Construction and First Analysis of a Corpus for the Evaluation and Training of Microblog/Twitter Geoparsers

Jan Oliver Wallgrün
GeoVISTA Center
Pennsylvania State University
University Park, PA 16802,
USA
wallgrun@psu.edu

Frank Hardisty
Department of Geography,
GeoVISTA Center
Pennsylvania State University
University Park, PA 16802,
USA
hardisty@psu.edu

Alan M. MacEachren
Department of Geography,
GeoVISTA Center
Pennsylvania State University
University Park, PA 16802,
USA
maceachren@psu.edu

Morteza Karimzadeh
Department of Geography,
GeoVISTA Center
Pennsylvania State University
University Park, PA 16802,
USA
karimzadeh@psu.edu

Yiting Ju
Department of Geography,
GeoVISTA Center
Pennsylvania State University
University Park, PA 16802,
USA
yvj5048@psu.edu

Scott Pezanowski
Department of Geography,
GeoVISTA Center
Pennsylvania State University
University Park, PA 16802,
USA
scottpez@psu.edu

ABSTRACT

This article presents an approach to place reference corpus building and application of the approach to a Geo-Microblog Corpus that will foster research and development in the areas of microblog/twitter geoparsing and geographic information retrieval. Our corpus currently consists of 6000 tweets with identified and georeferenced place names. 30% of the tweets contain at least one place name. The corpus is intended to support the evaluation, comparison, and training of geoparsers. We introduce our corpus building framework, which is developed to be generally applicable beyond microblogs, and explain how we use crowdsourcing and geovisual analytics technology to support the construction of relatively large corpora. We then report on the corpus building work and present an analysis of causes of disagreement between the lay persons performing place identification in our crowdsourcing approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Human Factors, Languages

Keywords

geoparsing, corpus building, microblogs, Twitter

SIGSPATIAL'14, November 04-07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2675354.2675701

1. INTRODUCTION

Geographical information is increasingly available, generated by a wide range of GPS enabled devices and other location-based sensors. Since Twitter added the capability for users to attach a location to their tweets, there has been substantial research attention to leveraging that information to support applications such as situational awareness for crisis events [1], mapping social activities related to elections and other events [13], and many other topics that people can be expected to tweet about. However, there is a potentially larger source of place-related information that often goes unused: information about place embedded in text messages and documents. The focus on geolocated tweets, for instance, misses a much larger source of place-relevant information: Only about 1-2% of tweets include geolocation [8], while 10% or more include references to places in the text of the tweet [11].

In more than a decade of research, the field of Geographic Information Retrieval (GIR) has focused on both search for documents that reference or are about places [6]. This research is complemented by work on place name entity recognition and extraction [5, 15] and more recent efforts to extend geoparsing methods to microblog and other more cryptic social media posts [3]. Nowadays a number of geoparsers and related software tools and services exist for identifying and geolocating places in different types of unstructured text ([4], see also [2] for a more exhaustive list), and in microblog or Twitter messages in particular [3, 7]. A key obstacle for progress in GIR in general and the recognition and disambiguation of place references in particular, however, is the current lack of adequate publically available ground-truth corpora of geoparsed text to facilitate the training, evaluation, and comparison of available computational methods and geoparsing systems. While the testing of toponym resolution algorithms based on corpora of news stories has been discussed in the literature [9, 10], the work reported on in this article is to our knowledge the first attempt to address this gap with a focus on microblogs and the goal of providing a publically available benchmarking corpus. The importance

of microblog data has increased drastically in the recent past in many different fields such as science, policy, and humanitarian relief. Microblog data also provides new challenges for geoparsing tools because geoparsing microblogs is harder than geoparsing news stories or other text documents due to the lack of direct context, the presence of abbreviations, special language and notations, etc. (see [3]).

In the research presented in this article, we are developing strategies and tools for geospatial corpus construction and then using the tools to build a ground-truth Geo-Microblog Corpus. It is one of our main aims to generate a corpus that will be freely available to the research community in the near future. In our corpus building work, we adopt a geovisual analytics approach to building a Geo-Microblog Corpus. The approach leverages crowdsourcing strategies through an Amazon Mechanical Turk visual interface to annotate place references in event-based microblog posts, and a visual-computational method to facilitate geographic disambiguation (toponym resolution) and geolocation. The corpus in its current form is based on a selection of 6000 tweets of which 30% contain at least one place name. Occurring place names are marked and annotated with a unique ID and geographic coordinates. We foresee the following applications of our corpus for researchers and developers working on (microblog) geoparsing approaches:

- **Training:** Many geoparsing approaches employ methods that can be improved by training them with a suitable data set. Since place names occurring in our corpus are identified, it can for instance be used to train Natural Entity Recognition (NER) or similar tools.
- **Evaluation:** The corpus provides researchers with a tool to evaluate the performance of algorithms, components, or geoparsing systems in their entirety.
- **Comparison:** Standard benchmarking data sets have significantly spurred research in other fields as they provide a way for quantitatively measuring and comparing the performance of existing systems.

