A Latent Variable Model for Geographic Lexical Variation

Jacob Eisenstein   Brendan O'Connor   Noah A. Smith   Eric P. Xing
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
`{jacobeis,brendano,nasmith,epxing}@cs.cmu.edu`

Abstract

The rapid growth of geotagged social media raises new computational possibilities for investigating geographic linguistic variation. In this paper, we present a multi-level generative model that reasons jointly about latent topics and geographical regions. High-level topics such as "sports" or "entertainment" are rendered differently in each geographic region, revealing topic-specific regional distinctions. Applied to a new dataset of geotagged microblogs, our model recovers coherent topics and their regional variants, while identifying geographic areas of linguistic consistency. The model also enables prediction of an author's geographic location from raw text, outperforming both text regression and supervised topic models.

1   Introduction

Sociolinguistics and dialectology study how language varies across social and regional contexts. Quantitative research in these fields generally proceeds by counting the frequency of a handful of previously-identified linguistic *variables*: pairs of phonological, lexical, or morphosyntactic features that are semantically equivalent, but whose frequency depends on social, geographical, or other factors (Paolillo, 2002; Chambers, 2009). It is left to the experimenter to determine which variables will be considered, and there is no obvious procedure for drawing inferences from the distribution of multiple variables. In this paper, we present a method for identifying geographically-aligned lexical variation directly from raw text. Our approach takes the form of a probabilistic graphical model capable of identifying both geographically-salient terms and coherent linguistic communities.

One challenge in the study of lexical variation is that term frequencies are influenced by a variety of factors, such as the topic of discourse. We address this issue by adding latent variables that allow us to model topical variation explicitly. We hypothesize that geography and topic interact, as "pure" topical lexical distributions are corrupted by geographical factors; for example, a sports-related topic will be rendered differently in New York and California. Each author is imbued with a latent "region" indicator, which both selects the regional variant of each topic, and generates the author's observed geographical location. The regional corruption of topics is modeled through a cascade of logistic normal priors—a general modeling approach which we call *cascading* topic models. The resulting system has multiple capabilities, including: (i) analyzing lexical variation by both topic and geography; (ii) segmenting geographical space into coherent linguistic communities; (iii) predicting author location based on text alone.

This research is only possible due to the rapid growth of social media. Our dataset is derived from the microblogging website Twitter,[1] which permits users to post short messages to the public. Many users of Twitter also supply exact geographical coordinates from GPS-enabled devices (e.g., mobile phones),[2] yielding *geotagged* text data. Text in computer-mediated communication is often more vernacular (Tagliamonte and Denis, 2008), and as such it is more likely to reveal the influence of geographic factors than text written in a more formal genre, such as news text (Labov, 1966).

We evaluate our approach both qualitatively and quantitatively. We investigate the topics and regions

---

[1] `http://www.twitter.com`
[2] User profiles also contain self-reported location names, but we do not use that information in this work.

that the model obtains, showing both common-sense results (place names and sports teams are grouped appropriately), as well as less-obvious insights about slang. Quantitatively, we apply our model to predict the location of unlabeled authors, using text alone. On this task, our model outperforms several alternatives, including both discriminative text regression and related latent-variable approaches.

2 Data

The main dataset in this research is gathered from the microblog website Twitter, via its official API. We use an archive of messages collected over the first week of March 2010 from the "Gardenhose" sample stream,[3] which then consisted of 15% of all public messages, totaling millions per day. We aggressively filter this stream, using only messages that are tagged with physical (latitude, longitude) coordinate pairs from a mobile client, and whose authors wrote at least 20 messages over this period. We also filter to include only authors who follow fewer than 1,000 other people, and have fewer than 1,000 followers. Kwak et al. (2010) find dramatic shifts in behavior among users with social graph connectivity outside of that range; such users may be marketers, celebrities with professional publicists, news media sources, etc. We also remove messages containing URLs to eliminate bots posting information such as advertising or weather conditions. For interpretability, we restrict our attention to authors inside a bounding box around the contiguous U.S. states, yielding a final sample of about 9,500 users and 380,000 messages, totaling 4.7 million word tokens. We have made this dataset available online.[4]

Informal text from mobile phones is challenging to tokenize; we adapt a publicly available tokenizer[5] originally developed for Twitter (O'Connor et al., 2010), which preserves emoticons and blocks of punctuation and other symbols as tokens. For each user's Twitter feed, we combine all messages into a single "document." We remove word types that appear in fewer than 40 feeds, yielding a vocabulary of 5,216 words. Of these, 1,332 do not appear in the English, French, or Spanish dictionaries of the

spell-checking program `aspell`.

Every message is tagged with a location, but most messages from a single individual tend to come from nearby locations (as they go about their day); for modeling purposes we use only a single geographic location for each author, simply taking the location of the first message in the sample.

The authors in our dataset are fairly heavy Twitter users, posting an average of 40 messages per day (although we see only 15% of this total). We have little information about their demographics, though from the text it seems likely that this user set skews towards teens and young adults. The dataset covers each of the 48 contiguous United States and the District of Columbia.

