Identifying Linguistic Areas for Geolocation

Tommaso Fornaciari, Dirk Hovy
Bocconi University, Milan, Italy
{fornaciari|dirk.hovy}@unibocconi.it

Abstract

Geolocating social media posts relies on the assumption that language carries sufficient geographic information. However, locations are usually given as continuous latitude/longitude tuples, so we first need to define discrete geographic regions that can serve as labels. Most studies use some form of clustering to discretize the continuous coordinates (Han et al., 2016). However, the resulting regions do not always correspond to existing linguistic areas. Consequently, accuracy at 100 miles tends to be good, but degrades for finer-grained distinctions, when different linguistic regions get lumped together. We describe a new algorithm, Point-to-City (P2C), an iterative k-d tree-based method for clustering geographic coordinates and associating them with towns. We create three sets of labels at different levels of granularity, and compare performance of a state-of-the-art geolocation model trained and tested with P2C labels to one with regular k-d tree labels. Even though P2C results in substantially more labels than the baseline, model accuracy increases significantly over using traditional labels at the fine-grained level, while staying comparable at 100 miles. The results suggest that identifying meaningful linguistic areas is crucial for improving geolocation at a fine-grained level.

1 Introduction

Predicting the location of a Social Media post involves first and foremost ways to identify the words that indicate geographic location. Secondly, and perhaps even more fundamentally, though, we also need to determine an effective notion of what a "location" is, i.e., what do our labels represent: a state, a city, a neighborhood, a street? In many NLP tasks, labels are ambiguous and open to interpretation (Plank et al., 2014). In geolocation, the information initially given is an unambiguous latitude/longitude pair, but this format captures a level of detail (precise down to a centimeter) that is both unnecessary and unrealistic for most practical applications. Collapsing coordinates to geographic categories is therefore a common step in geolocation. However, this discretization step is open to interpretation: what method should we choose?

Previous work includes three different approaches to discretizing continuous values into location labels (see also Section 2):

1.) Geodesic grids are the most straightforward, but do not "lead to a natural representation of the administrative, population-based or language boundaries in the region" (Han et al., 2012).

2.) Clustering coordinates prevents the identification of (nearly) empty locations and keeps points which are geographically close together in one location. Unfortunately, in crowded regions, clusters might be too close to each other, and therefore divide cultural/linguistic areas into meaningless groups.

3.) Predefined administrative regions, like cities, can provide homogeneous interpretable areas. However, mapping coordinates to the closest city can be ambiguous. Previous work typically considered cities with a population of at least 100K (Han et al., 2012, 2014). This approach has the opposite problem of clustering: different linguistic areas might be contained within a single administrative region.

Here, we propose Point-To-City (P2C), a new method mapping continuous coordinates to locations. It combines the strengths of the last two approaches, keeping coordinates which appear close to each other in the same location, while also representing them in terms of meaningful administrative regions, with adjustable granularity. We show that these two criteria also result in superior prediction performance for geolocation.

Relying on k-d trees (Maneewongvatana and Mount, 1999), P2C iteratively clusters points

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 231–236
Hong Kong, Nov 4, 2019. ©2019 Association for Computational Linguistics

within a specified maximum distance d, and maps them to the coordinates of the closest town with a minimum population size.

We evaluate P2C on two data sets commonly used for geolocation. We create three different conditions by using three different values for d as maximum distance between points, and compare the results to those obtained using k-d tree labels (as used in the W-NUT shared task (Han et al., 2016)). For all four labeling schemes, we train an attention-based convolutional neural network, and evaluate mean and median distance between target and predicted point, and accuracy within 161 km (Acc@161). We also show the standard accuracy score relative to the specific labels, usually much worse than Acc@161, and often not reported in the literature.

Our results show that P2C reliably produces Acc@161 performance which is comparable with state-of-the-art models. For exact accuracy, however, P2C labels always result in substantially better performance than previous methods, in spite of the larger set of classes. This suggests that P2C captures more meaningful location distinctions (backed up by a qualitative analysis), and that previous labels capture only broader, linguistically mixed areas. More generally, our results show that language reflects social and geographical distinctions in the world, and that more meaningful real-world labels help language-based prediction models to perform their task more efficiently.