Beyond these applications, such a corpus has applications outside of the main aim of ultimately benefiting the future development of geoparsing tools. It is also a valuable source for more linguistic and cognitive-oriented 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 crowdsourced lay persons. In our analysis of the crowdsourcing results, we encountered interesting causes of disagreement that we expect to have a significant impact on our future corpus building work.

In the remainder of this article, we focus on the underlying corpus building framework and the design decisions made in its development (Section 2). We provide a first analysis of the crowdsourcing-based place identification component of our framework, looking at reasons for disagreement between the persons performing the place identification for the same tweet (Section 3) by introducing our own classification scheme. We close with conclusions and insights gained as well as an outlook on future work in Section 4.

2. CORPUS BUILDING FRAMEWORK

The aim of the framework presented in this section is to build a corpus of tweets in which all occurrences of place names have been identified, and each such place name is annotated with (a) a unique ID to identify it, and (b) its coordinates. For the unique ID, it has been decided to adopt the IDs used by the online gazetteer Geonames[1] (a data source that is often used by geoparsing tools). Coordinates are given in terms of latitude and longitude in WGS 1984. Because of our approach for geocoding the place names (see Section 2.4), these coordinates also stem from Geonames.org and represent the centroids of the respective places. 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, but in different forms, each occurrence would be annotated with the information: (1) toponym: New York City; (2) geonamesID: 5128581; (3) latitude: 40.71427; and (4) longitude: -74.00597.

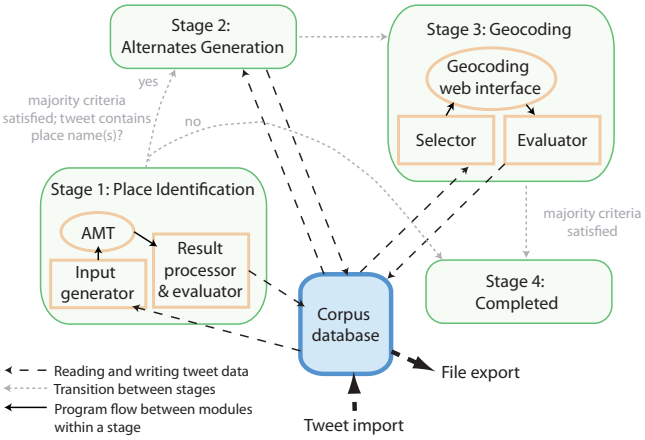

Figure 1: Architecture: In stage 1, place names in the tweet are tagged by AMT workers. In stage 2, for each identified place name a list of potential toponyms is created automatically. In stage 3, a visual interface is used to geocode the place names by selecting the right toponyms from the lists. With reaching stage 4, the ground-truth geoparsing result for a tweet has been completely determined.

To build up such a corpus, we have designed a framework that employs a crowdsourcing approach and different visual interfaces for the two main steps: (a) place name identification and (b) geocoding. The architecture of the framework is shown in Figure 1. A central database is used to store the tweets in the corpus with all required additional information. Each tweet goes through different stages (place identification, alternate generation, geocoding, processing completed). A status field in the database is used to keep track of the processing stage that a tweet currently is at. In the following, we describe the main components including the three main phases (stages 1-3) of corpus generation.

2.1 Tweet Collection and Sampling

Twitter provides an Application Programming Interface (API) through which developers can access and collect tweets. The API used limits the number of tweets accessible, and the textual content of the returned tweets must contain one of the search terms provided with the query. This API has been used to collect close to one billion tweets since January

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

1, 2014. The search terms used are selected based upon relatively current events, mainly related to crisis events. The tweets are stored in a PostgreSQL database. From this collection, a random sample of 6000 tweets was drawn by querying for particular keywords: 500 tweets were drawn for each of the 12 event-related terms dengue, malaria, earthquake, measles, fire, protest, flood, rebels, flu, riot, gun, tornado. These keywords were selected from three kinds of events, representing (a) natural disasters, (b) infectious disease, and (c) human threats/violence, that we expected to have a higher number of tweets that are place focused than in an "average" sample (e.g., those about celebrities). Since our corpus is for training and evaluating geoparsers, we were intentionally looking to achieve a higher number of place names and wanted to use our workers efficiently.

When the tweets were initially drawn from the database, many tweets were written primarily in languages other than English. Since our corpus is intended to focus on English tweets, we include the word "the"—the most common word in English and, thus, very unlikely to create any sort of bias—when drawing the tweets. This resulted in only a few exceptions that are not purely in English. All 6000 tweets were imported into the corpus database, making them available for stage 1, the place identification step.

