The role of geographic knowledge in sub-city level geolocation

Laura Di Rocco
Università di Genova, DIBRIS
Genova, Italy
laura.dirocco@dibris.unige.it

Davide Buscaldi
Universitè Paris 13, LIPN
Paris, France
davide.buscaldi@lipn.univ-paris13.fr

Michela Bertolotto
University College Dublin (UCD)
Dublin, Ireland
michela.bertolotto@ucd.ie

Barbara Catania
Università di Genova, DIBRIS
Genova, Italy
barbara.catania@unige.it

Giovanna Guerrini
Università di Genova, DIBRIS
Genova, Italy
giovanna.guerrini@unige.it

ABSTRACT

Geolocation of microblog messages has been largely investigated in the literature. Many solutions have been proposed that achieve good results at the city level. Existing approaches are mainly data-driven (i.e., they rely on a training phase). However, the development of algorithms for geolocation at sub-city level is still an open problem. In this paper, we investigate the role that external geographic knowledge can play in geolocation approaches. We show how different geographical data sources can be combined with a semantic layer within a knowledge base to achieve reasonably accurate sub-city level geolocation.

KEYWORDS

Geographic data source, geographic ontology, geolocation algorithm, microblog message

ACM Reference Format:
Laura Di Rocco, Davide Buscaldi, Michela Bertolotto, Barbara Catania, and Giovanna Guerrini. 2019. The role of geographic knowledge in sub-city level geolocation. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3297280.3297557

1 INTRODUCTION

Microblog message mining has recently gathered a lot of attention as a viable approach for identifying social trends, enabling emergency response applications, and even predicting physical and social phenomena [1, 4]. Many methods rely on the availability of already geotagged messages, which contain coordinates on where the user was located when the message was sent. Nevertheless, for either technical or privacy reasons, the majority of messages does not include spatial coordinates. However, it might be possible to infer them by analysing the content of the message (e.g., if it contains names of specific places). To exploit the intuition that users often mention places that are near their current location, several approaches to automatically geolocate non-geotagged messages using textual content have been developed [3–5, 10]. Most of these methods rely on a training phase, during which they construct language models, in order to probabilistically infer the location of unseen messages. These types of models can very accurately geolocate microblog messages at a city level [3, 5] but suffer from problems related to text noise (e.g., use of slang, links, mis-spellings). Moreover, during classification, the finer the grid used to geolocate is (i.e., the higher sub-city detail), the higher the number of options to choose from, and this negatively affects performance. Other methods (e.g., [10]) investigate the use of explicitly mentioned location information in microblog messages. Such methods rely on manually labeled messages in order to be trained. When such manually preprocessed datasets are not available, or a higher detail is required [6], data-driven methods fail. Our claim is that, in such cases, an external information source can be exploited. This is facilitated by the fact that many publicly available geographic information sources have recently been developed. Our second intuition is that an additional semantic level (e.g., an ontology) on top of these geographic sources will increase geolocation accuracy. The obtained geolocation methods, relying on such geographic knowledge, are called *knowledge-driven.*

In this work, we study the impact, in terms of geolocation accuracy, of exploiting diverse geographic information sources in a naive knowledge-driven geolocation algorithm. Specifically, we analyze both widely accepted (semi-authoritative) data sources, such as Geonames[1], and crowdsourced geographical data, such as OpenStreetMap[2] (OSM). While GeoNames contains an associated semantic level, this is not the case for crowdsourced data (e.g., OSM). For this reason, we consider a semantically enriched version of OSM, called LinkedGeoData [8] (LGD), as well as an external ontology for conceptualized cities called OpenStreetMap Facet Ontology [2].

2 ALGORITHMS

In order to support the intuition that an algorithm that exploits a geographical knowledge to geolocate a microblog message at sub-city level achieves a higher accuracy than a data-driven algorithm, we introduce in this section a naive Knowledge-Driven (KD) algorithm. Given a set of terms in a microblog message, KD identifies which terms are in the geographic knowledge and extracts the physical locations (points) associated with them. In order to infer the geographic position of the message, it calculates the average latitude and longitude of the obtained set of points.