3 Model

We develop a model that incorporates two sources of lexical variation: topic and geographical region. We treat the text and geographic locations as outputs from a generative process that incorporates both topics and regions as latent variables.[6] During inference, we seek to recover the topics and regions that best explain the observed data.

At the base level of model are "pure" topics (such as "**sports**", "**weather**", or "**slang**"); these topics are rendered differently in each region. We call this general modeling approach a *cascading* topic model; we describe it first in general terms before moving to the specific application to geographical variation.

3.1 Cascading Topic Models

Cascading topic models generate text from a chain of random variables. Each element in the chain defines a distribution over words, and acts as the mean of the distribution over the subsequent element in the chain. Thus, each element in the chain can be thought of as introducing some additional corruption. All words are drawn from the final distribution in the chain.

At the beginning of the chain are the priors, followed by unadulterated base topics, which may then be corrupted by other factors (such as geography or time). For example, consider a base "**food**" topic

---

that emphasizes words like *dinner* and *delicious*; the corrupted "**food-California**" topic would place weight on these words, but might place extra emphasis on other words like *sprouts*.

The path through the cascade is determined by a set of indexing variables, which may be hidden or observed. As in standard latent Dirichlet allocation (Blei et al., 2003), the base topics are selected by a per-token hidden variable z. In the geographical topic model, the next level corresponds to regions, which are selected by a per-author latent variable r.

Formally, we draw each level of the cascade from a normal distribution centered on the previous level; the final multinomial distribution over words is obtained by exponentiating and normalizing. To ensure tractable inference, we assume that all covariance matrices are uniform diagonal, i.e., $a\mathbf{I}$ with $a > 0$; this means we do not model interactions between words.

3.2 The Geographic Topic Model

The application of cascading topic models to geographical variation is straightforward. Each document corresponds to the entire Twitter feed of a given author during the time period covered by our corpus. For each author, the latent variable r corresponds to the geographical region of the author, which is not observed. As described above, r selects a corrupted version of each topic: the kth basic topic has mean $\boldsymbol{\mu}_k$, with uniform diagonal covariance σ_k^2; for region j, we can draw the regionally-corrupted topic from the normal distribution, $\boldsymbol{\eta}_{jk} \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2\mathbf{I})$.

Because $\boldsymbol{\eta}$ is normally-distributed, it lies not in the simplex but in $\mathbb{R}^W$. We deterministically compute multinomial parameters $\boldsymbol{\beta}$ by exponentiating and normalizing: $\boldsymbol{\beta}_{jk} = \exp(\boldsymbol{\eta}_{jk})/\sum_i \exp(\eta_{jk}^{(i)})$. This normalization could introduce identifiability problems, as there are multiple settings for $\boldsymbol{\eta}$ that maximize $P(\boldsymbol{w}|\boldsymbol{\eta})$ (Blei and Lafferty, 2006a). However, this difficulty is obviated by the priors: given $\boldsymbol{\mu}$ and σ^2, there is only a single $\boldsymbol{\eta}$ that maximizes $P(\boldsymbol{w}|\boldsymbol{\eta})P(\boldsymbol{\eta}|\boldsymbol{\mu},\sigma^2)$; similarly, only a single $\boldsymbol{\mu}$ maximizes $P(\boldsymbol{\eta}|\boldsymbol{\mu})P(\boldsymbol{\mu}|\boldsymbol{a},b^2)$.

The observed latitude and longitude, denoted $\boldsymbol{y}$, are normally distributed and conditioned on the region, with mean $\boldsymbol{\nu}_r$ and precision matrix Λ_r indexed by the region r. The region index r is itself drawn

from a single shared multinomial $\boldsymbol{\vartheta}$. The model is shown as a plate diagram in Figure 1.

Given a vocabulary size W, the generative story is as follows:

- **Generate base topics**: for each topic $k < K$
  - Draw the base topic from a normal distribution with uniform diagonal covariance: $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{a}, b^2\mathbf{I})$,
  - Draw the regional variance from a Gamma distribution: $\sigma_k^2 \sim \mathcal{G}(c,d)$.
  - **Generate regional variants**: for each region $j < J$,
    * Draw the region-topic $\boldsymbol{\eta}_{jk}$ from a normal distribution with uniform diagonal covariance: $\boldsymbol{\eta}_{jk} \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2\mathbf{I})$.
    * Convert $\boldsymbol{\eta}_{jk}$ into a multinomial distribution over words by exponentiating and normalizing: $\boldsymbol{\beta}_{jk} = \exp(\boldsymbol{\eta}_{jk})/\sum_i^W \exp(\eta_{jk}^{(i)})$, where the denominator sums over the vocabulary.

- **Generate regions**: for each region $j < J$,
  - Draw the spatial mean $\boldsymbol{\nu}_j$ from a normal distribution.
  - Draw the precision matrix Λ_j from a Wishart distribution.

- Draw the distribution over regions $\boldsymbol{\vartheta}$ from a symmetric Dirichlet prior, $\boldsymbol{\vartheta} \sim \text{Dir}(\alpha_\vartheta \mathbf{1})$.

