Exploiting Text and Network Context for Geolocation of Social Media Users

Afshin Rahimi,[1] **Duy Vu,**[2] **Trevor Cohn,**[1] and **Timothy Baldwin**[1]
[1]Department of Computing and Information Systems
[2]Department of Mathematics and Statistics
The University of Melbourne
arahimi@student.unimelb.edu.au
{duy.vu, t.cohn}@unimelb.edu.au
tb@ldwin.net

Abstract

Research on automatically geolocating social media users has conventionally been based on the text content of posts from a given user or the social network of the user, with very little crossover between the two, and no benchmarking of the two approaches over comparable datasets. We bring the two threads of research together in first proposing a text-based method based on adaptive grids, followed by a hybrid network- and text-based method. Evaluating over three Twitter datasets, we show that the empirical difference between text- and network-based methods is not great, and that hybridisation of the two is superior to the component methods, especially in contexts where the user graph is not well connected. We achieve state-of-the-art results on all three datasets.

1 Introduction

There has recently been a spike in interest in the task of inferring the location of users of social media services, due to its utility in applications including location-aware information retrieval (Amitay et al., 2004), recommender systems (Noulas et al., 2012) and rapid disaster response (Earle et al., 2010). Social media sites such as Twitter and Facebook provide two primary means for users to declare their location: (1) through text-based metadata fields in the user's profile; and (2) through GPS-based geotagging of posts and check-ins. However, the text-based metadata is often missing, misleading or imprecise, and only a tiny proportion of users geotag their posts (Cheng et al., 2010). Given the small number of users with reliable location information, there has been significant interest in the task of automatically geolocating (predicting lat/long coordinates) of users based on their publicly available posts, metadata and social network information. These approaches are built on the premise that a user's location is evident from their posts, or through location homophily in their social network.

Our contributions in this paper are: *a*) the demonstration that network-based methods are generally superior to text-based user geolocation methods due to their robustness; *b*) the proposal of a hybrid classification method that backs-off from network- to text-based predictions for disconnected users, which we show to achieve state-of-the-art accuracy over all Twitter datasets we experiment with; and *c*) empirical evidence to suggest that text-based geolocation methods are largely competitive with network-based methods.

2 Related Work

Past work on user geolocation falls broadly into two categories: text-based and network-based methods. Common to both methods is the manner of framing the geolocation prediction problem. Geographic coordinates are real-valued, and accordingly this is most naturally modelled as (multiple) regression. However for modelling convenience, the problem is typically simplified to classification by first pre-partitioning the regions into discrete sub-regions using either known city locations (Han et al., 2012; Rout et al., 2013) or a k-d tree partitioning (Roller et al., 2012; Wing and Baldridge, 2014). In the k-d tree methods, the resulting discrete regions are treated either as a flat list (as we do here) or a nested hierarchy.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1362–1367,
Denver, Colorado, May 31 – June 5, 2015.

2.1 Text-based Geolocation

Text-based approaches assume that language in social media is geographically biased, which is clearly evident for regions speaking different languages (Han et al., 2014), but is also reflected in regional dialects and the use of region specific terminology. Text based models have predominantly used bag of words features to learn per-region classifiers (Roller et al., 2012; Wing and Baldridge, 2014), including feature selection for location-indicative terms (Han et al., 2012).

Topic models have also been applied to model geospatial text usage (Eisenstein et al., 2010; Ahmed et al., 2013), by associating latent topics with locations. This has a benefit of allowing for prediction over continuous space, i.e., without the need to render the problem as classification. On the other hand, these methods have high algorithmic complexity and their generative formulation is unlikely to rival the performance of discriminative methods on large datasets.

2.2 Network-based Geolocation

Although online social networking sites allow for global interaction, users tend to befriend and interact with many of the same people online as they do off-line (Rout et al., 2013). Network-based methods exploit this property to infer the location of users from the locations of their friends (Jurgens, 2013; Rout et al., 2013). This relies on some form of friendship graph, through which location information can be propagated, e.g., using label propagation (Jurgens, 2013; Talukdar and Crammer, 2009). A significant caveat regarding the generality of these techniques is that friendship graphs are often not accessible, e.g., secured from the public (Facebook) or hidden behind a heavily rate-limited API (Twitter).