2.2 Crowdsourcing-based Place Identification

To be able to build a large corpus, we are employing a crowdsourcing approach that makes use of Amazon's Mechanical Turk (AMT) platform[2] for the place identification phase. AMT allows for distributing small tasks, referred to as Human Intelligence Tasks (HITs), over a large base of workers. Recruitment and payment is handled efficiently via the AMT platform such that a large number of results can be collected in a short amount of time.

Over the past 5-6 years, crowdsourcing has been used increasingly to annotate corpora and for corpus building in general. Potthast [12] reports on an AMT based project to create a corpus of PAN Wikipedia vandalism edits. In this project, each edit was judged as vandalism or not by three annotators. When 2/3 or more annotators did not agree, the edit was re-annotated by 3 more, and again until disagreements were resolved or until the number that remained undecided was small. These were then resolved manually by the corpus creators. In our approach we adopt best practices established in that article.

To use AMT for our purposes, we designed a web interface allowing an AMT worker to annotate the place names in a tweet. A worker who accepts one of our HITs is presented with a web page with detailed instructions and two tweets that she/he is supposed to tag. Figure 2 shows the central part of the web interface in which a tweet has to be annotated, before and after the annotating. In this interface, every tweet is split into words, each of which is treated as a token. Workers can easily select the token that they recognize as a place name by clicking on that token, which will be highlighted afterwards. As some place names are composed of multiple words/tokens, workers are required to merge these. Our GUI follows a direct manipulation style utilizing HTML5's Drag and Drop feature, which allows the workers to click and drag the text directly, rather than using radio buttons or checkboxes, which are the default UI elements available in the AMT environment. It reduces workload and increases workers' efficiency and productivity. As a result, we can retrieve more results of better quality. In addition to reducing workers' workload, using this interface can also avoid some common identification issues such as misspelling, as workers do not have to type. As an example, for the place name "New York", workers only need to click on the "New", drag it , and drop it onto the "York". Then these two words are merged, and "New York" is formed as one token. To undo the merging, workers just need to double-click on the token, and it will again be split into tokens.

In the instruction part of the web page that workers are presented with, we show an animated image to illustrate the highlighting and merging process. We also provide instructions (with examples) on what should be tagged and what not. Things that should be tagged are: (1) any named town, city, county, state, or country (e.g., Los Angeles, Jefferson County, NY, Italy); (2) named buildings (e.g., Eiffel Tower, Dodgers Stadium, Alcatraz, James J. Ferris High School); (3) named areas (e.g., Grand Canyon, Pacific Ocean, Washington Mall, Hyde Park); (4) street and highway names (Atherton Street, 1st Ave, Highway 1, I-70); and (5) place names inside hashtags (#newyork).

Things not to be tagged are: (1) businesses (Starbucks, Microsoft, Texas Steak House, Baltimore Ravens); (2) organizations (Lutheran Church, Red Cross, United Nations, Grand Canyon Historical Society); and (3) place names used as descriptors (U.S. dollar, Philadelphia cheesesteak).

We pay AMT workers 4 cents per assignment resulting in overall costs of $600 to collect 5 AMT results for each of the 6000 tweets. In addition to giving detailed instructions, we utilize the options provided by AMT to ensure a high quality of the results by only allowing for workers with a HIT acceptance rate of 95% or higher from previous HITs at the start, later increased to 97%. Moreover, each tweet is tagged by at least 5 different AMT workers allowing us to compare the results as further explained below.

HITs are uploaded to AMT in batches. A module in our framework (see again Figure 1) called Input Generator is responsible for selecting the tweets for the next batch based on their status which in turn is based on whether enough results have been collected for this tweet and whether there is a sufficient agreement between the results we get from the different workers. The newly created batch is then uploaded to AMT. Once results for all HITs in a batch have been submitted by the AMT workers, a second software module, the Result Processor & Evaluator, processes the result file downloaded from AMT: Each HIT result is stored in the central database and for each tweet in the batch, all results collected up to this point are compared. We then follow an approach similar to what we described about the work in [12] to decide whether the results satisfy our agreement criteria. We require the following two agreement criteria: (1) ≥ 5 results from different workers exist for this tweet; (2) for each word in the tweet, there is at least a 70% majority agreement on whether that word is part of a place name and if so together with which other words.

If the criteria are satisfied, a ground-truth annotation reflecting the majority vote for each place name/token over all results is created and stored. Stage 1 for this tweet is then finished and its status changed accordingly. If the ground-truth place identification solution for a tweet does not contain any place names, processing of this tweet is completed. Else, the tweet is moved on to stage 2 (alternate generation).

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

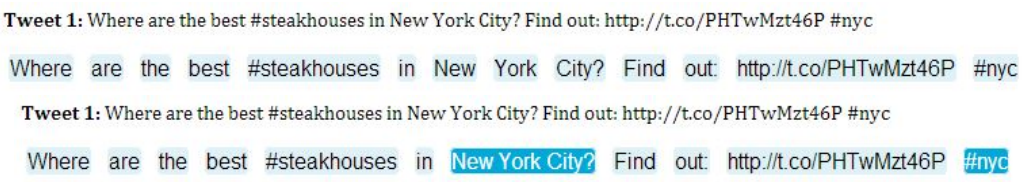

Figure 2: Place identification in the AMT web interface. (a): Tweet before the annotating. (b): Tweet after place names have been highlighted and words belonging to the same place name have been merged.