- **Generate text and locations**: for each document d,
  - Draw topic proportions from a symmetric Dirichlet prior, $\boldsymbol{\theta} \sim \text{Dir}(\alpha\mathbf{1})$.
  - Draw the region r from the multinomial distribution $\boldsymbol{\vartheta}$.
  - Draw the location $\boldsymbol{y}$ from the bivariate Gaussian, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\nu}_r, \Lambda_r)$.
  - For each word token,
    * Draw the topic indicator $z \sim \boldsymbol{\theta}$.
    * Draw the word token $w \sim \boldsymbol{\beta}_{rz}$.

4 Inference

We apply mean-field variational inference: a fully-factored variational distribution Q is chosen to minimize the Kullback-Leibler divergence from the true distribution. Mean-field variational inference with conjugate priors is described in detail elsewhere (Bishop, 2006; Wainwright and Jordan, 2008); we restrict our focus to the issues that are unique to the geographic topic model.
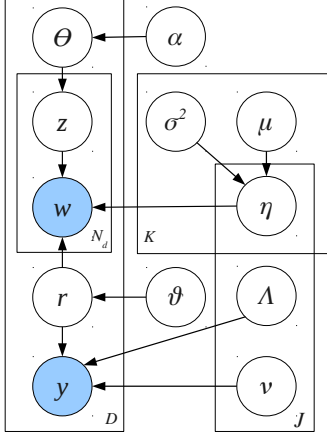
Symbol	Description
$\boldsymbol{\mu}_k$	log of base topic k's distribution over word types
σ_k^2	variance parameter for regional variants of topic k
$\boldsymbol{\eta}_{jk}$	region j's variant of base topic $\boldsymbol{\mu}_k$
$\boldsymbol{\theta}_d$	author d's topic proportions
r_d	author d's latent region
$\boldsymbol{y}_d$	author d's observed GPS location
$\boldsymbol{\nu}_j$	region j's spatial center
Λ_j	region j's spatial precision
z_n	token n's topic assignment
w_n	token n's observed word type
$\boldsymbol{\alpha}$	global prior over author-topic proportions
$\boldsymbol{\vartheta}$	global prior over region classes

Figure 1: Plate diagram for the geographic topic model, with a table of all random variables. Priors (besides α) are omitted for clarity, and the document indices on z and w are implicit.

We place variational distributions over all latent variables of interest: $\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{r}, \boldsymbol{\vartheta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{\nu}$, and Λ, updating each of these distributions in turn, until convergence. The variational distributions over $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ are Dirichlet, and have closed form updates: each can be set to the sum of the expected counts, plus a term from the prior (Blei et al., 2003). The variational distributions $q(z)$ and $q(r)$ are categorical, and can be set proportional to the expected joint likelihood—to set $q(z)$ we marginalize over r, and vice versa.[7] The updates for the multivariate Gaussian spatial parameters $\boldsymbol{\nu}$ and Λ are described by Penny (2001).

4.1 Regional Word Distributions

The variational region-topic distribution $\boldsymbol{\eta}_{jk}$ is normal, with uniform diagonal covariance for tractability. Throughout we will write $\langle x \rangle$ to indicate the expectation of x under the variational distribution Q. Thus, the vector mean of the distribution $q(\boldsymbol{\eta}_{jk})$ is written $\langle \boldsymbol{\eta}_{jk} \rangle$, while the variance (uniform across i) of $q(\boldsymbol{\eta})$ is written $\mathcal{V}(\boldsymbol{\eta}_{jk})$.

To update the mean parameter $\langle \boldsymbol{\eta}_{jk} \rangle$, we maximize the contribution to the variational bound L from the relevant terms:

$$L_{[\langle \eta_{jk}^{(i)} \rangle]} = \langle \log p(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{r}) \rangle + \langle \log p(\eta_{jk}^{(i)}|\mu_k^{(i)}, \sigma_k^2) \rangle, \tag{1}$$

---

[7]Thanks to the naïve mean field assumption, we can marginalize over $\boldsymbol{z}$ by first decomposing across all N_d words and then summing over $q(z)$.

with the first term representing the likelihood of the observed words (recall that β is computed deterministically from η) and the second term corresponding to the prior. The likelihood term requires the expectation $\langle \log \boldsymbol{\beta} \rangle$, but this is somewhat complicated by the normalizer $\sum_i^W \exp(\eta^{(i)})$, which sums over all terms in the vocabulary. As in previous work on logistic normal topic models, we use a Taylor approximation for this term (Blei and Lafferty, 2006a).

