A Multi-Indicator Approach for Geolocalization of Tweets

Axel Schulz[1,2], **Aristotelis Hadjakos**[2],
Heiko Paulheim[3], **Johannes Nachtwey**[1], and **Max Mühlhäuser**[2]

[1]SAP Research, Darmstadt, Germany
[2]Telecooperation Lab, Technische Universität Darmstadt, Germany
[3]Data and Web Science Group, University of Mannheim, Germany

Abstract

Real-time information from microblogs like Twitter is useful for different applications such as market research, opinion mining, and crisis management. For many of those messages, location information is required to derive useful insights. Today, however, only around 1% of all tweets are explicitly geotagged. We propose the first multi-indicator method for determining (1) the location where a tweet was created as well as (2) the location of the user's residence. Our method is based on various weighted indicators, including the names of places that appear in the text message, dedicated location entries, and additional information from the user profile. An evaluation shows that our method is capable of locating 92% of all tweets with a median accuracy of below 30km, as well as predicting the user's residence with a median accuracy of below 5.1km. With that level of accuracy, our approach significantly outperforms existing work.

Introduction

Twitter has become a very popular microblogging platform during the last years with more than 400 million tweets created per day.[1] Research has shown that tweets provide valuable real-time information, e.g., for opinion analysis preceding political elections (Tumasjan et al. 2010), for regional health monitoring (Aramaki 2011), or local emergency detection (Starbird et al. 2010). However, according to recent analyses (Hale and Gaffney 2012), only around 1% of all tweets are explicitly geotagged. Thus, without a possibility to predict the location of tweets, 99% of all tweets cannot be used for the above mentioned purposes.

Simple approaches to determine the location of a tweet are not applicable: The location cannot be estimated using the IP address of a user's device, as neither Twitter nor the telecommunication provider will allow access to that information for application programmers. Twitter's Search API, which provides spatial filters, relies solely on user profiles, which are often incomplete and incorrect (Hecht et al. 2011).

Extracting location information by other means is challenging: Place names (also called *toponyms*) have to be identified in the tweet message and in the tweet's metadata. Tweets are short (max. 140 characters) and often consist of non-standard language. This results in people using different names or abbreviations for locations, e.g., *LA*, *L. A.*, *City of Los Angeles*. Furthermore, for mapping toponyms to locations, two general problems have to be solved (Lieberman, Samet, and Sankaranarayanan 2010): First, a toponym can refer to multiple geographic locations (Geo/Geo disambiguation), e.g., *Paris* is referring to 23 cities in the USA. Second, a toponym can relate to entries that can refer to a spatial location but also a person or a thing (Geo/Non-geo disambiguation). E.g., *Vienna* may refer to a city as well as to a person; *as* is used as an adverb but may also refer to a city in Belgium; or *metro* may reference a city in Indonesia, a train, or a company. This disambiguation is called *toponym resolution* (Leidner 2004) and is one of the major challenges when dealing with location information in microblogs.

For resolving ambiguous toponyms, it is helpful to leverage different indicators. For example, for distinguishing the European country *Norway* from the equally named city in Australia, the time zone may be a helpful additional indicator. Thus, we propose a multi-indicator method to solve the disambiguation problem in microblogs. Our approach can be used for determining (1) the location where a tweet was created and (2) the user's residence. Our proposed method consists of four steps:

1. Detection of spatial indicators: Spatial indicators are location information that allow geolocalization. Our method spots spatial indicators in the text message and in the user profile of a tweeter.

2. Geographical interpretation: Each spatial indicator refers to (at least) one geographical area. We determine that area and represent it with a polygon.

3. Weighting: As some spatial indicators are more reliable than others, we attribute a variable height to each polygon. The height is computed based on weights determined using an optimization algorithm and the reported uncertainty of the spatial indicator for the currently analyzed case.

4. Stacking: By intersecting and stacking the 3D polygons over each other, a height map is built. The highest area in this height map is then used for geolocalization.

An evaluation shows that this method is capable of lo-

[1] http://www.theverge.com/2012/6/6/3069424/twitter-400-million-total-daily-tweets

cating 92% of all tweets with a median accuracy of below 30km, as well as predicting the user's residence with a median accuracy of below 5.1km. With that level of accuracy, our approach significantly outperforms existing work and is the only combined approach that allows user **and** tweet geolocalization.

The rest of this paper is structured as follows: An introduction to spatial indicators in tweets is given in Section 2. Related work is presented in Section 3. Our method is presented in Section 4. In Section 5, we evaluate our method and compare it to state-of-the-art approaches. An example application is shown in Section 6. We conclude with a discussion and future work in Section 7.

Spatial indicators in tweets

Spatial indicators are pieces of information that help us locating a tweet. Twitter users provide many spatial indicators in their messages and in their profile. The message text, user account information, website links, current time zone, a dedicated location field, and sometimes even accurate GPS coordinates determined by the user's mobile device may all be part of a tweet.

Tweet Message

The text message of a tweet is at most 140 characters long, is unstructured, and often written in non-standard language. Extracting location information from the message is difficult as proper place names are seldom used while abbreviations and nicknames are more common. Furthermore, the toponyms may or may not refer to the user's current location as he could write about a place he is on the way to or where he would like to be. A tweet might even include more than one location in the text, e.g., *I'd love go to Hawaii or Mauritius*.

Links included in tweets might reference geotagged pictures on Flickr or other links from location-based services. E.g., Foursquare allows to "check-in" at a venue resulting in the creation of a tweet with accurate location information. In our data set we found links to many location-based services such as Foursquare[2] and Ubersocial[3]. As these location-based services are commonly used to inform about the user's current location, the linked web pages can be used as spatial indicators.

Profile Information

Twitter users can maintain a personal profile. Furthermore, Twitter adds further information about the user and the tweet. All this information is available as metadata for each tweet, in particular:

Location field: Users can specify their home location(s) in the location field. The entries in the location field are heterogeneous; the user may, e.g., provide their home country or their state (Gelernter and Mushegian 2011). Furthermore, abbreviations are commonly used, like *NY* for *New York* or *MN*, which may stand for *Minnesota*, but also for the country *Mongolia*. Most of these location entries have a relatively large geographic scope, like *California* or *UK*. Besides real location information, the location field is also used for sarcastic comments or fake location information like *Middleearth* (which is an actual city, but mostly used as the fantasy place).

