REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
22-08-2013	Final Report	23-Aug-2012 - 22-May-2013

4. TITLE AND SUBTITLE
Geo-Coding for the Mapping of Documents and Social Media Messages

5a. CONTRACT NUMBER
W911NF-12-1-0419

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS
Gelernter

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Carnegie Mellon University
Office of Sponsored Programs
5000 Forbes Ave
Pittsburgh, PA 15213 -3815

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62308-NS-II.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Many places on the earth have the same name, so it is difficult to determine which written place is meant. This research aims to improve the precision of geo-coding by using natural language processing and machine learning techniques (SVM specifically). We used data that was already geo-coded the ACE Spatial ML data set, and a large tweet set in which tweets were selected for having GPS locations (that we could use to improve validity). The report details our methods for text and microtext.

15. SUBJECT TERMS
toponym resolution, toponym disambiguation, grounding, geocoding, georeferencing, Twitter, microtext, tweet

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Judith Gelernter
UU	UU	UU	UU		19b. TELEPHONE NUMBER 412-268-4788

Report Title

Geo-Coding for the Mapping of Documents and Social Media Messages

ABSTRACT

Many places on the earth have the same name, so it is difficult to determine which written place is meant. This research aims to improve the precision of geo-coding by using natural language processing and machine learning techniques (SVM specifically). We used data that was already geo-coded the ACE Spatial ML data set, and a large tweet set in which tweets were selected for having GPS locations (that we could use to improve validity). The report details our methods for text and microtext.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

A description of our geo-coding methods for text and microtext have been presented at informal meetings of:

* Topographic Engineering Center of the U.S. Army Research Engineering and Development Center, June 11, 2013.

* Technology Symposium, Digital Intelligence and Investigation Directorate, Software Engineering Institute, June 12, 2013

* University of Maryland Institute of Advanced Computing Studies, July 1, 2013

Number of Presentations: 3.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Paper

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Haibo Wang	0.41	
Wei Zhang	0.70	
FTE Equivalent:	**1.11**	
Total Number:	**2**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Judith Gelernter	0.30	No
FTE Equivalent:	**0.30**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	Discipline
Juan Acosta	0.03	Modern Languages
Vivian Chang	0.01	Social & Decision Sciences
Molly Cook	0.00	Humanities & Arts
Luis Marquina	0.02	Statistics
Ava Murphey	0.01	History
FTE Equivalent:	**0.07**	
Total Number:	**5**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	0.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	1.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):	0.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Foreword

Geocoding is determining the physical location corresponding to a physical location identifier such as postal code, zip code, address or place name. This enables the visualization of each location on a geographic map. The importance of the process is demonstrated by the variety of geo-coders that, given a postal address, can output geo-coordinates. Texas A&M , the Google, and the TeleAtlas geo-coder are just a few examples. Postal and zip codes are unique, but place names are not, which makes toponym resolution an area of research. The extent of the ambiguity for place names is demonstrated by these Leetaru statistics:

" …the United States has a 40% overlap in geographic names while the United Kingdom has a 9.7% overlap. The average for all countries is 31% and when all locations worldwide are pooled together, 33% of locations share their name with another location somewhere else in the world. Northern Africa and the Middle East have fairly low levels of name duplication, while North and South America and South-Eastern Asia have higher-than normal concentrations of name duplication, suggesting a higher level of potential error in geocoding locations there."

List of Tables and Illustrations

Single attachment in a Microsoft Word document consisting of both:

Fig. 1: Gold standard Twitter data for scoring the geo-coding algorithm

Table 1: Performance comparison of our algorithm and a state-of-the-art algorithm on the Spatial ML text data

Problem studied

We looked at toponym resolution in text and microtext (tweets) for the purpose of making geographical maps to visualize the data.

Research question_1: Can we find the candidate in the gazetteer that is the correct match for each toponym in text?

Research question_2: Can we find the candidate in the gazetteer that is the correct match for each toponym in microtext? -- an unstudied problem.

Method to solve the problem

We first studied the identification of location words in text and microtext (Non-Geo/Geo disambiguation), and then we identified which physical location should be associated with which location word (Geo/Geo disambiguation) because many places in the world have the same name.

