Evaluating hypotheses in geolocation on a very large sample of Twitter

Bahar Salehi and **Anders Søgaard**
bahar.salehi@gmail.com soegaard@di.ku.dk
Department of Computer Science
University of Copenhagen

Abstract

Recent work in geolocation has made several hypotheses about what linguistic markers are relevant to detect where people write from. In this paper, we examine six hypotheses against a corpus consisting of all geo-tagged tweets from the US, or whose geo-tags could be inferred, in a 19% sample of Twitter history. Our experiments lend support to all six hypotheses, including that spelling variants and hashtags are strong predictors of location. We also study what kinds of common nouns are predictive of location after controlling for named entities such as *dolphins* or *sharks*.

1 Introduction

Geolocation is interesting for several reasons. It has applications to personalization, event extraction, fraud detection, criminology, privacy, etc.; but it is also a method for studying how location affects language use, as well as how linguistic change interacts with geography.

The growth of social media has made large-scale geolocation studies possible, and most recent work on geo-location use social media data, primarily from Twitter. On Twitter about 1% of tweets are geo-tagged by the media users (Cheng et al., 2010), and while this is a tiny fraction of the full corpus, it enables us to query for millions of geo-tagged tweets.

Geolocation models rely on various intuitions about what linguistic constructions are predictive of location. Specifically, many authors have used city and country names as features for geolocation, as well as Twitter hashtags and spelling variations.

In this paper, we study names, hashtags, spelling variants, as well as a wide range of other features, and evaluate their usefulness in geolocation at a *very large scale*. We discuss different feature groups and show, for example, what common nouns are more predictive of location. One example of such a noun could be *earthquake*, which is a commonly used noun that refers to a natural disaster, but obviously within a given time frame, such natural disasters hit in very specific places, where people are more likely to tweet about them.

This paper does not introduce a novel geolocation model, but uses more data than previous studies to examine the research hypotheses that have guided recent work in the field.

Contributions (a) We evaluate common hypotheses about language and location on a much larger scale than previously done. (b) We show, as expected, that place names and hashtags are predictive of location. (c) We also show that spelling variation, out of vocabulary and non-standard words such as *feelinn* are indicative of location, even more so than the standard (in dictionary) words. This seems to hold for British spelling (in the US), abbreviations and phonologically motivated spelling. (d) We also analyze what classes of common nouns are indicative of location, discussing the problem of controlling for named entities that are frequent members of some of these classes. In social media, for example, animal words such as *dolphins* and *sharks*, may refer to cities' sport teams. Best predictors after controlling for named entities include natural phenomena, occupations, and organizations. (e) We show that the same findings also apply to geolocation of users around the world.

2 Related work

In text-based geolocation, researchers have used KL divergence between the distribution of a users words and the words used in geographic regions

Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 62–67
Copenhagen, Denmark, September 7, 2017.

(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., 2012a, 2014; Wing and Baldridge, 2014).

Han et al. (2012b) showed that information gain ratio is a useful metric for measuring how location-indicative words are. They used a sample of 26 million tweets in their study, obtained through the public Twitter API.

Salehi et al. (2017) evaluate nine name entity types. Using various metrics, they find that `GEO-LOC`, `FACILITY` and `SPORT-TEAM` are more informative for geolocation than other NE types.

Chi et al. (2016) specifically study the contributions of city and country names, hashtags, and user mentions, to geolocation. Their results suggested that a combination of city and country names, as well as hashtags, are good location predictors. They used a sample of 9 million tweets in their study, obtained through the public Twitter API.

Pavalanathan and Eisenstein (2015) investigate the potential demographic biases of using non-standard words and entities. They show that younger users are more likely to use geographically specific non-standard words, while old men tend to use more location revealing entities.

In this study, we use more data than previous studies to examine the research hypotheses and linguistic features established in previous works on geolocation. In addition, in order to examine the generalizability of our findings we examine them on a more geographically diverse dataset covering tweets from all around the world.

3 Gathering Data

Our dataset is a fraction of a 19% random sample of the entire history of Twitter (the union of two independent 10% samples), up until early 2016. We consider the fraction of geo-located tweets with US geo-coordinates, as well as tweets with low-entropy location strings, for which we infer geo-tagged fractional counts:

If a location string s is used more than n times, we compute its distribution $P(s \mid county)$ over US counties. Non-geotagged instances of tweets associated with s are then attributed to counties based on this distribution as fractional counts. The final corpus consists of roughly 120 billion tweets and around 450 million of these had geo-tags. Using the inference methodology, were able to attribute roughly 10 billion tweets.

We look at the distribution of words over US counties in this corpus, limiting ourselves to the most frequent 100,000 words. This is important to ensure support, but also makes geolocation harder, since rare words are generally more predictive of location. On the other hand, the fact that we rely on relatively frequent words makes our analysis more widely applicable.

