Huntsville, hospitals, and hockey teams: Names can reveal your location

Bahar Salehi[♣], **Dirk Hovy**[♣], **Eduard Hovy**[♢] and **Anders Søgaard**[♣]
[♣]Department of Computer Science, University of Copenhagen
[♢]School of Computer Science, Carnegie Mellon University
bahar.salehi@gmail.com dirk.hovy@di.ku.dk
hovy@cmu.edu soegaard@di.ku.dk

Abstract

Geolocation is the task of identifying a social media user's primary location, and in natural language processing, there is a growing literature on to what extent automated analysis of social media posts can help. However, not all content features are equally revealing of a user's location. In this paper, we evaluate nine name entity (NE) types. Using various metrics, we find that GEO-LOC, FACILITY and SPORT-TEAM are more informative for geolocation than other NE types. Using these types, we improve geolocation accuracy and reduce distance error over various famous text-based methods.

1 Introduction

Because social media such as Twitter are used in both research and industry to monitor trends and identify sudden changes in society, it is critical to be able to locate social media users. In Twitter, however, only about 1% of all tweets are geotagged, and the location specified by users in their profile is often noisy and unreliable (Cheng et al., 2010).

Geolocation is the task of identifying users' general or primary location, when this is not readily available. Accurate geolocation can improve scientific studies, as well as technologies such as event detection, recommender systems, sentiment analysis, and knowledge base population.

Since tweets contain at most 140 characters, geolocation of individual tweets is rarely feasible. Instead, most studies focus on predicting the primary location of a user by concatenating their entire tweet history. While this provides more context, it is still a noisy source with features of varying informativeness.

In this paper, we focus on named entities (NEs), a particular rich source of information, and investigate how much they can reveal about a user's primary location. Wing and Baldridge (2014) showed lists of predictive features for multiple cities, where we observe NEs among the top 20 features. This is due to the inherent localization of many NEs. E.g., top features for Los Angeles contain NEs such as names (Irvine, disneyland), parts of names (diego , angeles), and abbreviations (UCLA, SoCal (Southern California)). This observation motivates us to examine nine common NE types in social media, and their location predictiveness. Additionally, we find that using only the top three most informative types for geolocation improves accuracy and reduces the median distance error.

Contributions We study (1) the geographical informativeness of nine named entity types, and (2) explore their effect in a logistic regression model of text-based geolocation. Among the previous top text-based models, we obtain the best performance using the hidden location information of the top three NE types. This suggests that users who would like to maintain privacy should avoid using such names.

2 Related Work

Most previous studies use textual features as input. Some use KL divergence between the distribution of a users words and the words used in each region (Wing and Baldridge, 2011; Roller et al., 2012), regional topic distributions (Eisenstein et al., 2010; Ahmed et al., 2013; Hong et al., 2012), or feature selection/weighting to find words indicative of location (Priedhorsky et al., 2014; Han et al., 2012, 2014; Wing and Baldridge, 2014).

All these studies require relatively large training sets to fit the models, and can be heavily biased by

Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 116–121
Copenhagen, Denmark, September 7, 2017.

Type	Example	%
PERSON	Barack Obama	31
GEO-LOC	Southern California	18
FACILITY	Edward theater	14
COMPANY	IBM	12
MOVIE	The town	10
BAND	pink floyd	9
PRODUCT	microsoft office	7
TV-SHOW	family guy	4
SPORT-TEAM	Eagles	2
All		55

Table 1: NE types considered in this paper and percentage of users in training set who use at least one of these NEs in their tweets.

major events during the time of collection, such as an election or a disaster. In contrast to our work, most do not consider multi-word NEs.

Only few text-based studies consider NEs, and if so, focus on location names using gazetteers like GeoNames, limiting the methods to the completeness of these gazetteers. Since they usually also use other text-based models, it is hard to determine how much location names contribute. These approaches depend on a name-disambiguation phase, using Wikipedia, DBPedia, or OpenStreetMap, since location names can refer to multiple locations (Brunsting et al., 2016).

