Locating Place Names at Scale

Using Natural Language Processing to Identify Geographical Information

Abstract

Historical sources are often tagged with metadata about place such as where the object was created, acquired, or stored. Rich latent geographical information is frequently mentioned throughout textual documents as well. A challenge, though, is how to extract this spatial information at scale. For example, when a text mentions Paris, does the writer mean Paris, Texas, USA or Paris, France? Out of context, most would assume the reference is to capital of France, but it could also be the city in Texas. While close reading would provide an answer, this becomes a challenge when working with hundreds and thousands of documents. How might we be able to more accurately predict the exact location using the broader context?

Method

We start with a document of text that has been tagged with geospatial metadata. This may, for example, be where the document was written, published, or the primary location of interest. The goal is to detect and extract additional locations described in the text using the initial metadata as a reference point. This is done as follows:

- Use a georeference service to convert the spatial metadata into: (i) latitude and longitude; (ii) country and other political designations. This data becomes a *reference point* for geolocating the remainder of the text.
- Extract named entities from the raw text, and filter out those referring to locations or political entities.
- Pass the raw named entities through a georeference service.
- Entities with a unique georeference response are tagged as-is.
- For entities that are not unique, append the political division to the name of the place in order of specificity and re-reference.
- Select the unique entity attached to the least general political entity.
- In case of further ties, return the point closest to the reference.

Example

Consider a corpus of Tweets from news organizations on Twitter. We have a particular document (a single) tweet made by the user @bangordailynews – who is described as being located in "Bangor, Maine".

> *"A red flag" and "highly irregular": that's how a judge sees the sudden decision to hike fees on large ships coming and going from Portland 50 percent.*

Locating a record for the metadata location, we have

```
Bangor, Penobscot County, Maine, USA
```

Next, we use NER and locate the entity **Portland** in the text. Which Portland is the text talking about? There are a large number of entities called "Portland" across the world.

The search algorithm proceeds as follows:

1. Search for string "Portland"; returns several cities:
   - Portland, OR, Multnomah County, USA
   - Portland, ME, Cumberland County USA
   - Isle of Portland, Dorset, England, UK
   - Portland, Ontario, CAN
   - etc.
2. Search for string "Portland + USA"; returns two cities:
   - Portland, OR, Multnomah County, USA
   - Portland, ME, Cumberland County USA
3. Search for string "Portland + Maine + USA"; returns just one city:
   - Portland, ME, Cumberland County USA

   Which we use as our result.

Note that we can't jump straight to the most specific result, "Portland + Penobscot County + Maine + USA". This search gives the following result:

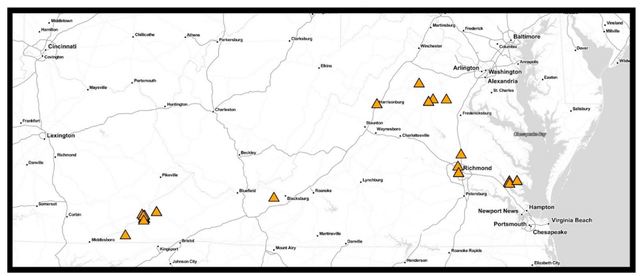
```
Portland St, Old Town, Penobscot County,
        Maine, USA
```

Application: Federal Writer's Project

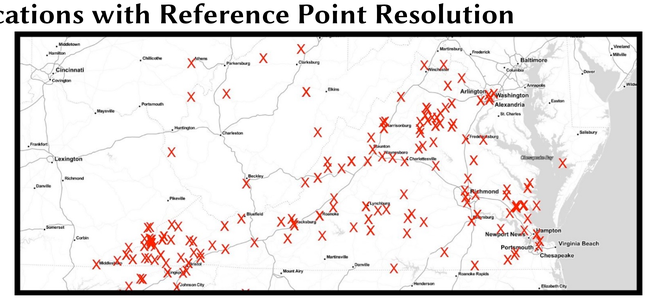
During the New Deal, thousands of life histories were written to capture the American experience. While the location of the interviews provides insight into the geographic expanse of the collection, the interviewees consistently spoke about places beyond the location of the physical interview.

We applied our method to identify the place names in the interviews. We are then able to identify and map the many different locations that interviewees mentioned. Across the interviews, we saw that many spoke of migration – whether their own or their kin – generating a more complex understanding of movement and place during the early 20th century in the United States.

Interview Locations (reference points)



Locations with Reference Point Resolution



Taylor Arnold
Mathematics & Computer Science
University of Richmond
@statsmaths

Lauren Tilton
Rhetoric & Communication Studies
University of Richmond
@nolauren

Courtney Rivard
English & Comparative Literature
University of North Carolina
Chapel Hill
@courtney_rivard

Code available at:
github.com/statsmaths/place-names

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNIVERSITY OF RICHMOND FOUNDED 1830