Discovering Geographical Topics In The Twitter Stream

Liangjie Hong[*] †, Amr Ahmed §, Siva Gurumurthy ¶, Alex Smola §, Kostas Tsioutsiouliklis ¶
† Dept. of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA
§ Yahoo! Research, Sunnyvale, CA, USA
¶ Twitter, San Francisco, CA, USA
lih307@cse.lehigh.edu, {amrahmed,smola}@yahoo-inc.com, {siva,kostas}@twitter.com

ABSTRACT

Micro-blogging services have become indispensable communication tools for online users for disseminating breaking news, eyewitness accounts, individual expression, and protest groups. Recently, Twitter, along with other online social networking services such as Foursquare, Gowalla, Facebook and Yelp, have started supporting location services in their messages, either explicitly, by letting users choose their places, or implicitly, by enabling geo-tagging, which is to associate messages with latitudes and longitudes. This functionality allows researchers to address an exciting set of questions: 1) How is information created and shared across geographical locations, 2) How do spatial and linguistic characteristics of people vary across regions, and 3) How to model human mobility. Although many attempts have been made for tackling these problems, previous methods are either complicated to be implemented or oversimplified that cannot yield reasonable performance.

It is a challenge task to discover topics and identify users' interests from these geo-tagged messages due to the sheer amount of data and diversity of language variations used on these location sharing services. In this paper we focus on Twitter and present an algorithm by modeling diversity in tweets based on topical diversity, geographical diversity, and an interest distribution of the user. Furthermore, we take the Markovian nature of a user's location into account. Our model exploits sparse factorial coding of the attributes, thus allowing us to deal with a large and diverse set of covariates efficiently. Our approach is vital for applications such as user profiling, content recommendation and topic tracking. We show high accuracy in location estimation based on our model. Moreover, the algorithm identifies interesting topics based on location and language.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, Theory

Keywords

Geolocation, Twitter, Topic models, User profiling, Language model, Graphical model, Latent variable inference

1. INTRODUCTION

Micro-blogging services such as Twitter, Tumblr and Weibo, have become very important tools for online users to share breaking news and interesting stories. They are even used for organizing flash mobs and protest groups. For example, Twitter was used extensively in a number of events and emergencies, ranging from elections, earthquakes and tsunamis to playing an instrumental role in facilitating political upheavals in the Middle East.

Key Questions: In addition to its use as a content sharing platform, micro-blogging services like Twitter, along with other location sharing services such as Foursquare, Gowalla, and Facebook Places are nowadays supporting location services. That is, users are able to specify their location in messages, either explicitly, by letting users choose their place, or implicitly, by enabling geo-tagging functionality. This presents an exciting opportunity to answer a range of questions:

1. How is information created and shared in different geographic locations? What is the inherent geographic variability of content?

2. What are the spatial and linguistic characteristics of people? How does this vary across regions?

3. What is a good model for human mobility? Can we discover patterns in users' usage of micro-blogging services.

There exists a considerable body of research addressing these issues [12, 15, 7, 5, 4]. However, the analysis of data still poses a considerable challenge due to its *size* and due to the *integration* of a range of different attributes. To our knowledge this is the first paper to address both scale, location and language modeling in an integrated fashion. That is, we customize the model to be sufficiently sparse to allow for a large scale in terms of users and locations. Furthermore, we design an accurate and scalable inference algorithm.

Our algorithm allows us to discover language patterns and to extract users' interests from geo-tagged messages. We

[*] This work was done when the first author was on an internship at Yahoo! Labs.

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

achieve this thanks to (and despite of) the sheer amount of data and the diversity of language variations used on Twitter. In addition, there are many factors to influence the language used in a tweet with a particular location. For example, words used in a tweet certainly depend on the author and the location where the tweet is written.

A user in New York City might be interested in entirely different matters compared to a user in Beijing. Moreover, the choice of words is clearly influenced by the topic of the tweet. Finally, location specific language will cause the same event to be reported quite differently in different locations (e.g. a soccer game between Brazil and Italy being reported quite differently in those two countries). Thus, different geographical regions have different language variations and topics have different chances of being discussed in these regions.