Geo-parsing as pre-requisite to geo-coding

We developed an algorithm to identify location words in text or tweet as part of a DARPA grant, Network Sciences Division , STTR A013015, Social Media in Strategic Communications Research . This algorithm relies both on string matching with a gazetteer and machine learning to determine based on sentence syntax which words are location words. Our geo-parser for the English language was used both to collect the data for the geo-coding problem, and as the first step to geo-coding.

Geo-coding

Our geo-coding algorithm uses machine learning to associate each location word extracted by the geoparser with candidate location words in the gazetteer, and to determine which gazetteer candidate is the most likely match with the word extracted. Classification algorithms are similar if not identical; it is the features used for the classification that make the algorithm perform adequately or optimally.

Classification features for geo-coding

Whether for text or tweet, geo-coding uses features of two types: those that indicate which gazetteer candidate is more probable, and those taken from text or tweet context that lend evidence to determine which physical location is signified by the location word extracted. Some of our features derive from heuristics known to be effective for geo-coding, as compiled in Leidner (2008).

Gazetteer features for disambiguating toponyms in text and tweets are three. When numerous gazetteer candidates are possible matches, prefer the gazetteer candidate that has a larger population, that has more alternate names in other languages, and is higher in the geo-spatial hierarchy.

Context features for disambiguating text and tweet provide additional clues as to what part of the world is being discussed so that the toponym will be resolved correctly. This is based on the assumption of spatial autocorrelation: that each location mentioned in a particular context is not independent, but is instead spatially correlated.

For text, the context consists of words in the same sentence or paragraph, and the entire document. For tweets, the context is other fields in the raw JSON file accompanying each tweet as distributed by the Twitter company. Relevant fields are time_zone (which is entered automatically by Twitter), and the user_location and user_description field which the user enters upon creating a Twitter account, and the GPS location of the tweet if the user has opted to include cell-phone coordinates with each tweet. Another way to gain insight into the geographic context of a particular tweet is to consult the geographic history of the user's past tweets. The hardware requirement for this in terms of memory, however, would be enormous, so we do not implement this idea.

Data set(s) to study the problem

Text. The Linguistic Data Consortium releases data sets to test algorithms. Spatial ML is a data set of news articles in which toponyms are identified, along with their latitude and longitude. We used version 2, which eliminates some errors and inconsistencies in the original version. Some sentences have toponyms or location adjectives (see "Canadian" in the passage below), and these have been hand-assigned latitude, longitude, with 4655 toponyms with valid latitude, longitude altogether. Control sentences have no toponyms. Bias in the data comes from the limited number of ambiguous toponyms in this data.

Annotations for the text. Here is an excerpt from the data:

Actually, it would not be advisable to accept Freddy's comments at face value. It is not acceptable policy and not legal to pay an honorarium or stipend to directors of charities. I checked with a lawyer familiar with <PLACE country="CA" form="NAM" gazref="IGDB:19437937" id="PI-14" latLong="60.000°N 96.000°W" predicative="true" type="COUNTRY">Canadian</PLACE> charity law, who explained that it is not permitted to remunerate board members of charities for their service as board members. The prohibition does not apply to directors of nonprofit organizations that are not charities.

We remove the < > tags for training and testing, because they contain the annotations we use for scoring.

Microtext. As there is no tweet set that has toponyms annotated along with the latitude and longitude, we are annotating a tweet set ourselves. We developed a toponym-rich data set by selecting tweets from January 1, 2013 to April 1, 2013 downloaded from the Carnegie Mellon University archive . We used JSON files for tweets that had GPS coordinates. We prepared the data by geo-parsing the tweet text field, and separately the user_location and user_description fields for toponyms, as well as the place fields. The place fields are the country and city name, the url for the city, and the coordinates for the bounding box of that city, as resolved automatically from the IP address from which the user entered the tweet. One annotator found 1678 entries for toponym1, the toponyms found in tweet text.

Annotations for the microtext. We gave the annotators tweets that had been pre-annotated by our geo-parser. That reduced the workload for each annotator since, instead of supplying annotations for each tweet, it was necessary in many cases only to verify that the geo-parser output was correct, or correct the geo-parser output if wrong. The annotators were asked also to add the geo-coordinates for each entry in the toponym1 field (location words in tweet text) and in the toponym2 field (location words in user_location, user_description and time_zone).