Hecht et al. (2011), who did the first in-depth study of the location field, showed that only 66% of the entered information have a valid geographic information. Furthermore, they showed the reflection of current trends like Justin Bieber as part of the location field (e.g. *Biebertown*). Also, 2.6% of the users enter multiple locations. GPS coordinates are part of the location field too, either in decimal or DMS (degrees/minutes/seconds) notation. Mostly, these GPS coordinates are provided by mobile devices or mobile applications. Besides correct coordinates, there are also parts of coordinates or IP addresses that could prevent easy parsing.

Websites: In their profile, twitter users may provide links to web pages which may, e.g., contain personal information. People provide links to Twitter, Facebook, or other social network pages as well as personal websites. Both the website's country code and the website's geocoded IP address are spatial indicators.

Time Zone: The time zone entries in the user's profile describes a region on earth that has a uniform standard time. The time zone is initially set by Twitter and can manually be adjusted by the user. It is typically represented by a city, which is often the capital city of the user's home country, e.g., *London*. On the other hand, the time zone can also describe a larger region without an explicit capital mentioned, e.g., *Eastern Time (USA&Canada)*.

UTC24-Offset: UTC is the time standard used for many World Wide Web standards. 24 main time zones on earth are computed as an offset from UTC, each time zone boundary being 15 degrees of longitude in width, with local variations. Therefore, the UTC offset only is a means for differentiate a location by longitude compared to the much more precise time zone information.

Coordinates and Place: Depending on the privacy settings, tweets may also contain location information as latitude/longitude coordinate pairs. The coordinates are set when the user tweets from a device with enabled GPS. These device locations are difficult to be changed and manipulated and can be seen as a very good approximation of the user's position when sending a tweet. Furthermore, Twitter provides an approximate location specified as a bounding box. For creating this bounding box, Twitter uses the user's IP address for creating the approximation.

Related Work

Identifying the geographical location of digital content is a field of extensive research. There are methods to identify the geographic location of digital text documents (Smith and Crane 2001), web pages (Zong et al. 2005), blogs and news pages (Lieberman, Samet, and Sankaranarayanan 2010), and Flickr tags (Popescu and Grefenstette 2010). The work focused on Twitter can be differentiated in three dimensions: The spatial indicators used, the techniques applied, and the localization focus.

[2] http://foursquare.com

[3] http://www.ubersocial.com

Table 1: Overview of related approaches. Spatial indicators and techniques marked with (X) were used for creating baselines or were part of the background analysis.

	Spatial Indicators				Techniques		Localization Focus		
	Text	Location Field	Social Network	Other	NLP	Gazetteer	User's Residence	Tweet Location	Message Focus
(Eisenstein et al. 2010)	X				X		X		
(Hecht et al. 2011)	X	(X)			X	X	X		
(Cheng, Caverlee, and Lee 2010)	X	(X)			X	X	X		
(Chandra, Khan, and Muhaya 2011)	X		X		X		X		
(Gelernter and Mushegian 2011)	X				X				X
(Sultanik and Fink 2012)	X				(X)	X			X
(Ikawa, Enoki, and Tatsubori 2012)	X				X			X	
(Kinsella and Murdock 2011)	X	(X)			X	(X)	X	X	
(Hong et al. 2012)	X				X			X	
(Paradesi 2011)	X					X		X	
(Hale and Gaffney 2012)	X	X		(X)		X		X	
(Abrol and Khan 2010)	(X)	(X)	X		X	(X)	X		
(Takhteyev, Gruzd, and Wellman 2012)		(X)	X		X	(X)	X		
(Clodoveu et al. 2011)			X			(X)	X		
(Mcgee, Caverlee, and Cheng 2011)		(X)	X			(X)	X		
(Gonzalez et al. 2011)			X			(X)	X		
(Sadilek, Kautz, and Bigham 2012)			X	(X)	X		X		
(Krishnamurthy and Arlitt 2006)				X	-	-	X		
(Bouillot, Poncelet, and Roche 2012)	X	X		X		X		X	
(MacEachren et al. 2011)	X	X		X		X		X	
Our Approach	X	X		X		X	X	X	

Spatial Indicators

Different information sources are used for geolocalization purposes. The message text is used most of the times, for instance, the approaches proposed in (Cheng, Caverlee, and Lee 2010), (Eisenstein et al. 2010), (Hecht et al. 2011), or (Kinsella and Murdock 2011) use language models based on the terms in the tweet message. Chandra, Khan, and Muhaya (2011) extend these approaches by taking the relationships of the users into account. Gelernter and Mushegian (2011), Sultanik and Fink (2012), and Paradesi (2011) apply named entity recognition to annotate tweet messages and preprocessing to handle the disambiguation problem. Ikawa, Enoki, and Tatsubori (2012) also use a language model, but in this case, they analyze only keywords from messages created by location-based services like Foursquare. The algorithm of Hong et al. (2012) is based on the words a user uses in his tweets. They show the advantage of identifying topical patterns for geographical regions. Furthermore, Hale et al. (2012) analyze if the language of the message text can be used for geolocalization. They conclude that the language is not an appropriate indicator.

Besides the message the location field is used for location estimation. Hecht et al. (2011) provide an in-depth analysis of the location field. As a result, they conclude that the location field alone does not provide enough information for geolocalization. Hale and Gaffney (2012) analyze different geocoders for identifying the location where a user is tweeting from based on the location field.

Instead of the directly usable information of the message or the location field, the relationships of the users are also useful for geolocalization. Abrol and Khan (2010) try to identify the location of a user based on his/her social activities. Takhteyev, Gruzd, and Wellman (2012), Clodoveu et al. (2011), and Mcgee, Caverlee, and Cheng (2011) analyze the relationship between a pair of users and the distance between the pair. Gonzalez et al. (2011) focus on the follower relationship and report that in countries like Brazil there is a high intra-country locality among users, while in English-speaking countries the external locality effect is higher. The approach of Sadilek, Kautz, and Bigham (2012) is also based on the relationship between users, but in this case, the GPS tags are also used for location inferencing.