The corpus has 4.5B tokens of the 100,000 most frequent words. So, on average we have 45,000 occurrences of each word. The minimum frequency is 612 tokens; the most frequent word occurs 138M times. The median is 1,742 occurrences.

4 Metrics

In this section, we introduce two metrics we use to examine the degree of location informativeness of words. Entropy and KL divergence.

KL divergence The Kullback-Leibler divergence (KLD) (also known as information gain) is used to measure the similarity between two distributions. We use KLD to measure the similarity between the distribution of a word (P) with the distribution of all words (Q) across counties.

$$\mathrm{KLD}(P_{word}, Q) = \sum_{c \in counties} P_{word}(c) \log \frac{P_{word}(c)}{Q(c)}.$$

Higher KLD shows less similarity to the distribution of all words and as a result higher location predictiveness.

Entropy In information theory, entropy measures the unpredictability, where low entropy indicates high predictability. In this study, we compute entropy of each word as below:

$$\mathrm{H}(word) = -\sum_{c \in counties} P_{word}(c) \log P_{word}(c)$$

where $P_{word}(c)$ is the probability of observing $word$ in the county c. This is computed by dividing the frequency of that word in that county by the total number of words in that county.

5 Experiments

5.1 Location Indicative Words

In this section, we examine the following 6 hypotheses using entropy and KLD metrics:

HYPOTHESIS (">" = "MORE PREDICTIVE THAN")	Ent1	Ent2	KLD1	KLD2	p-value
#0 Dictionary words > stopwords	**5.67**	7.74	**0.74**	0.12	< 0.001
#1 US English < British	5.59	**5.13**	0.63	**0.94**	< 0.05
#2 Dictionary words < geonames	5.67	**4.62**	0.74	**1.79**	< 0.001
#3 Dictionary words < OOV words	5.67	**4.61**	0.74	**1.52**	< 0.001
#4 Dictionary words < hashtags	5.67	**4.25**	0.74	**1.82**	< 0.001
#5 Dictionary words < emoticons	5.67	**5.07**	0.74	**1.03**	< 0.001
#6 Non-standard words > their normalized version	**5.27**	7.52	**0.92**	0.17	< 0.001

Table 1: Evaluating hypotheses by comparing average entropy/KLD scores of first group's words (Ent1/KLD1) with the words in second group (Ent2/KLD2). Lower entropy and higher KLD show higher predictability. P-value shows the significance of the differences.

0. Stopwords are not good predictors compared to other dictionary words.

1. British English is more location-specific than American English (in the US).

2. Geonames are better predictors compared to dictionary words.

3. Words not in dictionary (OOV) are better predictors than words in dictionary.

4. Hashtags are better predictors than dictionary words.

5. Emoticons are better predictors than dictionary words.

6. Non-standard spelling variants are better predictors than their standard spellings.

The results are shown in Table 1. The rest of this section investigates each of the hypotheses in more details.

Hypothesis #0: We use the NLTK stopword list for English and found 143 unique stopwords in our data.[1] As expected, stopwords are the least location predictive group of words.[2] Among the stopwords *ain*, *wasn* and *wouldn* are among the most predictive ones. Note that *wasn't* and *wouldn't* are also in our data, but they are not as location predictive.

Hypothesis #1: We also compare words with spelling variations in British and American English, using https://en.oxforddictionaries.com/usage/british-and-american-terms as our data source. Overall, in our data, we found 475 words that have different spellings in British and American English.

We observe that British spellings are more predictive. For example, while *harbor* and *harbour* are mostly observed in coastal areas, *harbour* (British) is more often observed in eastern coastal regions, while *harbor* (American) is distributed more diversely. However, the difference between British and American words is not significant.

Hypothesis #2: City names and country names are often said to be more predictive of location. In this experiment, we use GeoNames[3] to find city/country names including their alternative names. We found 23,701 geonames in our data. We observe that on average, geonames are significantly more location indicative than the rest of dictionary words.

Hypothesis #3: Both metrics show that OOV words are on average more predictive of location than the dictionary words. Note that such words are among the 100K most frequent words and are not considered as random noise. We consider words not found in WordNet for OOV words. Overall, we found 31,049 dictionary words and 68,951 OOV words in our data.

Hypothesis #4: Our experiments show that hashtags are significantly more predictive than the dictionary words as well as the rest of the examined OOV words. In our data, *#1* and *#fail* are among the least predictive hashtags, while the most predictive hashtags are mainly location names and events such as *#monett* (a city name), *#disney366* and *#zipsblackout*. In our data, we have 17,131 hashtags.

Hypothesis #5: Emoticons[4] are the last group

[1] http://www.nltk.org/data.html

[2] This experiment is more of a sanity check.