The annotator project took months, as the coding is time consuming and the coders were asked to stop their work rather than continue when tired. The same was true for the adjudication of the toponyms. A data set with mistakes will train the algorithm to make mistakes, so a gold standard data set is critical. The result is that this part of the project is on-going.

We gave annotators a definition of what constitutes a location, and a spreadsheet with toponyms to code for the tweet_text field, and for the other tweet fields that might contain toponyms. Two people coded the same tweets independently. We then had a third person adjudicate and make the final decision as to what the coding should be. Fig. 1 shows an example of the data with the Toponym1 annotations for training and then scoring the algorithm.

Results

Our score in toponym resolution in text surpassed state-of-the-art experiments on the same data set. The Lieberman et al., 2012 team on Spatial ML reported a precision of .99 and a recall of .70, with an F1 of .82. Our score for precision was .94 but our recall was even higher at .95, giving us an F1 of .94 (Table 1).

Preliminary geospatial error analysis shows that only .07% of algorithm output was the wrong country, with .4% of algorithm output in the correct country but the wrong state, and .07% of algorithm output imprecise while in the correct country and state.

We are updating the code, and these statistics might change somewhat. We do not yet have the scores for toponym resolution in tweet because the tweet annotations are still being completed.

Summary of Most Important Accomplishments

(1) End-to-end system to geo-parse and geo-code text in English
(2) End-to-end system to geo-parse and geo-code tweet text in English
(3) Creation of a data set that can be used to test tweet geo-coding
(4) Script to score precision and recall for the geo-coding algorithm
(5) Script to evaluate geospatial error for the geo-coding algorithm

(1-2) End-to-end system. The end to end system comprises a geo-parser and a geo-coder that disambiguates toponyms. The geo-parser was discussed in an earlier paper (Gelernter and Zhang, 2013), and is now in version 2.0.3. Strong toponym disambiguation results – that is, accurate geo-coding – is contingent upon strong results in identifying toponyms, or geo-parsing. The geo-coder uses the geo-parser output as input to find the best gazetteer match. The geo-coder is based on a machine learning algorithm (SVM) to determine, of the possible matches with the given toponym, which gazetteer toponym is the most likely match. A score based on the composite of the learning features is assigned to each gazetteer candidate. The candidate with the highest score is the algorithm output, along with the distance between the two candidates in kilometers and the top candidate's country and state/province.

(3) Data for testing. Our gold-standard data set for Twitter (Fig. 1) will be finished within a few weeks, and will be usable by other researchers, provided that the Twitter company imposes no legal restrictions.

(4) Algorithm to score geo-coding results. Geographic information retrieval is a subset of information retrieval, and as such, is evaluated with the same recall and precision metrics (Martins et al, 2005). We have built-on a script that score results, so that we can measure algorithm improvements easily, and so that other users can measure their results.

(5) Script to evaluate geo-coding results. The spatial element, if included in a geo-coding algorithm evaluation, is sometimes given as a physical measurement along the earth (for example, 366 km. as the difference between the actual and expected answer). However, the number of on-the-ground units between the actual and expected location is less important than whether the geospatial hierarchy was preserved, since a relatively small offset in Europe might mean that the toponym was resolved to the wrong country – a worse mistake that resolving to the right country but the wrong state. So while presenting our results in the standard format of precision and recall, we also introduce a method of evaluation that is more logical to the data than typical methods.

Bibliography

Agirre, E., and de Lacalle, O.L. (2007). UBC-ALM: Combining KNN with SVD for WSD. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czechoslovakia, Association for Computational Linguistics, 342–345.

Blessing, A., Kuntz, R. and Schütze, H.. (2007). Towards a context model driven German geo-tagging system. Proceedings of the 4th ACM workshop on Geographical information retrieval CIKM '07 Conference on Information and Knowledge Management Lisbon, Portugal, November 6-10, 2007, 25-30.

