Carmen: A Twitter Geolocation System with Applications to Public Health

Mark Dredze and **Michael J. Paul** and **Shane Bergsma** and **Hieu Tran**
Human Language Technology Center of Excellence
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21211
{*mdredze,mpaul19,htran21*}*@jhu.edu,shane.a.bergsma@gmail.com*

Abstract

Public health applications using social media often require accurate, broad-coverage location information. However, the standard information provided by social media APIs, such as Twitter, cover a limited number of messages. This paper presents Carmen, a geolocation system that can determine structured location information for messages provided by the Twitter API. Our system utilizes geocoding tools and a combination of automatic and manual alias resolution methods to infer location structures from GPS positions and user-provided profile data. We show that our system is accurate and covers many locations, and we demonstrate its utility for improving influenza surveillance.

Introduction

Social media, such as Twitter, are filled with millions of status updates by people around the world. By analyzing streams of social media data, one can automatically discover events across the globe. For example, researchers have shown that earthquakes can be quickly detected by analyzing Twitter messages ("tweets") (Sakaki, Okazaki, and Matsuo 2010), and the thoughts and sentiments of Twitter users have been correlated with opinion polls (O'Connor et al. 2010).

In the fields of public health and medical informatics, social media is becoming a valuable resource for learning about population health. Social media data can be analyzed to track the spread and prevalence of disease (Sadilek, Kautz, and Silenzio 2012b; Lamb, Paul, and Dredze 2013), learn about trends in tobacco and drug use (Prier et al. 2011; Paul and Dredze 2013), and understand pain and other ailments (Heaivilin et al. 2011; Paul and Dredze 2011). While these are promising applications of social media analysis, these applications can depend on knowing the geographic locations of users (Eke 2011).

Most Twitter health studies have focused on tweets tagged with the GPS position given by the user's device. While such information is accurate, geo-tagged tweets represent only about 2% of all tweets and 3% of Twitter users (Burton et al. 2012), severely limiting the potential of large-scale analyses. The 1% sample limitation on free public tweets only increases this problem. However, the location can often be inferred for tweets that are not geo-tagged, based on user profiles (Hecht et al. 2011) and language analysis (Roller et al. 2012; Wing and Baldridge 2011).

In this paper, we present Carmen[1], a geolocation system that infers structured location information – country, state, county, city – for Twitter users based on both geocoordinates and user profile information. We show that our system provides accurate location information for a large percentage of users, and we show how this information can be used to improve the accuracy of influenza tracking (Lamb, Paul, and Dredze 2013). The code and data behind our system is publicly available.[2]

Geolocation in Twitter

Our *geolocation* task is to infer the geographic location associated with a Twitter message based on the information we have about the message and user. The location information our system returns is a structured object containing multiple attributes, such as country and city; some of these attributes may be unknown (e.g. we may be able to infer the country but not city). Sometimes the location of a particular tweet can be identified (in cases where it was tagged with a location), and other times we can only associate messages with the inferred "home" location of the user.

Twitter provides various data APIs, and tweets come as JSON objects that include the tweet text along with metadata, such as the time, the location (coordinates) associated with the tweet (if provided by the user), and user profile information, which includes optional user-provided information such as the user's real name and location. These metadata can be used to geolocate the tweets. In particular, there are four primary[3] ways in which geolocation is commonly performed (Gonzalez, Figueroa, and Chen 2012; Oussalah et al. 2012) on Twitter users and messages:

1. **Place object from tweets** Some tweets delivered by the Twitter API include a JSON "Place" object which encodes a location associated with the tweet. These include fields such as the country and city associated with

[1] As in "Where in the world is Carmen Sandiego?"

[2] https://github.com/mdredze/carmen

[3] Others have explored using the social network structure (Backstrom, Sun, and Marlow 2010; Sadilek, Kautz, and Bigham 2012; Davis Jr et al. 2011).

the place, as well as geographic coordinates. Some Place types include finer-grained information such as business names and street addresses; the number of known Places grows as more are added by users. Users have the option to tag their tweets with a Place; the tagging can also be done automatically based on matches to the user's current GPS position, if the user allows this. For tweets containing Place objects, the geolocation has already been done by Twitter, although Places do not contain all attributes we care about – in particular, they do not contain county information. The user coverage is small, however: only 1% of tweets in our collection are associated with a Place.

