Twitter User Geolocation Using a Unified Text and Network Prediction Model

Afshin Rahimi, Trevor Cohn, and **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
arahimi@student.unimelb.edu.au
{t.cohn,tbaldwin}@unimelb.edu.au

Abstract

We propose a label propagation approach to geolocation prediction based on Modified Adsorption, with two enhancements: (1) the removal of "celebrity" nodes to increase location homophily and boost tractability; and (2) the incorporation of text-based geolocation priors for test users. Experiments over three Twitter benchmark datasets achieve state-of-the-art results, and demonstrate the effectiveness of the enhancements.

1 Introduction

Geolocation of social media users is essential in applications ranging from rapid disaster response (Earle et al., 2010; Ashktorab et al., 2014; Morstatter et al., 2013a) and opinion analysis (Mostafa, 2013; Kirilenko and Stepchenkova, 2014), to recommender systems (Noulas et al., 2012; Schedl and Schnitzer, 2014). Social media platforms like Twitter provide support for users to declare their location manually in their text profile or automatically with GPS-based geotagging. However, the text-based profile locations are noisy and only 1–3% of tweets are geotagged (Cheng et al., 2010; Morstatter et al., 2013b), meaning that geolocation needs to be inferred from other information sources such as the tweet text and network relationships.

User geolocation is the task of inferring the primary (or "home") location of a user from available sources of information, such as text posted by that individual, or network relationships with other individuals (Han et al., 2014). Geolocation models are usually trained on the small set of users whose location is known (e.g. through GPS-based geotagging), and other users are geolocated using the resulting model. These models broadly fall into two categories: text-based and network-based methods. Orthogonally, the geolocation task can be viewed as a regression task over real-valued geographical coordinates, or a classification task over discretised region-based locations.

Most previous research on user geolocation has focused either on text-based classification approaches (Eisenstein et al., 2010; Wing and Baldridge, 2011; Roller et al., 2012; Han et al., 2014) or, to a lesser extent, network-based regression approaches (Jurgens, 2013; Compton et al., 2014; Rahimi et al., 2015). Methods which combine the two, however, are rare.

In this paper, we present our work on Twitter user geolocation using both text and network information. Our contributions are as follows: (1) we propose the use of Modified Adsorption (Talukdar and Crammer, 2009) as a baseline network-based geolocation model, and show that it outperforms previous network-based approaches (Jurgens, 2013; Rahimi et al., 2015); (2) we demonstrate that removing "celebrity" nodes (nodes with high in-degrees) from the network increases geolocation accuracy and dramatically decreases network edge size; and (3) we integrate text-based geolocation priors into Modified Adsorption, and show that our unified geolocation model outperforms both text-only and network-only approaches, and achieves state-of-the-art results over three standard datasets.

2 Related Work

A recent spike in interest on user geolocation over social media data has resulted in the development of a range of approaches to automatic geolocation prediction, based on information sources such as the text of messages, social networks, user profile data, and temporal data. Text-based methods model the geographical bias of language use in social media, and use it to geolocate non-geotagged users. Gazetted expressions (Leidner and Lieberman, 2011) and geographical names (Quercini et

Rahimi, Afshin, Trevor Cohn and Timothy Baldwin (to appear) Twitter User Geolocation Using a Unified Text and Network Prediction Model, In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015), Beijing, China.

al., 2010) were used as feature in early work, but were shown to be sparse in coverage. Han et al. (2014) used information-theoretic methods to automatically extract location-indicative words for location classification. Wing and Baldridge (2014) reported that discriminative approaches (based on hierarchical classification over adaptive grids), when optimised properly, are superior to explicit feature selection. Cha et al. (2015) showed that sparse coding can be used to effectively learn a latent representation of tweet text to use in user geolocation. Eisenstein et al. (2010) and Ahmed et al. (2013) proposed topic model-based approaches to geolocation, based on the assumption that words are generated from hidden topics and geographical regions. Similarly, Yuan et al. (2013) used graphical models to jointly learn spatio-temporal topics for users. The advantage of these generative approaches is that they are able to work with the continuous geographical space directly without any pre-discretisation, but they are algorithmically complex and don't scale well to larger datasets. Hulden et al. (2015) used kernel-based methods to smooth linguistic features over very small grid sizes to alleviate data sparseness.