Contributions The contributions of this paper are the following: 1.) we propose P2C, a k-d tree based procedure to cluster geographic points associated with existing towns within a certain distance between town and cluster centroid. 2.) we show that P2C produces more meaningful, interpretable cultural and linguistic locations 3.) we show that P2C labels substantially improve model performance in exact, fine-grained classification

2 Related work

Geolocation prediction can, in principle, be modeled both as regression and as classification problem. In practice, however, given the difficulty of predicting continuous coordinate values, regression is often carried out in conjunction with the classification (Eisenstein et al., 2010; Lourentzou et al., 2017; Fornaciari and Hovy, 2019b). In general, however, the task is considered a classification problem, which requires solutions for the identification of geographic regions as labels.

Geodesic grids were used for the geolocation of posts on Flickr, Twitter and Wikipedia (Serdyukov et al., 2009; Wing and Baldridge, 2011).

Hulden et al. (2015) noticed that “using smaller grid sizes leads to an immediate sparse data problem since very few features/words are [selectively] observed in each cell”.

In order to enhance the expressiveness of the geographic cells, Wing and Baldridge (2014), constructed both flat and hierarchical grids relying on k-d tree, and testing their methods at different levels of granularity. The same labels were used in the study of Rahimi et al. (2018).

Han et al. (2012, 2014), who released TWITTER-WORLD, use the information provided by the Geoname dataset[1] in order to identify a set of cities around the world with at least 100K inhabitants. Then they refer their geo-tagged texts to those cities, creating easily interpretable geographic places. Cha et al. (2015) proposed a voting-based grid selection scheme, with the classification referred to regions/states in US.

Most works use deep learning techniques for classification (Miura et al., 2016). Often, they include multi-view models, considering different sources (Miura et al., 2017; Lau et al., 2017; Ebrahimi et al., 2018; Fornaciari and Hovy, 2019a). In particular, Lau et al. (2017) implemented a multi-channel convolutional network, structurally similar to our model. Rahimi et al. (2018) proposes a Graph-Convolutional neural network, though the text features are represented by a bag-of-words, while we rely on word embeddings.

The ability of the labels to reflect real anthropological areas, however, affects primarily the models which rely on linguistic data. This is the case of the studies of Han et al. (2012) and Han et al. (2014) who based their predictions on the so-called *Location-Indicative Words* (LIW). Recently, neural models have been built with the same purpose (Rahimi et al., 2017; Tang et al., 2019).

3 Methods

Data sets We apply our method to two widely used data sets for geolocation: TWITTER-US (Roller et al., 2012), and TWITTER-WORLD (Han et al., 2012). They are all collections of En-

[1] http://www.geonames.org

glish tweets aggregated by author and labeled with geographic coordinates. TWITTER-US and TWITTER-WORLD contain 450K and 1.39M texts, respectively. They are each divided into their own training, development, and test sets. Readers are referred to the respective papers for additional details. We round the coordinates to the second decimal number. A distance of 0.01 degrees corresponds to less than 1.1 km on the longitude axis (the distance is not constant on the latitude axis). Smaller distinctions are not relevant for any common NLP task.

Data set	d	labels	mean	median
TW.-US	.1	1554	7.07	3.81
	.25	914	9.10	5.64
	.5	418	15.54	12.21
	W-NUT	256	–	–
TW.-WORLD	.1	3047	0.45	0.00
	.25	2818	1.77	0.00
	.5	2350	3.28	2.39
	W-NUT	930	–	–

Table 1: Number of labels and mean/median distance in km between instances and the cluster town center. For W-NUT, distance can not be computed, as centroids are not close to meaningful places

Point-To-City (P2C) For the classification, we need to identify labels corresponding to existing cultural/linguistic areas, so that the geographic information conveyed through language can be fully exploited.To this end, P2C iteratively creates clusters of points, and afterwards associates the final clusters with specific towns.

