Text-based Geolocation of German Tweets

Johannes Gontrum **Tatjana Scheffler**
Department of Linguistics
University of Potsdam, Germany
gontrum,tatjana.scheffler@uni-potsdam.de

Abstract

We show a new, data-driven method for geolocating single tweets based on the geographical variance of their tokens. While more than half of German tweets do not contain reliable textual indicators of their location, our method can locate 40% of tweets very accurately, up to a distance of 7km (median) or 93km (mean).

1 Introduction

Twitter data is interesting for many NLP applications because of its abundant metadata. This includes geolocation data (GPS coordinates), indicating where the tweet's author was located at the time of writing. Geolocation information is important for the detection of regional events, the study of dialectal variation (Eisenstein, to appear 2015), and many other possible applications. However, not all users allow the public distribution of their location data, and in some language communities, geolocated tweets are very rare. For example, only about 1% of German tweets contain a location, and these come from an even smaller number of users that allow this feature (Scheffler, 2014).

In this paper we introduce an approach to recover a geolocation of origin for individual tweets using only the text of the tweet. This allows the enrichment of Twitter corpora that do not contain sufficient geo information, even for unseen users or users who never share their location. This is important since many users (e.g. in Germany) use made-up or false locations in their user profile field. We use geo-tagged tweets in order to derive a lexicon of regionally salient words, which can then be used to classify incoming tweets.

2 Related Work

Geolocation of Twitter messages can be based on the user's location as indicated in the profile, or a tweet's GPS location. Text-based geolocation does not take user information into account. Previous approaches however commonly aggregate all of a user's tweets (Cheng et al., 2010; Wing and Baldridge, 2014) or conversations including replies (Chandra et al., 2011) to determine one location. Some researchers have instead attempted to directly derive location-specific words or dialectal variation from geotagged tweets (Eisenstein et al., 2010; Eisenstein, to appear 2015; Gonçalves and Sánchez, 2014), using GPS locations or user profile locations.

(Pavalanathan and Eisenstein, 2015) compared the data sets obtained by user profile and GPS geolocation of tweets, respectively, and show that they differ significantly with respect to demographics and linguistic features. (Graham et al., 2014) show that user profile information is only rarely a reliable indicator of the location of the user, more than half of profiles containing empty location fields, unhelpful locations ("earth") or diverging user profile and GPS information.

In a previous paper (Scheffler et al., 2014), we first attempted to geolocate individual tweets based only on that tweet's text, using predefined "dialect" regions in Germany as our goal. In that work, we also discussed a thesaurus-based approach using an existing list of known dialectal words as seed words. That approach was vastly inferior to a method that automatically induces regionally salient words from geo-tagged tweets. The current paper shows a completely new, data-driven solution to that problem.

3 Approach

It is important to note that there are at least two distinct sources for regionally distinctive language in a tweet: (i) the current location of the author, which leads to the use of local event and place names, and (ii) the dialectal region of origin of the author, which yields regionally salient dialectal

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 28–32, University of Duisburg-Essen, September 28, 2015. ©2015 GSCL

expressions. In principle, these two sources are independent of each other (think of an Austrian travelling to Berlin). However, using current methods, neither we nor any of the previous work can systematically distinguish these two types of geographic origin of a tweet. In this work, we assume that for statistical purposes, most users are located close to their region of origin and thus do not address this problem further. However, this may lead to discrepancies in individual cases where a user is either travelling or writes about a distant location.

Further, for evaluation purposes we regard the GPS metadata information provided by Twitter as gold location data for our corpus. This is in line with previous approaches, but potentially biases the algorithm towards case (i) above – the current location of the tweet author. Dialect origin information is a lot harder to obtain, but could potentially be gathered through surveys or in an unsupervised or bootstrapping manner.

Data Our corpus consists of 65 mio. tweets that have been collected through the Twitter API between February and May 2015, by filtering the Twitter stream using a keyword list of common German words (Scheffler, 2014). Language identification was carried out using LangID (Lui and Baldwin, 2012). Further, we extracted only tweets that were geo-tagged and located in Germany, Switzerland or Austria. To remove bots we manually created lists of suspicious user ids and ignored messages containing the words 'nowplaying' or '4sq'. We tokenized the lower-cased tweets, removed numbers, URLs, user-mentions and most special characters. After removing the '#'-character, hashtags remain in the tweets since they can provide useful information about local events. Only 360k tweets (0.55% of all collected documents) fulfilled our criteria. We then randomly extracted 1000 messages each for testing and development.

