An NLP and Geospatial Workflow for Harvesting Local Place Names from Geotagged Social Web

Yingjie Hu[1], Huina Mao[2], and Grant McKenzie[3]

[1]Department of Geography, University of Tennessee, Knoxville, TN 37996, USA
[2]Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
[3]Department of Geographical Sciences, University of Maryland, College Park, MD 20742, USA

Abstract: Local place names are those used by local residents but not recorded in existing gazetteers. Some of them are colloquial place names, which are frequently referred to in conversations but not formally documented. Some are not recorded in existing gazetteers due to other reasons, such as their insignificance to a general gazetteer that covers a large geographic extent (e.g., the entire world). Yet, these local place names play important roles in many applications, from supporting public participation GIS to disaster response. This extended abstract describes our preliminary work in developing an automatic workflow for harvesting local place names from the geotagged Social Web. Specifically, we make use of the geotagged Craigslist posts in the apartments/housing section where people use local place names in their posts frequently. Our workflow consists of two major steps, a natural language processing (NLP) step and a geospatial step. The NLP step focuses on the textual contents of the posts, and extracts candidate place names by analysing the grammatical structure of the texts and applying a named entity recognition model. The geospatial step examines the geographic coordinates associated with the candidate place names, and performs multi-scale clustering to filter out the false positives (non-place names) included in the result of the first step. We ran a preliminary comparison between our initial result and a comprehensive gazetteer, GeoNames. Possible future steps are discussed.

Keywords: place name; gazetteer; natural language processing; geosocial; geospatial semantics.

1. Introduction

Place names are widely studied in GIScience. People generally use place names in their everyday conversations instead of numeric coordinates to refer to locations. As a result, a GIS often needs to be equipped with the capability of understanding the geographic meaning of place names. Digital gazetteers fill this gap by providing an organized collection of entries with place names, place types, and their spatial footprints (Hill, 2000, Goodchild and Hill, 2008). However, traditional gazetteers, such as the Geographic Names Information System (GNIS) and GEOnet Names Server (GNS), contain only standard place names specified by authorities. With the rise of volunteered geographic information (VGI), studies were also conducted on enriching authoritative gazetteers with colloquial and vague place entries. For example, Jones et al. (2008) proposed a computational approach to modelling vague places by harvesting related geographic entities (e.g., hotels) using a Web search engine. Keßler et al. (2009) used geotagged photo data, including Flickr, Panoramio, and Picasa, to approximate the geographic boundary of vague place names.

Existing studies, however, often focus on finding the spatial footprint of a given (vague) place name. In this work, we explore a different research question, namely, *given a geographic region, how can we extract the place names used by locals that are not recorded* *in existing gazetteers?* The answer to this question has important implications. For example, in disaster response, response teams may come from other cities, states, and even other countries (e.g., international humanitarian organizations), and may not be familiar with the local place names referred by residents in the disaster-affected area. In addition, having information about local place names can also facilitate the interactions between GIS and local users, and can help design more effective public participation GIS. In this extended abstract, we describe our preliminary study on developing a NLP and geospatial workflow for harvesting local place names from the geotagged Social Web.

2. Dataset

The dataset for this study was retrieved from Craigslist, a classified advertisement website where users can post various types of ads. Specifically, we collected the ad posts in the *apartments/housing* section. Several reasons make Craigslist and the posts in this section a suitable dataset. First, place names are frequently mentioned in the posts related to housing. This is due to the importance of location in housing choices, and post owners are fully motivated to provide details about the nearby places in their posts to attract the interests of readers. Second, local place names are often observed in these posts. This is because Craigslist websites are local-specific (e.g., there is one Craigslist website for Los Angeles county and a separate website for New York City), and most users are from the local community. Residents often use local place names to communicate with one another. Third, many housing ads are tagged with geographic coordinates, which can then be used to derive the spatial footprints of the related place names.

As a preliminary study, we have retrieved 7,500 posts from the Craigslist Los Angeles county website (https://losangeles.craigslist.org/) from Feb. 18, 2017 to Feb. 20, 2017. The retrieved posts are associated with an ID, a repost ID (if this is a repost from an earlier post), timestamp, longitude, latitude, and the textual content of the post. Among the 7,500 posts, 6,759 posts are geotagged, which represent a high percentage (around 90%). Yet, there also exist numerous reposts. By removing these reposts (based on the post IDs and repost IDs) and non-geotagged posts, we obtained 1,455 data records. Figure 1 shows the locations of the geotagged posts.