The parameter d controls the maximum spherical distance we allow between points assigned to the same cluster at the initialization step. We run P2C considering three values: 0.1, 0.25, and 0.5 coordinate decimal points, which correspond to 11.12 km (6.91 miles), 27.80 km (17.27 miles), and 55.60 km (34.55 miles) on the longitude axis. We use these values to explore the feasibility of finer (and more challenging) predictions than those usually accepted in the literature.

One of the most popular metrics in previous studies (see Section 2 and 4) is the accuracy of the predictions within 161 km, or 100 mi, from the target point. In contrast, we are interested in the accuracy relative to the precise prediction of the labels, and we want labels representing points aggregated according to a distance much smaller than 161 km/100 mi: even the highest value we choose for d, 0.5, is about one third the distance of accuracy at 161 km (Acc@161). However, since P2C iteratively creates clusters of clusters, it is possible that the original points belonging to different clusters are further apart than the threshold of d. For this reason, we selected values of d which are about three to fifteen times smaller than 161 km/100 mi.

Given d and a set of coordinate points/instances in the data set, P2C iterates over the following steps until convergence:

1. for each point, apply k-d trees to find clusters of points where each pair has a reciprocal distance less than d;
2. remove redundant clusters by ordering their elements (e.g., (A, B) vs. (B, A));
3. remove subsets of larger clusters (e.g. (A, B) vs. (A, B, C));
4. compute clusters' centroids as the average coordinates of all points belonging to the cluster;
5. assign the points which fall into more than one cluster to the one with the nearest centroid;
6. substitute each instance's coordinates with the centroid coordinates of the corresponding cluster.

The algorithm converges when the final number of points cannot be further reduced, since they all are farther apart from each other than the maximum distance d. After assigning each instance its new coordinates, we follow Han et al. (2012, 2014) in using the GeoNames data set to associate clusters with cities, by substituting the instance coordinates with those of the closest town center. In our case, however, rather than collecting cities with a population of at least 100K, we consider all towns with a population of at least 15K.

This last step further reduces the set of points associated with our instances. Table 1 shows the resulting number of labels, and the mean distance in km between the new instance coordinates and the respective town center.

This choice of 15K inhabitants is coherent with the settings of d: we aim to account for linguistic/social environments more specific than the broad and compound communities of densely populated cities. This is helpful for high resolution

method	model	# labels	Acc	Acc@161	mean	median
		Twitter-US				
Han et al. (2014)	NB + IGR	378	26%	45%	-	260
Wing and Baldridge (2014)	HierLR k-d	"fewer classes"	-	48%	687	191
Rahimi et al. (2017)	MLP + k-d tree	256	-	55%	581	91
	Att-CNN + k-d tree	256	26.17%	55.27%	580.7	93.02
	Att-CNN + P2C .1	1554	44.04%*	59.76%*	544.35*	47.19*
	Att-CNN + P2C .25	914	49.08%*	60.4%*	537.0*	39.71*
	Att-CNN + P2C .5	418	54.73%*	58.56%	537.79*	0*
		Twitter-World				
Han et al. (2014)	NB + IGR	3135	13%	26%	-	913
Wing and Baldridge (2014)	HierLR k-d	"fewer classes"	-	31%	1670	509
Rahimi et al. (2017)	MLP + k-d tree	930	-	36%	1417	373
	Att-CNN + k-d tree	930	18.35%	33.85%	1506.33	469.48
	Att-CNN + P2C .1	3047	22.57%*	39.41%*	1372.3*	328.42*
	Att-CNN + P2C .25	2818	26.68%*	39.94%*	1269.13*	299.04*
	Att-CNN + P2C .5	2350	32.64%*	41.8%*	1257.36*	292.09*

Table 2: Performance of prior work and of the proposed model with W-NUT and P2C labels. * : $p \leq 0.01$.

geolocation both in the case of crowded regions and of areas with low density of inhabitants. However, we found that in spite of qualified information, such as the annual Worlds Cities report of the United Nations, it is actually difficult to set an optimal threshold. In fact, not even that document provides a detailed profile of small towns at a global level. Therefore we rely on the format of the information offered by Geonames.