The prior on $\boldsymbol{\eta}$ is normal, so the contribution from the second term of the objective (Equation 1) is $-\frac{1}{2\langle \sigma_k^2 \rangle} \langle (\eta_{jk}^{(i)} - \mu_k^{(i)})^2 \rangle$. We introduce the following notation for expected counts: $N(i, j, k)$ indicates the expected count of term i in region j and topic k, and $N(j, k) = \sum_i N(i, j, k)$. After some calculus, we can write the gradient $\partial L / \partial \langle \eta_{jk}^{((i))} \rangle$ as

$$N(i, j, k) - N(j, k)\langle \beta_{jk}^{(i)} \rangle - \langle \sigma_k^{-2} \rangle (\langle \eta_{jk}^{(i)} \rangle - \langle \mu_k^{(i)} \rangle), \tag{2}$$

which has an intuitive interpretation. The first two terms represent the difference in expected counts for term i under the variational distributions $q(z, r)$ and $q(z, r, \beta)$: this difference goes to zero when $\beta_{jk}^{(i)}$ perfectly matches $N(i, j, k)/N(j, k)$. The third term penalizes $\eta_{jk}^{(i)}$ for deviating from its prior $\mu_k^{(i)}$, but this penalty is proportional to the expected inverse variance $\langle \sigma_k^{-2} \rangle$. We apply gradient ascent to maximize the objective L. A similar set of calculations gives the gradient for the variance of $\boldsymbol{\eta}$; these are described in an forthcoming appendix.

4.2 Base Topics

The base topic parameters are $\boldsymbol{\mu}_k$ and σ_k^2; in the variational distribution, $q(\boldsymbol{\mu}_k)$ is normally distributed and $q(\sigma_k^2)$ is Gamma distributed. Note that $\boldsymbol{\mu}_k$ and σ_k^2 affect only the regional word distributions $\boldsymbol{\eta}_{jk}$. An advantage of the logistic normal is that the variational parameters over $\boldsymbol{\mu}_k$ are available in closed form,

$$\langle \mu_k^{(i)} \rangle = \frac{b^2 \sum_j^J \langle \eta_{jk}^{(i)} \rangle + \langle \sigma_k^2 \rangle a^{(i)}}{b^2 J + \langle \sigma_k^2 \rangle}$$
$$\mathcal{V}(\boldsymbol{\mu}_k) = (b^{-2} + J\langle \sigma_k^{-2} \rangle)^{-1},$$

where J indicates the number of regions. The expectation of the base topic μ incorporates the prior and the average of the generated region-topics—these two components are weighted respectively by the expected variance of the region-topics $\langle \sigma_k^2 \rangle$ and the prior topical variance b^2. The posterior variance $\mathcal{V}(\boldsymbol{\mu})$ is a harmonic combination of the prior variance b^2 and the expected variance of the region topics.

The variational distribution over the region-topic variance σ_k^2 has Gamma parameters. These parameters cannot be updated in closed form, so gradient optimization is again required. The derivation of these updates is more involved, and is left for a forthcoming appendix.

5 Implementation

Variational scheduling and initialization are important aspects of any hierarchical generative model, and are often under-discussed. In our implementation, the variational updates are scheduled as follows: given expected counts, we iteratively update the variational parameters on the region-topics $\boldsymbol{\eta}$ and the base topics $\boldsymbol{\mu}$, until convergence. We then update the geographical parameters $\boldsymbol{\nu}$ and Λ, as well as the distribution over regions ϑ. Finally, for each document we iteratively update the variational parameters over $\boldsymbol{\theta}, \boldsymbol{z}$, and r until convergence, obtaining expected counts that are used in the next iteration of updates for the topics and their regional variants. We iterate an outer loop over the entire set of updates until convergence.

We initialize the model in a piecewise fashion. First we train a Dirichlet process mixture model on the locations $\boldsymbol{y}$, using variational inference on the truncated stick-breaking approximation (Blei and Jordan, 2006). This automatically selects the number of regions J, and gives a distribution over each region indicator r_d from geographical information alone. We then run standard latent Dirichlet allocation to obtain estimates of $\boldsymbol{z}$ for each token (ignoring the locations). From this initialization we can compute the first set of expected counts, which are used to obtain initial estimates of all parameters needed to begin variational inference in the full model.

The prior $\boldsymbol{a}$ is the expected mean of each topic $\boldsymbol{\mu}$; for each term i, we set $a^{(i)} = \log N(i) - \log N$, where $N(i)$ is the total count of i in the corpus and $N = \sum_i N(i)$. The variance prior b^2 is set to 1, and the prior on σ^2 is the Gamma distribution $\mathcal{G}(2, 200)$, encouraging minimal deviation from the base topics. The symmetric Dirichlet prior on $\boldsymbol{\theta}$ is set to $\frac{1}{2}$, and the symmetric Dirichlet parameter on ϑ is updated from weak hyperpriors (Minka, 2003). Finally, the geographical model takes priors that are linked to the data: for each region, the mean is very weakly encouraged to be near the overall mean, and the covariance prior is set by the average covariance of clusters obtained by running K-means.

6 Evaluation

For a quantitative evaluation of the estimated relationship between text and geography, we assess our model's ability to predict the geographic location of unlabeled authors based on their text alone.[8] This task may also be practically relevant as a step toward applications for recommending local businesses or social connections. A randomly-chosen 60% of authors are used for training, 20% for development, and the remaining 20% for final evaluation.

6.1 Systems

We compare several approaches for predicting author location; we divide these into latent variable generative models and discriminative approaches.

---

[8]Alternatively, one might evaluate the attributed regional memberships of the words themselves. While the Dictionary of American Regional English (Cassidy and Hall, 1985) attempts a comprehensive list of all regionally-affiliated terms, it is based on interviews conducted from 1965-1970, and the final volume (covering Si–Z) is not yet complete.