2.3 Alternates Generation

Since in the geocoding stage (see next section), coders will be allowed to select the correct toponym for each place name in a tweet on a map interface, we need to generate a list of alternative toponyms for each place name, each with its Geonames ID and WGS 1984 coordinates. To generate this list, we use the relevance ranking and geocoding component from our own geoparser GeoTxt [7]. Essentially, this component queries a local Solr index originally created from a Geonames data dump and finds the 100 most likely toponyms for a given place name using our ranking algorithm based on name similarity, population and geographic prominence. The lists are ordered in decreasing order of likelihood. Once these lists have been generated for all place names in a tweet, they are stored with that tweet and the tweet is moved to the geocoding stage.

2.4 Geocoding

In corpus generation, geocoding is the process of annotating the previously identified place names in text with the corresponding geographic coordinates of those places. We use the alternative toponym lists created in stage 2 and an interactive map interface (see Figure 3) to facilitate the lookup process for geocoding. An advantage of using Geonames as the gazetteer for looking up the alternative toponyms is that it is used by many geoparsing tools and the Geonames ID is becoming the defacto standard for linking toponyms in other databases such as DBpedia[3]. However, it has to be noted that the Geonames data is not static. Changes made over time may concern the annotations in our corpus such that a suitable strategy will have to be devised to ensure that the published corpus will remain applicable.

For this stage of corpus building, we use experts, currently PhD students or university faculty members in geography. We have three reasons for this choice: (1) only tweets that have a place name in them (30% of the initial sample) will reach this stage, thus making this approach using individuals with geographic expertise feasible; (2) geocoding place names requires either local knowledge of the place [20] or general geographical knowledge and the willingness to spend extra time to research the existence and location of such a place and its most likely toponym candidate, typically using different web sources; (3) in case of disagreement in geocoding between annotators, we want them to be able to communicate and collectively make decisions, rather than assigning the same task to a new person over and over again without establishing a fundamental understanding of the reasons for disagreement. The communication among expert geocoders has the potential to provide insights on the challenges of geographic disambiguation, including kinds of ambiguity that are common in microblog (or other text) references to place. Therefore, we believe experts are a better choice for the geocoding stage compared to a wider crowdsourcing approach, at least until a set of rules and instructions are generated out of experts' experience with geocoding that would serve as guidelines for the more general public to perform geocoding as well. Once such a rule base is established, a set of well-formed instructions for crowdsourcing has the potential to be used to build larger corpora, a goal that we are pursuing in future research.

For each tweet, the interactive map component of the user interface displays all the places mentioned in the tweet, with the automatically determined toponym with the highest relevance as the default choice for each place name (Figure 3 top). If the automatically assigned toponym is not correct, the annotator is able to click on any of the toponyms to see the alternative candidates for that place name and then pick the correct one (Figure 3 bottom). If none of the presented toponyms is the correct one, this can be indicated by the annotator as well.

Because we are entrusting geocoding to experts, we expect the results of geocoding to be precise. However, to eliminate human error as much as possible, each tweet is assigned to two experts. If their annotation is in agreement, then the resulting annotated tweet will be recorded as a corpus artifact. If there is a disagreement between the two annotations, the tweet will be sent back to both of the annotators to see if they have made a mistake. If both annotators insist on annotations that do not match, that tweet will be flagged and saved for future investigation and guideline establishment.

Using the framework described in this section, we have successfully finished the place identification stage (stage 2) for the 6000 tweets. Since the geocoding part (stage 3) is still in progress, we concentrate our analysis in the next section on the results of the AMT-based place identification stage.

3. A FIRST ANALYSIS OF THE CORPUS

In this section, we provide a first analysis of the corpus collected, focusing on the results from the place identification stage. We are considering the questions of how often place names occur in our corpus and what the sources of disagreement between AMT workers in the place identification stage are. After proposing a classification scheme for the sources of disagreement, we discuss what we can learn from these disagreements about place conceptualization generally and more specifically, how we can improve the components of our corpus building framework and resolve different kinds of disagreement to make the results usable for our corpus.