Network-based geolocation models, on the other hand, utilise the fact that social media users interact more with people who live nearby. Jurgens (2013) and Compton et al. (2014) used a Twitter reciprocal mention network, and geolocated users based on the geographical coordinates of their friends, by minimising the weighted distance of a given user to their friends. For a reciprocal mention network to be effective, however, a huge amount of Twitter data is required. Rahimi et al. (2015) showed that this assumption could be relaxed to use an undirected mention network for smaller datasets, and still attain state-of-the-art results. The greatest shortcoming of network-based models is that they completely fail to geolocate users who are not connected to geolocated components of the graph. As shown by Rahimi et al. (2015), geolocation predictions from text can be used as a backoff for disconnected users, but there has been little work that has investigated a more integrated text- and network-based approach to user geolocation.

3 Data

We evaluate our models over three pre-existing geotagged Twitter datasets: (1) GEOTEXT (Eisenstein et al., 2010), (2) TWITTER-US (Roller et al., 2012), and (3) TWITTER-WORLD (Han et al., 2012). In each dataset, users are represented by a single meta-document, generated by concatenating their tweets. The datasets are pre-partitioned into training, development and test sets, and rebuilt from the original version to include mention information. The first two datasets were constructed to contain mostly English messages.

GEOTEXT consists of tweets from 9.5K users: 1895 users are held out for each of development and test data. The primary location of each user is set to the coordinates of their first tweet.

TWITTER-US consists of 449K users, of which 10K users are held out for each of development and test data. The primary location of each user is, once again, set to the coordinates of their first tweet.

TWITTER-WORLD consists of 1.3M users, of which 10000 each are held out for development and test. Unlike the other two datasets, the primary location of users is mapped to the geographic centre of the city where the majority of their tweets were posted.

4 Methods

We use label propagation over an @-mention graph in our models. We use k-d tree descretised adaptive grids as class labels for users and learn a label distribution for each user by label propagation over the @-mention network using labelled nodes as seeds. For k-d tree discretisation, we set the number of users in each region to 50, 2400, 2400 for GEOTEXT, TWITTER-US and TWITTER-WORLD respectively, based on tuning over the development data.

Social Network: We used the @-mention information to build an undirected graph between users. In order to make the inference more tractable, we removed all nodes that were not a member of the training/test set, and connected all pairings of training/test users if there was any path between them (including paths through non training/test users). We call this network a "collapsed network", as illustrated in Figure 1. Note that a celebrity node with n mentions connects $n(n-1)$ nodes in the collapsed network. We experiment with both binary and weighted edge (based on the number of mentions connecting the given users) networks.

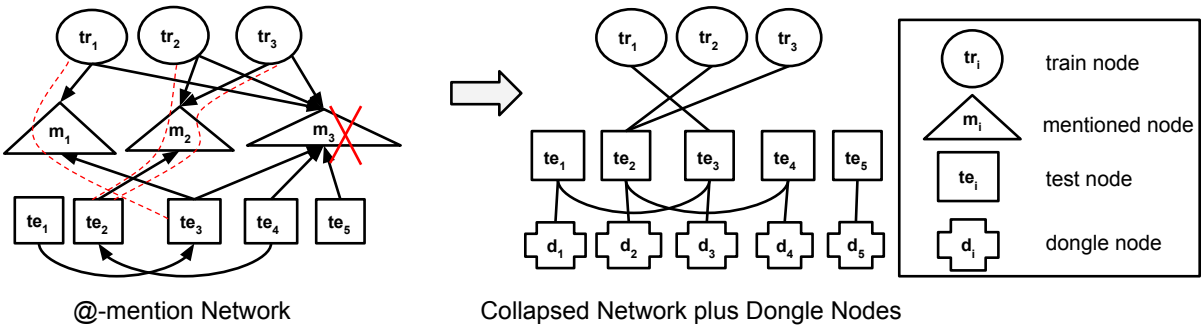

Figure 1: A collapsed network is built from the @-mention network. Each mention is shown by a directed arrow, noting that as it is based exclusively on the tweets from the training and test users, it will always be directed from a training or test user to a mentioned node. All mentioned nodes which are not a member of either training or test users are removed and the corresponding training and test users, previously connected through that node, are connected directly by an edge, as indicated by the dashed lines. Mentioned nodes with more than T unique mentions (celebrities, such as m_3) are removed from the graph. To each test node, a dongle node that carries the label from another learner (here, text-based LR) is added in MADCEL-B-LR and MADCEL-W-LR.