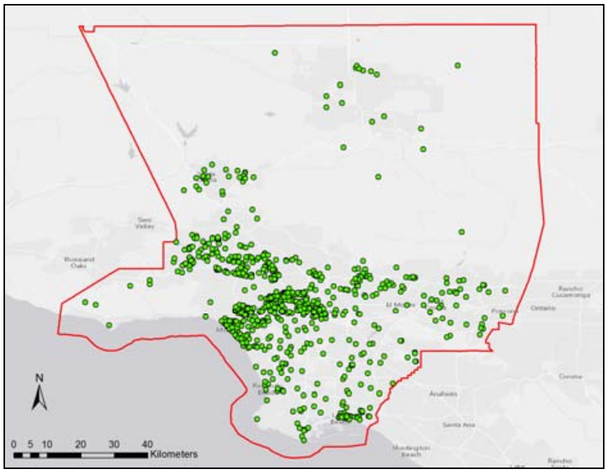


Figure 1. Locations of the geotagged posts in Los Angeles county (with reposts removed).

3. A NLP and Geospatial Workflow

The workflow we propose consists of two major steps: a natural language processing step and a geospatial step. The NLP step focuses on the textual content of the posts, and aims at extracting candidate place names. Two approaches were used previously to extract place names: natural language parsing (Vasardani et al., 2013) and named entity recognition (NER) (Gelernter and Mushegian, 2011). In our preliminary experiment, we tried both. The Stanford Dependency Parser is used for natural language parsing. Figure 2 shows the grammatical structure extracted from the sentence “We are located in Hollywood, close to West Hollywood, the Larchmont district, and Mid city.” Place names extracted include “Hollywood”, “West Hollywood”, “Larchmont district”, and “Mid city”.

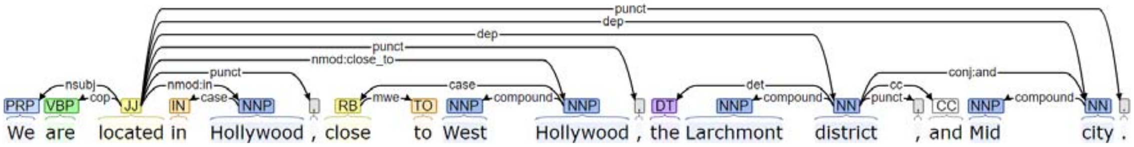
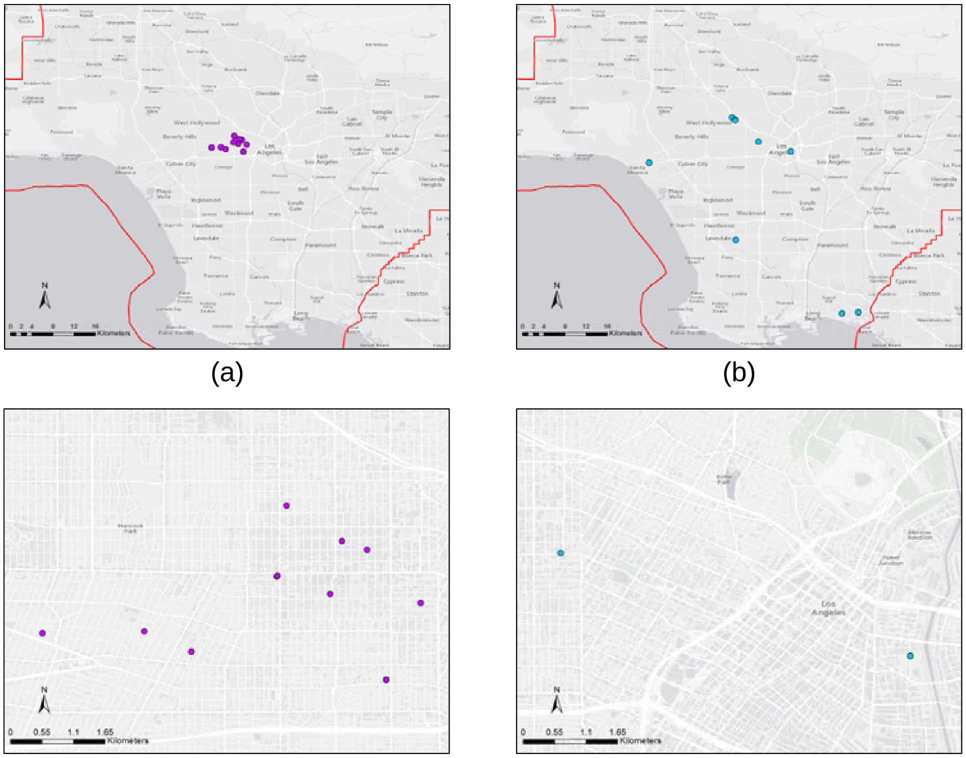

punct
dep
dep
punct
nmod:close_to
punct
nsubj
cop
nmod:in
case
case
mwe
compound
det
compound
cc
punct
conj:and
compound
PRP VBP JJ IN NNP . RB TO NNP NNP . DT NNP NN . CC NNP NN .
We are located in Hollywood , close to West Hollywood , the Larchmont district , and Mid city .