3.1 Frequency of Place Names

One of the objectives when starting to create the corpus

[3]http://dbpedia.org

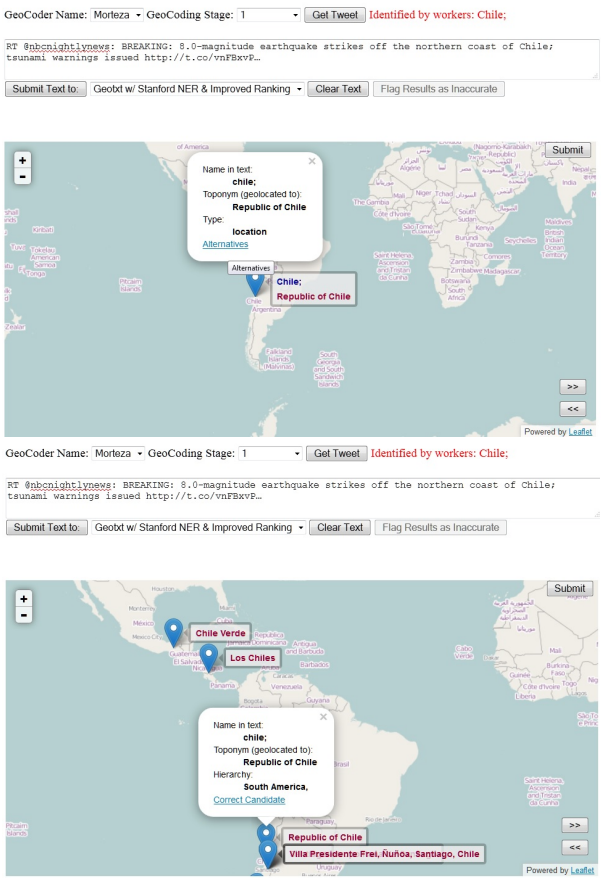

Figure 3: Interactive geocoding interface. Top: Initial situation with automatically assigned toponyms (here 'Chile"). Bottom: Alternative toponyms are displayed allowing the expert to pick the right one.

was to achieve a higher frequency of place names than in a completely random sample. Overall, it turns out that of the event-based sample of tweets making up our corpus 70.0% contain no place names. 22.3% contain one place name. 5.7% contain two place names and 1.3% more than two place names. Thus, the percentage with place names is much higher than the 10% tweets with place names that we obtained in previous work by querying the twitter API for about 150 terms that had greater diversity.

3.2 Place Identification Disagreement

While our corpus collection framework automatically creates more HITs for tweets for which the agreement criteria in the place identification phase have not been satisfied, we blocked further processing of tweets where results did not seem to converge after getting more results than the initial five. With this threshold, agreement for place identification has been reached for 91.6% of the 6000 tweets in our corpus.

Disagreement between the workers can have many reasons and as argued by Wong and Lee [14] is potentially due to ambiguity in natural language, thus recording situations of disagreement can provide important information about that ambiguity. Wong and Lee report on a project directed to dealing with annotator disagreement in corpus construction. Specifically, they outline current approaches to dealing with disagreement that include: providing detailed annotation guidelines (thus making the annotators more expert), using expert adjudication (by a subject matter expert), discussion among annotators, removal of entities that agreement cannot be reached on, relaxing criteria that allow slightly different judgments to be counted as the same, and crowd wisdom (often used with AMT, in which it is possible to add annotators until some predetermined level of agreement is reached). With this background, Wong and Lee argue that the approach in most corpus building to remove all disagreement and keep only entities that sufficient agreement has been reached on is flawed as it fails to consider "legitimate" disagreement due to ambiguity in natural language

In the following, we take a look at the reasons for disagreement in place identification by establishing a novel classification scheme. After inspecting the tweets, we came up with the following categories of causes of disagreement between AMT workers, grouped into three main classes: tweeter errors, worker errors, and the most interesting category of disagreement based on differing definitions of place name.

3.2.1 *Tweeter Errors (TWE)*

This category refers to cases in which the disagreement is rooted in typos, etc. in the original tweet leading to uncertainty on the side of the AMT workers on whether something constitutes a place name that should be marked or not.

Example: *"Its been 20 years since the north ridge earthquake? Wow time flies"*

In this example the tweeter erroneously added a space in the name Northridge (and did not capitalize it). We believe that this is the reason why several workers did not tag it.

3.2.2 *Worker Errors*

This category refers to cases where workers make mistakes that clearly were against the instructions given to them. We distinguish the following subgroups:

Familiarity / unfamiliarity (FAM): In some cases, it was clear that some less well-known places were not recognized by some of the workers, resulting in disagreement.

Example: *"... help us find missing white large dog Leechan http://t.co/pxZY5wmpov Sendai #Japan. one of ..."*

We believe that those AMT workers who did not highlight Sendai in this tweet were not familiar with that name.

Misidentified tokens (MTN): Workers sometimes identified tokens as place names when the tweeter was not intending to refer to a place name.

Example: *"@lakeline @HockeenightsCT great. Now youve done it CT. Youve ..."*

The facts that "CT" appears in a user handle and the "Now youve done it CT" make it clear that CT refers to a person's name, rather than the state of Connecticut in the U.S. In the context of tweets it can sometimes be tricky or even impossible to decide about such a case.

Non-Merging errors (NME): Despite our instructions and the animated examples on the web site, several workers did not get the idea of having to merge the words making up a single place name or were unable to do so.