Krishnamurthy and Arlitt (2006) use the UTC offset information to get a user's local time and thereby an approximate longitude. They compare their results to the top-level domains of the URL of a user. Users with URL in the *.com* domain are distributed around the world, while the rest of the UTC data is lined up with the domain information.

Several approaches propose the combination of different information sources. Bouillot, Poncelet, and Roche (2012) propose an approach based on different aspects of user information, like the message, the location field as well as the language for homonym differentiation. MacEachren et al. (2011) developed an application that leverages the geocoded location field, the timezone, hashtags and named entities from the tweet for geolocalization and geovisual analytics of tweets in crisis management. None of these approaches provide quantitative evaluation results for geolocalization.

Techniques

The approaches can be divided into methods mainly based on natural language processing (NLP) that do not use external information, and approaches based on geographical dictionaries (gazetteers). NLP-based approaches are especially used to estimate the location using language models and context information about the user. On the other side, the gazetteer approaches use geocoders to determine the place that is being referred to. This approach cannot find information easily that is not present in the gazetteer, but needs no training data and is much simpler. Gazetteers have also been used several times by the NLP-based approaches on the location field for creating a baseline or training the models.

Localization Focus

All analyzed approaches pursue different goals: As some try to predict the home location of the user, other approaches predict the location where the tweet was sent or the location of what the user is tweeting about. This differentiation is clearly necessary depending on the use case: in emergency management, it is relevant what place a message is related to, for location-based services, the location where a Tweet is created is relevant, and for market research, we rather focus on the user's home location.

Discussion

Table 1 provides an overview on the related approaches. Language models are used for localization of the user as well as for localization of the tweet. On the other side, the social network is only used for predicting the user's residence. Except the language model of Kinsella and Murdock (2011), there is no approach for detecting both the user's residence and the location where the tweet was sent. The advantage of using different information sources at once, e.g., language information as well as place names from the location field and the message, has been shown several times by (Bouillot, Poncelet, and Roche 2012; MacEachren et al. 2011; Hecht et al. 2011).

Our method is innovative in several aspects compared to related work. To our knowledge, it is the first multi-indicator approach using a vast variety of spatial indicators to solve geolocalization problems on tweets. We are able to determine the location of the tweet as well as the location of the user's residence by taking the message, the location field and further metadata into account.

Multi-Indicator Approach

In order to estimate the location of a tweet, we use a variety of spatial indicators. This section presents how spatial indicators are combined to form a single geolocation estimate and discusses how the spatial indicators are extracted from tweets.

Combining Spatial Indicators

Properties of spatial indicators: Usually one or a multitude of spatial indicators can be extracted from a single tweet. In order to successfully combine the spatial indicators, it is necessary to understand their basic properties:

- **Contradiction:** The spatial indicators extracted from a tweet can coincide (e.g., location field: *Paris*, message: *Nice weather in Paris*) or they can be contradictory (e.g., location field: *Paris*, message: *Nice weather in Athens*).
- **Scale:** The spatial indicators can relate to areas of different scale. Consider for instance the spatial indicators *France* and *Eiffel Tower* that may occur together in a twitter message and which represent geographical areas of vastly different size.
- **Ambiguity:** As discussed above, spatial indicators are ambiguous, such as the different cities called *Paris*. Further ambiguity may come because of spelling errors, the use of abbreviations, incomplete information, and slang. Gazetteers usually provide a list of different geographical interpretations of a geographical name with ratings of their uncertainty, e.g., based on the edit distance of a misspelled city name.

Polygon mapping: Simple solutions to combine the spatial indicators into a single location estimate, such as computing the average of the coordinates given by each spatial indicator, are bound to fail due to problems with contradiction, scale, and ambiguity. In order to get a good combined estimate we adopted the approach of Woodruff and Plaunt (1994) for localizing bibliographical text documents, which is based on intersecting the geographical outlines of the geographical areas that the spatial indicators refer to. These geographical outlines are represented by polygons. The mapping from spatial indicators to polygons is either done directly by the resolution method itself or indirectly using coordinate pairs that are provided by the resolution method and mapping to an appropriate surrounding area in a spatial database (see below).

Polygon height: To arrive at a uniform prediction, a height is attributed to each polygon, making it a three-dimensional shape. The height allows for modeling the uncertainty that may come with a spatial indicator. This uncertainty can be an outcome from the method itself, which may sometimes make wrong prediction, or from inherent inaccuracy of a spatial indicator, e.g., the time zone indicator. Therefore, the final polygon height is determined based on two factors: First, it is based on the quality of the resolution method that was used. Based on our evaluation results, we assign a quality factor Q_{ext} to each method based on how well it contributes to predict the tweets location. The value Q_{ext} is determined using the simplex method of Nelder and Mead (see below). In addition to this "external" quality measure, many methods also provide an internal assessment of the quality when more than one alternative is suggested.

The internal quality measure $Q_{\text{int}}(x)$ provides an estimation for the quality of the x-th alternative. Since different resolution methods have vastly different scales when reporting this internal quality measure, they are normalized to a $[0, 1]$ interval. Resolution methods returning only one result are rated with $Q_{\text{int}} = 1$. The height h of the polygon representing the x-th alternative is then computed as $h(x) = Q_{\text{ext}} \cdot Q_{\text{int}}(x)$.

Polygon stacking: Once all three-dimensional polygon shapes are determined, they are stacked one over the other and form a height profile (see Figure 1). The highest area in that height profile is then found and its polygon outline is determined as the intersection of the contributing polygons. The geolocation is estimated as the geometric center of that area as a coordinate pair.