Chi et al. (2016) explicitly study the contributions of city and country names, hashtags, and user mentionings, to geolocation. Their results suggested that a combination of city and country names, as well as hashtags, are good location predictors. Pavalanathan and Eisenstein (2015) suggest that non-standard words are more location-specific, and also, more likely to occur in geo-tagged tweets. In contrast to this paper, none of the previous works study how much various NE types reveal about the user location. Similarly, Salehi and Søgaard (2017) evaluate common hypotheses about language and location. However, they do not explicitly study named entities.

3 Resources

Data We use the WORLD dataset (Han et al., 2012), which covers 3,709 cities worldwide and consists of tweets from 1.4M users. Han et al. (2012) hold out 10,000 users as development and 10,000 as test set. For each user with at least 10 geotagged tweets, the user's location is set to be the city in which the majority of their tweets are from. We also use Han et al. (2012)'s method to extract the nearest city to a given latitude-longitude coordinate.

NER We use TwitterNLP (Ritter et al., 2011) to extract the nine most common NE types in Twitter. Table 1 shows the percentage of users in our training data who use at least one NE in their tweets. Overall, 55% of the users use at least one NE, with PERSON, GEO-LOC and FACILITY as the most popular types.

Twitter corpus In order to measure the geographical diversity of NEs, we construct a corpus from tweets posted one week before the WORLD dataset was collected (14 Sep, 2011 to 20 Sep, 2011). We remove all non-English and non-geo-tagged tweets from this corpus. This leaves us with 0.5M tweets. This corpus covers 167 countries and 2263 cities/regions around the world.[1] The most frequent countries are USA, Great Britain, Indonesia, Canada, Malaysia, Philippine and Australia, and the most frequent cities are London, Los Angeles, Chicago, Manhattan, Atlanta, Jakarta and Singapore. Using this corpus, we obtain the distribution of NEs over the cities of the world.

4 NE types and Geolocation

In Table 1, we have seen the *general distribution* of NE types, with PERSON, GEO-LOC and FACILITY as top three. In this section, we focus on the *predictiveness* of NEs (as features) for geolocation. Later, in Section 5, we will propose a method to improve geolocation by putting more emphasis on the top NEs and their hidden location information.

We conduct three experiments to quantify predictiveness of NEs. In the first, we measure the geographic distribution of each NE type, and measure their entropy. In the second experiment, we conduct feature selection via randomized logistic regression, and, in the third experiment, we establish a baseline by using majority classes for all types.

Geographic diversity We first measure the geographic distribution of each type. We extract all NEs in the WORLD training set and use the Tweet corpus to measure entropy and mean pairwise distance (in kilometers) between tweets that contain the same NEs. We compute unpredictability as entropy:

$$\mathrm{H}(x) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

[1] We map the latitude and longitude coordinates to cities/regions based on Han et al. (2012).

	Entropy		Avg. pairwise distance	
	city-level ↓	country-level ↓	in kilometers ↓	LR ↑
GEO-LOC	**2.581**	**0.756**	**3982.077**	**0.831**
FACILITY	**2.774**	**0.798**	**4368.122**	**0.851**
SPORT-TEAM	3.002	**0.806**	**4127.404**	**0.729**
MOVIE	**2.980**	1.110	5524.074	0.492
TV-SHOW	3.090	0.906	4713.947	0.465
PERSON	3.351	1.106	5157.701	0.544
BAND	3.519	1.199	5261.419	0.535
PRODUCT	4.119	1.358	5481.787	0.498
COMPANY	5.562	1.646	5814.398	0.611

Table 2: Average geographical variation/sparsity of each NE type in Twitter and average randomized logistic regression (LR) weights. ↓ = lower values are better, ↑ = higher values are better. The top three types in each column are shown in **BOLD**.

where the entropy of NE x is measured by computing $P(x_i)$, which is the probability that x is referring to the i^{th} city/country, based on the frequency. We measure the entropy in both city and country level, shown in Table 2.

For example, suppose *CMU* is found in four tweets from Pittsburgh and one from San Francisco, and *IBM* is found in one tweet each from Pittsburgh, San Francisco, Melbourne, and New York. In this case, the entropy for *CMU* will be lower than for *IBM*. This would indicate that *IBM* is less predictive than *CMU* for geolocation. To compute the entropy of an NE type, we average over the entropies of all NEs of that type.