While the raw accuracies reported for network-based methods (e.g., Jurgens (2013) and Rout et al. (2013)) are generally higher than those reported for text-based methods (e.g., Wing and Baldridge (2014) and Han et al. (2014)), they have been evaluated over different datasets and spatial representations, making direct comparison uninformative. Part of our contribution in this paper is direct comparison between the respective methods over standard datasets. In this, we propose both text- and network-based methods, and show that they achieve state-of-the-art results on three pre-existing Twitter geolocation corpora. We also propose a new hybrid method incorporating both textual and network information, which also improves over the state-of-the-art, and outperforms the text-only or network-only methods over two of the three datasets.

3 Data

We evaluate on three Twitter corpora, each of which uses geotagged tweets to derive a geolocation for each user. Each user is represented by the concatenation of their tweets, and is assumed to come from a single location.

GEOTEXT: around 380K tweets from 9.5K users based in contiguous USA, of which 1895 is held out for development and testing (Eisenstein et al., 2010); the location of each user is set to the GPS coordinates of their first tweet.

TWITTER-US: around 39M tweets from 450K users based in the contiguous USA. 10K users are held out for each of development and testing (Roller et al., 2012); again users' locations are taken from their first tweet.

TWITTER-WORLD: around 12M English tweets from 1.4M users based around the world, of which 10K users are held out for each of development and testing (Han et al., 2012); users are geotagged with the centre of the closest city to their tweets.

In each case, we use the established training, development and testing partitions, and follow Cheng et al. (2010) and Eisenstein et al. (2010) in evaluating based on: (1) accuracy at 161km ("Acc@161"); (2) mean error distance, in kilometres ("Mean"); and (3) median error distance, in kilometres ("Median").

4 Methods

4.1 Text-based Classification

Our baseline method for text based geolocation is based on Wing and Baldridge (2014), who formulate the geolocation problem as classification using k-d

trees. In summary, their approach first discretises the continuous space of geographical coordinates using a k-d tree such that each sub-region (leaf) has similar numbers of users. This results in many small regions for areas of high population density and fewer larger regions for country areas with low population density. Next, they use these regions as class labels to train a logistic regression model ("LR"). Our work is also subject to a sparse l_1 regularisation penalty (Tibshirani, 1996). In their work, Wing and Baldridge (2014) showed that hierarchical logistic regression with a beam search achieves higher results than logistic regression over a flat label set, but in this research, we use a flat representation, and leave experiments with hierarchical classification to future work.

For our experiments, the number of users in each region was selected from $\{300, 600, 900, 1200\}$ to optimise median error distance on the development set, resulting in values of 300, 600 and 600 for GEOTEXT, TWITTER-US and TWITTER-WORLD, respectively. The l_1 regularisation coefficient was also optimised in the same manner.

As features, we used a bag of unigrams (over both words and @-mentions) and removed all features that occurred in fewer than 10 documents, following Wing and Baldridge (2014). The features for each user were weighted using tf-idf, followed by per-user l_2 normalisation. The normalisation is particularly important because our 'documents' are formed from all the tweets of each user, which vary significantly in size between users; furthermore, this adjusts for differing degrees of lexical variation (Lee, 1995). The number of features was almost 10K for GEOTEXT and about 2.5M for the other two corpora. For evaluation we use the median of all training locations in the sub-region predicted by the classifier, from which we measure the error against a test user's gold standard location.

