Changes in Tweet Geolocation over Time: A Study with Carmen 2.0

Jingyu Zhang, Alexandra DeLicia, Mark Dredze
Johns Hopkins University
Introduction

Twitter Geolocation tools are useful for demographic studies in various topics

- Civil unrest
- Natural disasters
- Disease spread

Existing tools identify the location of tweets base on tweet metadata, tweet content, and social networks
Problem Statement

While widely used, geolocation tools tend to be English-centric and are often not evaluated for global coverage or performance across time and language.

We assess the following factors’ impact on geolocation tool Carmen:

- Language
- Country
- Time
Carmen: A Review

Introduced in Dredze et al. (2013), Carmen is a metadata-based geolocation tool that resolve locations from:

- Embedded coordinates in the **Geo object**
- Matching the **Place object** to internal location database
- Mapping **user profile location** string to internal location database
Carmen 2.0

- Compatible with Twitter API v2
- Performance optimization (25x faster geocode resolver)
- Expanded database with GeoNames
  - GeoNames Only
  - GeoNames + Carmen Original
Database Comparison

Carmen Original Database
- 7K location entries
- Inferred from tweets between May 2009 and Aug 2012 (primarily English tweets from US)
- Does not align with external knowledge base

Carmen 2.0
- 73K entries extracted from the GeoNames database
- Alternative names in many languages
- Hierarchical (CITY, ADMIN, COUNTRY) structure compatible with GeoNames
Evaluating Geotagging Performance

Geotagging tools should be able to accurately cover a wide range of locations:

- **Coverage**: for what portion of data can the geotagger propose a location
- **Accuracy**: how well the proposed locations compare to ground truth

We develop *multiple* metrics tailored to geotagging performance
Metrics for Geotagger

Coverage: percentage of data successfully mapped to a location

Accuracy:

- **Match Ratio** of level $L$: percentage of resolved tweets that is correct on level $L$. $L$ is one of \{country, admin, city\}
- **Distance**: geodesic distance between resolved and ground truth location
- **Acc@$K$**: percentage of resolved tweets such that the distance error does not exceed $K$ miles.
Experiments
Ground Truth Data

We introduce Twitter-Global, a new geolocation evaluation dataset collected from multiple Twitter API streams

- 15.3M geotagged tweets
- Collected from 2013 to 2021
- Covers a wide range of languages and countries
Performance across Language

We create two subsets of English and Non-English data from Twitter-Global.

| Language   | Database          | Coverage | \(mr_{country}\) | \(mr_{admin}\) | \(mr_{city}\) | \(d\)  | Acc@10 | Acc@100 | Acc@1000 |
|------------|-------------------|----------|-------------------|-----------------|---------------|--------|--------|---------|----------|
| English    | GeoNames-Only     | 49.58%   | 99.42%            | 95.63%          | 47.49%        | 853.9  | 0.81   | 0.85    | 0.86     |
|            | GeoNames-combined | 49.63%   | 99.43%            | 94.36%          | 47.69%        | 58.7   | 0.81   | 0.91    | 0.99     |
|            | Original          | 48.14%   | 99.35%            | 94.94%          | 48.90%        | 46.4   | 0.78   | 0.91    | 1.00     |
| Non-English| GeoNames-Only     | 41.77%   | 99.36%            | 66.50%          | 20.13%        | 482.3  | 0.84   | 0.88    | 0.88     |
|            | GeoNames-combined | 41.78%   | 99.35%            | 66.83%          | 20.27%        | 105.3  | 0.84   | 0.90    | 0.99     |
|            | Original          | 32.27%   | 98.95%            | 75.61%          | 14.22%        | 106.2  | 0.67   | 0.87    | 0.99     |

On Non-English data, GeoNames

- Substantially increased coverage
- Moderate increased accuracy-based metrics
Performance across Countries

We create two subsets of US and Non-US data from **Twitter-Global**

| Origin | Database       | Coverage | $mr_{country}$ | $mr_{admin}$ | $mr_{city}$ | $d$     | Acc@10 | Acc@100 | Acc@1000 |
|--------|----------------|----------|----------------|--------------|-------------|---------|---------|---------|----------|
| US     | GeoNames-only  | 50.56%   | 99.37%         | 99.87%       | 53.66%      | 994.2   | 0.79    | 0.84    | 0.84     |
|        | GeoNames-combined| 50.60%  | 99.37%         | 99.87%       | 53.81%      | 23.6    | 0.79    | 0.91    | 1.00     |
|        | Original       | 51.03%   | 99.93%         | 99.96%       | 55.33%      | 23.7    | 0.79    | 0.91    | 1.00     |
| non-US | GeoNames-only  | 42.63%   | 99.37%         | 99.37%       | 61.51%      | 18.73%  | 439.3   | 0.84    | 0.89     |
|        | GeoNames-combined| 42.65%  | 99.37%         | 99.37%       | 60.81%      | 18.88%  | 121.2   | 0.84    | 0.90     |
|        | Original       | 32.89%   | 98.45%         | 98.45%       | 66.11%      | 11.10%  | 118.0   | 0.67    | 0.87     |

On Non-US data, GeoNames

- Substantially increased coverage
- Achieved comparable accuracy with original database
Performance over Time

We create subsets of Twitter-Global for each year between 2013-2021

Main findings

- Due to change in metadata availability, Coverage dropped significantly after 2014
- GeoNames provide slightly better coverage regardless of metadata availability
Summary

- Introduced **Carmen 2.0**, an updated version of geolocation tool Carmen backed by an open-source gazetteer, GeoNames
- **Twitter-Global** is a Twitter geolocation evaluation dataset for language, country, and time ablation studies
- Significant **difference in performance in the ablation**, with higher performance for English and US-based tweets
- Geolocation tools should be **robust to language, country of origin**, and available metadata
- More **work is needed for a fine-grained study** on individual languages and countries
Thank you!

Analysis Code: https://github.com/AADeLucia/carmen-wnut22-submission

Carmen: https://github.com/mdredze/carmen-python
Supplementary
Database Statistics

|       | Original |       | GeoNames |       |
|-------|----------|-------|----------|-------|
|       | Count    | Percent | Count    | Percent |
| City  | 4401     | 62.51%  | 24568    | 33.24% |
| County| 1995     | 28.33%  | 45154    | 61.08% |
| State | 461      | 6.55%   | 3947     | 5.34%  |
| Country| 184   | 2.61%   | 252      | 0.34%  |
| Total | 7041     | 73921  |

Table 1: The statistics of city, county, state, and country-level locations in the original Carmen location database and the new GeoNames database versions developed for Carmen 2.0. The GeoNames-augmented databases have more than 10 times the number of location entries than Original. Percentage refers to portion of the database dedicated to each granularity.