Inferring the location of authors from words in their texts

Max Berggren & Jussi Karlgren
Gavagai, Stockholm
{max, jussi}@gavagai.se

Robert Östling & Mikael Parkvall
Dept of Linguistics, Stockholm University
{robert, parkvall}@ling.su.se

Abstract

For the purposes of computational dialectology or other geographically bound text analysis tasks, texts must be annotated with their or their authors' location. Many texts are locatable but most have no explicit annotation of place. This paper describes a series of experiments to determine how positionally annotated microblog posts can be used to learn location indicating words which then can be used to locate blog texts and their authors. A Gaussian distribution is used to model the locational qualities of words. We introduce the notion of placeness to describe how locational words are.

We find that modelling word distributions to account for *several locations* and thus several Gaussian distributions per word, defining a filter which picks out words with high placeness based on their *local distributional context*, and aggregating locational information in a *centroid* for each text gives the most useful results. The results are applied to data in the Swedish language.

1 Text and Geographical Position

Authors write texts in a location, about something in a location (or about the location itself), reside and conduct their business in various locations, and have a background in some location. Some texts are personal, anchored in the here and now, where others are general and not necessarily bound to any context. Texts written by authors reflect the above facts explicitly or implicitly, through explicit author intention or incidentally. When a text is locational, it may be so because the author mentions some location or because the author is contextually bound to some location. In both cases, the text may or may not have explicit mentions of the context of the author or mention other locations in the text.

For some applications, inferring the location of a text or its author automatically is of interest. We present in this paper how establishing the location of a text can be done by the locational qualities of the terminology used by its author. Here, we investigate the utility of doing so for two distinct use cases.

Firstly, for detecting regional language usage for the purposes of real-time dialectology. The issue here is to find differences in term usage across locations and to investigate whether terminological variation differs across regions. In this case, the ultimate objective is to collect sizeable text collections from various regions of a linguistic area to establish if a certain term or turn of phrase is used more or less frequently in some specific region. The task is then to establish where the author of a text originally is from. This has hitherto been investigated by manual inspection of text collections. (Parkvall 2012, e.g.)

Secondly, for monitoring public opinion of e.g. brands, political issues, or other topic of interest. In this case the ultimate objective is to find whether there is a regional variation for the occurrence of opinionated mentions for the topic or topical target under consideration. The task is then to establish the location where a given text is written, or, alternatively, what location the text refers to.

In both cases, the system is presented with a body of text with the task of assigning a likely location to it. In the former task, typically the body of text is larger and noisier (since authors may refer to other locations than their immediate context); in the second task, the text may be short and have little evidence to work from. Both tasks, that of identifying the location of an author, or that of a text, have been addressed by recent experiments with various points of departure: knowledge-based, making use of recorded points of interest in a location, modelling the geographic distribution of topics, or using social network analysis to find additional information about the author.

This set of experiments focuses on the text itself and on using distributional semantics to refine the set of terms used for locating a text.

2 Location and words as evidence of locations

Most words contribute little or not at all to positioning text. Some words are dead giveaways: an author may mention a specific location in the text. Frequently, but not always, this is reasonable evidence of position. Some words are less patently locational, but contribute incidentally, such as the name of some establishment or some characteristic feature of a location.

Some locational terms are polysemous; some inspecific; some are vague. As indicated in Figure 1, the term *Falköping* unambiguously indicates a town in *Southern Sweden*, which in turn is a vague term without a clear and well defined border to other bits of Sweden. The term *Södermalm* is polysemous and refers to a section of town in several Swedish towns; the term *spårvagn* ("tram") is indicative of one of several Swedish towns with tram lines. We call both of these latter types of term *polylocational* and allow them to contribute to numerous places simultaneously.

Other words contribute variously to location of a text. Some words are less patently locational than named places, but contribute incidentally, such as the name of some establishment, some characteristic feature of a location, some event which takes place in some location, or some other topic the discussion of which is more typical in one location than in another. We will estimate the *placeness* of words in these experiments.