SAC '19, April 8–12, 2019, Limassol, Cyprus
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5933-7/19/04.
https://doi.org/10.1145/3297280.3297557

[1] http://geonames.org
[2] http://openstreetmap.org

In our experiments we compare KD with Geoloc [5], a state of the art text-based data-driven geolocation algorithm. In order to correctly infer the position of a message, it relies on a geographic grid. We choose Geoloc as it is one of the newest algorithms that have the same input (microblog messages) as KD. Moreover, the authors performed a comprehensive comparative evaluation with previous algorithms.

3 DATASETS AND METRICS

We considered two datasets of tweets. The first, GeoText[3], was described for the first time in [3] and later used for comparison and evaluation of several approaches [5, 7, 9]. The dataset was retrieved from the official Twitter Streaming API[4] in the first week of March 2010, by keeping only messages associated with coordinates, i.e., geotagged messages. The dataset was preprocessed to be used as input for a topic extraction algorithm. To this aim, only tweets of users that wrote at least 20 messages in the considered period, follow less than 1000 other users, and have less than 1000 followers were taken into account. The second, FollowTheHashtag[5], contains geotagged tweets retrieved over 167h that correspond to 7 days from 14/04/2016 to 21/04/2016, after removing retweets. No further preprocessing was applied to the retrieved messages.

Since our aim is to increase geolocation accuracy at sub-city level, working on a given target area, we selected from GeoText only tweets geolocated in New York City, NY, USA and from FollowTheHashtag tweets geolocated in Greater London, UK. Hereafter, the retrieved datasets are denoted by NY and London, respectively. Table 1 shows the number of respective tweets.

We selected the information related to these two cities also in the semantic gazetteers. For *GeoNames* we downloaded Great Britain and USA information and filtered only data related to London and NY. On LGD we executed two SPARQL queries (one for each target area) on the LGD endpoint. Due to server limitations, each result was limited to 50,000 entry. For OSM, we first downloaded all the data in London and NY and then filtered the entire OSM dataset with OSM facet ontology obtaining a different subset of OSM data[6] w.r.t. LGD. Table 1 shows the difference between the number of entries in each semantic gazetteers and the number of terms in each semantic gazetteers for NY and London as well as other statistics.

To evaluate the results of KD and Geoloc, we start from datasets where each tweet m is associated with a location $loc_r(m)$. This is our ground-truth for the evaluation. For the analysis, we choose a commonly adopted distance-based evaluation metric: Accuracy Distance Error (ADE). The metric is defined in terms of the Distance Error $DE(m)$, computed for each tweet m and defined as the Euclidean distance d between the location originally associated with m, $\mathrm{loc}_r(m)$, and the inferred location, $\mathrm{loc}(m)$. ADE is defined as the ratio between the number of tweets with distance error lower than a given threshold *dist* and the total number of tweets. Moreover, we analyze the number of tweets that a KD can geolocate providing a measure that we call GeoTweet Percentage (GTP). GTP is defined as the ratio between the localizable tweets, i.e., the messages that contain at least a term that exists in the semantic gazetteer, and the total number of tweets present in the dataset. We finally provide an *ambiguity value measure*, in order to compare semantic gazetteers, defined as: $\alpha = \frac{\text{nuumber of entities}}{\text{number of terms}}$. This value represents the number of instances containing the same toponym in the semantic gazetteer. As we can see in Table 1 , the data for London is less ambiguous than NY (on average).

Table 1: Summary of semantic gazetteers information.