Figure 2. Grammatical structure of an example sentence

In the NER approach, we employ a deep learning neural network (ConvNet) developed by Yuan (2016) to identify location entities in the text. To train and test our model, we use the OntoNote 5.0 dataset (https://catalog.ldc.upenn.edu/LDC2013T19) which contains 18 classes of entities. We focus on location entities, and all other classes are considered as non-locations. The training, validation, and testing datasets include *75,187*, *9,603*, and *9,479* sentences from OntoNote. The trained model shows a high F1 score on the testing set: 83%. We then apply the trained model to Craigslist posts, and have successfully extracted a number of location names including streets, buildings, neighbourhoods, and regions.

The NLP step has extracted many candidate place names, but it has also included some false positives, such as “bedroom” and “Monday”. Thus, in the geospatial step, we identify the true place names from the false ones. While we have been experimenting different approaches, a pattern is observed for place and non-place names, as demonstrated in Figure 3.



(a)

(b)

(c)

(d)

Figure 3. (a)(c): locations of the posts related to the term “Koreatown”; (b)(d) locations of the posts related to the term “Monday”.

As is shown, a true place name like "Koreatown" will cluster at certain geographic scales, such as in Figure 3(a). However, true place names do not always cluster at any scale, such as in Figure 3(c). In contrast, a non-place term, such as "Monday", is less likely to cluster at any geographic scale within the bounds of the dataset, as shown in Figure 3(b) and 3(d). Thus, a potential algorithm could perform a test by clustering the data at multiple geographic scales simultaneously.

4. Preliminary Result

We have manually compared a sample of extracted place names with the place names recorded in a comprehensive gazetteer, GeoNames, within Los Angeles county. The results are promising. Specifically, we list the following three place names which are not recorded in GeoNames but are valid place names (by checking Wikipedia).

- **Silicon Beach**: The Westside region of the Los Angeles metropolitan area where more than 500 technology start-up companies are located.
- **Noho Arts District**: A community in North Hollywood area which is home to many contemporary theaters, art galleries, cafes, and shops.
- **Arclight Theater**: A 14-screen multiplex located on Sunset Boulevard in Hollywood. It is also called Arclight Cinema or Arclight Hollywood.

It is notable that the first two place names represent sub regions and are not specific buildings or points of interest. This is often true of colloquial place names. There are also other local place names in our result in addition to the above three, and we plan to do a comprehensive comparison between our result and multiple existing gazetteers.

5. Conclusions & Future Work

In this extended abstract, we presented a preliminary study on harvesting local place names from geotagged Social Web posts. The extracted local place names can enrich existing gazetteers, and can be applied to disaster response and developing public participation GIS. For the next steps, we will continue developing our workflow on both the NLP and geospatial steps, and will perform a comprehensive comparison with existing gazetteers.

References

GELERNTER, J. & MUSHEGIAN, N. 2011. Geo□parsing Messages from Microtext. *Transactions in GIS,* 15**,** 753-773.

GOODCHILD, M. F. & HILL, L. L. 2008. Introduction to digital gazetteer research. *International Journal of Geographical Information Science,* 22**,** 1039-1044.

HILL, L. L. Core elements of digital gazetteers: placenames, categories, and footprints. International Conference on Theory and Practice of Digital Libraries, 2000. Springer, 280-290.

JONES, C. B., PURVES, R. S., CLOUGH, P. D. & JOHO, H. 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science,* 22**,** 1045-1065.

KEßLER, C., MAUÉ, P., HEUER, J. T. & BARTOSCHEK, T. Bottom-up gazetteers: Learning from the implicit semantics of geotags. International Conference on GeoSpatial Sematics, 2009. Springer, 83-102.

VASARDANI, M., TIMPF, S., WINTER, S. & TOMKO, M. From descriptions to depictions: A conceptual framework. International Conference on Spatial Information Theory, 2013. Springer, 299-319.

YUAN, J. 2016. Automatic Building Extraction in Aerial Scenes Using Convolutional Networks. *arXiv preprint arXiv:1602.06564.*