Background Our method is based on the observation that tokens are not used at all locations with the same frequency. Hence, there must exist a function that describes the probability that a token is used in a tweet at given coordinates. Additionally, we assume that these probabilities are distributed around a specific location at which the probability of the token is the highest.

We have discovered that in contrast to common words that are used uniformly throughout, regional words like city names are used in an area with a diameter of 50-150km by many users. The highest level of information is provided by local terms denoting for example local events or street names that are only used a few times, but at a very narrow location. This distinction can be observed by printing the probability density function (PDF) of the tokens in tweet (1), see Figure 1.

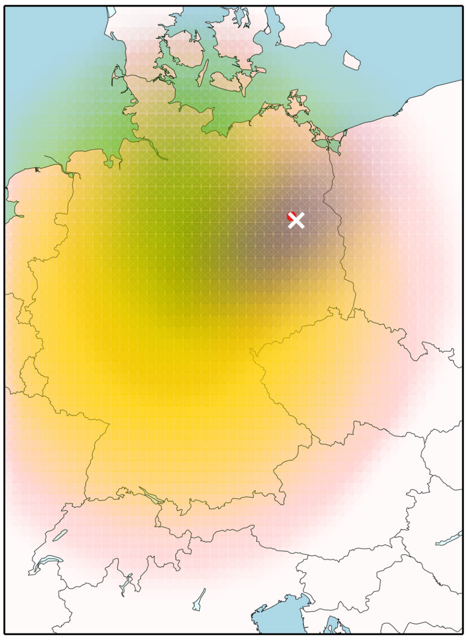

Figure 1: PDF of tokens in tweet (1): regional words *berlin* (blue), *hhwahl* (green); highly local word *nordbahnhof* (red); common words (yellow). Position of the tweet marked by a white cross.

(1) *balken gucken und so hhwahl pa nordbahnhof in berlin*

The common tokens (*balken, gucken, und, so, pa, in*) are drawn in yellow. They are so widely distributed that they cover the whole of Germany and are not providing any local information that could help classify the tweet. The regional word *hhwahl*, denoting an election in Hamburg, is illustrated in green and the density function of the other regional word *berlin* is drawn in blue around the location of the city. Finally, the word *nordbahnhof* has only been observed close to that station in Berlin and is therefore a local word (red). In fact, the tweet was sent within a distance of only 4km.

Classification method The tweet in (1) illustrates the importance of finding a parameter to distinguish common and widespread words from regional and local tokens. Additionally, we need a method to weight the remaining tokens so that highly local words are given more significance than less concrete regional words. We use the variance of the probability distribution of a token as a score that can be used to solve both our problems.

Since the variance describes how widespread the data points are, regional or local words that appear only in a small area will have a low variance, while variance is high for common words or even low-frequent words like typos that are not regionally biased. First, we use the variance as a threshold to remove common words from tweets and calculate the geographical midpoint of the remaining tokens. We found that the median for a token position outperforms the mean especially for infrequent terms, since it marks an actual coordinate where the token was used.

An analysis of our data reveals the importance of low-variance local terms. If a tweet contains one of these highly local tokens, the tweet's position is almost entirely determined by that token's median position and any influence of other tokens would worsen our score. Secondly, we therefore weight the individual tokens by their inverse variance σ^{-1}, so that very local tokens receive an extremely high score and overshadow all other words. If a tweet on the other hand contains exclusively regional words, their inverse variance is not too high so all of them have an influence on the position.

Algorithm The median position and the variance for each token in a tweet is calculated based on the coordinates of all tweets in the training corpus in which they are used. Note that we are converting the longitude and latitude information provided by Twitter to three-dimensional Cartesian coordinates. Since longitude and latitude are projections on a sphere, the calculation of midpoints and distances becomes less error prone this way. Therefore, we are from now on regarding median and variance values as vectors.