Example: *"... Earthquake Shakes Mexicos Pacific Coast: A strong earthquake has shaken the southern Pacific... "*

In this tweet, we had two workers that highlighted "Mexicos Pacific Coast" and "southern Pacific" but without any

word merging; each word was tagged as a place name by itself. This makes it pretty clear that they were not aware that they had to or not able to merge the words.

Skipped repeats (REP): We found cases in which the same place appears several times in the same tweet and some workers only marked one of these occurrences. Since we instructed the workers to "mark all place names", we feel that these cases belong into the worker errors category.

Example: *"... Text: No measles outbreak in Cavite - TRECE MARTIRES CITY, Cavite There is still no ..."*

In this example, some workers only highlighted either the first or second of the two occurrences of "Cavite".

3.2.3 Disagreement Based on Differing Definitions

This category refers to cases where reasonable people may disagree about what words refer to a place. They are of special interest for both practical and theoretical reasons. Practically, they contain the most common sources of disagreement, and so we need to understand them to maximise our yield of identifications for our corpus. Theoretically, examining these sources of disagreement can give us insight into how people refer to and recognize places, and help us make progress on the overall question, "what is a place?"

Adjectival usage (ADJ): A place name is used in an adjectival way to describe a different object rather than to refer to a place by itself.

Example: *"S'Sudan rebels reinforce captured city defences: South Sudan's rebels are strengthening ..."*

"South Sudan's rebels" is not a place and neither "S'Sudan", the way it is used here. In this tweet, the place name is used as an adjective to modify an object. While our instructions say that "Place names used as descriptors" should not be tagged, some workers did not take this into account.

Organization with a place in name (ORG): In these cases, a place name does not explicitly refer to a place but instead is part of the name of an organization or business.

Example: *"#Patriots How the Indianapolis Colts Defense Can Keep Tom Brady in Check http://t.co/TxBtIhHlSS"*

In this example, some workers highlighted Indianapolis despite our instructions saying that places in organization or business names should not be tagged including the example "Baltimore Ravens" which is very similar to this one.

Place hierarchy related issues (PHR): In our set of tweets with high disagreement, we find quite a few examples in which a place name further up in the place hierarchy (parent) is used to unambiguously identify another name (child). Typically, the place names appear separated by a comma, for instance <city>, <state> (e.g. "Hartford, CT").

Example: *"RT @NewEarthquake: 5.6 earthquake, 297km ENE of Grytviken, South Georgia and the ..."*

In the submissions of the AMT workers we find several versions of how such a case is handled. Some highlight both places, child and parent, both individually meaning they are not merged together. Some highlight both and merge them together presumably under the logic that it is actually one place that is referred to here. We also find workers that only highlight the child (Grytviken) most likely also thinking that this is the place referred to and the parent only occurs to clearly identify it. Lastly, some workers also only highlighted the parent (South Georgia). The reasoning behind this is less clear; it is possible that these are simply cases of unfamiliarity with the place name of the child (FAM).

Non-proper name (NPN): Places are often referred to without a proper place name in the designation. This results in these references being marked by only part of the workers.

Example: *"... Flash flood watch has now been trimmed back to just include the south shore through 9am."*

Here "south shore", while referring to a particular location, does not contain a proper name. Typically, in such situations it is extremely challenging from a natural language interpretation perspective to decide what specific place is referred to. Without more context, the reference refers to a class of places rather than a particular place.

Vaguely qualified place name (VQP): This describes a situation where there is a proper place name with an addition in front of it that narrows things down spatially, but at the same time introduces some vagueness.

Example 1: *"... wildfires in south #California: At least two structures burned to the ground an..."*

Example 2: *"... Theres an outbreak of measles in the Bronx and upper Manhattan, officials warn: ..."*

The qualifiers ("south" and "upper") narrow things down spatially to a smaller area within the spatial extension of the properly named place. It also becomes less clear which exact area is being referred to. Workers either consider the place name to consist of the complete construct ("south #California" in the first example, "upper Manhattan" in the second) or highlight just the proper name without the qualifier.

Inclusion or exclusion of kind of place (KND): We often find the name of a place in a combination with a noun describing its kind, such as in City of London, Island of Saint Louis, South Georgia Island, Ganges River, etc.

Example: *"... How Adelia, 9, and the island of Sabang, #Indonesia, triumphed over #malaria ..."*

Part of the AMT workers here highlight and combine "island of Sabang" and the other part only Sabang. This reflects the uncertainty on whether the kind noun is part of the place name or not. For many physical features such as lakes, mountains, etc., it is clear that the noun has to be considered part of the proper name. This is particularly important from a geoparsing perspective where otherwise it would be impossible to distinguish the physical feature from a similarly named populated place, e.g. Lake Michigan vs. the state Michigan. For other entities such as cities, villages, or states, it is often difficult to make a decision even for geographic experts, and it requires some background research.

