A Stacking-based Approach to Twitter User Geolocation Prediction

Bo Han,[♠♡] **Paul Cook,**[♡] and **Timothy Baldwin**[♠♡]
♠ NICTA Victoria Research Laboratory
♡ Department of Computing and Information Systems, The University of Melbourne
hanb@student.unimelb.edu.au, paulcook@unimelb.edu.au,
tb@ldwin.net

Abstract

We implement a city-level geolocation prediction system for Twitter users. The system infers a user's location based on both tweet text and user-declared metadata using a stacking approach. We demonstrate that the stacking method substantially outperforms benchmark methods, achieving 49% accuracy on a benchmark dataset. We further evaluate our method on a recent crawl of Twitter data to investigate the impact of temporal factors on model generalisation. Our results suggest that user-declared location metadata is more sensitive to temporal change than the text of Twitter messages. We also describe two ways of accessing/demoing our system.

1 Introduction

In this paper, we present and evaluate a geolocation prediction method for Twitter users.[1] Given a user's tweet data as input, the task of user level geolocation prediction is to infer a primary location (i.e., "home location": Mahmud et al. (2012)) for the user from a discrete set of pre-defined locations (Cheng et al., 2010). For instance, President Obama's location might be predicted to be Washington D.C., USA, based on his public tweets and profile metadata.

Geolocation information is essential to location-based applications, like targeted advertising and local event detection (Sakaki et al., 2010; MacEachren et al., 2011). However, the means to obtain such information are limited. Although Twitter allows users to specify a plain text description of their location in their profile, these descriptions tend to be ad hoc and unreliable (Cheng et al., 2010). Recently, user geolocation prediction based on a user's tweets has become popular (Wing and Baldridge, 2011; Roller et al., 2012), based on the assumption that tweets implicitly contain locating information, and with appropriate statistical modeling, the true location can be inferred. For instance, if a user frequently mentions *NYC*, *JFK* and *yankees*, it is likely that they are from New York City, USA.

In this paper, we discuss an implementation of a global city-level geolocation prediction system for English Twitter users. The system utilises both tweet text and public profile metadata for modeling and inference. Specifically, we train multinomial Bayes classifiers based on location indicative words (LIWs) in tweets (Han et al., 2012), and user-declared location and time zone metadata. These base classifiers are further stacked (Wolpert, 1992) using logistic regression as the meta-classifier. The proposed stacking model is compared with benchmarks on a public geolocation dataset. Experimental results demonstrate that our stacking model outperforms benchmark methods by a large margin, achieving 49% accuracy on the test data. We further evaluate the stacking model on a more recent crawl of public tweets. This experiment tests the effectiveness of a geolocation model trained on "old" data when applied to "new" data. The results reveal that user-declared locations are more variable over time than tweet text and time zone data.

2 Background and Related Work

Identifying the geolocation of objects has been widely studied in the research literature over target objects including webpages (Zong et al., 2005), search queries (Backstrom et al., 2008), Flickr images (Crandall et al., 2009) and Wikipedia editors (Lieberman and Lin, 2009). Recently, a considerable amount of work has been devoted to extending geolocation prediction for Twitter

[1] We only use public Twitter data for experiments and exemplification in this study.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 7–12,
Sofia, Bulgaria, August 4-9 2013.

users (Cheng et al., 2010; Eisenstein et al., 2010). The geolocations are usually represented by unambiguous city names or a partitioning of the earth's surface (e.g., grid cells specified by latitude/longitude). User geolocation is generally related to a "home" location where a user regularly resides, and user mobility is ignored. Twitter allows users to declare their home locations in plain text in their profile, however, this data has been found to be unstructured and ad hoc in preliminary research (Cheng et al., 2010; Hecht et al., 2011).

While popular for desktop machine geolocation, methods that map IP addresses to physical locations (Buyukokkten et al., 1999) cannot be applied to Twitter-based user geolocation, as IPs are only known to the service provider and are non-trivial to retrieve in a mobile Internet environment. Although social network information has been proven effective in inferring user locations (Backstrom et al., 2010; Sadilek et al., 2012; Rout et al., 2013), we focus exclusively on message and metadata information in this paper, as they are more readily accessible.