Equation (2) shows the calculation of the location of a tweet t with tokens $t_0, ..., t_n$, their variance values $\vec{\sigma_0}, .., \vec{\sigma_1}$ and their median $\vec{m_0}, .., \vec{m_1}$.

$$Loc(t) = \frac{\sum_{i=0}^{n} \vec{\sigma_i}^{-1} * \vec{m_i}}{\sum_{i=0}^{n} \vec{\sigma_i}^{-1}} \qquad (2)$$

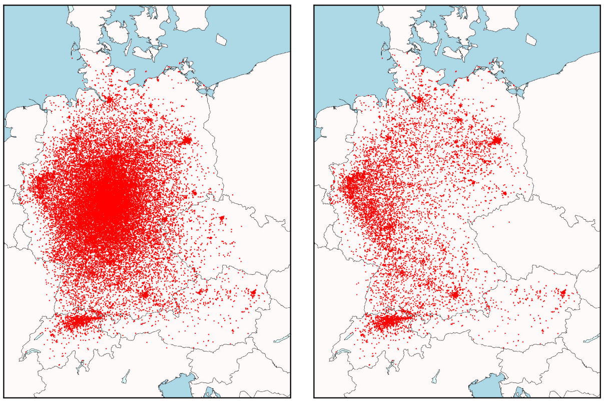

Figure 2: Left: Mean coordinates of all tokens. Right: Only regional tokens under the assumption that 25% of all tokens are regionally salient.

4 Results and Discussion

Filtering Step It is clear that some tweets are unsuitable for geolocation using only their text. This is due to the fact that a majority of tokens are so common that they carry no information about any location whatsoever. As a consequence, the original position of tweets that contain only these irrelevant tokens cannot be recovered from the text alone. To make things worse, any attempt to do so will lead to unjustified confidence in the calculated position and will result in an unreliable algorithm.

Figure 2 shows the mean coordinates for all tokens in the corpus on the left, while in the right graphic only the top 25% of tokens (by lowest variance) remain. The blob in the center of Germany are those meaningless tokens that are removed with a decreasing variance threshold. For this reason, we are deliberately filtering a number of tokens that are lacking reliable information and consequently accept a high amount of unclassifiable tweets for the sake of accuracy.

Experiments The determination of a variance threshold for common words can be seen as an estimate of the ratio of regional tokens in the corpus. For example, a threshold of 30% means that we regard the 30% of the tokens with the lowest variance as regional and remove all other words. Figure 3 displays these scores for different parameter estimates of the percentage of regional words (x-axis). As expected, the geolocation error (measured in distance to the true location) decreases

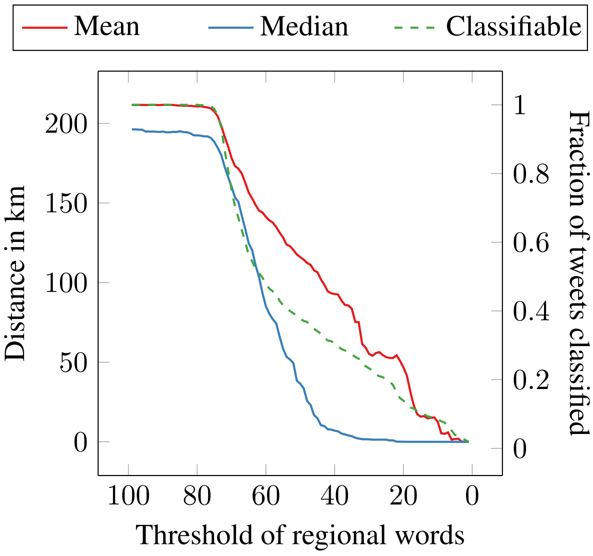

Figure 3: The mean and median distance in *km* between the predicted and the true metadata coordinates of a tweet.

Threshold	Mean	Median	#Tweets
100	212km	196km	1000
75	207km	188km	988
50	116km	36km	377
40	93km	7km	306
30	55km	1.56km	233
20	47km	0.06km	139
10	12km	0.00km	84

Table 1: Results of geolocation algorithm for different variance estimates: “Threshold”=ratio of ‘regional’ words (by variance), error distances to the true location, and number of classified tweets (N=1000) are given.

with a stronger threshold, as the amount of unclassifiable tweets grows. We can make out three stages that correspond to our classification of tokens: The first notable improvement of the score happens when the most frequent of the common words are removed at about 70%. In the next stage widespread regional words are gradually removed and at about 30%-40%, most tweets rely exclusively on local words.