The code for computing P2C is available at github.com/Bocconi-NLPLab.

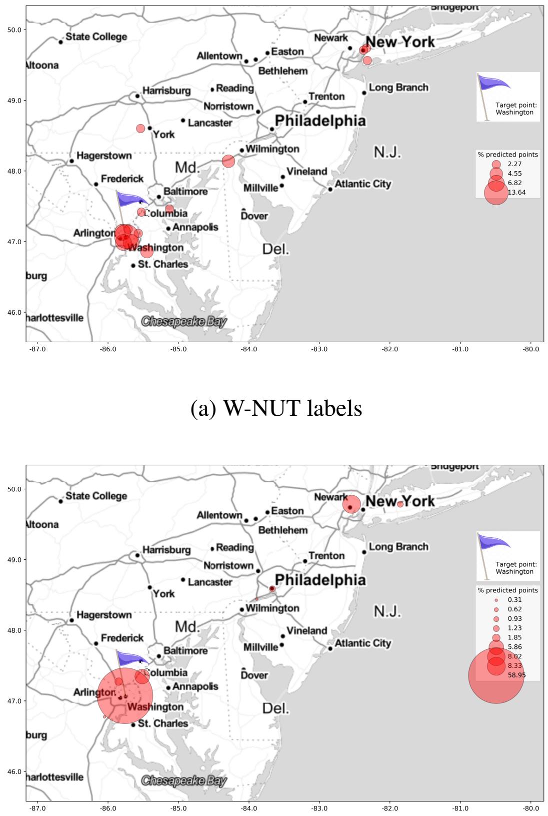
Feature selection The two corpora have very different vocabulary sizes. Despite fewer instances, Twitter-US contains a much richer vocabulary than Twitter-World: 14 vs. 6.5 millions words. This size is computationally infeasible. In order to maximize discrimination, we filter the vocabulary with several steps.

In order not to waste the possible geographic information carried by the huge amount of low frequency terms, we use replacement tokens as follows: We again take only the training data into account. First, we discard the hapax legomena, that is the words with frequency 1, as there is no evidence that these words could be found elsewhere. Then, we discard words with frequency greater than 1, if they appear in more than one place. We replace low frequency terms which appear uniquely in on place with a replacement token specific for that place, i.e., label. Finally, we substitute these words with their replacement token in the whole corpus, including development and test set. Since the word distribution follows the Zipf curve (Powers, 1998) we are able to exploit the geographic information of millions of words using only a small number of replacement tokens. The use of this information is fair, as it relies on the information present in the training set only. In terms of performance, however, the effect of the replacement tokens is theoretically not different from that resulting from the direct inclusion of the single words in the vocabulary.The benefit is in terms of noise reduction, for the selective removal of geographically ambiguous words, and computational affordability.

Following Han et al. (2012), we further filter the vocabulary via Information Gain Ratio (IGR), selecting the terms with the highest values until we reach a computationally feasible vocabulary size: here, 750K and 460K for Twitter-US and Twitter-World.

Attention-based CNN For classification, we use an attention-based convolutional neural model. We first train our own word embeddings for each corpus, and feed the texts into two convolutional channels (with window size 4 and 8) and max-pooling, followed by an overall attention mechanism, and finally a fully-connected layer with softmax activation for prediction.

For evaluation, as discussed in Section 3, we use the common metrics considered in literature: acc@161, that is the accuracy within 161 km (100 mi) from the target point, and mean and median distance between the predicted and the target points. We are also interested in the exact accuracy. This metric is often not shown in literature, but is important for the geolocation in real case scenarios. We evaluate significance via bootstrap sampling, following Søgaard et al. (2014).



(a) W-NUT labels

(b) P2C labels

Figure 1: Example of cumulative point accuracy with the two label sets for gold label Washington DC (flag). Circles are predictions, diameter represents percentage of predictions on that point.