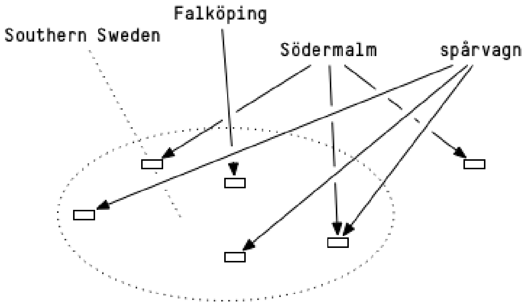


Figure 1: Some terms are polylocational

3 Mapping from a continuous to a discrete representation

We, as has been done in previous experiments, collect the geographic distribution of word usage through collecting microblog posts, some of which have longitude and latitude, from Twitter. Posts with location information are distributed over a map in what amounts to a continuous representation. The words from posts can be collected and associated with the positions they have been observed in.

First experiments which use similar training data to ours have typically assigned the posts and thus the words they occur in directly to some representation of locations - a word which occurs in tweets at $[N59.35, E18.11]$ and $[N59.31, E18.05]$ will have both observations recorded to be in the same city (Cheng et al. 2010, Mahmud et al. 2012). An alternative and later approach by e.g. Priedhorsky et al. (2014) is to aggregate all observations of a word over a map and assign a named location to the distribution, rather than to each observation, deferring the labeling to a point in the analysis where more understanding of the term distribution is known.

Another approach is to model *topics* as inferred from vocabulary usage in text across their geographical distribution, and then, for each text, to assess the topic and thus its attendant location visavi the topic model most likely to have generated the text in question (Eisenstein et al. 2010, Yin et al. 2011, Kinsella et al. 2011, Hong et al. 2012). We have found that topic models as implemented are computationally demanding, do not add accuracy to prediction, and have little explanatory value to aid the understanding of localised language use.

In these experiments we will compare using a list of known places with a model where we aggregate the locational information provided by words (and potentially other linguistic items such as constructions) trained on longitude and latitude either by letting the words vote for place or by averaging the information on a word-by-word basis. The latter model defers the mapping to place until some analysis has been performed; the former assigns place to the words earlier in the process.

4 Test Data

These experiments have focused on Swedish-language material and on Swedish locations. Most Swedish-speakers live in Sweden; Swedish is mainly written and spoken in Sweden and in Finland. Sweden is a roughly rectangular country of about 450 000 km^2 as shown in Figure 2. Sweden has since 1634 been organised into 22 counties or *län* of between 3 000 km^2 and 100 000 km^2. The median size of a county is 10 545 km^2 which would, assuming quadratic counties, give a side of 100 *km* for a typical county.

We measure accuracy of textual location using the *Haversine distance*, the great-circle distance between two points on a sphere. We report averages, both mean and median, as well as percentage of texts we have located within 100 *km* from their known position.

Our test data set is composed of social media texts. Firstly, 18 GB of blog text from major Swedish blog and forum sites, with self-reported location by author - variously, home town, municipality, village, or county. The texts are mainly personal texts with authors of all ages but with a preponderance of pre-teens to young adults. The data are from 2001 and onward, with more data from the latest years. The data are concatenated into one document per blog, totalling to 154 062 documents from unique sources. Somewhat more than a third, 35%, have more than 10k characters.

Secondly, 37 GB of blog text without any explicit indication of location. A target task for these experiments is to enrich these 37 GB of non-located data with predicted location, in order to address data sparsity for unusual dialectal linguistic items.

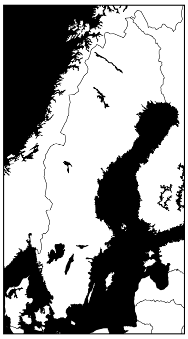


Figure 2: Map of Sweden

5 Baseline: the GAZETTEER model

For a list of known places we used a list[1] of 1 956 Swedish cities and 2 920 towns and villages as defined by Statistics Sweden[2] in 2010.