6.1.1 Latent Variable Models

Geographic Topic Model This is the full version of our system, as described in this paper. To predict the unseen location $\boldsymbol{y}_d$, we iterate until convergence on the variational updates for the hidden topics $\boldsymbol{z}_d$, the topic proportions $\boldsymbol{\theta}_d$, and the region r_d. From r_d, the location can be estimated as $\hat{\boldsymbol{y}}_d = \arg\max_{\boldsymbol{y}} \sum_j^J p(\boldsymbol{y}|\boldsymbol{\nu}_j, \Lambda_j)q(r_d = j)$. The development set is used to tune the number of topics and to select the best of multiple random initializations.

Mixture of Unigrams A core premise of our approach is that modeling topical variation will improve our ability to understand geographical variation. We test this idea by fixing $K = 1$, running our system with only a single topic. This is equivalent to a Bayesian mixture of unigrams in which each author is assigned a single, regional unigram language model that generates all of his or her text. The development set is used to select the best of multiple random initializations.

Supervised Latent Dirichlet Allocation In a more subtle version of the mixture-of-unigrams model, we model each author as an admixture of regions. Thus, the latent variable attached to each author is no longer an index, but rather a vector on the simplex. This model is equivalent to supervised latent Dirichlet allocation (Blei and McAuliffe, 2007): each topic is associated with equivariant Gaussian distributions over the latitude and longitude, and these topics must explain both the text and the observed geographical locations. For unlabeled authors, we estimate latitude and longitude by estimating the topic proportions and then applying the learned geographical distributions. This is a linear prediction

$$f(\bar{\boldsymbol{z}}_d; \boldsymbol{a}) = (\bar{\boldsymbol{z}}_d^{\mathsf{T}} \boldsymbol{a}^{\text{lat}},\ \bar{\boldsymbol{z}}_d^{\mathsf{T}} \boldsymbol{a}^{\text{lon}})$$

for an author's topic proportions $\bar{\boldsymbol{z}}_d$ and topic-geography weights $\boldsymbol{a} \in \mathbb{R}^{2K}$.

6.1.2 Baseline Approaches

Text Regression We perform linear regression to discriminatively learn the relationship between words and locations. Using term frequency features $\boldsymbol{x}_d$ for each author, we predict locations with word-geography weights $\boldsymbol{a} \in \mathbb{R}^{2W}$:

$$f(\boldsymbol{x}_d; \boldsymbol{a}) = (\boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{a}^{\text{lat}},\ \boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{a}^{\text{lon}})$$

Weights are trained to minimize the sum of squared Euclidean distances, subject to L_1 regularization:

$$\sum_d (\boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{a}^{\text{lat}} - y_d^{\text{lat}})^2 + (\boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{a}^{\text{lon}} - y_d^{\text{lon}})^2$$
$$+ \lambda_{\text{lat}} ||\boldsymbol{a}^{\text{lat}}||_1 + \lambda_{\text{lon}} ||\boldsymbol{a}^{\text{lon}}||_1$$

The minimization problem decouples into two separate latitude and longitude models, which we fit using the `glmnet` elastic net regularized regression package (Friedman et al., 2010), which obtained good results on other text-based prediction tasks (Joshi et al., 2010). Regularization parameters were tuned on the development set. The L_1 penalty outperformed L_2 and mixtures of L_1 and L_2.

Note that for both word-level linear regression here, and the topic-level linear regression in SLDA, the choice of squared Euclidean distance dovetails with our use of spatial Gaussian likelihoods in the geographic topic models, since optimizing $\boldsymbol{a}$ is equivalent to maximum likelihood estimation under the assumption that locations are drawn from equivariant circular Gaussians centered around each $f(\boldsymbol{x}_d; \boldsymbol{a})$ linear prediction. We experimented with decorrelating the location dimensions by projecting $\boldsymbol{y}_d$ into the principal component space, but this did not help text regression.

K-Nearest Neighbors Linear regression is a poor model for the multimodal density of human populations. As an alternative baseline, we applied supervised K-nearest neighbors to predict the location $\boldsymbol{y}_d$ as the average of the positions of the K most similar authors in the training set. We computed term-frequency inverse-document frequency features and applied cosine similarity over their first 30 principal components to find the neighbors. The choices of principal components, IDF weighting, and neighborhood size $K = 20$ were tuned on the development set.