Hashtag related issues (HAS): While we give clear instruction that "Place names in hashtags" should be tagged using the example #newyork, it turns out that there are many more cases not clearly covered by these instructions where the place name is only part of the hashtag.

Example: *"... everyone please pray for the people there. #prayforindonesia"*

Since in our current tool there is no way that would allow workers to highlight part of a token, some workers highlighted the hashtag #prayforindonesia because of the presence of Indonesia, while others did not highlight it. In addition to that, we can find in hashtags many of the other cases like organization names containing place names in the hashtag (ORG), adjectival usage (ADJ), etc. In principle, this category can occur in combination with any of the others.

Below we summarize for each category roughly how many of the contentious tweets had submissions falling into that category (several categories per tweet are possible).

TWE	FAM	MTN	NME	REP	ADJ
1%	32%	5%	20%	6%	23%
ORG	PHR	NPN	VQP	KND	HAS
16%	11%	5%	10%	6%	16%

3.3 Resolving the Disagreements

Resolving the disagreements of the AMT workers to proceed with the geocoding step and include these tweets in the corpus is crucial for avoiding an undesired bias. Many of these tweets are particularly interesting cases for evaluating or training geoparsing tools as they tend to contain several place names, often in a spatial relationship, e.g., category PHR. There are different possibilities of how to deal with annotator disagreement (see again [14]). We here consider the following three options, whose suitability depends on the disagreement categories introduced previously: (1) manual resolution of the disagreements by experts that decide (ideally collectively) what are correct submissions (according to our own rules) and the ground truth result for the controversial tweets; (2) automatic resolution via algorithms that can correct or discard incorrect submissions (according to our own rules plus the set of entity designations by workers); and (3) introducing new rules to instruct workers more clearly and collect more results via AMT.

These measures can also be combined and probably will have to be. New instructions alone will probably not be sufficient in all cases, as we found errors in the current results that violate some of the instructions already given to the workers. Discarding submissions that are incorrect (according to our own rules), either manually or where possible automatically, can significantly reduce the number of new submissions needed to achieve the specified majority.

Starting with the tweeter errors (TWE), a decision has to be made on whether misspelled place names should count and be highlighted. We believe that including misspelled names is an appropriate strategy because robustness against typing errors is a desirable property in geoparsing tools. On the other hand, these errors are sometimes extremely difficult to detect automatically. In these cases, where autocorrection of spelling is impractical, manual resolution and/or new instructions to AMT workers are the best options.

In the case of worker errors, unfamiliarity reasons (category FAM) and misidentified tokens (category MTN) can obviously not be resolved by improved instructions. Automatic resolution based on a geographic gazetteer is also problematic as this is a form of geoparsing and could lead to a bias in the corpus. Hence, manual resolution by experts seems to be the only viable approach in these cases. As we discussed, we already provide explicit instructions that should prevent non-merging errors (category NME) and skipped repeats (category REP) but we do not completely achieve these goals. On the positive side, these are categories of worker errors where automatic resolution seems to be a promising approach. In the case of NME errors, a sequence of tokens that is merged and highlighted by the majority of workers who processed this tweet and that was highlighted but not merged by the rest can be easily detected and the latter treated as if they were merged.

This leaves the categories based on different definitions of what constitutes a place. For adjectival usage (category ADJ) and for organization with a place in the name (category ORG) we already provide instructions that say that these should not be marked as place names, as it is not the place described by that name that is being referred to in the tweet. The instructions and examples for these cases could potentially be improved but there are hard cases where it is difficult to decide and where simple rules will not always work. Since automatic resolution in these cases is extremely challenging from an algorithmic and natural language processing perspective, manual resolution seems to be the most suitable approach to get results quickly. For place hierarchy issues (category PHR), we have made the decision that we want both place names (child and parent) highlighted separately without merging. Fortunately, automatic resolution techniques are applicable here as these cases are rather easy to detect and resolve. Moreover, we expect that giving explicit instructions about this case to the AMT workers will lead to a substantial reduction in these errors.

The remaining categories have in common that it is not straightforward to decide what the preferred worker behavior is with regard to whether something should be tagged or not. Expressions such as “south shore” or simply “South West” falling into the NPN category refer to a place, but not with a proper place name. In the context of geoparsing, we believe it is more reasonable to restrict oneself to proper place names. We, therefore, do not want these references to be tagged. Unfortunately, detecting NPN cases automatically is extremely challenging. However, we expect that providing clear instructions with examples that tell AMT workers to only highlight proper place names should reduce the frequency in which this case occurs significantly.