As the most obvious baseline, we identify all tokens found in the gazetteer. Each such token is converted to a position through the Geoencoding API offered by Google[3]. The position with largest observed frequency of occurrence in the text is assumed to be the position of the text. Other approaches have taken this as a useful approach for identifying features such as Places of Interest mentioned in texts (Li et al. 2014). We call this approach the GAZETTEER approach.

6 Training Data

As a basis for learning how words were used we used geotagged microblog data from Twitter. About 2% of Swedish Twitter posts have latitude and longitude explicitly given,[4] typically those that have been posted from a mobile phone. We gathered data from Twitter's streaming API[5] during the months of May to August of 2014, saving posts with latitude and longitude and with Sweden explicitly given as point of origin. This gave us 4 429 516 posts of about 630 MB.

7 Polylocational Gaussian Mixture Models

Given a set of geographically located texts, we record for each linguistic item – meaning word, in these experiments – the locations from the metadata of every text it occurs in. This gives each word a mapped geographic distribution of latitude-longitude pairs. We model these observed distributions using Gaussian 2-D functions, as defined by Priedhorsky et al. (2014). A 2-D Gaussian function will assume a peak at some position and allow for a graceful inclusion of hits at nearby positions into the model in a bell-like distribution.

In contrast to the original definition and and other similar following approaches, we want to be able to handle polylocational words. After testing various models on a subset of our data we find that fitting more than one Gaussian function—in effect, assuming that locationally interesting words refer to several locations–yields better results than fitting all locational data into one distribution. After some initial parameter exploration as shown in Figure 3, we settle on three Gaussian functions as a reasonable model: words with more than three distributional peaks are likely to be of less utility for locating texts. We consequently fit each word with three Gaussian functions to allow a word to contribute to many locations for the texts it is observed in.

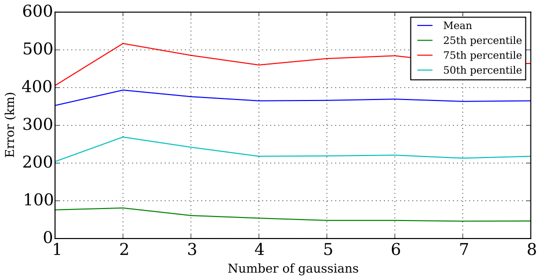


Figure 3: Effect of allowing polylocational representations

8 The notion of placeness

In keeping with previous research on geolocational terms such as Han et al. (2014), we rank candidate words for their locational specificity. From the Gaussian Mixture Model representation, we take the log probability ρ in the mean of the Gaussian and transform it into a *placeness* score by $p = e^{\frac{100}{-\rho}}$. This is done for every word, for all three Gaussians. The score is then used to rank words for locational utility.

		Gaussian		
		1st	2nd	3d
Falköping		58	9	9
Stockholm		37	10	10
spårvagn	*"tram"*	36	18	15
och	*"and"*	16	15	9

Table 1: Example words and their log placeness

Table 1 shows the placeness of the three Gaussians for some sample words. The two sample named locations have high placeness for their first Gaussians,

[1] http://en.wikipedia.org/wiki/List_of_urban_areas_in_Sweden One named location ("När") was removed from the list since it is homographic to the adverbials corrresponding to the English *near* and *when*, causing a disproportionate amount of noise.

[2] *A locality consists of a group of buildings normally not more than 200 metres apart from each other, and must fulfil a minimum criterion of having at least 200 inhabitants. Delimitation of localities is made by Statistics Sweden every five years.* [http://www.scb.se]

[3] https://developers.google.com/.../geocoding/

[4] Determined by listening to Twitter's streaming API for about a day.

[5] The "garden hose": https://dev.twitter.com/streaming/public

indicating that they have locational utility. “Stockholm”, the capital city, which is frequently mentioned in conversations elsewhere has less placeness than has “Falköping”, a smaller city. The word “tram” has lower placeness than the two cities, and the word “and” with a log placeness score of 16 can not be considered locational at all. Inspecting the resulting list as given in Table 2 which shows some examples from the top of the list, we find that words with high placeness frequently are non-gazetteer locations (“Slottsskogen”), user names, hash tags – frequently referring to events (“#lundakarneval”), and other local terms, most typically street names (“Holgersgatan”), spelling variants (“Ståckhålm”), or public establishments.