Even though the distance median drops below 10km at 43%, the mean distance stays relatively high. We explain this gap by a few tweets whose predicted location is hundreds of kilometers away from their true metadata position. As discussed above, this can happen either when tweets mention distant events or locations, or when people travel away from their dialect regions and use dialectal expressions in tweets. Since we compare the predicted location with the GPS metadata from Twitter (our “gold” data), our method cannot avoid these problems. On the other hand, some tokens are wrongly classified as local or regional due to their infrequent appearance in our small training corpus and therefore the accuracy will increase with a bigger data set. Table 1 shows the geolocation errors as well as the number of classified tweets for different variance parameter thresholds of regional words.

We have also analyzed which tokens are classified as local or regional for certain cities, as shown in Table 2. In Berlin and Essen for example, mostly street or district names are revealing, while in Zurich dialectal words are dominating.

Finally, we created a score to compare our results to the ones from a previous paper (Scheffler et al., 2014), where the German speaking area was manually divided into seven regions, and success was measured by the percentage of tweets correctly classified into these regions. To achieve a rough comparison, we used a clustering algorithm on randomised data to create seven regions that cover an equally large area. In (Scheffler et al., 2014) a threshold was used to remove common words and only 20% of all tweets were classified, resulting in 53% correctly classified tweets. When adjusting our method to this threshold, we accurately classify 86% of tweets into the correct region, a large improvement. However, since the previous paper used a different dataset, the results are still not directly comparable.

In summary, this paper introduces a new, language independent, highly accurate approach to geolocating single tweets based on the geographical variance of words in the corpus. The method can be further augmented by user-oriented approaches in order to improve recall.

5 Future Work

The task opens up many avenues for future research. Most importantly, the differentiation of the two essentially distinct sub-tasks – identifying the location and dialect origin of the author – must be addressed, although this will require

Berlin	**Zurich**	**Essen**
kadewe	tagi	rheinische
kudamm	uf	hattingen
alexanderplatz	het	herne
friedrichshain	isch	westfalen
brandenburg	scho	ddorf
fernsehturm	au	ruhr
dit	zuerichsee	thyssenkrupp
morjen	gseh	duisburg

Table 2: Notable local tokens with low variance and high frequency in Berlin, Zurich, and Essen.

more complex models. A resource for location words such as OpenStreetMap might help here. Another obvious improvement, also suggested by a reviewer, is the training of the words' significance weights by machine-learning methods (instead of fixing them to the variance). Finally, it is still unclear how much the algorithm overfits to certain frequent and predictable tweeters, like bots. Frequently-tweeting bots may on the one hand hurt performance, since the model falsely associates all its words with the bot's location. On the other hand, this may also help if the test data also includes tweets from the same source. This behavior can be tested by evaluating the system on sufficiently different material (e.g., from a different point in time (Rehbein, p.c.)), and mitigated by developing methods to exclude non-natural tweets during preprocessing.

Acknowledgments

The authors would like to thank Matthias Wegel and Steve Wendler for their involvement in previous versions of this work. We are grateful to Ines Rehbein and the anonymous reviewer for their detailed comments on our work and suggestions for future improvements.

References

S Chandra, L Khan, and F B Muhaya. 2011. Estimating Twitter User Location Using Social Interactions–A Content Based Approach. In *IEEE Third International Conference on Social Computing (SocialCom)*, pages 838–843, October.

Z. Cheng, J. Caverlee, and K. Lee. 2010. You are where you tweet. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 759.

J. Eisenstein, B. O'Connor, N. Smith, and E.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.

J. Eisenstein. to appear 2015. Identifying regional dialects in online social media. In *Handbook of Dialectology*.

B. Gonçalves and D. Sánchez. 2014. Crowdsourcing dialect characterization through twitter. *PLOS One*, 9(11):1–10.

M. Graham, S. Hale, and D. Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*.

M. Lui and T. Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Republic of Korea.

U. Pavalanathan and J. Eisenstein. 2015. Confounds and Consequences in Geotagged Twitter Data.

T. Scheffler, J. Gontrum, M. Wegel, and S. Wendler. 2014. Mapping German tweets to geographic regions. In *Proceedings of NLP4CMC workshop at the 12th KONVENS*.

T. Scheffler. 2014. A German Twitter snapshot. In N. Calzolari et al., editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

B. Wing and J. 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.