Text data tends to contain salient geospatial expressions that are particular to specific regions. Attempts to leverage this data directly have been based on analysis of gazetted expressions (Leidner and Lieberman, 2011) or the identification of geographical entities (Quercini et al., 2010; Qin et al., 2003). However these methods are limited in their ability to capture informal geospatial expressions (e.g. *Brissie* for *Brisbane*) and more non-geospatial terms which are associated with particular locations (e.g. *ferry* for Seattle or Sydney).

Beyond identifying geographical references using off-the-shelf tools, more sophisticated methods have been introduced in the social media realm. Cheng et al. (2010) built a simple generative model based on tweet words, and further added words which are local to particular regions and applied smoothing to under-represented locations. Kinsella et al. (2011) applied different similarity measures to the task, and investigated the relative difficulty of geolocation prediction at city, state, and country levels. Wing and Baldridge (2011) introduced a grid-based representation for geolocation modeling and inference based on fixed latitude and longitude values, and aggregated all tweets in a single cell. Their approach was then based on lexical similarity using KL-divergence. One drawback to the uniform-sized cell representation is that it introduces class imbalance: urban areas tend to contain far more tweets than rural areas. Based on this observation, Roller et al. (2012) introduced an adaptive grid representation in which cells contain approximately the same number of users, based on a KD-tree partition. Given that most tweets are from urban areas, Han et al. (2012) consider a city-based class division, and explore different feature selection methods to extract "location indicative words", which they show to improve prediction accuracy. Additionally, time zone information has been incorporated in a coarse-to-fine hierarchical model by first determining the time zone, and then disambiguating locations within it (Mahmud et al., 2012). Topic models have also been applied to the task, in capturing regional linguistic differences (Eisenstein et al., 2010; Yin et al., 2011; Hong et al., 2012).

When designing a practical geolocation system, simple models such as naive Bayes and nearest prototype methods (e.g., based on KL divergence) have clear advantages in terms of training and classification throughput, given the size of the class set (often numbering in the thousands of classes) and sheer volume of training data (potentially in the terabytes of data). This is particularly important for online systems and downstream applications that require timely predictions. As such, we build off the text-based naive Bayes-based geolocation system of Han et al. (2012), which our experiments have shown to have a good balance of tractability and accuracy. By selecting a reduced set of "location indicative words", prediction can be further accelerated.

3 Methodology

In this study, we adopt the same city-based representation and multinomial naive Bayes learner as Han et al. (2012). The city-based representation consists of 3,709 cities throughout the world, and is obtained by aggregating smaller cities with the largest nearby city. Han et al. (2012) found that using feature selection to identify "location indicative words" led to improvements in geolocation performance. We use the same feature selection technique that they did. Specifically, feature selection is based on information gain ratio (IGR) (Quinlan, 1993) over the city-based label set for each word.

In the original research of Han et al. (2012),

only the text of Twitter messages was used, and training was based exclusively on geotagged tweets, despite these accounting for only around 1% of the total public data on Twitter. In this research, we include additional non-geotagged tweets (e.g., posted from a non-GPS enabled device) for those users who have geotagged tweets (allowing us to determine a home location for the user).

In addition to including non-geotagged data in modeling and inference, we further take advantage of the text-based metadata embedded in a user's public profile (and included in the JSON object for each tweet). This metadata is potentially complementary to the tweet message and of benefit for geolocation prediction, especially the user-declared location and time zone, which we consider here. Note that these are in free text rather than a structured data format, and that while there are certainly instances of formal place name descriptions (e.g., *Edinburgh, UK*), they are often informal (e.g., *mel* for Melbourne). As such, we adopt a statistical approach to model each selected metadata field, by capturing the text in the form of character 4-grams, and training a multinomial naive Bayes classifier for each field.

To combine together the tweet text and metadata fields, we use stacking (Wolpert, 1992). The training of stacking consists of two steps. First, a multinomial naive Bayes base classifier (*L0*) is learned for each data type using 10-fold cross validation. This is carried out for the tweet text (TEXT), user-declared location (MB-LOC) and user-declared time zone (MB-TZ). Next, a meta-classifier (*L1* classifier) is trained over the base classifiers, using a logistic regression learner (Fan et al., 2008).

4 Evaluation and Discussion

In this section, we compare our proposed stacking approach with existing benchmarks on a public dataset, and investigate the impact of time using a recently collected dataset.