The performance of the predictive models introduced below can be improved by excluding words with low placeness from the centroid. This exclusion threshold is referred to as T below.

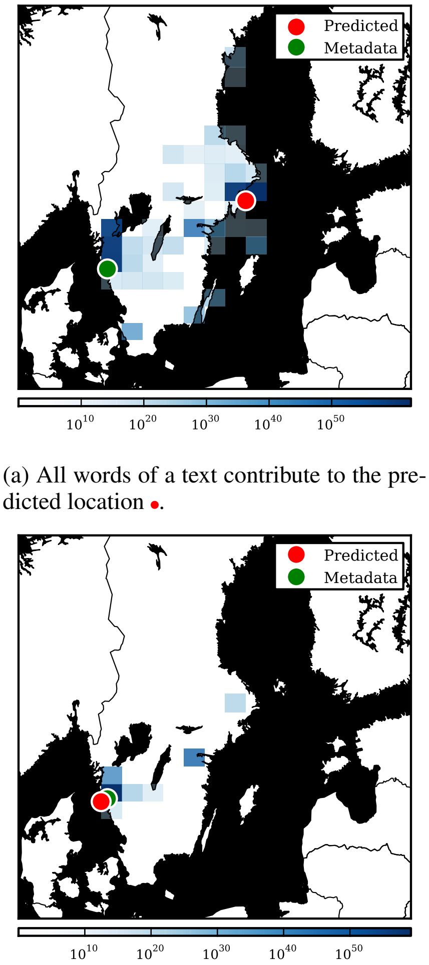
known places	hash tags	other
hogstorp	#lundakarneval	holgersgatan
nyhammar	#bishopsarms	margretegärdeparken
sjuntorp	#gothenburg	uddevallahus
tyringe	#westpride14	kampenhof
slottsskogen	#swedenlove1dday	ståckhålm
storvik	#sverigemotet	gullmarsplan
charlottenberg	#sthlmtech	tvärbanan

Table 2: Example words with high placeness

9 Experimental settings: the TOTAL and FILTERED models

We run one experimental setting with all words of a set, only filtered for placeness. We call this approach the TOTAL approach.

As a more informed model, we filter the words in the feature set to find the most locationally appropriate terms, in order to reduce noise and computational effort, but above all, in keeping with our hypothesis that the locational signal is present in only part of the texts. Backstrom et al. (2008) and following them, Cheng et al. (2010), using similar data as we do, also limit their analyses to “local” rather than “non-local” words in the text matter they process, modeling word locality through observed occurrences, modulated with some geographical smoothing. To find the most appropriate localised linguistic items, we bootstrap from the gazetteer and collect the most distinctive distributional contexts of gazetteer terms. For this, we used context windows of six words before $(6+0)$, around $(3+3)$, and after $(0+6)$ each target word. These context windows were tabulated and the most frequently occurring constructions[6] are then ranked based on their ability to return words with high placeness. For each construction, the percentage of words returned with $\log T > 20$ is used as a ranking criterion. Using this ranking, the top 150 constructions are retained as a paradigmatic filter to generate usefully locational words. Constructions such as `lives in <location>` will be at the top of the list. Examples are given in Figure 8.



(a) All words of a text contribute to the predicted location ●.

(b) Only words filtered through the distributional model contribute votes to yield a prediction ● very close to the correct position ●.