Extracting Spatial Indicators

The different spatial indicators described in Section are extracted from tweets as follows:

Tweet Message: From the text, we extract entities using DBpedia Spotlight (Mendes et al. 2011), a named entity recognition service which identifies entities in texts and maps them to DBpedia (Bizer et al. 2009) entities (**SP**). For example, in the tweet *yeah, watching muse at fedex field!!!*,

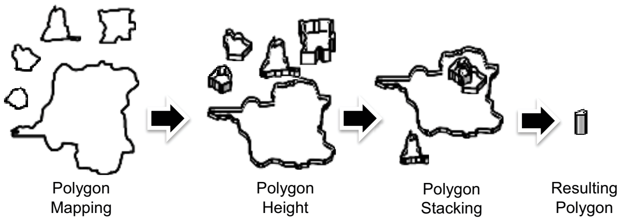

Figure 1: The height profile is determined by stacking the three-dimensional polygon shapes over each other.

the text *muse* is recognized as a named entity and mapped to http://dbpedia.org/resource/Muse_(Band), as well as *fedex field* is mapped to http://dbpedia.org/resource/FedEx_Field. As most geographic entities (such as *FedEx Field*) have coordinates in DBPedia, we use those coordinates for polygon mapping. Entities without coordinates (such as *Muse*) are discarded. For calculating $Q_{\text{int}}(x)$ we use the confidence values provided by Spotlight.

For processing information from location-based services (**LBS**), we analyzed our dataset for the occurrences of the most common services. In this case, we extract coordinates from UeberSocial, TrendsMap, Flickr, Roketatchi and Foursquare based on the information provided on the web page. For every location-based service we identify the coordinates based on predefined patterns, e.g. for Foursquare, we use the meta tags referencing the venues and corresponding location information.

Location Field: For toponym resolution in the location field we use Geonames[4] (**GN**). Geonames is a gazeteer that contains more than 10 million entries about geographic entities in different languages. This includes countries, cities as well as building and street names. Using the full text search, Geonames returns a list of possible results with a confidence score, which we use for calculating $Q_{\text{int}}(x)$. As Geonames is not able to resolve all location field entries directly, we preprocess the entries in different steps if no results are returned:

- Geonames has problems processing unaligned text segments like *Paris, France*. We solve this by text preprocessing (**GN-1**).
- We extract several toponyms from the location entry (**GN-2**). First, as lot of the location field entries contain separators like '—', e.g., *Salvador — Bahia — Brasil*, we split this entry into a list of entities. Furthermore, more general location information is often provided in brackets, e.g., *Berlin, Germany (Europe)*. In this case, we extract the content of the bracket and try to resolve the first comma group and the bracket itself in Geonames.
- As gazeteers often have problems with city-level entities like local places and their nicknames, we use DBPedia Spotlight to annotate the entry in the location field (**GN-3**). In this case, commonly used nicknames like *The Big Apple* can be retrieved.
- As a last means for extracting toponyms, we split the whole location entry into a list of words (**GN-4**). Every word is then sent to Geonames.

As previously mentioned, coordinates are also part of the location field. For extracting these, we use regular expressions to identify them in decimal or the DMS notation (**COD**). As location-based services do not follow a common pattern for setting coordinate entries, regular expressions have been adapted to match most of the common cases. For instance, analyzing entries of location fields in DMS notation show that numbers are set before the cardinal direction as well as behind.

Website: To handle the website entries, we follow a twofold approach. First, we extract the top-level domain using a regular expression (**WS-1**). The top-level domains are then matched against country codes using a manually created mapping of country codes and the corresponding country names. *.com*, *.net*, *.org* are not processed in this case, as they do not provide any helpful location information (Krishnamurthy and Arlitt 2006). To provide estimations for these cases, we also extract the IP addresses using the host names (**WS-2**). Coordinates are then retrieved using IPinfoDB[5].

Time Zone: Our analysis of the different time zone entries has shown that these are mostly provided in a standardized format stating the capital of the home country. Besides these kinds of entries, United States and Canadian time zone entries are also present like *Central Time (USA&Canada)*. The provided time zone entries can be used directly as they are machine-generated (**TZ**).

Mapping to Polygons

To enable the mapping of geocoordinates to polygons we built a spatial database with polygons suitable for every spatial indicator.

Tweet Message and Location Field: For mapping the coordinates retrieved from the message and the location field, we use polygons of the world's administrative areas. E.g. the Bronx can be retrieved as part of the administrative districts of New York City allowing us to narrow down our estimation as good as possible. The polygons used for this were retrieved from the GADM database of Global Administrative Areas.[6] For mapping coordinates retrieved from location-based services, we use a circle of 100m radius around the position.

Website: As the website entries might relate to the home country of the user, but not the home town, we use country polygons for mapping the website entries. In this case the polygons are retrieved from ThematicMapping.[7] The extracted country names from the top-level domains are then matched to the polygons representing the world borders.

Time Zone: For mapping the time zones, we use polygons retrieved from the IANA Time Zone Database.[8] In this case, the polygons for the time zones of the US, Canada, Russia and China have been aggregated manually, as they

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

[5] http://ipinfodb.com/

[6] www.gadm.org

[7] http://thematicmapping.org/downloads/world_borders.php

[8] http://efele.net/maps/tz/world/

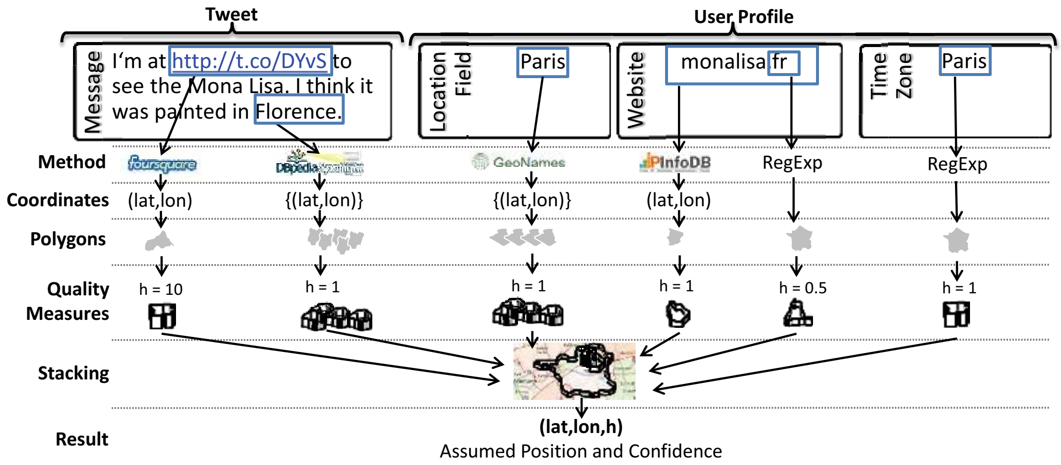

Figure 2: Example pipeline for our approach: Spatial indicators are identified based on the methods described. The results are either a pair of coordinates *(lat,lon)* or a set of coordinates and quality measures. The coordinates are mapped to the corresponding polygons. Then the external quality measures are applied before conducting the stacking. As a result we estimate the location of the tweet with a confidence value.

are not present in the initial dataset. Furthermore, the polygons for the time zones spanning multiple countries like the *Central Standard Time* (CST) or *Pacific Standard Time* (PST) were created manually based on the regions contained in the corresponding time zone.