4.1 Evaluation Measures

In line with other work on user geolocation prediction, we use three evaluation measures:

- **Acc** : The percentage of correct city-level predictions.
- **Acc@161** : The percentage of predicted locations which are within a 161km (100 mile) radius of the home location (Cheng et al., 2010), to capture near-misses (e.g., Edinburgh UK being predicted as Glasgow, UK).
- **Median** : The median distance from the predicted city to the home location (Eisenstein et al., 2010).

Methods	Acc	Acc@161	Median
KL	.117	.277	793
MB	.126	.262	913
KL-NG	.260	.487	181
MB-NG	.280	.492	170
MB-LOC	.405	.525	92
MB-TZ	.064	.171	1330
STACKING	.490	.665	9

Table 1: Results over WORLD

4.2 Comparison with Benchmarks

We base our evaluation on the publicly-available WORLD dataset of Han et al. (2012). The dataset contains 1.4M users whose tweets are primarily identified as English based on the output of the `langid.py` language identification tool (Lui and Baldwin, 2012), and who have posted at least 10 geotagged tweets. The city-level home location for a geotagged user is determined as follows. First, each of a user's geotagged tweets is mapped to its nearest city (based on the same set of 3,709 cities used for the city-based location representation). Then, the most frequent city for a user is selected as the home location.

To benchmark our method, we reimplement two recently-published state-of-the-art methods: (1) the KL-divergence nearest prototype method of Roller et al. (2012) based on KD-tree partitioned grid cells, which we denote as KL; and (2) the multinomial naive Bayes city-level geolocation model of Han et al. (2012), which we denote as MB. Because of the different class representations, Acc numbers are not comparable between the benchmarks. To remedy this, we find the closest city to the centroid of each grid cell in the KD-tree representation, and map the classification onto this city. We present results including non-geotagged data for users with geotagged messages for the two methods, as KL-NG and MB-NG, respectively. We also present results based on the user-declared location (MB-LOC) and time zone (MB-TZ), and finally the stacking method (STACKING) which combines MB-NG, MB-LOC and MB-TZ. The results are shown in Table 1.

The approximate doubling of Acc for KL-NG and MB-NG over KL and MB, respectively, demonstrates the high utility of non-geotagged data in tweet text-based geolocation prediction. Of the two original models, we can see that MB is comparable to KL, in line with the findings of Han et al. (2012). The MB-LOC results are by far the highest of all the base classifiers. Contrary to the suggestion of Cheng et al. (2010) that user-declared locations are too unreliable to use for user geolocation, we find evidence indicating that they are indeed a valuable source of information for this task. The best overall results are achieved for the stacking approach (STACKING), assigning almost half of the test users to the correct city-level location, and improving more than four-fold on the previous-best accuracy (i.e., MB). These results also suggest that there is strong complementarity between user metadata and tweet text.

4.3 Evaluation on Time-Heterogeneous Data

In addition to the original held-out test data (WORLD$_{test}$) from WORLD, we also developed a new geotagged evaluation dataset using the Twitter Streaming API.[2] This new LIVE$_{test}$ dataset is intended to evaluate the impact of time on predictive accuracy. The training and test data in WORLD are time-homogeneous as they are randomly sampled from data collected in a relatively narrow time window. In contrast, LIVE$_{test}$ is much newer, collected more than 1 year later than WORLD. Given that Twitter users and topics change over time, an essential question is whether the statistical model learned from the "old" training data is still effective over the "new" test data?

The LIVE$_{test}$ data was collected over 48 hours from 2013/03/03 to 2013/03/05. By selecting users with at least 10 geotagged tweets and a declared language of English, 55k users were obtained. For each user, their recent status updates were aggregated, and non-English users were filtered out based on the language predictions of `langid.py`. For some users with geotagged tweets from many cities, the most frequent city might not be an appropriate representation of their home location for evaluation. To improve the evaluation data quality, we therefore exclude users who have less than 50% of their geotagged tweets originating from a single city. After filtering, 32k users were obtained, forming the final LIVE$_{test}$ dataset. In the final LIVE$_{test}$, the smallest class has only one test user, and the largest class has 569 users. The mean users per city is 27.76.