It turns out that users tend to appear only in a handful of geographic locations [5]. This is useful in improving location accuracy in estimates. The arising challenge is how to best integrate all these strands of information into a single model.

Prior work falls into two groups: Some work only models certain aspects of the problem described above while ignoring the remainder. For instance [17] investigated how location information can be used to better understand patterns in social photo sharing services. A Gaussian mixture model and a probabilistic topic model are combined to learn clusters of locations and latent topics. However, no regional language models are learned and user preferences are also not taken into account. Thus, models developed for such data are usually limited and cannot easily be applied to content-rich social media. Similarly [5] proposed a two component Gaussian mixture model to study the mobility of users in a number of location sharing services. However, their model does not incorporate content at all.

At the other end of the spectrum we find rather complex models, however, without the ability to scale to industrial size. For instance [7] propose a model to predict locations of users in Twitter. Their model has a global topic matrix and each region has different variation of this matrix. However, the inference algorithm is complex. Furthermore, the problem of over-parametrization makes it nontrivial to perform inference accurately. Furthermore, previous models ignore user preferences.

Our Contribution: We propose a model that is both flexible enough to embed all reasonable components of content and geographical locations, as well as user preference modeling. Moreover, it scales to real-world datasets to handle millions of documents and users.

In this paper, we address the problem of modeling geographical topical patterns on Twitter by introducing a novel sparse generative model. It utilizes both statistical topic models and sparse coding techniques to provide a principled method for uncovering different language patterns and common interests shared across the world. Our approach is vital for applications such as user profiling, content recommendation and topic tracking and the method can be easily extended in a number of ways. We show that interesting topics can be identified by the model and we demonstrate its effectiveness on the task of predicting locations of new messages and outperform non-trivial baselines. The main contributions are as follows:

- An additive generative model of content and locations that incorporates multiple facets of micro-blogging environments in an integral fashion.
- Sparse coding techniques and Bayesian treatments are smoothly embedded in our modeling, resulting in an efficient and effective implementation.
- Our model outperforms several state-of-the-art algorithms in the task of location predictions and it demonstrates interesting patterns in real-world datasets.

The paper is organized as follows. In Section 2 we will briefly discuss some recent related work in terms of geographical modeling in micro-blogging environments. In Section 3 we proceed with detailed description of the proposed model with implementation notes. In Section 4 we compare our model with several state-of-the-art algorithms in a number of tasks and demonstrate its effectiveness. Finally, we conclude in Section 5 with discussions and future work.

2. RELATED WORK

We briefly review two lines of related research. The first is a range of papers which use geographical language modeling in general while the second is a set of works which are specifically tuned for Twitter data. We are particularly interested in models and approaches that combine geographical modeling and language modeling to discover topics from geographical regions. We summarize some of representative work here:

- Mei et al. [12] propose a model based on Probabilistic Latent Semantic Indexing (PLSA) [11]. It assumes that each word is either drawn from a universal background topic or from a location and time dependent language model. Inference is performed via EM. However, the mixture coefficients between the background topic and other spatio-temporal topics ones is tuned manually. Since the model uses PLSA, no prior distribution is (or could be) assumed. Evaluation is carried out by showing anecdotal results.
- Later, Wang et al. [15] introduce a fully Bayesian generative model to incorporate locations. Rather than working with real latitudes and longitudes, they have a fixed number of region labels and they assume that each term is associated with a location label. For each word in a document, a topic assignment is first generated according to a multinomial distribution. Then the term and the location are generated dependent on this topic assignment, according to two different multinomial distributions. The inference is performed by Variational EM. Again the evaluation is limited to anecdotal results.
- Sizov [13] propose a similar model to [15]. Rather than using a multinomial distribution to generate locations they replace it with two Gaussian distributions for generating latitude and longitude respectively. For inference, this work uses Gibbs Sampling and the evaluation is done by showing anecdotal results, by measuring Deviation Information Criteria (a model complexity criterion similar to BIC), as well as classification accuracy using manually labeled data. One of the drawbacks of the work is that they only use data from Flickr restricted to the greater London area.