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

[4] There was no emojis in our list of most frequent 100K

Synsets	Examples	After elimination Synsets	Examples
indian.n.01	yuma, muskogee	unit.n.03	usaf, sss
amerindian.n.01	yuma, muskogee,	natural_phenomenon.n.01	whiteout, earthquake
wood.n.01	hazelwood, tupelo	alcohol.n.01	homebrew, oktoberfest,
agency.n.01	usaf, sss	phenomenon.n.01	whiteout, earthquake,
extremity.n.04	terminus, skyline	occupation.n.01	engineering, internship
plant_material.n.01	hazelwood, tupelo	symbol.n.01	emmys, phd,
traveler.n.01	trespasser, tourists	region.n.01	aerospace, rooftop
person_of_color.n.01	yuma, muskogee	worker.n.01	esthetician, hairstylist
fish.n.01	sharks, marlins	implement.n.01	poker, nutcracker
administrative_unit.n.01	usaf, sss	inhabitant.n.01	peruvian, hoosiers
american.n.01	hoosiers, tarheels	organization.n.01	friendlys, usaf
geological_formation.n.01	seaside, canyon	liquid.n.01	cocktails, espresso

Table 2: Most location predictive synsets before and after eliminating the location and sport team names.

	Median distance error	Accuracy (city)	Accuracy (country)
In dictionary	860	10.85	77.93
OOV	667	14.96	79.95
Geonames	698	14.43	79.60
All	510	17.41	84.04

Table 3: Geolocation results on WORLD dataset

of OOV words in our analysis. We found 196 emoticons in our data. According to Table 1, emoticons on average are more predictive than dictionary words. Yet, they are among the least predictive ones in the group of OOV words. Our further analysis shows that *:)* and *;)* are the least predictive emoticons, while *(^-^)* and *=)* are among the most predictive ones showing that emoticons can also be location predictive. This is in line with the work of (Park et al., 2013), where they observed that people in Eastern countries prefer vertical emoticons (based on eye shape style), while Western countries prefer horizontal ones (based on mouth style).

Hypothesis #6: Among the words not found in dictionary, there exist non-standard words, which are typos, ad hoc abbreviations, unconventional spellings and phonetic substitutions (Han et al., 2012a), such as *2mrw* (i.e., *tomorrow*). Here, we use (Han et al., 2012a) to compare these nonstandard words with their normalized versions. Overall, we found 4,795 non-standard words, as well as their normalized version in our data.

Using entropy and KLD, we show that the normalized versions are not very location indicative, yet, the non-standard words are significantly more predictive than their normalized versions. This shows that preferred styles to write words in a non-standard way have implicit location information.

words.

5.2 Semantic Classes

In this section, we examine the semantic categories that are most location indicative using WordNET. For each word, we extract all the possible hypernymes. The synsets with less than 10 samples are removed. For each synset, the median entropy and KLD of the respective samples are calculated.

The synsets observed in both the top 20 synsets using entropy and the top 20 synsets using KLD are shown in Table 2. We noticed that the name of sports teams and locations are among the top categories. For example, the samples of *Wood.n.01*, such as *hazelwood* and *tupelo*, are also part of the name of locations in the United States, and *sharks* and *marlins* from *fish.n.01* are part of sports team names. Therefore, we removed the words which are part of the names of US teams using DBPedia and locations using geonames. This resulted in a different top 20 categories, which are shown under *after elimination* column. After eliminating named entities of cities, countries and sport teams, we observe that the best predictors are mostly natural phenomena, occupations, and organizations.

6 Geolocation

We also evaluated the above hypotheses in the context of a geolocation experiment using the geographically diverse more dataset, WORLD (Han et al., 2012c). The WORLD dataset covers 3,709 cities worldwide and consists of tweets from 1.4M

users, where 10,000 users are held out as development set and 10,000 as test set. The task is to predict the primary location of a new user based on that person's tweet history.

We use logistic regression as classifier to predict the users location, following Rahimi et al. (2015). The results (median distance error, city accuracy and country accuracy) are shown in Table 3.

Similar to our findings in our analysis above, we see that OOV words are better features than dictionary words. Also geonames features, alone, have high performance, even better than dictionary words. The rest of the examined groups are not performing as good, individually. The combination of all words (shown as All) results in the best performance.

7 Conclusion

In this paper, we examined six hypotheses about location-specific language use. We confirmed that OOV words are more predictive of location than dictionary words. Moreover, we showed that spelling variants and hashtags are strong predictors for location. Finally, we showed that our findings are also applicable to geolocation of users around the world.

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.

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. 2012a. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, pages 421–432.

Bo Han, Paul Cook, and Timothy Baldwin. 2012b. 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. 2012c. 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.

Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. 2013. Emoticon style: Interpreting differences in emoticons across cultures. In *ICWSM*.

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.

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, Dirk Hovy, Eduard Hovy, and Anders Søgaard. 2017. Huntsville, hospitals, and hockey teams: Names can reveal your location. 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.