Vaguely qualified place names (VQP) such as “south California” or “upper Manhattan” is another case where one can argue about what the best approach is. One can either take the standpoint that geoparsers should essentially process the proper name without the vague qualification, which is extremely difficult to interpret in a geometric sense anyway, or attempt to deal with the entire phrase in a reasonable way, e.g. by indicating or visualizing that the identified proper place has been qualified further. We are currently leaning towards focusing on the proper place name for training and evaluation purposes and treat the processing of additional qualifications as an extra feature of a geoparser whose evaluation is outside of the scope of this corpus. Fortunately, it is straightforward to provide clear instructions on the desired worker behavior in future collections as well as to automatically resolve (most of) these cases. This also opens up the option to introduce a special labeling of these cases distinguishing the proper name and the qualification component such that the corpus could also be used specifically to benchmark this aspect in a geoparser.

As we already briefly discussed, the inclusion or exclusion of kind of place (KND) such as “City of London” or “Lake Michigan” is another difficult case even experts might struggle with. We believe that for physical features (mountains, rivers, etc.) these should be tagged as part of the proper place name but we are less certain about populated places (“city of”, “village of”, etc.) and similar cases. A reasonable approach might be to instruct workers to include them but allow for some flexibility when comparing the output of a geoparser with an instance from the corpus in a benchmark-

ing scenario. It should be possible to resolve the majority of these cases automatically and the same holds true for implementing some special treatment of these for training.

In the case of place names in hashtags (category HAS), it is our opinion that they should be tagged if they contain place names, unless they fall into one of the categories discussed above where we argued these should not be tagged. That means that in cases where several place names are combined in a single hashtag, several places need to be associated with a single token as in #PrayForSerbiaAndBosnia. As a future refinement it would make sense to provide workers with a way to mark which parts of a hashtag constitute a place name. Improving the instructions to cover the case that a hashtag consists of more than just a place name should improve the results we get from the workers. Moreover, it should be possible to detect and resolve at least some of these cases automatically in the current corpus.

4. CONCLUSIONS & OUTLOOK

We have presented our initiative of constructing a GeoMicroblog Corpus that we plan to publish in the near future to facilitate the training, evaluation, and comparison of geoparsing tools and promote GIR research. The details of releasing the corpus are still under discussion due to legal issues and other considerations that need to be taken into account. The corpus building framework we described adopts geovisual analytics and crowdsourcing components and we expect it to play an important role in further corpus building work, for instance for other kinds of text or for building more domain specific or localized corpora.

The analysis of submissions in the place identification stage brought several interesting reasons for disagreement to light. Based on the insights gained, we plan on evaluating three strategies for making tweets with high disagreement useful for our corpus and increasing the yield of usable place name references in future collections. These are manual decision making by experts, improving instructions and collecting more results, and automatic resolution methods.

Furthermore, one of the next steps will be to apply the corpus to quantitatively evaluate the performance of our own geoparsing web service GeoTxt [7]. In addition, we plan to use the results from the place identification stage to train the NER tools in our system. We expect new insights from these activities which should lead to improvements of the corpus and corpus building framework. Finally, we are going to look into ways to actively disseminate the corpus in the research community and discuss options for establishing a benchmarking initiative or competition in the field.

5. ACKNOWLEDGMENTS

We thank the anonymous reviewers for very valuable feedback and suggestions. This work was partially funded by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009- ST-061-CI0003.

6. REFERENCES

[1] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17:124–147, 2012.

[2] C. D'Ignazio, R. Bhargava, E. Zuckerman, and L. Beck. Cliff-clavin: Determining geographic focus for news. In *NewsKDD: Data Science for News Publishing, at KDD 2014*, 2014.

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

[4] R. Guillén. Geoparsing web queries. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152, pages 781–785. Springer, 2008.

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

[6] C. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–388, 2002.

[7] M. Karimzadeh, W. Huang, S. Banerjee, J. O. Wallgrün, F. Hardisty, S. Pezanowski, P. Mitra, and A. M. MacEachren. GeoTxt: A web API to leverage place references in text. In C. Jones and R. Purves, editors, *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pages 72–73, 2013.

[8] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5), 2013.

[9] J. L. Leidner. An evaluation dataset for the toponym resolution task. In *Computers, Environment and Urban Systems*, volume 30, 2006.

[10] J. L. Leidner. *Toponym Resolution in Text*. PhD thesis, School of Informatics, University of Edinburgh, 2007.

[11] A. MacEachren, A. Jaiswal, A. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: GeoTwitter analytics support for situational awareness. In S. Miksch and M. Ward, editors, *IEEE Conference on Visual Analytics Science and Technology*, 2011.

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

[13] M.-H. Tsou, J.-A. Yang, D. Lusher, S. Han, B. Spitzberg, J. Gawron, D. Gupta, and L. An. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing). *Cartography and Geographic Information Science*, 40:1–12, 2013.

[14] B. Wong and S. Lee. Annotating legitimate disagreement in corpus construction. In *Sixth International Joint Conference on Natural Language Processing*, 2013.

[15] D. Woodward, J. Witmer, and J. Kalita. A comparison of approaches for geospatial entity extraction from Wikipedia. In *IEEE Fourth International Conference on Semantic Computing (ICSC)*, pages 402–407, 2010.