6.2 Metrics

Our principle error metrics are the mean and median distance between the predicted and true location in kilometers.[9] Because the distance error may be difficult to interpret, we also report accuracy of classi-

---

[9]For convenience, model training and prediction use latitude and longitude as an unprojected 2D Euclidean space. However, properly measuring the physical distance between points on the

System	Regression		Classification accuracy (%)	
	Mean Dist. (km)	Median Dist. (km)	Region (4-way)	State (49-way)
Geographic topic model	**900**	**494**	**58**	24
Mixture of unigrams	947	644	53	19
Supervised LDA	1055	728	39	4
Text regression	948	712	41	4
K-nearest neighbors	1077	853	37	2
Mean location	1148	1018		
Most common class			37	**27**

Table 1: Location prediction results; lower scores are better on the regression task, higher scores are better on the classification task. Distances are in kilometers. Mean location and most common class are computed from the test set. Both the geographic topic model and supervised LDA use the best number of topics from the development set (10 and 5, respectively).

fication by state and by region of the United States. Our data includes the 48 contiguous states plus the District of Columbia; the U.S. Census Bureau divides these states into four regions: West, Midwest, Northeast, and South.[10] Note that while major population centers straddle several state lines, most region boundaries are far from the largest cities, resulting in a clearer analysis.

6.3 Results

As shown in Table 1, the geographic topic model achieves the strongest performance on all metrics. All differences in performance between systems are statistically significant ($p < .01$) using the Wilcoxon-Mann-Whitney test for regression error and the χ^2 test for classification accuracy. Figure 2 shows how performance changes as the number of topics varies.

Note that the geographic topic model and the mixture of unigrams use identical code and parametrization – the only difference is that the geographic topic model accounts for topical variation, while the mixture of unigrams sets $K = 1$. These results validate our basic premise that it is important to model the interaction between topical and geographical variation.

Text regression and supervised LDA perform especially poorly on the classification metric. Both methods make predictions that are averaged across
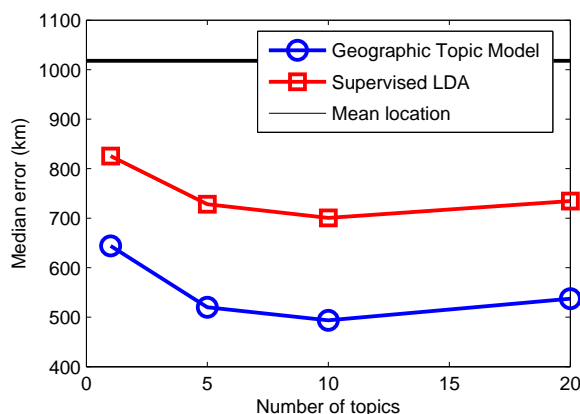

Figure 2: The effect of varying the number of topics on the median regression error (lower is better).

each word in the document: in text regression, each word is directly multiplied by a feature weight; in supervised LDA the word is associated with a latent topic first, and then multiplied by a weight. For these models, all words exert an influence on the predicted location, so uninformative words will draw the prediction towards the center of the map. This yields reasonable distance errors but poor classification accuracy. We had hoped that K-nearest neighbors would be a better fit for this metric, but its performance is poor at all values of K. Of course it is always possible to optimize classification accuracy directly, but such an approach would be incapable of predicting the exact geographical location, which is the focus of our evaluation (given that the desired geographical partition is unknown). Note that the geographic topic model is also not trained to optimize classification accuracy.

Earth's surface requires computing or approximating the great circle distance – we use the Haversine formula (Sinnott, 1984). For the continental U.S., the relationship between degrees and kilometers is nearly linear, but extending the model to a continental scale would require a more sophisticated approach.

[10] http://www.census.gov/geo/www/us_regdiv.pdf

	"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	*ese* exam suttin sippin
N. California	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	*pues* hella koo SAN fckn	hella flirt hut iono OAKLAND
New York	NETS KNICKS	BRONX	iam cab	oww	wasssup nm
Los Angeles	#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af *papi* raining th bomb coo HOLLYWOOD	wyd coo af *nada* tacos messin fasho bomb
Lake Erie	CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	;d blvd BIEBER hve OHIO	foul WIZ salty excuses lames officer lastnight

Table 2: Example base topics (top line) and regional variants. For the base topics, terms are ranked by log-odds compared to the background distribution. The regional variants show words that are strong compared to both the base topic and the background. Foreign-language words are shown in *italics*, while terms that are usually in proper nouns are shown in SMALL CAPS. See Table 3 for definitions of slang terms; see Section 7 for more explanation and details on the methodology.
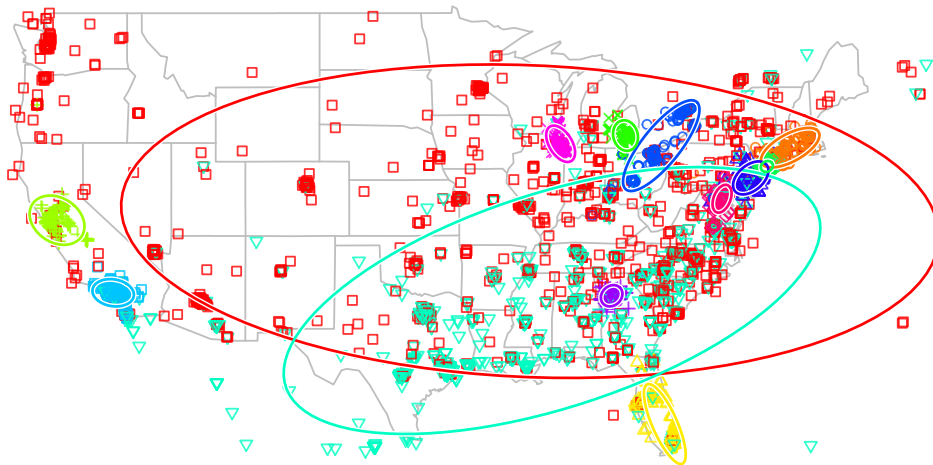

Figure 3: Regional clustering of the training set obtained by one randomly-initialized run of the geographical topic model. Each point represents one author, and each shape/color combination represents the most likely cluster assignment. Ellipses represent the regions' spatial means and covariances. The same model and coloring are shown in Table 2.