Blessing, A. and Schütze, H. (2008). Automatic acquisition of vernacular places. Proceedings of iiWAS2008, November 24-26, 2008, Linz, Austria, 662-665.

Buscaldi , D., Magnini, B. (2010) Grounding toponyms in an Italian local news corpus. Proceedings of the 6th Workshop on Geographic Information Retrieval, February 18-19, 2010, Zurich, Switzerland, [5 p.]

Gale, W., Church, K., and Yarowsky, D. (1992). One sense per discourse. In Proceedings of the 4th DARPA Speech and Natural Language Workshop. pp. 233-237, 1992. Retrieved June 17, 2013 from http://acl.ldc.upenn.edu/H/H92/H92-1045.pdf

Gelernter, J. and Zhang, W. (2013). A geo-parser for Spanish. Under review for ACM SigSpatial.

Gimpel, E., Schneider, N., O'Connor, B., Das, D., Mills, D. Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J, Smith, N.A. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. Proceedings of the Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, June 2011, 42-47. Http://www.ark.cs.cmu.edu/TweetNLP

Goldberg, D. W., Cockburn, M. G. (2010a). Improving Geocode Accuracy with Candidate Selection Criteria. Transactions in GIS, 2010, 14(s1), 149–176.

Goldberg, D. W., Wilson, J. P., & Cockburn, M. G. (2010b). Toward quantitative geocode accuracy metrics. Proceedings of the

Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, July 20-23, Leicester, U.K., 329-332.

Hall, M.M. and Jones, C.B. (2008). Quantifying spatial prepositions: an Experimental Study. ACM GIS '08, November 5-7, 2008, Irvine, CA, [4 p.].

Hosokawa, Y. (2012). Improving vertical geo/geo disambiguation by increasing geographical feature weights of places. Proceeding of RACS '12 Proceedings of the 2012 ACM Research in Applied Computation Symposium, October 23-26, 2012, San Antonio, TX, USA, 92-99.

Ireson, N. and Ciravegna, F. (2010). Toponym resolution in social media. In P.F. Patel-Schneider et al (Eds.) ISWC 2010, part I, LNCS 6496, 370-385.

Jameel, M. S. and Chingtham, T.S. (2009). Compounded uniqueness level: geo-location indexing using address parser. International Journal of Computer Theory and Engineering, IJCTE, 1 (1), April 2009

Leetaru, K. H. (2012). Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. D-Lib Magazine 18(9-10), http://www.dlib.org/dlib/september12/leetaru/09leetaru.html

Leidner, J. L. (2008) Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. Dissertation, University of Edinburgh.

Leveling, J. and Hartrumpf, S. (2007). Inferring location names for geographic information retrieval. C. Peters et al (Eds.) CLEF 2007, LNCS 5152, 773-780.

Li, H., Srihari, R. K., Niu, C. and Li, W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References-Volume 1. May 31, 2003, Edmonton, Alberta, Association for Computational Linguistics, 2003 (pp. 39-44).

Lieberman, M. D., and Samet, H. (2012). Adaptive context features for toponym resolution in streaming news. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012, August 12-16, 2012, Portland, OR, USA, 731-740.

Martins, B., Silva, M. J. and Chaves, M.S. (2005) Challenges and Resources for evaluating geographical IR. GIR'05, November 4, 2005, Bremen, Germany. [5 p.]

Nothman, J., Honnibal, M., Hachey, B., & Curran, J. R. (2012). Event linking: Grounding event reference in a news archive. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 10 July 2012, Jeju Island, Korea, 228-232.

Purves, R., Clough, P, Joho, J. (2005). Identifying imprecise regions for geographic information retrieval using the web. In Proceedings of the GIS Research UK 13th Annual Conference (2005), pp. 313–318. Retrieved June 17, 2013 from http://www.dcs.gla.ac.uk/~hideo/pub/gisuk05/gisuk05.pdf

Roberts, K., Bejan, C. A., & Harabagiu, S. (2010). Toponym disambiguation using events. In Twenty-Third International Florida Artificial Intelligence Research Society Conference Proc. FLAIRS, May 19-21, Daytona Beach, Florida, 271-276.