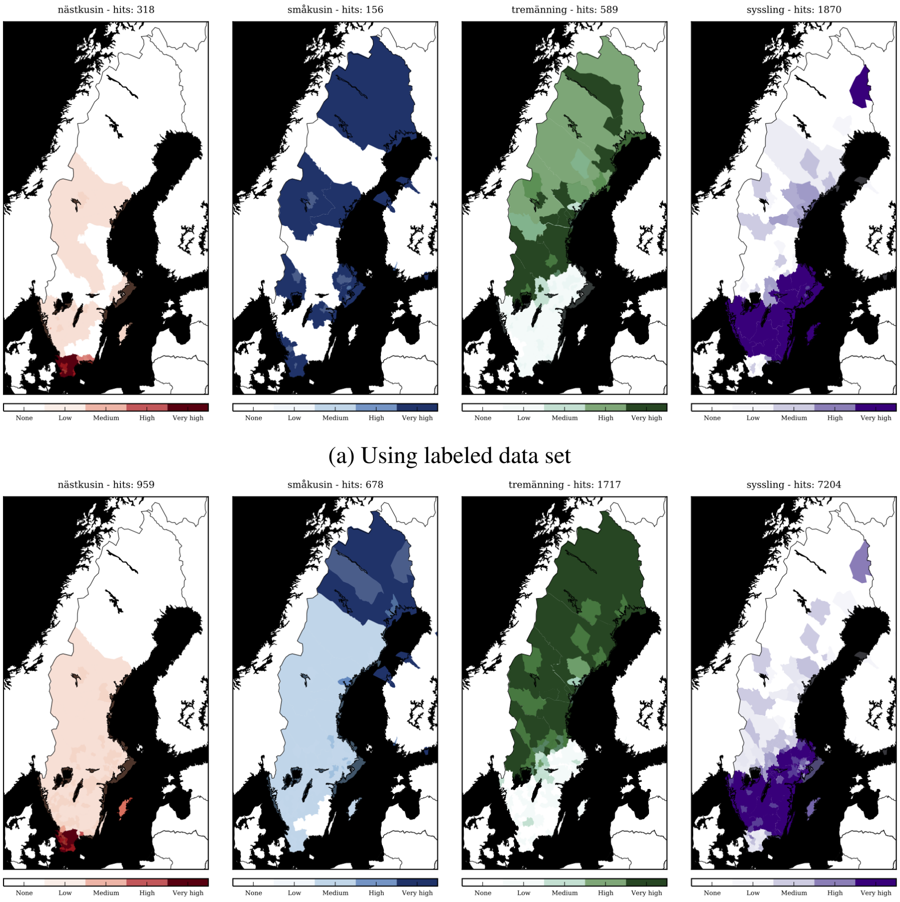
Figure 4: Comparison illustrating the grid and showing how the grammar transforms the result.

Words found in the `<location>` slot of the constructions are frequency filtered with respect to N, the length of the text under analysis, with thresholds set by experimentation to $0.00008 \times N \leq f_{wd} \leq N/300$. This reduces the number of Gaussian models to evaluate drastically. Each text under consideration was then filtered to only include words found through the above procedure, reducing the size of the texts to about 6% of the original.

10 Aggregating the locational information for filtered texts

The filtered texts are now processed in two different ways. Every unique word token in the Twitter dataset has a Gaussian mixture model i based on its observed

[6] In these experiments, the 900 most frequent constructions are used.



(a) Using labeled data set

(b) Using enriched data set increases the data

Figure 5: Regional terminology for "second cousin"

```
<location> mellan
varit i <location>
bor i <location>
var i <location>
vi till <location>
in till <location>
ska till <location>
<location> centrum
av till <location>
det av till <location>
hemma i <location>
till <location>
upp till <location>
```

(a) In Swedish

```
<location> between
been in <location>
live(s) in <location>
was in <location>
we to <location>
in to <location>
going to <location>
<location> centre
off to <location>
go to <location>
home in <location>
to <location>
up to <location>
```

(b) Translated to English

Figure 8: Examples of locational constructions

occurrences, as shown in Section 8. This is represented by the three mean coordinates $\overline{\mu}^i$ and their corresponding *placenesses* $\overline{p}^i$.

$$\overline{\mu}^i = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}^i \qquad \overline{p}^i = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}^i$$

We compute a centroid for these coordinates, as an average best guess for geographic signal for a text. We do this with an arithmetic weighted mean. Given n words:

$$M = \frac{\sum\limits_{i=1}^{n} \overline{\mu}^n \cdot \overline{p}^n}{\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{3} p_j^n}$$

Where $\overline{\mu}^n \cdot \overline{p}^n$ is the dot product[7]. We call this model FILTERED CENTROID

Alternatively, we do not average the coordinates, but select by weighted majority vote. We divide Sweden into a grid of roughly 50x50km cells. The placeness score of every locational word in a text is added to its cell. The centerpoint of the cell with highest score is assigned to the text as a location. We call this model FILTERED VOTE.