2. **Coordinates from tweets** Some tweets are geotagged with the coordinates (latitude and longitude) of the user when the message was written, based on the user's GPS position. For these tweets, we know the exact location of the tweet, but because they have not been resolved to Places, we are not provided high-level data about the location such as the city and country. Details can be obtained by reverse geocoding using APIs from Google or Bing Maps. There is a larger set of tweets containing geo-coordinates than those that have been resolved to Places, but such tweets represent a tiny percentage of all tweets.
3. **Location from user profile** Many users publicly provide a location in their profile, a free-form field with values such as "NYC" or "Baltimore, MD". Such strings can often be resolved to structured locations by existing map APIs. These locations are coarse-grained and are mostly static, corresponding to the user's primary location rather than the location at the time of the message posting, which may be different if the user is traveling. Many more users have profile locations than geo-coordinates. However, users may lie or provide nonsensical locations such as "Candy Land" (Hecht et al. 2011; Hale, Gaffney, and Graham 2012).
4. **Content-based geolocation** Geolocation can also be done on a message or set of messages based on the textual content of the messages (Eisenstein et al. 2010; Roller et al. 2012; Wing and Baldridge 2011; Cheng, Caverlee, and Lee 2010). A user's primary location can be detected based on their dialect or the mention of regional issues like sports teams, for example, as well as the mention of landmarks. Such methods can be used to geolocate users who do not provide explicit location information, although this is a more involved approach, and many user messages may be needed to do this accurately.

As described above, others have investigated more sophisticated methods of geolocation based on the content of a message and the social network of the user. These methods, while they may be slower to run in practice, give state of the art results for geolocation. However, our primary goal is the development of an easy to use, freely available, and very fast system for geolocating tweets that can be used by the wider community. Therefore, our system utilizes methods 1–3: we perform geolocation using location information from tweet metadata and user profiles, but not message content. This allows us to process a stream of tweets without a reliance on a complex user model, knowledge of the social network, or reliance on a specific language. A detailed description of our system is given in the next section.

Carmen

The goal of Carmen is to assign a location to each tweet from a database of struckctured location information. Each location is represented by a unique identifier, a set of coordinates, and the name of the city, county, state and country, including `null` depending on the resolution of the location (from country to specific city). This 4-tuple uniquely identifies each location. For example, the location for "United Kingdom" will contain only a country, while the location for "Busan, South Korea" will lack a county. Additionally, the system organizes locations into a hierarchy: Earth → Country → State → County → City. The hierarchy facilitates aggregation of locations for computing statistics of interest, such as counting all the tweets in Canada.

The system uses a combination of the first 3 approaches from the previous section. We describe each stage of Carmen in the order that a tweet is processed.

Places The city, state and country fields of the structured Twitter place are queried against the database, returning a match if found. During system development, we used frequency statistics to include as many Twitter places as possible in the database by obtaining full location information from Yahoo's PlaceFinder API.[4] Additionally, Twitter does not use a consistent naming scheme for places, e.g. both "yhdistynyt kuningaskunta" and "united kingdom," "polnia" and "poland." We manually created an alias list (see below).

Coordinates If the tweet has coordinates but no place, we look up a database location within 25 miles, a user configurable option. Unfortunately, the database does not contain location size, so while 25 miles may be a reasonable distance for New York City, it's too narrow for the United States. We are exploring adding bounding box information in future releases.

Location from user profile For most tweets, we rely on the user profile's location. This string is matched (case insensitive) against aliases that map to known locations. We normalize the string and extract names as follows.

- Remove certain punctuation (`: ()_/-.!#;?`).
- Remove all punctuation characters and extra spacing.
- Regex for state or country names: `.+,\\s*(\\w+)` and check match against known US states, countries, and abbreviations for each.