Determining External Quality Measures

Since not all indicators are expected to perform equally well, we assign a quality measure to each method based on how well it contributes to predict the tweets location. To determine a good external quality measure of each approach we use the downhill simplex method of Nelder and Mead (1965). To apply the method, we regard the weight of each method as a variable of our objective function. For tweet geolocalization the objective function is the mean squared error of all distance estimations compared to the device location. With the optimization method, we are able to calculate a local optimum for minimizing the objective function.

To calculate the optimal solution, we use a hold-out sample set of 10,000 randomly chosen tweets from our test set (see below) and calculate their distances. The external quality measures for good spatial indicators like the coordinates and the location-based services indicators are high (LBS=4.26, COD=2.72). The first three Geonames optimizations are valuable compared to the plain Geonames approach (GN=1.51, GN-1=2.01, GN-2=1.67, GN-3=1.96). The fourth optimization, which processes every word in the location field is not useful for geolocalization (GN-4=-0.54). Processing the time zone as well as the top-level domains is also contributing to the overall result (TZ=1.12, WS-1=1.07) as well as the message processing based on DBPedia Spotlight (SP=0.87). Using the IP addresses does not provide valuable estimations (WS-2=-2.32). The final weights are shown in Table 2.

Discussion

Figure 2 illustrates how an example tweet consisting of several spatial indicators is processed using our approach. As a result of this process, we estimate the location of the tweet with a confidence value.

Compared to other approaches, our way to stack 3D polygon shapes over each other differs significantly. Our method is a multi-indicator approach that allows combining vastly different spatial indicators such as the user's time zone and check-ins from location-based services. Due to the variety of indicators used, our method is less vulnerable to missing or incomplete data, which is rather commonplace in tweets. E.g., the location field may be blank or the text message might not contain any information about the user's location, but our method can in most cases find a reasonable estimate of the tweet's location based on the other spatial indicators.

In contrast to other approaches, we combine a variety of indicators using *optimized weights* obtained by a thorough evaluation on a large and diverse test set. Furthermore, since none of the techniques relies on English as an input language (although, e.g., DBpedia Spotlight has to be specifically configured and deployed to work on other languages), our approach is applicable to arbitrary languages in principle.

Evaluation

We conduct the evaluation of our method on the publicly available Twitter feed. First, we evaluate the performance of each single spatial indicator for geolocalization of the tweet. Second, we measure the performance of the overall approach for geolocalization of a tweet as well as for determining the user's residence.

Test Data: From September 2011 to February 2012 we crawled around 80 million tweets from the *Spritzer* stream using the Twitter Streaming API[9]. The Spritzer stream is a

[9] https://dev.twitter.com/docs/streaming-api/methods

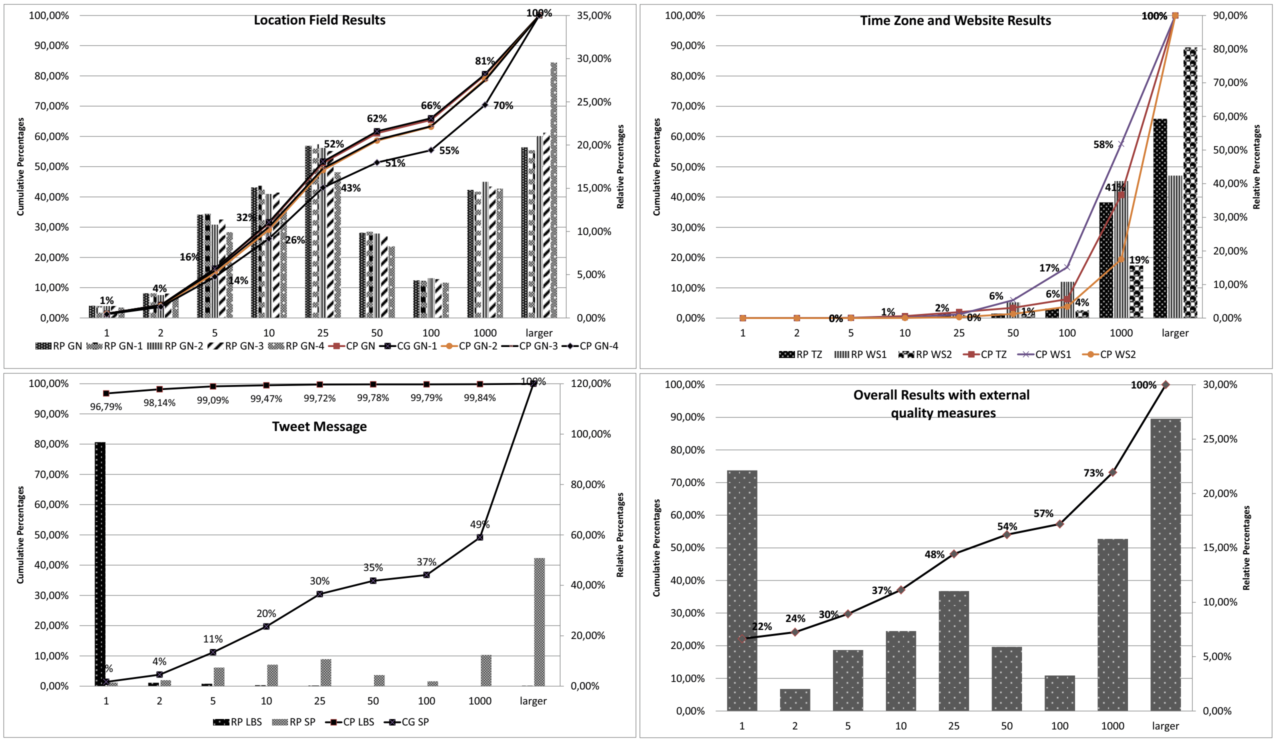

Figure 3: Overview of the evaluation results for the location field approaches (top left), the time zone and website results (top right), the approaches based on the message (bottom left) and the overall results with external quality measures (bottom right). The right y-axis of each chart provides the relative percentages while the left one shows the cumulative percentages. The x-axis shows the distance in km.

feed of 1-2% tweets of a random selection of all public messages. From these tweets, we extracted 1.03 million messages with device locations to use them for our evaluation. No further preprocessing has been applied on the sample set to keep it representative for a real-world scenario. For implementing our approach, we used 10% randomly selected tweets from the dataset with device locations for tuning the identification of spatial indicators and 90% for testing.