- Hao et al. [10] propose a model built upon Wang et al. [15]. However, they introduce the notion of global topics and local topics where more general terms are grouped into global topics and terms related to local events going to local topics. The inference is performed by Gibbs Sampling. Hao et al. [10] evaluate their model based on anecdotal results and some heuristic measurements.

- Yin et al. [17] propose a model is similar in spirit to Eisenstein et al. [7]. The terms and the location of a particular document are generated by a latent region. The location is generated from a region by a normal distribution and the region is sampled from a multinomial distribution. The prior is also placed into the model, however the inference is done by MAP-style EM rather than a fully Bayesian fashion. The model is evaluated using perplexity and by showing anecdotal results.

- Wing and Baldridge [16] use an even simpler approach where documents are assigned to geodesic grids and thus a supervised learning method is utilized, essentially yielding to build naïve Bayes classifiers on geodesic grids.

Although there exists such attempts of modeling language patterns and geographical locations, *most prior work does not consider users at all.*

A second line of work covers models directly designed to work on Twitter data. For instance, Eisenstein et al. [7] propose a model utilizing the correlations between global and local topics. In their model, each author is assigned a latent region variable and an observed GPS location. Terms and the actual GPS location are both conditioned on the latent region variable. The topics to generate terms are local topics, which are derived from global topics. The inference is done by Variational EM and the evaluation is done by measuring the accuracy of predicted location and showing anecdotal results. Finally, Cho et al. [5] studied the problem of human mobility in location sharing services. Their findings include that users tend to appear in a very limited number of places (e.g., office and home). They demonstrated that it might be effective enough to use a two component Gaussian mixture model to estimate users' locations.

It has been an active research area to incorporate different information sources into topic modeling. For example, Chemudugunta et al. [3] propose a method to combine corpus-wide topics and document-specific language patterns together by using a "switch" variable for each term in the document, becoming a popular scheme in topic modeling literature. We use a "switch-free" approach in this work and therefore reduce the number of variables used in the model. Last, for general patterns and analysis of social location sharing services, please refer to Cheng et al. [4].

3. MODEL

We now introduce our model that addresses the problems raised in the previous sections. We start with an overview of the basic components in Section 3.1 by discussing generative models without explicit switch variables. This allows us to describe the basic aspects of our model in Section 3.2. In order to learn more discriminative features, in Section 3.3, we impose L_1 penalty on certain parts of our model, resulting in a sparse modeling approach. For geographical modeling, non-informative prior distributions are discussed in Section 3.5. More implementation details follow in Section 3.6.

Table 1: Notation

Symbol	Size	Usage
$\boldsymbol{\eta}^0$	$1 \times \mathbb{R}$	global region distribution
$\boldsymbol{\eta}^{\text{user}}$	$\mathbb{U} \times \mathbb{R}$	user-dependent region distribution
$\boldsymbol{\theta}^0$	$1 \times \mathbb{K}$	global topic distribution
$\boldsymbol{\theta}^{\text{geo}}$	$\mathbb{R} \times \mathbb{K}$	region-dependent topic distribution
$\boldsymbol{\theta}^{\text{user}}$	$\mathbb{U} \times \mathbb{K}$	user-dependent topic distribution
$\boldsymbol{\phi}^0$	$1 \times \mathbb{V}$	global term distribution
$\boldsymbol{\phi}^{\text{geo}}$	$\mathbb{R} \times \mathbb{V}$	region-dependent term distribution
$\boldsymbol{\Pi}$	$\mathbb{K} \times \mathbb{V}$	a global topic matrix
$\boldsymbol{\mu}$	$\mathbb{R}^2$	mean location of a latent region
$\boldsymbol{\Sigma}$	$\mathbb{R}^{2\times 2}$	covariance matrix of a latent region

3.1 Preliminaries

Our model is closely related to the Sparse Additive Generative model (SAGE). The basic idea of the SAGE model is that the outcome variable is generated by the mixture of all components without any explicit indicator variable. The key difference to traditional mixture models is that the mixture occurs not in terms of the expectation parameters (i.e. the distribution) but in terms of the natural parameters of the exponential family model. Such a model has the advantage that it can easily take a large number of aspects into account without having to infer a complex indicator variable distinguishing the set of causes.