Baseline: Our baseline geolocation model ("MAD-B") is formulated as label propagation over a binary collapsed network, based on Modified Adsorption (Talukdar and Crammer, 2009). It applies to a graph $G = (V, E, W)$ where V is the set of nodes with $|V| = n = n_l + n_u$ (where n_l nodes are labelled and n_u nodes are unlabelled), E is the set of edges, and W is an edge weight matrix. Assume C is the set of labels where $|C| = m$ is the total number of labels. Y is an $n \times m$ matrix storing the training node labels, and $\hat{Y}$ is the estimated label distribution for the nodes. The goal is to estimate $\hat{Y}$ for all nodes (including training nodes) so that the following objective function is minimised:

$$C(\hat{Y}) = \sum_l \Big[\mu_1 (Y_l - \hat{Y}_l)^T S (Y_l - \hat{Y}_l) + \mu_2 \hat{Y}_l^T L \hat{Y}_l \Big]$$

where μ_1 and μ_2 are hyperparameters;[1] L is the Laplacian of an undirected graph derived from G; and S is a diagonal binary matrix indicating if a node is labelled or not. The first term of the equation forces the labelled nodes to keep their label (prior term), while the second term pulls a node's label toward that of its neighbours (smoothness term). For the first term, the label confidence for training and test users is set to 1.0 and 0.0, respectively. Based on the development data, we set μ_1 and μ_2 to 1.0 and 0.1, respectively, for all the experiments. For TWITTER-US and TWITTER-WORLD, the inference was intractable for the default network, as it was too large.

There are two immediate issues with the baseline graph propagation method: (1) it doesn't scale to large datasets with high edge counts, related to which, it tends to be biased by highly-connected nodes; and (2) it can't predict the geolocation of test users who aren't connected to any training user (MAD-B returns Unknown, which we rewrite with the centre of the map). We redress these two issues as follows.

Celebrity Removal To address the first issue, we target "celebrity" users, i.e. highly-mentioned Twitter users. Edges involving these users often carry little or no geolocation information (e.g. the majority of people who mention Barack Obama don't live in Washington D.C.). Additionally, these users tend to be highly connected to other users and generate a disproportionately high number of edges in the graph, leading in large part to the baseline MAD-B not scaling over large datasets such as TWITTER-US and TWITTER-WORLD. We identify and filter out celebrity nodes simply by assuming that a celebrity is mentioned by more than T users, where T is tuned over development data. Based on tuning over the development

[1] In the base formulation of MAD-B, there is also a regularisation term with weight μ_3, but in all our experiments, we found that the best results were achieved over development data with $\mu_3 = 0$, i.e. with no regularisation; the term is thus omitted from our description.

	GeoText			Twitter-US			Twitter-World		
	Acc@161	Mean	Median	Acc@161	Mean	Median	Acc@161	Mean	Median
MAD-B	50	683	146	×××	×××	×××	×××	×××	×××
MADCEL-B	56	609	76	54	709	117	70	936	**0**
MADCEL-W	58	586	60	54	705	116	71	976	**0**
MADCEL-B-LR	57	608	65	**60**	533	**77**	**72**	**786**	**0**
MADCEL-W-LR	**59**	**581**	**57**	**60**	**529**	78	**72**	802	**0**
LR (Rahimi et al., 2015)	38	880	397	50	686	159	63	866	19
LP (Rahimi et al., 2015)	45	676	255	37	747	431	56	1026	79
LP-LR (Rahimi et al., 2015)	50	653	151	50	620	157	59	903	53
Wing and Baldridge (2014) (uniform)	—	—	—	49	703	170	32	1714	490
Wing and Baldridge (2014) (k-d)	—	—	—	48	686	191	31	1669	509
Han et al. (2012)	—	—	—	45	814	260	24	1953	646
Ahmed et al. (2013)	???	???	298	—	—	—	—	—	—
Cha et al. (2015)	???	**581**	425	—	—	—	—	—	—