4.2 Network-based Label Propagation

Next, we consider the approach of Jurgens (2013) who used label propagation ("LP"; Zhu and Ghahramani (2002)) to infer user locations using social network structure. Jurgens (2013) defined an undirected network from interactions among Twitter users based on @-mentions in their tweets, a mechanism typically used for conversations between friends. Consequently these links often correspond to offline friendships, and accordingly the network will exhibit a high degree of location homophily. The network is constructed by defining as nodes all users in a dataset (train and test), as well as other external users mentioned in their tweets. Unlike Jurgens (2013) who only created edges when both users mentioned one another, we created edges if either user mentioned the other. For the three datasets used in our experiments, bi-directional mentions were too rare to be useful, and we thus used the (weaker) uni-directional mentions as undirected edges instead. The edges between users are undirected and weighted by the number of @-mentions in tweets by either user.[1]

	GEOTEXT	TWITTER-US	TWITTER-WORLD
User mentions	109K	3.63M	16.8M
Disconnected test users:	23.5%	27.7%	2.36%

Table 1: The graph size and proportion of test users disconnected from training users for each dataset.

The mention network statistics for each of our datasets is shown in Table 1.[2] Following Jurgens (2013), we ran the label propagation algorithm to update the location of each non-training node to the weighted median of its neighbours. This process continues iteratively until convergence, which occurred at or before 10 iterations.

4.3 A Hybrid Method

Unfortunately many test users are not transitively connected to any training node (see Table 1), meaning that LP fails to assign them any location. This can happen when users don't use @-mentions, or when a set of nodes constitutes a disconnected component of the graph.

In order to alleviate this problem, we use the text for each test user in order to estimate their location, which is then used as an initial estimation during label propagation. In this hybrid approach, we first

[1] As our datasets don't have tweets for external users, these nodes do not contribute to the weight of their incident edges.

[2] Note that @-mentions were removed in the published TWITTER-US and TWITTER-WORLD datasets. To recover these we rebuilt the corpora from the Twitter archive.

	GeoText			Twitter-US			Twitter-World		
	Acc@161	Mean	Median	Acc@161	Mean	Median	Acc@161	Mean	Median
LR (text-based)	38.4	880.6	397.0	50.1	686.7	159.2	**63.8**	**866.5**	**19.9**
LP (network-based)	45.1	676.2	255.7	37.4	747.8	431.5	56.2	1026.5	79.8
LP–LR (hybrid)	**50.2**	**653.9**	**151.2**	**50.2**	**620.0**	**157.1**	59.2	903.6	53.7
Wing and Baldridge (2014) (uniform)	—	—	—	49.2	703.6	170.5	32.7	1714.6	490.0
Wing and Baldridge (2014) (k-d)	—	—	—	48.0	686.6	191.4	31.3	1669.6	509.1
Han et al. (2012)	—	—	—	45.0	814	260	24.1	1953	646
Ahmed et al. (2013)	???	???	298	—	—	—	—	—	—

Table 2: Geolocation accuracy over the three Twitter corpora comparing Logistic Regression (LR), Label Propagation (LP) and LP over LR initialisation (LP–LR) with the state-of-the-art methods for the respective datasets ("—" signifies that no results were published for the given dataset, and "???" signifies that no results were reported for the given metric).

estimate the location for each test node using the LR classifier described above, before running label propagation over the mention graph. This iteratively adjusts the locations based on both the known training users and guessed test users, while simultaneously inferring locations for the external users. In such a way, the inferred locations of test users will better match neighbouring users in their sub-graph, or in the case of disconnected nodes, will retain their initial classification estimate.

5 Results

Table 2 shows the performance of the three methods over the test set for the three datasets. The results are also compared with the state of the art for Twitter-US and Twitter-World (Wing and Baldridge, 2014), and GeoText (Ahmed et al., 2013).

Our methods achieve a sizeable improvement over the previous state of the art for all three datasets. LP–LR performs best over GeoText and Twitter-US, while LR performs best over Twitter-World; the reduction in median error distance over the state of the art ranges from around 40% to over 95%. Even for Twitter-World, the results for LP–LR are substantially better than the best-published results for that dataset.