We constructed our alias list using two steps. First, we extracted common profile locations over several million tweets, yielding both real places (e.g. "London," "New York") and spurious places (e.g. "Earth", "Neverland"). A combination of automated filters and a manual review removed invalid locations and merged duplicates (e.g. "New York City" and "NYC".) The resulting list was geolocated using Yahoo's PlaceFinder API. Since the API returns a location for even spurious queries we further pruned the list

[4] The PlaceFinder API provides geocoding services, returning structured information about a place based on a provided place name string. http://developer.yahoo.com/boss/geo/

and merged aliases based on the returned location. The final list contains 4811 unique places. We refer to this as the "Human Curated" list.

Next, we added aliases based on a new resource from Bergsma et al. (2013): observed attributes of hundreds of millions of Twitter users clustered based on the observed social network. Clusters include first names, last names, and user-provided locations. We use the same process described in §7 of Bergsma et al. (2013) to augment our alias list. This process can discover that unknown, user-provided locations on Twitter such as *balto* or *bmore* are aliases for the known location *baltimore*. This discovery is based on the observation that users with locations *balto* or *bmore* frequently communicate with users with the location *baltimore*. We refer to this as the "Automatically Extended" list.

Software Our geolocation system is implemented in Java and is available as an open source standalone library. Including disk IO time, we were able to process about 27,000 tweets a second, or about .04 milliseconds per tweet.

System Evaluation

We measured both coverage (recall) and accuracy (precision) of geolocations.We begin with accuracy. Using tweets containing places or coordinates, we created development (10,000 tweets) and test (56,167 tweets) sets. These tweets were randomly selected from the Twitter public stream and are the same as those used by (Bergsma et al. 2013)[5] We assume locations based on places and coordinates are correct and compare with geolocations based on locations from user profiles using several metrics:[6]

- Country: predicted and true location match countries.
- State/Country: compares country and state (if available.)
- Accuracy@K: locations match if distance $< K$ miles.

We note several evaluation limitations. First, usefulness depends on application: some may require country but not state or city. Second, locations are represented by their center so measuring distance is less useful for large areas (e.g. distance to the center of the United States.)[7] Additionally, we explored topological relationships, but these can be misleading (e.g. many users list a city name in their profile but live in a nearby suburb.) Still, we believe this is a useful approximation of system quality, and we provide additional application specific evidence below.

We evaluated the human curated and automatically extended alias lists. Parameters for building the extended list were tuned on the development data to obtain a similar accuracy to the human curated list but with increased coverage.

List	Place	Coordinates	Profile
Human Curated	0.9%	4.2%	94.9%
Automatically Extended	0.8%	3.6%	95.6%

Table 2: Percentage of geolocations attributed to each field.

Resolution	Human Curated	Automatically Extracted
City	57.9%	63.4%
County	0.9%	1.0%
State	13.6%	12.5%
Country	27.5%	23.0%

Table 3: Percentage of geolocations at each resolution.

Our system predicts the correct country for a tweet based on the location from the user provided more than 90% of the time (Table 1.) Even when location accuracy is measured to a resolution of 25 miles, our system achieves over 50% accuracy. The extended list increases coverage by 6%.

We evaluated the full geolocation system, including stages that use Twitter location information, on a large sample of tweets: 1% of public tweets from the first 9 days of March 2013 (43,656,388 tweets.)

First, we observed that about 1.3% of tweets contained a place field and 1.2% a coordinates field, while 56% of tweets had a non-empty user profile location. Using this information, the system with the human curated lists resolved 19.2% of the tweets while the automatically extended list increased this to 22.3%. Table 2 shows how often each tweet field was used for the geolocaton. Finally, we measured the resolution of the provided locations to a city, count, state or country. Table 3 shows that the majority of geolocations provide a city, which may be especially helpful for some applications.