	NY			London		
	LGD	GeoNames	OSM	LGD	GeoNames	OSM
Number of terms	5,951	2,397	10,523	6,448	2,452	18,393
Number of entities	1,048,576	10,360	55,506	283,830	10,076	115,920
Ambiguity α	176.20	4.32	5.27	44.01	4.10	6.30
Tot tweets in area		95,775			44,152	
GTP	0.04	0.007	0.16	0.86	0.33	0.58

4 EXPERIMENTAL RESULTS

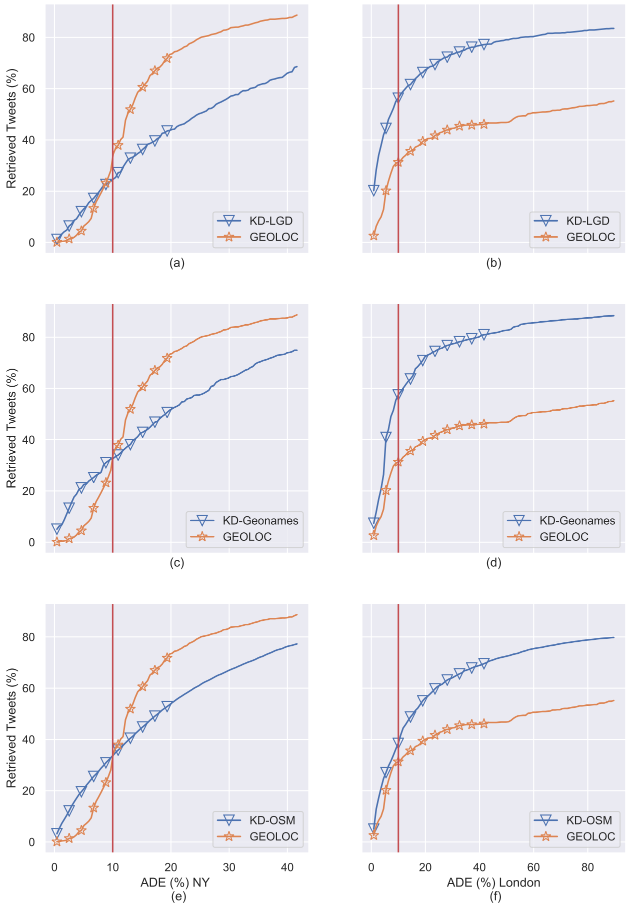

Figure 1: ADE results using three semantic gazetteers.

In our experiments, we use a threshold equal to 10% of the total geolocalizable area, and we consider it to be the maximum acceptable value that represent the sub-city level. We investigate how the percentage of tweets that we can correctly geo-localize changes as a function of the error with which we geolocate it. The results are

[3] www.cs.cmu.edu/~ark/GeoText/

[4] developer.twitter.com/en/docs

[5] www.followthehashtag.com/datasets/170000-uk-geolocated-tweets-free-twitter-dataset/

[6] From now on, we call the geographic knowledge obtained with OSM facet ontology as OSM for readability.

shown in Figure 1, where we compare KD with Geoloc in terms of the percentage of correctly geolocated tweets, while increasing the geolocation error. Figures 1(a)(c)(e) refer to NY and Figures 1(b)(d)(f) refer to London. Moreover, Figures 1(a)(b) correspond to the LGD semantic gazetteer, while Figures 1(c)(d) are related to GeoNames and Figures 1(e)(f) to OSM. With the red vertical line, we highlight the accuracy threshold we are interested in. We can immediately see that in all cases, at the sub-city level, a knowledge-driven approach is better than a data-driven one.

Accuracy. In Figure 2, we see how the percentage of retrieved tweets changes as we increase the error threshold. The first insight that we get is that there is no globally optimal semantic gazetteer, as different gazetteers provide higher retrieval ratios across different datasets. In NY (Figure 2 left) we see that the best performing semantic gazetteer is OSM, while in London (Figure 2 right) the best one is LGD. We also see that this ranking changes according to the threshold at which we are interested in. For example, in London, the ranking between GeoNames and LGD changes as the retrieval percentage for GeoNames crosses over the one of LGD, for ADE values larger than 10%. Finally, Table 1 shows that GTP differs across the three semantic gazetteers. In order to find a semantic gazetteer that works best for a given setting, our goal is to identify a good trade-off between accuracy and GTP. LGD and OSM are the best performing solutions since they exhibit the best values for such trade-off. Their equivalence in quality is not surprising, as we already know that they are derived from the same data source, and just filtered in different ways.