Comparing LR and LP, no strong conclusion can be drawn — the text-based LP actually outperforms the network-based LR for two of the three datasets, but equally, the combination of the two (LP–LR) performs better than either component method over two of the three datasets. For the third (Twitter-World), LR outperforms LP–LR due to a combination of factors. First, unlike the other two datasets, the label set is pre-discretised (everything is aggregated at the city level), meaning that LP and LR use the same label set.[3] This annuls the representational advantage that LP has in the case of the other two datasets, in being able to capture a more fine-grained label set (i.e., all locations associated with training users). Second, there are substantially fewer disconnected test users in Twitter-World (see Table 1), meaning that the results for the hybrid LP–LR method are dominated by the empirically-inferior LP.

Although LR is similar to Wing and Baldridge (2014), we achieved large improvements over their reported results. This might be due to: (a) our use of @-mention features; (b) l_1 regularisation, which is essential to preventing overfitting for large feature sets; or (c) our use of l_2 normalisation of rows in the design matrix, which we found reduced errors by about 20% on GeoText, in keeping with results from text categorisation (Lee, 1995). Preliminary experiments also showed that lowering the term frequency threshold from 10 can further improve the LR results on all three datasets.

LP requires few hyper-parameters and is relatively robust. It converged on all datasets in fewer than 10 iterations, and geolocates not only the test users but all nodes in the mention graph. Another advantage of LP over LR is the relatively modest amount of memory and processing power it requires.

[3] For consistency, we learn a k-d tree for Twitter-World and use the merged representation for LR, but the k-d tree largely preserves the pre-existing city boundaries.

6 Conclusion

We proposed a series of approaches to social media user geolocation based on: (1) text-based analysis using logistic regression with regularisation; (2) network-based analysis using label propagation; and (3) a hybrid method based on network-based label propagation, and back-off to text-based analysis for disconnected users. We achieve state-of-the-art results over three pre-existing Twitter datasets, and find that, overall, the hybrid method is superior to the two component methods.The LP−LR method is a hybrid approach that uses the LR predictions as priors. It is not simply a backoff from network information to textual information in the sense that it propagates the LR geolocations through the network. That is, if a test node is disconnected from the training nodes but still has connections to other test nodes, the geolocation of the node is adjusted and propagated through the network. It is possible to add extra nodes to the graph after applying the algorithm and to geolocate only these nodes efficiently, although this approach is potentially less accurate than inferencing over the full graph from scratch.

Label propagation algorithms such as Modified Adsorption (Talukdar and Crammer, 2009) allow for different levels of influence between prior/known labels and propagated label distributions. These algorithms require a discretised output space for label propagation, while LP can work directly on continuous data. We leave label propagation over discritised output and allowing different influence levels between prior and propagated label distributions to future work.

There is no clear consensus on whether text- or network-based methods are empirically superior at the user geolocation task. Our results show that the network-based method (LP) is more robust than the text-based (LR) method as it requires a smaller number of hyper-parameters, uses less memory and computing resources, converges much faster and geolocates not only test users but all mentioned users. The drawback of LP is that it fails to geolocate disconnected test users. So for connected nodes – the majority of test nodes in all our datasets – LP is more robust than LR. Text-based methods are very sensitive to the regularisation settings and the types of textual features. That said, with thorough parameter tuning, they might outperform network-based method in terms of accuracy.

In future work, we hope to look at different types of network information for label propagation, more precise propagation methods to deal with non-local interactions, and also efficient ways of utilising both textual and network information in a joint model.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. This work was funded in part by the Australian Research Council.

References

Amr Ahmed, Liangjie Hong, and Alexander J Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 25–36.

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-where: geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768.

Paul Earle, Michelle Guy, Richard Buckmaster, Chris Ostrum, Scott Horvath, and Amy Vaughan. 2010. OMG earthquake! can Twitter improve earthquake response? *Seismological Research Letters*, 81(2):246–251.

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*, pages 1277–1287.

Bo Han, Timothy Baldwin, and Paul Cook. 2012. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers*, pages 1045–1062.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research (JAIR)*, 49:451–500.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013)*, pages 273–282.

Joon Ho Lee. 1995. Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–188.

Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 2012. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510.

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 *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning (ECML-PKDD) 2009*, pages 442–457.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Benjamin P Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University.