The first three columns of Table 2 show that GEO-LOC and FACILITY are the least diverse location-wise. NEs of type PERSON are the most frequent NEs (see Table 1), and occur in more diverse locations. On the other hand, NEs of type SPORT-TEAM, the least frequent NEs, have low location diversity. PRODUCT and COMPANY are the least predictive types.

Feature evaluation In our second experiment, we use L_1 randomized logistic regression (Ng, 2004) on the training set to get the most predictive features. It measures how often a feature is predictive under varying conditions, by fitting hundreds of L_1-regularized models on subsets of the data. Each feature is assigned a weight between 0 and 1 based on their predictiveness. For example, the weights for countries and city names are high (on average 0.831) showing that they are very predictive. Yet, some examples of features with zero weight are *web*, *today* and *t.v.* showing that these features are not predictive at all. Table 2 (under LR column) shows the resulting prevalence for each type. These are compatible with the previous two metrics, showing GEO-LOC, FACILITY and SPORT-TEAM as the most predictive types.

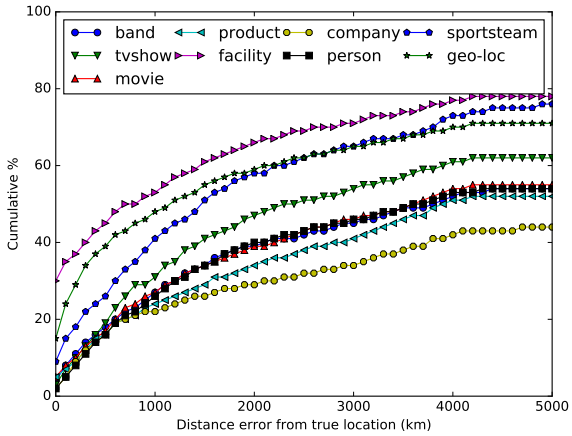

Figure 1: Distance error (test set)

Majority Vote Accuracy In the third experiment, we measure accuracy and distance error (in kilometers) using majority voting for each NE on the Tweet corpus. E.g., if we see *CMU* in four tweets from Pittsburgh and one tweet from San Francisco, we label the user's location as Pittsburgh. Figure 1 shows the percentage of the test set with a distance error from the true location less than K kilometers (also known as ACC@K). The three top types are again GEO-LOC, FACILITY and SPORT-TEAM, showing their higher impact on revealing the location of users.

5 NE Impact on Geolocation

Having established the informativeness of the various NE types, in this section, we examine the impact of NEs and their hidden location information for geolocation. To extract the hidden location information of each NE, we collect the locations of all tweets in our tweet corpus that contain that

Example	Me, my friend and the Eiffel tower
baseline	`Me my friend and the Eiffel tower`
Only NE	`[Eiffel tower]`
Baseline without NE	`ME my friend and the`
Our Method	`Me my friend and the Eiffel tower Paris Paris Paris Paris [Las Vegas] [Las Vegas]`

Table 3: Examples and features of methods in Section 5

	Accuracy			Distance	
Method	city↑	country↑	@161↑	Median↓	Mean↓
Baseline	17.6	**83.6**	33.6	515	**1727**
Only NE	9.3	53.6	17.7	2186	5317
Baseline without NE	14.8	82.2	29.9	612	1885
Our Method$_{allNEs}$	17.5	83.3	33.7	520	1769
Our Method$_{top3}$	**17.8**	**83.6**	**34.0**	**495**	1735
Previous studies					
Wing and Baldridge (2014)	–	–	31.0	509	1669
Han et al. (2012)	10.3	–	24.1	646	1953

Table 4: Accuracy and distance results for various methods. – indicates no report in respective paper

NE. To divide the world into regions with roughly the same number of users, we use a k-d tree approach proposed by Roller et al. (2012). As a result, we will cover larger regions when the population density is low and vice versa. Each region is then considered as a label to train the classifiers. The approach of using k-d tree is also used in Rahimi et al. (2015); Han et al. (2012) and Wing and Baldridge (2014).

See Table 3 for an example of the following methods. All use logistic regression as classifier, following Rahimi et al. (2015).