Metrics: To evaluate our approach, we compare the coordinate pair estimation of our approach with the device location. As error metrics we provide the *Average Error Distance* (AED), the *Median Error Distance* (MED), and the *Mean Squared Error* (MSE) to ensure comparability to related work. We also report the *Recall*, which is the number of tweets with identified spatial indicators compared to the amount of all tweets.

Results: Single Spatial Indicator

We investigated the performance of our approach for geolocalization of the tweet. In this case, we evaluated the different approaches for every spatial indicator itself before combining the approaches. The results are shown in Table 2.

Tweet Message: The method SP for tagging the messages identifies toponyms in 5.13% of the tweets. The overall estimation with a median error distance of 1100km is not suitable for location estimation. DBPedia Spotlight retrieves good estimations on messages mentioning the current location as toponyms in the text, which are created by location-based services. Furthermore, @-mentions like '@Bryant Park' provide good estimations. On the other side, DBpedia Spotlight has some problems with the non-standard language in Tweets, resulting, e.g., in regular words are identified as toponyms. In this case, identifying methods for detecting the relevant named entities is necessary. In contrast, using the LBS method, we get a high precision using the links created by location-based services with about 97% within a 1km radius, which makes this a suitable source for estimations. The recall of the LBS method is rather low with 18.25%.

Location Field: Using only the coordinates (COD) provided in the location field results in a low recall of 7.73%, which is the result of only having a few coordinates in the entry. The precision of this approach is very high, as 77% are within a 25km radius and 31% within 10km. In this case, some outliers result from large differences between the coordinates in the location field entry and the real position. These outliers might be a result of late updates of the device position, e.g., during long-distance flights.

Adding the plain Geonames approach GN, we get a good recall of 65.82%. Furthermore, the median error distance of 23,30km is a result of estimating 62% within a 50km radius and even 52% within 25km. Errors are the result of location field entries with multiple toponyms, with brackets and other unparsable combinations. With the first optimization GN-1,

Table 2: Results of the individual indicator approaches (in km) and external quality measures of the indicators.

	COD	GN	GN-1	GN-2	GN-3	GN-4	SP	LBS	TZ	WS-1	WS-2
MSQ	1670	3402	3432	3539	3631	4618	5939	403	4229	4896	7230
AED	349	1354	1320	1380	1459	2188	3689	15.41	2600	2618	5529
MED	9.25	23.3	22.65	22.63	25.46	41.40	1100	0.01	1543	494	3287
Recall	7.73%	63.55%	65.82%	69.03%	71.64%	83.29%	5.13%	18.25%	81.22%	6.46%	34.40%
Recall (all)	4.54%	68.67%	70.16%	72.51%	74.74%	82.32%	5.66%	22.19%	96.24%	17.43%	79.15%
$\mathbf{Q}_{\text{ext}}$	2.72	1.51	2.01	1.67	1.96	-0.54	0.87	4.26	1.12	1.07	-2.32

we can increase the recall by 2% while further increasing the median error distance by 0.7km. The second optimization GN-2 further increases the recall by 3% without a significant loss of precision. In contrast, the third optimization GN-3 results in a loss of precision while further increasing the recall by 2%. The fourth optimization GN-4 increases the recall to 83%, but also reduces the accuracy considerably. It is still possible to estimate 51% of the tweets in a 50km radius, but the median error distance with more than 1000km is much higher.

As an overall result, the location field handler extracts toponyms quite well. It still needs more discriminators for better precision, e.g., *loading...* is mapped to *Port Bonython Loading Terminal*. In this case, further information about the user has to be used to identify these cases. Our analysis of the location field further has further shown that people enter IP addresses, dates, as well as incomplete coordinates. These types of entries are discarded.

Time Zone and Website: The estimation based on the time zone approach TZ geolocalizes 81.22% of the tweets. This approach results in a low precision, because we use a polygon with the size of the whole country. The same applies for both website handling approaches. The first website approach WS-1 has a low recall of 6.46%, because website information is either not provided or related to a top-level domain, which we do not extract. Using the IP addresses in approach WS-2 is also imprecise, but the recall is 34.40%, because all websites are used. In this case, the precision is even lower than the top-level domain approach. Same as the time zone approach, the two website approaches have low precision, because we use the country wide polygons. All of these approaches are good estimators for smaller countries such as the Netherlands, but loose precision on large countries like the US. Nevertheless, the provided information can be valuable to differentiate toponyms extracted from the other approaches.

Results: Geolocalization of a tweet

For the overall evaluation, we discard the fourth Geonames optimization GN-4 as well as the approach based on the IP addresses WS-2, because their external quality measures are less than zero. As an overall result, we are able to create estimations for 92% of the tweets in our dataset with a median error distance of 29.66km. We are able to estimate 54% of the tweets within a 50km radius (cf. Figure 3, bottom right). The use of optimized quality measures (QM) drastically increases our estimation (cf. Table 3), with a small reduction in recall that results from disposing the two mentioned approaches.