Table 1: Geolocation results over the three Twitter corpora, comparing baseline Modified Adsorption (MAD-B), with Modified Adsorption with celebrity removal (MADCEL-B and MADCEL-W, over binary and weighted networks, resp.) or celebrity removal plus text priors (MADCEL-B-LR and MADCEL-W-LR, over binary and weighted networks, resp.); the table also includes state-of-the-art results for each dataset ("—" signifies that no results were published for the given dataset; "???" signifies that no results were reported for the given metric; and "×××" signifies that results could not be generated, due to the intractability of the training data).

set of GeoText and Twitter-US, T was set to 5 and 15 respectively. For Twitter-World tuning was very resource intensive so T was set to 5 based on GeoText, to make the inference faster. Celebrity removal dramatically reduced the edge count in all three datasets (from 1×10^9 to 5×10^6 for Twitter-US and from 4×10^{10} to 1×10^7 for Twitter-World), and made inference tractable for Twitter-US and Twitter-World. Jurgens et al. (2015) report that the time complexity of most network-based geolocation methods is $\mathcal{O}(k^2)$ for each node where k is the average number of vertex neighbours. In the case of the collapsed network of Twitter-World, k is decreased by a factor of 4000 after setting the celebrity threshold T to 5. We apply celebrity removal over both binary ("MADCEL-B") and weighted ("MADCEL-W") networks (using the respective T for each dataset). The effect of celebrity removal over the development set of Twitter-US is shown in Figure 2 where it dramatically reduces the graph edge size and simultaneously leads to an improvement in the mean error.

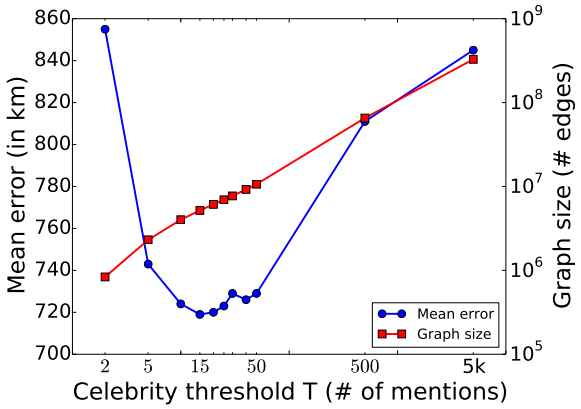

Figure 2: Effect of celebrity removal on geolocation performance and graph size. For each T performance is measured over the development set of Twitter-US by MADCEL-W.

A Unified Geolocation Model To address the issue of disconnected test users, we incorporate text information into the model by attaching a labelled dongle node to every test node (Zhu and Ghahramani, 2002; Goldberg and Zhu, 2006). The label for the dongle node is based on a text-based l_1 regularised logistic regression model, using the method of Rahimi et al. (2015). The dongle nodes with their corresponding label confidences are added to the seed set, and are treated in the same way as other labelled nodes (i.e. the training nodes). Once again, we experiment with text-based labelled dongle nodes over both binary ("MADCEL-B-LR") and weighted ("MADCEL-W-LR") networks.

5 Evaluation

Following Cheng et al. (2010) and Eisenstein et al. (2010), we evaluate using the mean and median error (in km) over all test users ("Mean" and "Median", resp.), and also accuracy within 161km of the actual location ("Acc@161"). Note that higher numbers are better for Acc@161, but lower numbers are better for mean and median error, with a lower bound of 0 and no (theoretical) upper bound.

To generate a continuous-valued latitude/longitude coordinate for a given user from the k-d tree cell, we use the median coordinates of all training points in the predicted region.

6 Results

Table 1 shows the performance of MAD−B, MADCEL−B, MADCEL−W, MADCEL−B−LR and MADCEL−W−LR over the GEOTEXT, TWITTER-US and TWITTER-WORLD datasets. The results are also compared with prior work on network-based geolocation using label propagation (LP) (Rahimi et al., 2015), text-based classification models (Han et al., 2012; Wing and Baldridge, 2011; Wing and Baldridge, 2014; Rahimi et al., 2015; Cha et al., 2015), text-based graphical models (Ahmed et al., 2013), and network–text hybrid models (LP−LR) (Rahimi et al., 2015).

Our baseline network-based model of MAD−B outperforms the text-based models and also previous network-based models (Jurgens, 2013; Compton et al., 2014; Rahimi et al., 2015). The inference, however, is intractable for TWITTER-US and TWITTER-WORLD due to the size of the network.