LIVE$_{test}$	Acc	Acc@161	Median
MB-NG	.268 (−.012)	.510 (−.018)	151 (−19)
MB-LOC	.326 (−.079)	.465 (−.060)	306 (+214)
MB-TZ	.065 (+.001)	.160 (−.011)	1529 (+199)
STACKING	.406 (−.084)	.614 (−.051)	40 (+31)

Table 2: Results over LIVE$_{test}$, and the absolute fluctuation over the results for WORLD$_{test}$

The results over LIVE$_{test}$, and the difference in absolute score over WORLD$_{test}$, are shown in Table 2. The stacked model accuracy numbers drop 5–8% on LIVE$_{test}$, and the median error distance increases moderately by 31km. Overall, the numbers suggest inference on WORLD$_{test}$, which is time-homogenous with the training data (taken from WORLD), is an easier classification than LIVE$_{test}$, which is time-heterogeneous with the training data. Training on "old" data and testing on "new" data is certainly possible, however. Looking over the results of the base classifiers, we can see that the biggest hit is for MB-LOC classifier. In contrast, the accuracy for MB-NG and MB-TZ is relatively stable (other than the sharp increase in the median error distance for MB-TZ).

5 Architecture and Access

In this section, we describe the architecture of the proposed geolocation system, as well as two ways of accessing the live system.[3] The core structure of the system consists of two parts: (1) the interface; (2) the back-end geolocation service.

We offer two interfaces to access the system: a Twitter bot and a web interface. The Twitter bot account is: @MELBLTFSD. A daemon process detects any user mentions of the bot in tweets via keyword matching through the Twitter search API. The screen name of the tweet author is extracted and sent to the back-end geolocation service, and the predicted user geolocation is sent to the Twitter user in a direct message, as shown in Figure 1.

Web access is via `http://hum.csse.unimelb.edu.au:9000/geo.html`. Users can input a Twitter user screen name through the web interface, whereby a call is made to the back-end geolocation service to geolocate that user. The geoloca-

[2] `https://dev.twitter.com/docs/api/1.1/get/statuses/sample`

[3] The source code is available from `https://github.com/tq010or/acl2013`

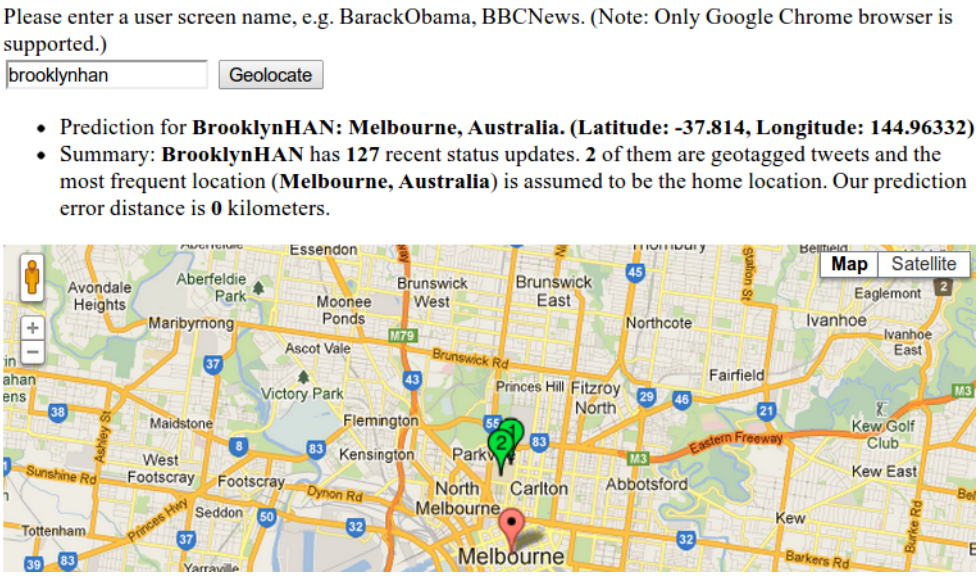

Figure 2: Web interface for user geolocation. The numbered green markers represent geotagged tweets. These coordinates are utilised to validate our predictions, and are not used in the geolocation process. The red marker is the predicted city-based user geolocation.