Baseline We use (Rahimi et al., 2015)'s bag-of-words model over tweets as baseline, which is also the state-of-the-art text-based method on the publicly available WORLD dataset.

Baseline without NE Here, we remove all NEs, to observe the influence of NEs in the bag-of-words model.

Only NEs In this approach, we consider *only* NEs and discard all other words in the tweets.

Our method We consider NEs and their inherent location information in addition to the bag-of-words model. The inherent location information for each NE is extracted from our Twitter corpus.[2] Suppose, for example, that *Eiffel tower* is found in four tweets from `Paris` and two tweets from `Las Vegas`. In this case, we add `Paris` (four times) and `Las Vegas` (twice) to the input text. The repetition is used to put more emphasis/weight based on frequency.[3] In order to measure the effectiveness of the three top NE *types* discovered in Section 4, we experiment with (1) considering all NE types (shown as **Our Method**$_{allNEs}$ in Table 4), and (2) the three most useful types (shown as **Our Method**$_{top3}$).

Evaluation metrics We use the same evaluation metrics as previous studies: accuracy depending on location granularity (city and country), accuracy within the distance of 100 miles/161km (ACC@161)[4], and median and mean error (in kilometers).

6 Results and Discussion

The results of applying each of the methods introduced in Section 5 are shown in Table 4. The baseline follows Rahimi et al. (2015), but does not use network information, to isolate the effect of NEs. They also add additional data, whereas we *only* consider the WORLD training set to be comparable with Wing and Baldridge (2014) and Han et al. (2012). Our baseline results are therefore lower than what Rahimi et al. (2015) report. Using only NEs results in a large performance drop with respect to the baseline. However, ignoring NEs (baseline-NE) also decreases the geolocation predictability by 15% (city level), indicating the importance of NEs in revealing the location of users.

Our proposed method, using all NE types

[2] As mentioned in Section 3, our Twitter corpus is the collection of tweets posted one week before the WORLD dataset was collected. This way we make sure that we are not training on test data.

[3] We also tried weighing features and samples according to their entropy, but we found repetition to perform better.

[4] ACC@161 measures near-miss predictions (Cheng et al., 2010)

according to their hidden location information, comes close, but does not improve over the baseline. However, when we consider only the top three NE types (GEO-LOC,FACILITY and SPORT-TEAM) from Section 4, performance increases, indicating that other NE types add noisy information.

Our error analysis shows that PERSON is very frequent, yet diverse, including politicians, athletes, and more general names. Since SPORT-TEAM is one of the most indicative types, we assume that athlete names can be useful as well. We leave this aspect for future work.

7 Conclusion

We compare the predictiveness of various named entity types for geolocation. We consider entropy, pairwise distance, feature selection weights, and the effect of the NEs on accuracy and error distance, and find that GEO-LOC, FACILITY and SPORT-TEAM are more predictive of location than other NE types.

Our results show that using the inherent localized information of NEs can improve geolocation accuracy. The results also suggest that users could obfuscate geolocation by avoiding these types.

Acknowledgments

This work was supported by the Data Transparency Lab.

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*. ACM, pages 25–36.

Shawn Brunsting, Hans De Sterck, Remco Dolman, and Teun van Sprundel. 2016. Geotexttagger: High-precision location tagging of textual documents using a natural language processing approach. *arXiv preprint arXiv:1601.05893* .

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*. ACM, pages 759–768.

Lianhua Chi, Kwan Hui Lim, Nebula Alam, and Christopher J Butler. 2016. Geolocation prediction in twitter using location indicative words and textual features. *WNUT 2016* page 227.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1277–1287.

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

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*. ACM, pages 769–778.

Andrew Y. Ng. 2004. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04, pages 78–85.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged twitter data. *arXiv preprint arXiv:1506.02275* .

Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, pages 1523–1536.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015. 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 the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL2015)*. The Association for Computational Linguistics, pages 630–636.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *EMNLP*.

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*. Association for Computational Linguistics, pages 1500–1510.

Bahar Salehi and Anders Søgaard. 2017. Evaluating hypotheses in geolocation on a very large sample of twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text (WNUT)*. Copenhagen, Denmark.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *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*. Association for Computational Linguistics, pages 955–964.