GeoTxt: A Web API to Leverage Place References in Text

Morteza Karimzadeh[1,2], Wenyi Huang[3], Siddhartha Banerjee[1,3], Jan Oliver Wallgrün[1,2], Frank Hardisty[1,2], Scott Pezanowski[1,2], Prasenjit Mitra[1,3] and Alan M. MacEachren[1,2]

1) GeoVISTA Center, 302 Walker Building, University Park, PA, 16802 +1-814-865-3433 {karimzadeh, wallgrun, hardisty}@psu.edu

2) Department of Geography, The Pennsylvania State University, 302 Walker Building, University Park, PA 16802 +1-814-865-3433 {scottpez,maceachren}@psu.edu

3) College of Information Sciences and Technology, 332 Info Science and Tech University Park, PA 16802, USA +1-814-865-3528 {wzh112,sub253,pmitra}@ist.psu.edu

ABSTRACT

Associating place name mentions in unstructured text with their actual references in geographic space is vital to enable spatial queries and analysis. In this paper, we introduce GeoTxt, a web API plus human-usable web tool designed and implemented to tackle three components of place-reference processing from text, namely: extraction, disambiguation, and geolocation of place names mentioned in unstructured text. Current GeoTxt development is focused particularly on support for processing short microblog posts.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithm, Design, Performance

Keywords

Geocoding, Geographic Information Retrieval, Natural Language Processing, Geographic Information Systems

1. INTRODUCTION

While the volume of textual data with explicit geo-location is increasing rapidly due to GPS-enabled devices and sensors of many kinds, an even larger source of place-based information exists in text artifacts ranging from microblogs, through news stories and press releases, to scientific publications. The GeoTxt API has been designed to support extraction, disambiguation, and geolocation of place entities in text submitted to the API from other applications. The primary focus is on extracting place references from microblog posts, partly similar to the goals of [2; 6; 8] as opposed to most other efforts such as [3-5; 9] that address similar issues but focus on longer and more grammatically correct text artifacts. This is a challenging task due to the limited context in these short posts (e.g., 140 character limit in Twitter), the related use of abbreviations, and the non-standard syntax often used (e.g., words are often not capitalized as they would be in standard text). Nevertheless, users will be able to indicate the nature of query text to get the best results for either microblog posts or longer text articles. Also, GeoTxt includes a human-usable interface for processing individual text artifacts and testing the methods.

Below, we outline the GeoTxt API approach and system specifications, detail how it works, and point to future work.

GIR'13, November 05 2013, Orlando, FL, USA

ACM 978-1-4503-2241-6/13/11$15.00.
http://dx.doi.org/10.1145/2533888.2533942

2. GEOTXT ARCHITECTURE AND CAPABILITIES

GeoTxt has been designed and implemented as an easy to use RESTful Web API. It identifies mentions of place names in unstructured text, and assigns geographic coordinates to those place names. Trusted applications are able to query the service with HTTP GET requests and receive the responses as GeoJSON FeatureCollection objects containing geocoded place names along with persons and organizations identified in free form text.

Figure 1 shows the schematic architecture of the GeoTxt API. GeoTxt, written in Java, processes input text in two separate steps, which work independently in the current release. At the first step, Named Entity Recognition is performed to extract place names. Second, those place names found in text are disambiguated and geolocated to their respective geographic coordinates by the GeoCoder module.

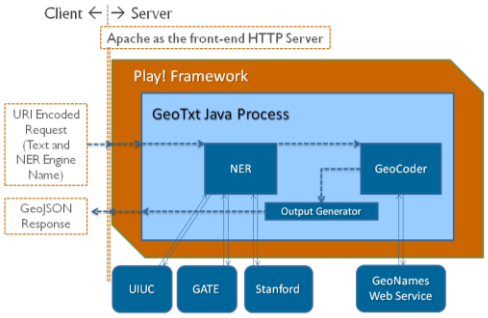

Figure 1. GeoTxt API Architecture.

The Illinois Named Entity tagger [7] (depicted as UIUC in Figure 1) has been tested and integrated with the system; however, its current beta release has been considerably slower than GATE ANNIE [1] and Stanford NER in terms of computation time. Because GeoTxt is designed as a web API backend for big data processing, the two faster NER engines i.e. Stanford NER and Gate ANNIE are used in the current release.