Figure 1: Twitter bot interface. When the Twitter bot is mentioned in a tweet, that user is sent a direct message with the predicted geolocation.

tion results are rendered on a map (along with any geotagged tweets for the user) as in Figure 2.[4]

The back-end geolocation service crawls recent tweets for a given user in real time,[5] and word and *n*-gram features are extracted from both the text and the user metadata. These features are sent to the *L0* classifiers (TEXT, MB-LOC and MB-TZ), and the *L0* results are further fed into the *L1* classifier for the final prediction.

6 Summary and Future Work

In this paper, we presented a city-level geolocation prediction system for Twitter users. Over a public dataset, our stacking method — exploiting both tweet text and user metadata — substantially outperformed benchmark methods. We further evaluated model generalisation on a newer, time-heterogeneous dataset. The overall results decreased by 5–8% in accuracy, compared with numbers on time-homogeneous data, primarily due to the poor generalisation of the MB-LOC classifier.

In future work, we plan to further investigate the cause of the MB-LOC classifier accuracy decrease on the new dataset. In addition, we'd like to study differences in prediction accuracy across cities. For cities with reliable predictions, the system can be adapted as a preprocessing module for downstream applications, e.g., local event detection based on users with reliable predictions.

[4]Currently, only Google Chrome is supported. `https://www.google.com/intl/en/chrome/`

[5]Up to 200 tweets are crawled, the upper bound of messages returned per single request based on Twitter API v1.1.

Acknowledgements

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence programme.

References

Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. 2008. Spatial variation in search engine queries. In *Proc. of WWW*, pages 357–366, Beijing, China.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. of WWW*, pages 61–70, Raleigh, USA.

Orkut Buyukokkten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, and Narayana Shivakumar. 1999. Exploiting geographical location information of web pages. In *ACM SIGMOD Workshop on The Web and Databases*, pages 91–96, Philadelphia, USA.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. of CIKM*, pages 759–768, Toronto, Canada.

David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. 2009. Mapping the world's photos. In *Proc. of WWW*, pages 761–770, Madrid, Spain.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proc. of EMNLP*, pages 1277–1287, Cambridge, MA, USA.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proc. of COLING*, pages 1045–1062, Mumbai, India.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proc. of SIGCHI*, pages 237–246, Vancouver, Canada.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proc. of WWW*, pages 769–778, Lyon, France.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'm eating a sandwich in glasgow": modeling locations with tweets. In *Proc. of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 61–68, Glasgow, UK.

Jochen L. Leidner and Michael D. Lieberman. 2011. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11.

Michael D. Lieberman and Jimmy Lin. 2009. You are where you edit: Locating wikipedia contributors through edit histories. In *ICWSM*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proc. of the ACL*, pages 25–30, Jeju Island, Korea.

Alan M. MacEachren, Anuj Jaiswal, Anthony C. Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. 2011. Senseplace2: Geotwitter analytics support for situational awareness. In *IEEE Conference on Visual Analytics Science and Technology*, pages 181–190, Rhode Island, USA.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2012. Where is this tweet from? inferring home locations of twitter users. In *Proc. of ICWSM*, Dublin, Ireland.

Teng Qin, Rong Xiao, Lei Fang, Xing Xie, and Lei Zhang. 2003. An efficient location extraction algorithm by leveraging web contextual information. In *Proc. of SIGSPATIAL*, pages 55–62, San Jose, USA.

Gianluca Quercini, Hanan Samet, Jagan Sankaranarayanan, and Michael D. Lieberman. 2010. Determining the spatial reader scopes of news sources using local lexicons. In *Proc. of the 18th International Conference on Advances in Geographic Information Systems*, pages 43–52, San Jose, USA.

John Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proc. of EMNLP*, pages 1500–1510, Jeju Island, Korea.

Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. Where's @wally?: a classification approach to geolocating users based on their social ties. In *Proc. of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20, Paris, France.

Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. 2012. Finding your friends and following them to where you are. In *Proc. of WSDM*, pages 723–732, Seattle, USA.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of WWW*, pages 851–860, Raleigh, USA.

Benjamin P. Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proc. of ACL*, pages 955–964, Portland, USA.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proc. of WWW*, pages 247–256, Hyderabad, India.

Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. 2005. On assigning place names to geography related web pages. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 354–362.