Ambiguity issues. Our results show that crowdsourced knowledge can be used to provide more accurate geolocation of microblog messages. Looking at both Table 1 and Figure 2, while also considering that LGD and OSM contain information from the same geographic data source, we see that different sources can have varying levels of ambiguity. Looking at our results for NY, we see that LGD has the highest ambiguity and the worst results. However, OSM and GeoNames, which have similar and smaller ambiguity, have similar and more accurate results. We also highlight an important difference in terms of GTP. Table 1 shows that GeoNames can geolocate only a very small percentage of the microblog messages w.r.t. OSM. This behavior is the same for London: OSM has an ambiguity level closer to GeoNames but they are very different in terms of GTP. In London, GeoNames has the worst GTP but produces better accuracy.

Random test. Our results are compared with the ones obtained using a random geolocation approach (Random), in order to demonstrate that our results are free of bias., i.e., the small targeted area (a city) does not have an impact on the higher accuracy of the results. Finally, we highlight that, as expected, it is not possible to infer a location within a city by just randomly picking a point in it.

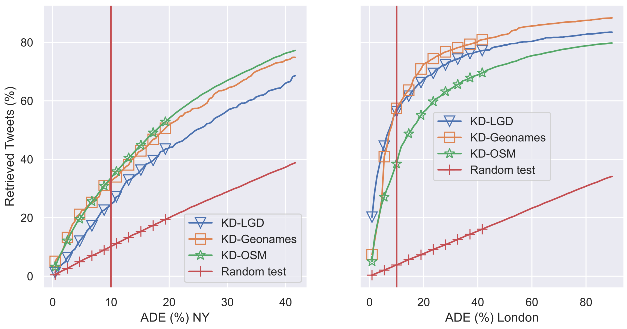

Figure 2: Comparison among the three semantic gazetteers: results for NY (left), results for London (right).

5 CONCLUSIONS AND FUTURE WORK

Our work shows the role of semantic gazetteers in geolocation algorithms and the impact of different gazetteers. It also demonstrates that knowledge-driven approaches work better at sub-city level then data-driven algorithms. We do not propose a general new solution for geolocating microblog messages, but rather another level of geolocation, where data-driven algorithms cannot achieve good results. We use a knowledge-driven approach in a pipeline after a data-driven algorithm. Data-driven algorithms, indeed, achieve very good results at city level. We are currently developing an improved KD algorithm that overcomes limitations of the naive approach presented here and better exploits semantics to further improve accuracy.

ACKNOWLEDGMENTS

We would like to thank the program “Investissements d’Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL) for the support given to this work.

REFERENCES

[1] Carlos Castillo. 2016. *Big Crisis Data: Social Media in Disasters and Time-critical Situations.* Cambridge University Press.

[2] Laura Di Rocco. 2013. Semantic Enhancement of Volunteered Geographical Information. *Master Thesis, University of Genova, Italy.* www.dibris.unige.it/en/di-rocco-laura

[3] Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric P Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 1277–1287.

[4] Judith Gelernter, Gautam Ganesh, Hamsini Krishnakumar, and Wei Zhang. 2013. Automatic Gazetteer Enrichment with User-geocoded Data. *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD ’13* (2013), 87–94.

[5] Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel Density Estimation for Text-Based Geolocation.. In *AAAI.* 145–150.

[6] Sheila Kinsella, Vanessa Murdock, and Neil O’Hare. 2011. "I’M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC ’11).* ACM, New York, NY, USA, 61–68.

[7] Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. 2014. Inferring the Origin Locations of Tweets with Quantitative Confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing.* ACM, 1523–1536.

[8] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. 2012. Linkedgeodata: A core for a web of spatial open data. *Semantic Web* 3, 4 (2012), 333–354.

[9] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Who, Where, When and What: Discover Spatio-temporal Topics for Twitter Users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 605–613.

[10] Wei Zhang and Judith Gelernter. 2014. Geocoding Location Expressions in Twitter Messages: A Preference Learning Method. *Journal of Spatial Information Science* 2014, 9 (2014), 37–70.