Ikawa et al. (2012) report a precision of 17% in a 10km radius and 25% in a 30km radius. Compared to this, we exceed their results with 37% on 10km and 48% on 25km. Kinsella and Murdock (2011) report a precision of 13.9% on zip code level and 29.8% on town level. If we assume a precision of 1km precision as zip code precision and 10km precision as town level, we also exceed their results as we are able to estimate 22% on a 1km radius and 37% on a 10km radius[10]. We omit a comparison to Paradesi (2011) and Hale and Gaffney (2012) as they restricted their datasets beforehand to a non-worldwide set. Furthermore, Hong et al. (2012), Bouillot et al. (2012), MacEachren et al. (2011) do not provide quantitative results. Summarized, we significantly outperform current state-of-the-art on tweet localization.

Table 3: Results of the overall geolocalization approach for tweets with and without external quality measures.

	MSE	AED	MED	Recall
w/o QM	4159km	1931km	64.46km	95.10%
with QM	3310km	1408km	29.66km	92.26%

Since our test set consists of those tweets whose coordinates we know, and this selection may be biased, we have also tested our approach to detect spatial indicators on a random sample of 10,000 tweets from the whole Spritzer dataset. In this case, no quality measures or mappings to polygons have been applied. The results show that our approach would also perform well on a dataset with and without device locations (cf. Table 2, Recall (all)). Even a suspected decrease of recall in the location-based services indicator could not be found. Though the use of LBS indicators might appear as skewing the results, since they are trivial to locate, 22.19% of all tweets in a representative sample are LBS related tweets, thus, taking that information into account is a valid approach. Also only 1% of all location-based services indicators are correlated with coordinate entries in the location field. Nevertheless, the differences in recall indicate that our approach can be tuned to match yet unknown cases, e.g., previously unknown top-level domains. Furthermore, spatial indicators could be detected that are currently not mapped to polygons, which is a result of imprecise location information in the different approaches we apply.

Results: User's residence

To show the applicability of our approach for estimating the user's residence, we also applied it on a smaller sample set. Since there is no dataset with home locations, we identified all tweets for users with more than 20 tweets in our test set

[10] We compare to the evaluation based on the worldwide FIREHOSE feed in this case, as the Spritzer feed was restricted to 10 towns by Kinsella and Murdock.

Table 4: Results of the overall approach for estimating the user's residence.

	MSE	AED	MED	Recall
with QM	2281km	751km	5.05km	100%

and manually geocoded the residence of 500 randomly selected users. In this case, the location field of the last tweet from every user was used as ground truth, as we suppose it describes the user's actual home location. For estimating the quality of our approach, we compared these geocodes with our estimations. The estimations were created based on the spatial indicators extracted from all tweets of a user, which is different compared to the geolocalization of a tweet where only spatial indicators of one tweet were used.

The most relevant work from Chandra et al. (2011) report that their approach is able to estimate in 22% of the cases the user's residence in a 100mi radius. Furthermore, they achieve an average error distance of 1044.28mi. Their dataset contains data from the US with 10,584 training and 540 test users. We cannot reuse this dataset, as it only contains the tweets and no user profiles. With our approach we are able to estimate 79% of the user's residences in a 100mi radius on our data set. Furthermore, with 751km (466mi) our approach has a much lower average error distance with a median error distance of 5.05km (3.24mi) (see Table 4). We admit that we are working on a relatively small dataset, but our evaluation shows promising results towards good estimations for the user's residence even on a worldwide dataset.

Example application

To demonstrate how our approach can be applied to practical scenarios, we crawled 150,000 tweets related to *Hurricane Sandy* in October 2012. Our approach was applied for geolocalization of the tweets. Furthermore, we applied mood analysis from uClassify[11] to differentiate them into the two classes *happy* and *upset*. Using both spatial and mood indicators allows for creating maps showing how the attitude of people towards a topic or event is distributed in different regions.

Figure 4 shows such a map for the U.S. east coast region affected by hurricane Sandy. In the example application, the amount of tweets in each administrative area is represented by the transparency level of the polygons. The mood classes are displayed by the color of the polygon.

At first glance, the results look quite surprising, since people in the directly affecting coast regions and Manhattan apparently show positive moods towards the Hurricane. However, having a closer look at the tweets of Manhattan reveals a lot of people being happy to be safe after the hurricane, having got notice that no one in their family was hurt, etc. Another portion of tweets reveals that people are happy that they do not have to go to work or school the next day. On the other hand, less directly affected people in the areas far from the coast rather express their regret and compassion.

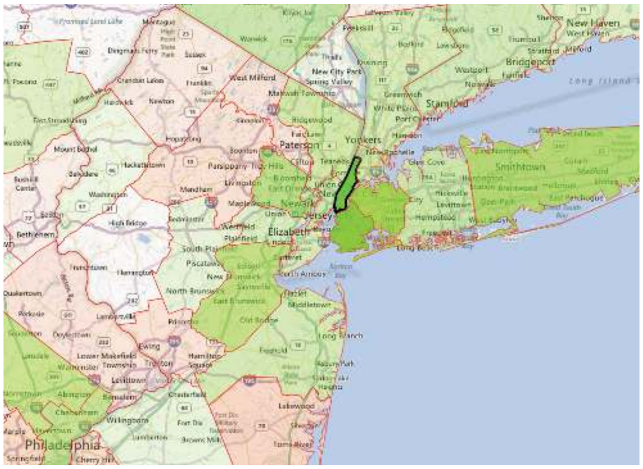

Figure 4: Example application of our approach based on tweets related to *Hurricane Sandy* showing the amount of tweets for each administrative area (represented by transparency level) as well as the mood classes (color of the polygon).

The example application shows how geolocated tweets may be used on a large scale, e.g., to do research on people's attitudes towards a topic. Furthermore, in crisis situations such the hurricane example, precisely located tweets may be used for creating rather exact maps of affected areas, and provide drill-down details when looking at the tweets that were aggregated for the map view.

Conclusion

This paper contributes the first multi-indicator approach that combines vastly different spatial indicators from the user's profile and the tweet's message for estimating the location of a tweet **as well as** the user's residence. In contrast to other works, our method uses a large variety of indicators and is thus less vulnerable to missing or incomplete data. We are able to create estimations for the location of a tweet for 92% of the tweets in our dataset with a median of 29.66km. Furthermore, we are able to predict the user's residence with a median accuracy of below 5.1km. Both predictions significantly outperform the state of the art. To achieve this degree of precision, we conducted an in-depth analysis of different spatial indicators that can be retrieved from tweets and determined their value for geolocalization problems based on an optimization algorithm.