Application: Influenza Surveillance

As a final demonstration of our system, we show how it can be used to improve a public health application: disease surveillance. Health officials track disease infection rates to prevent and manage outbreaks. Traditional systems rely on patient clinical visits, which take up to two weeks to publish. Recent work has demonstrated that Twitter can provide realtime infection rates (Collier 2012; Signorini, Segre, and Polgreen 2011). Strategies for Twitter influenza surveillance include supervised classification (Culotta 2010b; 2010a; Aramaki, Maskawa, and Morita 2011), unsupervised models for disease discovery (Paul and Dredze 2011), keyword counting[8], tracking geographic propagation (Sadilek, Kautz, and Silenzio 2012b), and combining tweet contents with the social network (Sadilek, Kautz, and Silenzio 2012a) and with location information (Asta and Shalizi 2012).

We use the system of Lamb, Paul, and Dredze (2013), a state of the art system for influenza surveillance in the United States. We evaluated trends using their infection classifier and compared against government data for both the

[5] We note that this may be a biased sample of tweets since users who geocode may not be representative of general users in terms of location or willingness to provide accurate user profiles.

[6] We did not consider city, as coordinates often gave suburban locations (coordinates resolved to Cambridge and user profile location was Boston.)

[7] This is an obvious place in which we could improve the system. However, doing so requires knowledge of either bounding boxes for locations or the approximate radius of a location. This information is currently unavailable in the system.

[8] The DHHS competition relied solely on keyword counting. http://www.nowtrendingchallenge.com/

			Accuracy@K				
List	Country	State/Country	25	50	100	250	Coverage
Human Curated	92.69%	65.92%	54.54%	59.65%	66.06%	76.62%	38.55%
Automatically Extended	90.74%	64.72%	54.51%	59.80%	65.77%	75.27%	44.45%

Table 1: Accuracy and coverage of the human curated and automatically extended alias lists.

	US	UK	US	UK
Location	2009		2011	
All Twitter	.9604	.5138	.6993	.6010
US Only	**.9714**	.1982	**.7792**	**.6312**
UK Only	.9231	**.8827**	.6277	.5123

Table 4: Correlations against government ILI data, from the CDC (US): Aug 2009–Aug 2010, Dec 2011–Aug 2012; HPA (UK): Jun 2009–May 2010, Dec 2011–Jun 2012.

United States[9] and the United Kingdom[10]), with Pearson correlations computed separately for 2009 and 2011.[11] We compare two trends: all tweets and only those as geolocated by our system (either US or UK.)

We observe significant improvements in trend correlations (Table 4), even for the earlier 2009 data which is mostly from the US anyway. The 2009 pandemic had different trajectories in the US and UK, which is apparent in the results: the correlation between UK data and UK tweets is .88, but only .20 with US tweets. On the other hand, the UK tweets actually have a worse correlation than US tweets in the 11-12 season. One possible reason is that, as Lamb, Paul, and Dredze (2013) noted, the 11-12 season was very mild and contained fewer tweets, so all systems performed worse. In this situation, it is possible that having more data (e.g. from a more populous country such as the US) may be better than having cleaner data (i.e. from the correct country).

Limitations and Future Work

The system has several limitations that suggest future work. The current system geolocates each tweet individually, although aggregation around each user may allow geotagged tweets to inform others. Geolocation information for members of a user's social network could also be exploited (Backstrom, Sun, and Marlow 2010; Sadilek, Kautz, and Bigham 2012). We could enable dynamic location additions to our database based on a location API and structured information from Twitter. While our system ignores tweet content, some have found this helpful in geolocation (Roller et al. 2012). Other resources could be used to augment our location database and alias lists, such as mining query results from search engines, and public resources (e.g. OpenStreetMaps, Wikimapia, Wikipedia, etc.) Finally, our system provides city resolution, but geocodes and some locations in user profiles are finer grained, including neighborhoods or specific points of interest. We plan to explore some of these ideas in future versions of the system. Additionally, our system may be applicable to other social media that include location strings.

While there has been promising work on application of Twitter to public health problems, most of these approaches have been limited by the amount of available geolocated data. While the volume of this data is increasing, it has a long way to go before it reflects a significant percentage of Twitter traffic. Carmen greatly aids this effort by increasing more than twentyfold the number of geolocated tweets. By distributing Carmen as an open source project and documenting its abilities, others can incorporate geolocation tools into their social media and health systems, increasing the ability of those systems to track geographic trends.