7 Analysis

Our model permits analysis of geographical variation in the context of topics that help to clarify the significance of geographically-salient terms. Table 2 shows a subset of the results of one randomly-initialized run, including five hand-chosen topics (of 50 total) and five regions (of 13, as chosen automatically during initialization). Terms were selected by log-odds comparison. For the base topics we show the ten strongest terms in each topic as compared to the background word distribution. For the regional variants, we show terms that are strong both regionally and topically: specifically, we select terms that are in the top 100 compared to both the background distribution and to the base topic. The names for the topics and regions were chosen by the authors.

Nearly all of the terms in column 1 ("**basketball**") refer to sports teams, athletes, and place names—encouragingly, terms tend to appear in the regions where their referents reside. Column 2 contains several proper nouns, mostly referring to popular music figures (including PEARL from the band Pearl Jam).[11] Columns 3–5 are more conversational. Spanish-language terms (*papi, pues, nada, ese*) tend to appear in regions with large Spanish-speaking populations—it is also telling that these terms appear in topics with emoticons and slang abbreviations, which may transcend linguistic barriers. Other terms refer to people or subjects that may be especially relevant in certain regions: *tacos* appears in the southern California region and *cab* in the New York region; TUPAC refers to a rap musician from Los Angeles, and WIZ refers to a rap musician from Pittsburgh, not far from the center of the "Lake Erie" region.

A large number of slang terms are found to have strong regional biases, suggesting that slang may depend on geography more than standard English does. The terms *af* and *hella* display especially strong regional affinities, appearing in the regional variants of multiple topics (see Table 3 for definitions). Northern and Southern California use variant spellings *koo* and *coo* to express the same meaning.

---

[11]This analysis is from an earlier version of our dataset that contained some Twitterbots, including one from a Boston-area radio station. The bots were purged for the evaluation in Section 6, though the numerical results are nearly identical.

term	definition	term	definition
af	as fuck (very)	jk	just kidding
coo	cool	jp	just playing (kidding)
dl	download		
fasho	for sure	koo	cool
gna	going to	lol	laugh out loud
hella	very	nm	nothing much
hr	hour	od	overdone (very)
iam	I am	omw	on my way
ima	I'm going to	smh	shake my head
imm	I'm	suttin	something
iono	I don't know	wassup	what's up
lames	lame (not cool) people	wyd	what are you doing?

Table 3: A glossary of non-standard terms from Table 2. Definitions are obtained by manually inspecting the context in which the terms appear, and by consulting `www.urbandictionary.com`.

While research in perceptual dialectology does confirm the link of *hella* to Northern California (Bucholtz et al., 2007), we caution that our findings are merely suggestive, and a more rigorous analysis must be undertaken before making definitive statements about the regional membership of individual terms. We view the geographic topic model as an exploratory tool that may be used to facilitate such investigations.

Figure 3 shows the regional clustering on the training set obtained by one run of the model. Each point represents an author, and the ellipses represent the bivariate Gaussians for each region. There are nine compact regions for major metropolitan areas, two slightly larger regions that encompass Florida and the area around Lake Erie, and two large regions that partition the country roughly into north and south.

8 Related Work

The relationship between language and geography has been a topic of interest to linguists since the nineteenth century (Johnstone, 2010). An early work of particular relevance is Kurath's (1949) *Word Geography of the Eastern United States*, in which he conducted interviews and then mapped the occurrence of equivalent word pairs such as *stoop* and *porch*. The essence of this approach—identifying variable pairs and measuring their frequencies—remains a dominant methodology in both dialec-

tology (Labov et al., 2006) and sociolinguistics (Tagliamonte, 2006). Within this paradigm, computational techniques are often applied to post hoc analysis: logistic regression (Sankoff et al., 2005) and mixed-effects models (Johnson, 2009) are used to measure the contribution of individual variables, while hierarchical clustering and multidimensional scaling enable aggregated inference across multiple variables (Nerbonne, 2009). However, in all such work it is assumed that the relevant linguistic variables have already been identified—a time-consuming process involving considerable linguistic expertise. We view our work as complementary to this tradition: we work directly from raw text, identifying both the relevant features and coherent linguistic communities.

An active recent literature concerns geotagged information on the web, such as search queries (Backstrom et al., 2008) and tagged images (Crandall et al., 2009). This research identifies the geographic distribution of individual queries and tags, but does not attempt to induce any structural organization of either the text or geographical space, which is the focus of our research. More relevant is the work of Mei et al. (2006), in which the distribution over latent topics in blog posts is conditioned on the geographical location of the author. This is somewhat similar to the supervised LDA model that we consider, but their approach assumes that a partitioning of geographical space into regions is already given.