We see further optimization potential of our approach. The indicators discussed in this paper may be refined, e.g., with respect to accuracy and internal quality measures, as well as new indicators may be integrated in our model. For instance, Sadilek, Kautz, and Bigham (2012) show promising results towards using the social network of a user for location inferencing. Furthermore, it would be beneficial to compute an overall confidence score for our estimations.

With an example application, we have demonstrated a use case for tweet localization for which our method provides sufficient accuracy: An analysis of the tweets created

[11] http://uclassify.com/

shortly after the hurricane Sandy struck the east coast of the U.S. reveals interesting and non-trivial information that could not have been generated without precise tweet location. This provides some initial evidence that the accuracy of our method is sufficient for analyzing microblog entries in various application scenarios.

Acknowledgments

This work has been partly funded by the German Federal Ministry for Education and Research (BMBF, 13N10711, 01jS12024), by the German Science Foundation (DFG, FU 580/2), and by EIT ICT Labs (12113 Emergent Social Mobility, 13064 CityCrowdSource).

References

Abrol, S., and Khan, L. 2010. Tweethood: Agglomerative Clustering on Fuzzy k-Closest Friends with Variable Depth for Location Mining. In *Proceedings of SOCIALCOM '10*, 153–160.

Aramaki, E. 2011. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In *Proceedings of EMNLP '11*, 1568–1576.

Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; and Hellmann, S. 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics - Science Services and Agents on the World Wide Web* 7(3):154–165.

Bouillot, F.; Poncelet, P.; and Roche, M. 2012. How and why exploit tweet's location information? In *Proceedings of AGILE'12*, 3–7.

Chandra, S.; Khan, L.; and Muhaya, F. B. 2011. Estimating Twitter User Location Using Social Interactions - A Content Based Approach. In *Proceedings of SocialCom2011*, 838–843.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users. In *Proceedings of CIKM '10*, 759–768.

Clodoveu, A. D. J.; Pappa, G. L.; de Oliveira, D. R. R.; and de L. Arcanjo, F. 2011. Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS* 15(6):735–751.

Eisenstein, J.; O'Connor, B.; Smith, N. N. A.; and Xing, E. P. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of EMNLP '10*, 1277–1287.

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

Gonzalez, R.; Cuevas, R.; Cuevas, A.; and Guerrero, C. 2011. Where are my followers? Understanding the Locality Effect in Twitter. *CoRR* abs/1105.3.

Hale, S., and Gaffney, D. 2012. Where in the world are you? Geolocation and language identification in Twitter. In *Proceedings of ICWSM'12*, 518–521.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Biebers Heart: The Dynamics of the "Location" Field in User Profiles. In *Proceedings of CHI '11*, 237.

Hong, L.; Ahmed, A.; Gurumurthy, S.; Smola, A. J.; and Tsioutsiouliklis, K. 2012. Discovering Geographical Topics In The Twitter Stream. In *Proceedings of WWW '12*, 769–778.

Ikawa, Y.; Enoki, M.; and Tatsubori, M. 2012. User Location Inference using Microblog Messages. In *Proceedings of WWW '12*, 687–690.

Kinsella, S., and Murdock, V. 2011. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proceedings of SMUC '11*, 61–68.

Krishnamurthy, B., and Arlitt, M. 2006. A Few Chirps About Twitter. In *Proceedings of WOSN '08*, 19–24.

Leidner, J. L. 2004. Toponym Resolution in Text: "Which Sheffield is it?". In *Proceedings of SIGIR '04*, 602–602.

Lieberman, M. D.; Samet, H.; and Sankaranarayanan, J. 2010. Geotagging with Local Lexicons to Build Indexes for Textually-Specified Spatial Data. In *Proceedings of ICDE'10*, 201–212.

MacEachren, A. M.; Jaiswal, A.; Robinson, A. C.; Pezanowski, S.; Savelyev, A.; Mitra, P.; Zhang, X.; and Blanford, J. 2011. SensePlace2: GeoTwitter analytics support for situational awareness. In *2011 Proceedings of VAST '11*, 181–190.

Mcgee, J.; Caverlee, J.; and Cheng, Z. 2011. A Geographic Study of Tie Strength in Social Media. In *Proceedings of CIKM 2011*, 2333–2336.

Mendes, P. N.; Jakob, M.; García-Silva, A.; and Bizer, C. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of I-Semantics*.

Nelder, A., and Mead, R. 1965. A Simplex Method for Function Minimization. *The Computer Journal* 7(4):308–313.

Paradesi, S. 2011. Geotagging tweets using their content. In *Proceedings of FLAIRS '11*, 355–356.

Popescu, A., and Grefenstette, G. 2010. Mining User Home Location and Gender from Flickr Tags. In *Proceedings of ICWSM'10*, 307–310.

Sadilek, A.; Kautz, H.; and Bigham, J. P. 2012. Finding Your Friends and Following Them to Where You Are. In *Proceedings of WSDM'12*, 723-732.

Smith, D. A., and Crane, G. 2001. Disambiguating Geographic Names in a Historical Digital Library. In *Proceedings of ECDL '01*, 127 – 136.

Starbird, K.; Palen, L.; Hughes, A. L.; and Vieweg, S. 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of CSCW '10*, 241–250.

Sultanik, E. A., and Fink, C. 2012. Rapid Geotagging and Disambiguation of Social Media Text via an Indexed Gazetteer. In *Proceedings of ISCRAM '12*, 1–10.

Takhteyev, Y.; Gruzd, A.; and Wellman, B. 2012. Geography of Twitter networks. *Social Networks* 34(1):73–81.

Tumasjan, T.; Sprenger, O.; Sandner, P.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of ICWSM'10*.

Woodruff, A. G., and Plaunt, C. 1994. GIPSY : Automated Geographic Indexing of Text Documents Previous Work in Georeferencing of Text Documents. *Journal of the American Society for Information Science* 45(9):645–655.

Zong, W.; Wu, D.; Sun, A.; Lim, E.-P.; and Goh, D. H.-L. 2005. On assigning place names to geography related web pages. In *Proceedings of JCDL '05*, 354–362.