Sano, T., Nobesawa, S.H., Okamoto, H., Susuki, H., Matsubara, M, Saito, H. (2009). Robust Toponym Resolution Based on Surface Statistics. IEICE Transactions on Information and Systems 92 (12), 2313-2320.

Schilder, F., Versley, Y. and Habel, C. (2004) Extracting spatial information: grounding, classifying and linking spatial expressions. Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004. [3 p.]

Schlieder, C. and Henrich, A. (2011) Spatial grounding with vague place models. SigSpatial 3(2), 20-23.

Smith, D, and Crane, G. (2001). Disambiguating geographic names in a historical digital library. Research and Advanced Technology for Digital Libraries (pp. 127-136). http://www.ccs.neu.edu/home/dasmith/geodl01.pdf

Speriosu, M. and Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013) Sofia, Bulgaria, August 4-9, 2013 [10 p.]

Wang, X., Zhang, Y., Chen, M., Lin, X., Yu, H., & Liu, Y. (2010, June). An evidence-based approach for Toponym Disambiguation. In Geoinformatics, 2010 18th International Conference on IEEE, 18-20 June, Beijing, China, 1-7.

Wing, B. P., and Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. 2011 (pp. 955-964).

Yarowsky, D. (1993) One sense per collocation. In Proceedings, ARPA Human Language Technology Workshop. March 21-24, Plainsboro, NJ, 266-271. http://acl.ldc.upenn.edu/H/H93/H93-1052.pdf

Yuan, Y. (2011). Extracting spatial relations from document <sic> for geographic information retrieval. 19th International Conference on Geoinformatics, 24-26 June, Shanghai, China, 1-5.

Technology Transfer

Tweet text	Toponym1
Damn I miss Cebu City! Good morning Pinas!	tp{**Pinas**[14,121]}tptp{**Cebu City**[10,124]}tp
Ringing in the new year with two of my new best friends from 2012 at @republicMN in Minneapolis :)	tp{**Minneapolis**[45,-93]}tp

Fig. 1 Two tweets and the toponyms extracted by our geo-parser and verified by an annotator, with the annotator-added geo-coordinates of each toponym included.

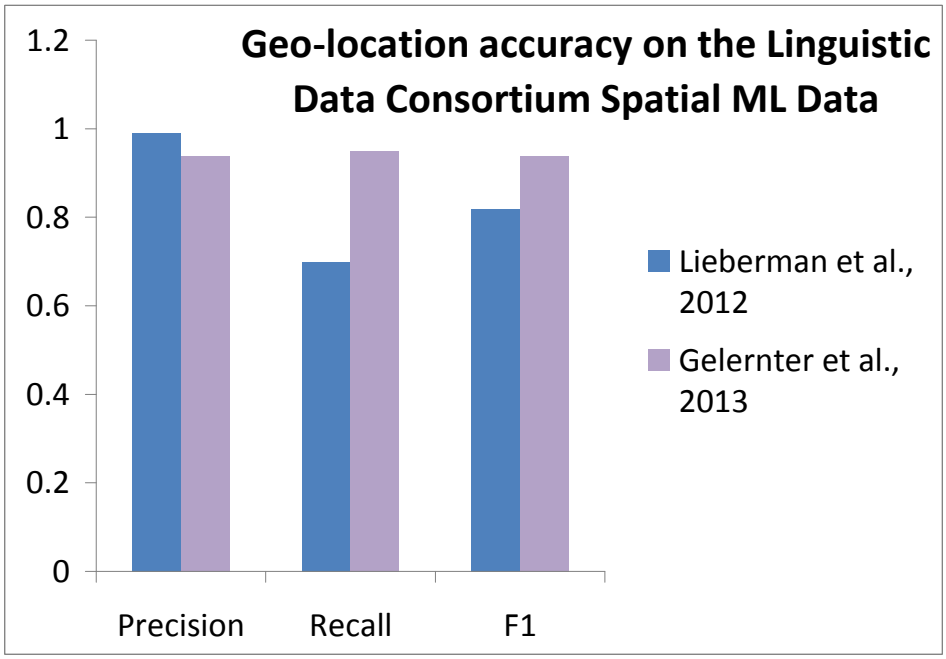

Table 1: Performance comparison on the Spatial ML data.