Figure 4 shows how filtering improves results, here illustrated by the FILTERED VOTE model. The top map shows how every word of a text contributes votes, weighted by their placeness, to give a prediction (●). The bottom map shows how when only words filtered through the distributional model are used, the voting yields a correct result in comparison with the gold standard (●) given by the metadata.

[7] $\overline{\mu}^i \cdot \overline{p}^i = \mu_1^i p_1^i + \mu_2^i p_2^i + \mu_3^i p_3^i$ for this specific case.

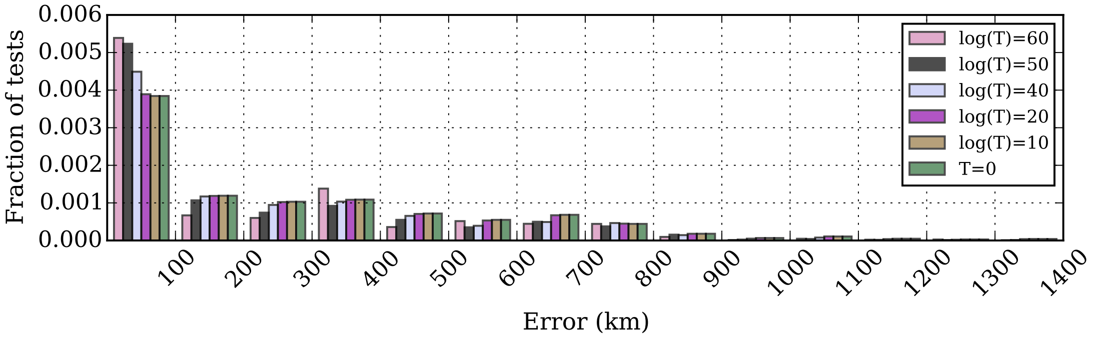


Figure 6: Comparing placeness thresholds for the FILTERED CENTROID model.

		Placeness	Error (km)		Percentile (km)			$e < 100\ km$	
		$\log T$	$\tilde{e}$	$\bar{e}$	25 %	50 %	75 %	Precision	Recall
FILTERED CENTROID	▮┅	—	204	365	45	204	464	0.38	0.38
FILTERED CENTROID	▮┅	10	204	365	45	204	464	0.38	0.38
FILTERED CENTROID	▮┅	20	200	365	44	200	460	0.38	0.38
FILTERED CENTROID	▮┅	40	145	333	32	145	396	0.44	0.32
FILTERED CENTROID	▮┅	50	90	286	22	90	321	0.52	0.23
FILTERED CENTROID	▮┅	60	70	271	13	70	330	0.53	0.04

Table 3: Comparing placeness thresholds for the FILTERED CENTROID model.

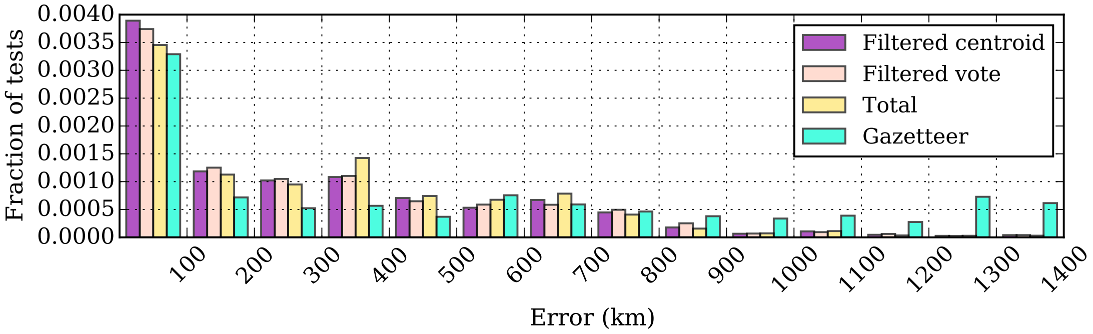


Figure 7: Comparing models with placeness threshold at $\log T = 20$.