After the identification of place names mentioned in text, GeoTxt uses the GeoNames geographic database (http://geonames.org) to

help geolocate those names. Although the GeoNames database is open source and available for free, its ranking mechanism is not public. Also, the GeoNames Search Web Service does not always come up with the ranking a human agent expects, e.g. it ranks *Colorado springs* and *San Luis Río Colorado* higher than *Colorado* (the State) when it is queried with the text *Colorado*. Therefore, we use (a) the geographic level, e.g. country, province, city of the place name in text when provided by the NER engine as best guess to initially rank and distinguish between candidates in the database, and then (b) the Levenshtein Distance of the name mentioned in text and the candidate's name (which indicates how close the two strings are) to choose the candidate with the least distance, and when multiple candidates have the same shortest distance to the name in question (c) the population of potential candidates with higher priority given to places with higher population to choose the best candidate.

The biggest challenge in geolocation is to disambiguate between places with identical names. The current implementation of GeoTxt leverages spatial logic to overcome such ambiguity in case more than one place name is mentioned within the same document. In such circumstances, GeoTxt retrieves the top five candidates for each place name mentioned in text, and for each candidate, retrieves all entities higher up in the geographical hierarchy. Then, for each name, the candidate for which an entity higher in its hierarchy matches any other place name mentioned in the text is picked. For example, in the tweet "*Finally landing in London. I love Canada*!", London will be geolocated to London, Ontario instead of London, UK; although the latter has higher population and stands higher in GeoNames ranking.

Figure 2. GeoTxt web User Interface.

Human users of GeoTxt have access to the same system through a visual web interface (see Figure 2). Users are able to paste in a piece of text, select the NER engine, and see the raw GeoJSON response in a text box and also geocoded locations overlayed on a base map. Each location is labeled with both the name that appears in text and the one picked by the GeoCoder module (in case they are different, whether due to a mistake or due to name abbreviation etc.) to help the user detect inaccurate results. Users are then able to flag any erroneous results as inaccurate. Such results are being used to monitor possible mistakes and improve the performance of GeoTxt. In the current release of GeoTxt, users (application and human) are able to select between two Named Entity Recognition engines of Stanford NER and GATE ANNIE to extract locations, people and organizations. But, the system has been designed to allow for additional NER engines to be added and compared, and also for versions of individual NER engines trained on different kinds of text to be selected. Enabling user feedback and the comparison of multiple NER engines are two unique features of GeoTxt compared to other efforts such as [3; 5; 9].

Play! Framework (http://www.playframework.com/) is used to expose GeoTxt functionality as a web API and to render User Interface and documentation web pages. The API provides versioned web services to guarantee backward compatibility.

3. FUTURE WORK

Plans for the future include customizing methods and NER tools to more comprehensibly utilize context in text and also context specific to microblog platforms (e.g. Twitter generated metadata) for the purpose of disambiguation, experimenting with ensemble approaches that combine the best of multiple methods, and utilizing detailed user feedback to improve the results. We are also starting to build a corpus of hand annotated microblogging posts to train the NER tools on such posts and to assess the overall system.

4. ACKNOWLEDGMENTS

This material is based in part upon work supported by the U.S. Department of Homeland Security under Award #2009-ST-061-CI0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

5. REFERENCES

[1] Cunningham, H., 2002. GATE, a general architecture for text engineering. *Computers and the Humanities 36*, 2, 223-254.

[2] Kitamoto, A. and Sagara, T., 2012. Toponym-based geotagging for observing precipitation from social and scientific data streams. In *Proceedings of the Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia* (Nara, Japan2012), ACM, 2390799, 23-26. DOI= http://dx.doi.org/10.1145/2390790.2390799.

[3] Leetaru, K.H., 2012. Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. *D-Lib Magazine 18*, 9, 5.

[4] Leidner, J.L., 2007. Toponym resolution in text.

[5] Lieberman, M.D. and Samet, H., 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* ACM, 731-740.

[6] Lingad, J., Karimi, S., and Yin, J., 2013. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on World Wide Web companion* International World Wide Web Conferences Steering Committee, 1017-1020.

[7] Ratinov, L. and Roth, D., 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* Association for Computational Linguistics, 147-155.

[8] Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., and Muhlhäuser, M., 2013. A Multi-Indicator Approach for Geolocalization of Tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[9] Speriosu, M. and Baldridge, J. Text-Driven Toponym Resolution using Indirect Supervision.