GIS*, 17, 124–147.

D'Ignazio, C., *et al.*, 2014. CLIFF-CLAVIN: Determining Geographic Focus for News. *In*: *NewsKDD: Data Science for News Publishing, at KDD 2014*.

Dredze, M., *et al.*, 2013. Carmen: A twitter geolocation system with applications to public health. *In*: *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, 20–24.

Fort, K., Ehrmann, M., and Nazarenko, A., 2009. Towards a Methodology for Named Entities Annotation. *In*: *Proceedings of the Third Linguistic Annotation Workshop, LAW 2009* The Association for Computer Linguistics, 142–145.

Gelernter, J. and Balaji, S., 2013. An algorithm for local geoparsing of microtext.. *GeoInformatica*, 17 (4), 635–667.

Gelernter, J. and Zhang, W., 2013. Cross-lingual Geo-parsing for Non-structured Data. *In*: *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, Orlando, Florida New York, NY, USA: ACM, 64–71.

Ghosh, D. and Guha, R., 2013. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System.. *Cartography and Geographic Information Science*, 40, 90–102.

Hu, Y. and Ge, L., 2007. A supervised machine learning approach to toponym disambiguation. *The Geospatial Web – How geobrowsers, social software and the Web 2.0 are shaping the network society*. Springer, 117–128.

Jones, C., *et al.*, 2002. Spatial information retrieval and geographical ontologies. *In*: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 387–388.

Karimzadeh, M., *et al.*, 2013. GeoTxt: A web API to leverage place references in text. *In*: C. Jones and R. Purves, eds. *Proceedings of the 7th Workshop on Geographic Information Retrieval*, 72–73.

Katz, P. and Schill, A., 2013. To Learn or to Rule: Two Approaches for Extracting Geographical Information from Unstructured Text. *In*: *Eleventh Australasian Data Mining Conference (AusDM 2013), Canberra.*

Leetaru, K., *et al.*, 2013. Mapping the global Twitter heartbeat: The geography of Twitter.. *First Monday*, 18 (5).

Leidner, J.L., 2006. An evaluation dataset for the toponym resolution task. *In*: *Computers, Environment and Urban Systems*, Vol. 30.

Leidner, J.L., 2007. Toponym Resolution in Text. Thesis (PhD). School of Informatics, University of Edinburgh.

Lieberman, M.D. and Samet, H., 2012. Adaptive Context Features for Toponym Resolution in Streaming News. *In*: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, Portland, Oregon, USA New York, NY, USA: ACM, 731–740.

MacEachren, A.M., 2014. Place Reference in Text as a Radial Category: A Challenge to Spatial Search, Retrieval, and Geographical. *In*: *2014 Specialist Meeting — Spatial*

Search, Santa Barbara, CA: UCSB Center for Spatial Studies.

MacEachren, A.M., *et al.*, 2013. *Report on New Methods for Representing and Interacting with Qualitative Geographic Information, Stage 2: Task Group 1 Core Re-engineering and Place-based Use Case.* Technical report, Pennsylvania State University.

MacEachren, A., *et al.*, 2011. SensePlace2: GeoTwitter Analytics Support for Situational Awareness.. *In*: S. Miksch and M. Ward, eds. *IEEE Conference on Visual Analytics Science and Technology.*

Mandl, T., *et al.*, 2009. *In*: *GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview.*, 808–821 Springer Berlin Heidelberg.

Markert, K. and Nissim, M., 2002. Towards a Corpus Annotated for Metonymies: the Case of Location Names. *In*: *Third International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain*, 1385–1392.

Moncla, L., *et al.*, 2014. Geocoding for Texts with Fine-grain Toponyms: An Experiment on a Geoparsed Hiking Descriptions Corpus. *In*: *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, Dallas, Texas New York, NY, USA: ACM, 183–192.

Potthast, M., 2010. Crowdsourcing a Wikipedia vandalism corpus. *In*: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 789–790.

Ritter, A., *et al.*, 2011. Named Entity Recognition in Tweets: An Experimental Study. *In*: *EMNLP'11* Association for Computational Linguistics, 1524–1534.

Sabou, M., *et al.*, 2014. Corpus annotation through crowdsourcing. *In*: *Proc. LREC.*

Shneiderman, B., 1982. The future of interactive systems and the emergence of direct manipulation. *eractive systems and the emergence of direct manipulation". Behaviour & Information Technology*, 1 (3), 237–256.

Speriosu, M. and Baldridge, J., 2013. Text-Driven Toponym Resolution using Indirect Supervision. *In*: *51st Annual Meeting of the Association for Computational Linguistics*, 1466–1476.

Starbird, K., *et al.*, 2014. Social Media, Public Participation, and the 2010 BP Deepwater Horizon Oil Spill. *Human and Ecological Risk Assessment: An International Journal.*

Stock, K., *et al.*, 2013. Creating a corpus of geospatial natural language. *In*: *International Conference on Spatial Information Theory*, 279–298.

Tapia, A.H., *et al.*, 2011. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. *In*: *Proceedings of the 8th International ISCRAM Conference*, 1–10.

Tsou, M.H., *et al.*, 2013. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing). *Cartography and Geographic Information Science*, 40, 1–12.

Wallgrün, J.O., *et al.*, 2014. Construction and first analysis of a corpus for the evaluation and training of microblog/twitter geoparsers. *In*: R. Purves and C.B. Jones, eds. *Proceedings of the 8th Workshop on Geographic Information Retrieval, GIR 2014, Dallas/Fort Worth, TX, USA, November 4-7, 2014* ACM, 4:1–4:8.

Wong, B. and Lee, S., 2013. Annotating Legitimate Disagreement in Corpus Construction. *In*: *Sixth International Joint Conference on Natural Language Processing.*

Woodward, D., Witmer, J., and Kalita, J., 2010. A Comparison of Approaches for Geospatial Entity Extraction from Wikipedia. *In*: *IEEE Fourth International Conference on Semantic Computing (ICSC)*, 402–407.

Yosef, M.A., *et al.*, 2011. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4 (12), 1450–1453.