4 Results

The model performance is shown in table 2. When applied to the W-NUT labels, our model replicates the results of Rahimi et al. (2017): in TWITTER-US the values correspond perfectly, in TWITTER-WORLD the Att-CNN performance is slightly lower. Compared to the W-NUT labels, the P2C labels are much more granular in every condition and, in spite of their apparent greater difficulty, they help to reach better performance in all metrics, with very high levels of significance. Such differences are surprisingly wide with respect to the accuracy: in TWITTER-US, for P2C with $d = .5$, the performance is more than doubled compared to the same model with the W-NUT k-d tree labels (54% vs. 26%).

Figure 1 shows the coordinates of the W-NUT (1a) and of the P2C cluster centroids (1b). The diameter of the circles represent the rate correct prediction for those points. As can be seen, P2C identifies a unique linguistic region around Washington, while different W-NUT labels cover more or less the same area. P2C labels also allow a much better concentration of predictions in the same administrative/linguistic area.

5 Conclusion

P2C is a method for geographic labeling that dynamically clusters points and links them to specific towns. The aims are 1) to gather the points belonging to the same linguistic areas; 2) to associate such areas with distinct, existing administrative regions; 3) to improve the models' effectiveness, training them with texts showing consistent linguistic patterns. Compared to the W-NUT k-d tree labels, P2C leads to remarkably higher performance in all metrics, and in particular in the accuracy, even in spite of the higher number of labels identified. This suggests that techniques like P2C might be particularly useful when high performance at high levels of granularity is required.

Acknowledgments

The authors would like to thank the reviewers of the various drafts for their comments. Both authors are members of the Bocconi Institute for Data Science and Analytics (BIDSA) and the Data and Marketing Insights (DMI) unit. This research was supported by a GPU donation from Nvidia, as well as a research grant from CERMES to set up a GPU server, which enabled us to run these experiments.

References

Miriam Cha, Youngjune Gwon, and HT Kung. 2015. Twitter geolocation and regional classification via sparse coding. In *Ninth International AAAI Conference on Web and Social Media*.

Mohammad Ebrahimi, Elaheh ShafieiBavani, Raymond Wong, and Fang Chen. 2018. A unified neural network model for geolocating twitter users. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 42–53.

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. Association for Computational Linguistics.

Tommaso Fornaciari and Dirk Hovy. 2019a. Dense Node Representation for Geolocation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (WNUT)*.

Tommaso Fornaciari and Dirk Hovy. 2019b. Geolocation with Attention-Based Multitask Learning Models. In *Proceedings of the 5th Workshop on Noisy User-generated Text (WNUT)*.

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

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

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217.

Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel density estimation for text-based geolocation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jey Han Lau, Lianhua Chi, Khoi-Nguyen Tran, and Trevor Cohn. 2017. End-to-end network for twitter geolocation prediction and hashing. *arXiv preprint arXiv:1710.04802*, pages 744–753.

Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2017. Text-based geolocation prediction of social media users with neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 696–705. IEEE.

Songrit Maneewongvatana and David M Mount. 1999. It's okay to be skinny, if your friends are fat. In *Center for Geometric Computing 4th Annual Workshop on Computational Geometry*, volume 2, pages 1–8.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2016. A simple scalable neural networks based model for geolocation prediction in twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 235–239.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1260–1272.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers)*, pages 507–511.

David MW Powers. 1998. Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pages 151–160. Association for Computational Linguistics.

Afshin Rahimi, Trevor Cohn, and Tim Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. *arXiv preprint arXiv:1804.08049*, pages 2009–2019.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*, pages 209–216.

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. Association for Computational Linguistics.

Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. 2009. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso. 2014. What's in a p-value in nlp? In *Proceedings of the eighteenth conference on computational natural language learning*, pages 1–10.

Haina Tang, Xiangpeng Zhao, and Yongmao Ren. 2019. A multilayer recognition model for twitter user geolocation. *Wireless Networks*, pages 1–6.

Benjamin 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)*, pages 336–348.

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*, pages 955–964. Association for Computational Linguistics.