		Placeness	Error (km)		Percentile (km)			$e < 100\ km$	
		$\log T$	$\tilde{e}$	$\bar{e}$	25 %	50 %	75 %	Precision	Recall
GAZETTEER	▮┅	20	450	626	62	450	964	0.31	0.31
TOTAL	▮┅	20	256	380	51	256	516	0.34	0.34
FILTERED CENTROID	▮┅	20	200	365	44	200	460	0.38	0.38
FILTERED VOTE	▮┅	20	208	377	58	208	467	0.37	0.36

Table 4: Comparing models: $\tilde{e}$ is the median error and $\bar{e}$ is the mean error in km.

11 Results

As shown in Table 4 and Figure 7, the Gaussian models FILTERED CENTROID and FILTERED VOTE outperform the GAZETTEER model handily. Filtering words distributionally, in addition to reducing processing, improves results further. The FILTERED CENTROID model is slightly better than the FILTERED VOTE model, providing support for late discretization of locational information. A closer look at the effect, shown in Table 3 and in Figure 6, of feature selection with the placeness threshold shows the precision-recall tradeoff contingent on reducing the number of accepted locational words.

These results are well comparable with the results reported by others: while direct comparison with other linguistic and geographic areas is difficult, Cheng et al. (2010) set a 100-mile ($\approx$ 160 km) success criterion for a similar task of geo-locating microblog authors (not single posts). They find that about 10% of microblog users can be localised within their 100-mile radius. Eisenstein et al. (2010) found they could on average achieve a 900 km accuracy for texts or a 24% accuracy on a US state level.

12 Regional variation

Returning to our use case we now use the FILTERED CENTROID model to position and thus enrich a further 38% of our unlabeled blog collection with a location tag (setting the placeness threshold $\log T = 20$). This gives a noticeably better resolution for studying regional word usage as shown in Figure 5: the term for “second cousin” varies across dialects, and given the enriched data set we are able to gain better frequencies and a more distinct image of usage.

13 Conclusions

We find that

- modelling geographical distribution of linguistic items with multiple (in this case, three) peaks proved useful;
- filtering locationally indicative linguistic items using distributional constructions proved useful;
- modelling the placeness of locational linguistic items for thresholding proved useful;
- training a locational model on positionally annotated microblog posts was a useful bootstrap for assigning location to texts of an entirely different genre;
- we are able to detect and explore regional variation in terminological usage.

Acknowlegdments

This work was in part supported by the grant SINUS (Spridning av innovationer i nutida svenska) from Vetenskapsrådet, the Swedish Research Council.

References

Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on the WWW*, ACM, 2008.

Zhiyyan Cheng, James Caverlee, and Kyumin Lee. 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* (CIKM) ACM, 2010.

Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (EMNLP). ACL, 2010.

Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research (JAIR)*, 49. 2014.

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

Sheila Kinsella, Vanessa Murdock, and Neil O’Hare. I’m eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 2011.

Guoliang Li, Jun Hu, Jianhua Feng, and Kian-lee Tan. Effective location identification from microblogs. In *30th IEEE International Conference on Data Engineering (ICDE)*. IEEE, 2014.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? Inferring home locations of Twitter users. In *Proceedings of the 6th International AAAI Conference on Web and Social Media*, 2012.

Mikael Parkvall. Här går gränsen. *Språktidningen*, October 2012. ISSN 1654-5028.

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

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on the WWW*, ACM, 2011.