Celebrity removal in MADCEL−B and MADCEL−W has a positive effect on geolocation accuracy, and results in a 47% reduction in Median over GEOTEXT. It also makes graph inference over TWITTER-US and TWITTER-WORLD tractable, and results in superior Acc@161 and Median, but slightly inferior Mean, compared to the state-of-the-art results of LR, based on text-based classification (Rahimi et al., 2015).

MADCEL−W (weighted graph) outperforms MADCEL−B (binary graph) over the smaller GEOTEXT dataset where it compensates for the sparsity of network information, but doesn't improve the results for the two larger datasets where network information is denser.

Adding text to the network-based geolocation models in the form of MADCEL−B−LR (binary edges) and MADCEL−W−LR (weighted edges), we achieve state-of-the-art results over all three datasets. The inclusion of text-based priors has the greatest impact on Mean, resulting in an additional 26% and 23% error reduction over TWITTER-US and TWITTER-WORLD, respectively. The reason for this is that it provides a user-specific geolocation prior for (relatively) disconnected users.

7 Conclusions and Future Work

We proposed a label propagation method over adaptive grids based on collapsed @-mention networks using Modified Adsorption, and successfully supplemented the baseline algorithm by: (a) removing "celebrity" nodes (improving the results and also making inference more tractable); and (b) incorporating text-based geolocation priors into the model.

As future work, we plan to use temporal data and also look at improving the text-based geolocation model using sparse coding (Cha et al., 2015). We also plan to investigate more nuanced methods for differentiating between global and local celebrity nodes, to be able to filter out global celebrity nodes but preserve local nodes that can have high geolocation utility.

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 (WWW 2013)*, pages 25–36, Rio de Janeiro, Brazil.

Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. Tweedr: Mining Twitter to inform disaster response. In *Proceedings of The 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2014)*, pages 354–358, University Park, USA.

Miriam Cha, Youngjune Gwon, and HT Kung. 2015. Twitter geolocation and regional classification via sparse coding. In *Proceedings of the 9th International Conference on Weblogs and Social Media (ICWSM 2015)*, pages 582–585, Oxford, UK.

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 Information and Knowledge Management (CIKM 2010)*, pages 759–768, Toronto, Canada.

Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *Proceedings of the IEEE International Conference on Big Data (IEEE BigData 2014)*, pages 393–401, Washington DC, USA.

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 (EMNLP 2010)*, pages 1277–1287, Boston, USA.

Andrew B Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the 1st Workshop on Graph Based Methods for Natural Language Processing (TextGraphs 2006)*, pages 45–52, New York, USA.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1045–1062, Mumbai, India.

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

Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel density estimation for text-based geolocation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-2015)*, pages 145–150, Austin, USA.

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 9th International Conference on Weblogs and Social Media (ICWSM 2015)*, pages 188–197, Oxford, UK.

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, Boston, USA.

Andrei P Kirilenko and Svetlana O Stepchenkova. 2014. Public microblogging on climate change: One year of Twitter worldwide. *Global Environmental Change*, 26:171–182.

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.

Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. 2013a. Understanding twitter data with tweetxplorer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2013)*, pages 1482–1485, Chicago, USA.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013b. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013)*, pages 400–408, Boston, USA.

Mohamed M Mostafa. 2013. More than words: Social networks text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251.

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 *Proceedings of the International Conference on Privacy, Security, Risk and Trust and Social Computing (SOCIALCOM-PASSAT 2012)*, pages 144–153, Amsterdam, Netherlands.

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

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT 2015)*, Denver, 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 *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural*

Language Processing and Computational Natural Language Learning (EMNLP-CONLL 2012), pages 1500–1510, Jeju, Korea.

Markus Schedl and Dominik Schnitzer. 2014. Location-aware music artist recommendation. In *Proceedings of the 20th International Conference on MultiMedia Modeling (MMM 2014)*, pages 205–213, Dublin, Ireland.

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, Bled, Slovenia.

Benjamin P Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (ACL-HLT 2011)*, pages 955–964, Portland, USA.

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)*, pages 336–348, Doha, Qatar.

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 (SIGKDD 2013)*, pages 605–613, Chicago, USA.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University.