References

Aramaki, E.; Maskawa, S.; and Morita, M. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *EMNLP*.

Asta, D., and Shalizi, C. 2012. Identifying influenza trends via twitter. In *NIPS Workshop on Social Network and Social Media Analysis: Methods, Models and Applications*.

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*.

Bergsma, S.; Dredze, M.; Durme, B. V.; Wilson, T.; and Yarowsky, D. 2013. Broadly improving user classification via communication-based name and location clustering on twitter. In *NAACL*.

Burton, S. H.; Tanner, K. W.; Giraud-Carrier, C. G.; West, J. H.; and Barnes, M. D. 2012. ” right time, right place” health communication on twitter: Value and accuracy of location information. *Journal of medical Internet research* 14(6).

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*.

Collier, N. 2012. Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health* 7(7):731–749.

Culotta, A. 2010a. Towards detecting influenza epidemics by analyzing Twitter messages. In *ACM Workshop on Soc.Med. Analytics*.

Culotta, A. 2010b. Detecting influenza epidemics by analyzing twitter messages. arXiv:1007.4748v1 [cs.IR].

Davis Jr, C. A.; 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.

[9] The Centers for Disease Control and Prevention (CDC) weekly estimates of the U.S. influenza-like illness (ILI) from the outpatient surveillance network http://www.cdc.gov/flu/weekly/overview.htm#Outpatient

[10] We use the Royal College of General Practitioners Weekly Returns Service for England and Wales: http://www.hpa.org.uk/webw/HPAweb&Page&HPAwebAutoListNameDesc/Page/1317135329846

[11] The 2009 data is a 10% sample of Twitter. To increase data coverage, for 2011 we collected Tweets mentioning health keywords and then normalized by the public stream counts. For our analysis, we excluded days that were missing data.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *EMNLP*.

Eke, P. 2011. Using social media for research and public health surveillance. *Journal of Dental Research* 90(9):1045–1046.

Gonzalez, R.; Figueroa, G.; and Chen, Y.-S. 2012. Tweolocator: a non-intrusive geographical locator system for twitter. In *LBSN*.

Hale, S. A.; Gaffney, D.; and Graham, M. 2012. Where in the world are you? geolocation and language identification in twitter. Technical report, Working paper.

Heaivilin, N.; Gerbert, B.; Page, J. E.; and Gibbs, J. L. 2011. Public health surveillance of dental pain via Twitter. *Journal of Dental Research* 90(9):1047–1051.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI*.

Lamb, A.; Paul, M. J.; and Dredze, M. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *NAACL*.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From Tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.

Oussalah, M.; Bhat, F.; Challis, K.; and Schnier, T. 2012. A software architecture for twitter collection, search and geolocation services. *Knowledge-Based Systems*.

Paul, M. J., and Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. In *ICWSM*.

Paul, M. J., and Dredze, M. 2013. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *NAACL*.

Prier, K. W.; Smith, M. S.; Giraud-Carrier, C.; and Hanson, C. L. 2011. Identifying health-related topics on Twitter: An exploration of tobacco-related Tweets as a test topic. In *SBP*.

Roller, S.; Speriosu, M.; Rallapalli, S.; Wing, B.; and Baldridge, J. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *EMNLP*.

Sadilek, A.; Kautz, H.; and Bigham, J. P. 2012. Finding your friends and following them to where you are. In *WSDM*.

Sadilek, A.; Kautz, H.; and Silenzio, V. 2012a. Modeling spread of disease from social interactions. In *ICWSM*.

Sadilek, A.; Kautz, H.; and Silenzio, V. 2012b. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI*.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*.

Signorini, A.; Segre, A.; and Polgreen, P. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza a H1N1 pandemic. *PLoS One* 6(5):e19467.

Wing, B. P., and Baldridge, J. 2011. Simple supervised document geolocation with geodesic grids. In *ACL*.