Methodologically, our cascading topic model is designed to capture multiple dimensions of variability: topics and geography. Mei et al. (2007) include sentiment as a second dimension in a topic model, using a switching variable so that individual word tokens may be selected from either the topic or the sentiment. However, our hypothesis is that individual word tokens reflect both the topic and the geographical aspect. Sharing this intuition, Paul and Girju (2010) build topic-aspect models for the cross product of topics and aspects. They do not impose any regularity across multiple aspects of the same topic, so this approach may not scale when the number of aspects is large (they consider only two aspects). We address this issue using cascading distributions; when the observed data for a given region-topic pair is low, the model falls back to the base topic. The use of cascading logistic normal distri-

butions in topic models follows earlier work on dynamic topic models (Blei and Lafferty, 2006b; Xing, 2005).

9 Conclusion

This paper presents a model that jointly identifies words with high regional affinity, geographically-coherent linguistic regions, and the relationship between regional and topic variation. The key modeling assumption is that regions and topics interact to shape observed lexical frequencies. We validate this assumption on a prediction task in which our model outperforms strong alternatives that do not distinguish regional and topical variation.

We see this work as a first step towards a unsupervised methodology for modeling linguistic variation using raw text. Indeed, in a study of morphosyntactic variation, Szmrecsanyi (2010) finds that by the most generous measure, geographical factors account for only 33% of the observed variation. Our analysis might well improve if non-geographical factors were considered, including age, race, gender, income and whether a location is urban or rural. In some regions, estimates of many of these factors may be obtained by cross-referencing geography with demographic data. We hope to explore this possibility in future work.

References

L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. 2008. Spatial variation in search engine queries. In *Proceedings of WWW*.

C. M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

D. M. Blei and M. I. Jordan. 2006. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144.

D. M. Blei and J. Lafferty. 2006a. Correlated topic models. In *NIPS*.

D. M. Blei and J. Lafferty. 2006b. Dynamic topic models. In *Proceedings of ICML*.

D. M. Blei and J. D. McAuliffe. 2007. Supervised topic models. In *NIPS*.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3:993–1022.

M. Bucholtz, N. Bermudez, V. Fung, L. Edwards, and R. Vargas. 2007. Hella Nor Cal or totally So Cal? the perceptual dialectology of California. *Journal of English Linguistics*, 35(4):325–352.

F. G. Cassidy and J. H. Hall. 1985. *Dictionary of American Regional English*, volume 1. Harvard University Press.

J. Chambers. 2009. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Blackwell.

D. J Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. Mapping the world's photos. In *Proceedings of WWW*, page 761770.

J. Friedman, T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

D. E. Johnson. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1):359–383.

B. Johnstone. 2010. Language and place. In R. Mesthrie and W. Wolfram, editors, *Cambridge Handbook of Sociolinguistics*. Cambridge University Press.

M. Joshi, D. Das, K. Gimpel, and N. A. Smith. 2010. Movie reviews and revenues: An experiment in text regression. In *Proceedings of NAACL-HLT*.

H. Kurath. 1949. *A Word Geography of the Eastern United States*. University of Michigan Press.

H. Kwak, C. Lee, H. Park, and S. Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of WWW*.

W. Labov, S. Ash, and C. Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. Walter de Gruyter.

W. Labov. 1966. *The Social Stratification of English in New York City*. Center for Applied Linguistics.

Q. Mei, C. Liu, H. Su, and C. X Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of WWW*, page 542.

Q. Mei, X. Ling, M. Wondra, H. Su, and C. X. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW*.

T. P. Minka. 2003. Estimating a Dirichlet distribution. Technical report, Massachusetts Institute of Technology.

J. Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1).

B. O'Connor, M. Krieger, and D. Ahn. 2010. TweetMotif: Exploratory search and topic summarization for twitter. In *Proceedings of ICWSM*.

J. C. Paolillo. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. CSLI Publications.

M. Paul and R. Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of AAAI*.

W. D. Penny. 2001. Variational Bayes for d-dimensional Gaussian mixture models. Technical report, University College London.

D. Sankoff, S. A. Tagliamonte, and E. Smith. 2005. Goldvarb X: A variable rule application for Macintosh and Windows. Technical report, Department of Linguistics, University of Toronto.

R. W. Sinnott. 1984. Virtues of the Haversine. *Sky and Telescope*, 68(2).

B. Szmrecsanyi. 2010. Geography is overrated. In S. Hansen, C. Schwarz, P. Stoeckle, and T. Streck, editors, *Dialectological and Folk Dialectological Concepts of Space*. Walter de Gruyter.

S. A. Tagliamonte and D. Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83.

S. A. Tagliamonte. 2006. *Analysing Sociolinguistic Variation*. Cambridge University Press.

M. J. Wainwright and M. I. Jordan. 2008. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers.

E. P. Xing. 2005. On topic evolution. Technical Report 05-115, Center for Automated Learning and Discovery, Carnegie Mellon University.