To be more concrete, we take language modeling as an example. Suppose we have a vocabulary $\mathcal{V}$ where each term v is generated by a background language model $\boldsymbol{\phi}_0$, a per-user background language model $\boldsymbol{\phi}_u$ and a regional language model $\boldsymbol{\phi}_g$. A conventional mixture model would attempt to represent the joint influence of the three components by a linear combination of the associated densities. Denote by $p(v|\boldsymbol{\phi})$ an exponential family model of the form

$$p(v|\boldsymbol{\phi}) = \exp\left(\boldsymbol{\phi}_v - g(\boldsymbol{\phi})\right) \text{ where } g(\boldsymbol{\phi}) = \log \sum_v \exp\left(\boldsymbol{\phi}_v\right)$$

Here $g(\boldsymbol{\theta})$ is often referred to as the log-partition function as it ensures that the distribution is properly normalized. In particular for the discrete distribution $\phi(v|\boldsymbol{\phi})$ is well-defined for all choices of $\boldsymbol{\phi}$. We now combine the factors via

$$P(v|\boldsymbol{\phi}_0, \boldsymbol{\phi}_u, \boldsymbol{\phi}_g) := p(v|\boldsymbol{\phi}_0 + \boldsymbol{\phi}_u + \boldsymbol{\phi}_g) \quad (1)$$

Unlike in traditional topic models, the formalism above does not require an indicator variable to specify which component to use in generating v. In addition to additive modeling, different language models can be constructed in such a way as to incorporate more discriminative terms. More specifically, in our model we choose $\boldsymbol{\phi}_0$ to denote the (baseline) log frequency of v in the dataset while other components are used to model the differences between the baseline and the background model. This idea is explored in [18, 6] to model topics. Here, we extend it to model regions and topics jointly and to propose an efficient inference procedure.

3.2 Model Description

We start the discussion with some notations in our model. Each tweet $d = \{\mathbf{w}_d, \mathbf{l}_d, u_d\}$ consists of three parts: Here

$\mathbf{w}_d$ is the word vector for the tweet, following a simple bag of word assumption, $\mathbf{l}_d$ is a real-valued pair $\mathbf{l}_d = \{l_0, l_1\}$, representing the latitude and longitude where this tweet is written and u_d is the user id for the author of the tweet. For simplicity, we assume that all the tweets in our dataset are generated by a fixed vocabulary $\mathcal{V}$ and a fixed user base $\mathcal{U}$. Moreover, we assume that the geographical locations have been clustered into R latent regions. Each region $r \in R$ is characterized by a mean location $\boldsymbol{\mu}_r$ and a covariance matrix $\boldsymbol{\Sigma}_r$. We assume that there are three types of language models: a) a background language model $\boldsymbol{\phi}^0$, b) a per-region language model $\boldsymbol{\phi}^{\text{geo}}$ and c) a topical language model $\boldsymbol{\Pi}$. All these language models are over the vocabulary $\mathcal{V}$. Each tweet is influenced by these three factors simultaneously. Before describing the generative process of our model, on a high level, our model encodes the following intuitions:

- Words used in a tweet depend on both the location and topic of the tweet.
- Different geographical regions have different language variations. Topics have different chances to be discussed in different regions (e.g. bullfights in India are unlikely to occur; likewise Spaniards are unlikely to discuss Divali).
- Users tend to appear in a handful geographical locations.

For each tweet, the model generates the location, the topic and terms in the tweet consecutively. In our model, all locations are categorized into R latent regions. For each tweet, we first choose from which latent region this tweet is written. To generate the region index r, we utilize a multinomial model as follows:

$$P\left(r|\boldsymbol{\eta}^0, \boldsymbol{\eta}_u^{\text{user}}\right) = p\left(r|\boldsymbol{\eta}^0 + \boldsymbol{\eta}_u^{\text{user}}\right) \tag{2}$$

Here $\boldsymbol{\eta}_0$ is a global distribution over latent regions and $\boldsymbol{\eta}_u$ is a user dependent distribution over latent regions for user u. Each location $\mathbf{l}_d$ is drawn from a latent region r by a region-dependent multivariate normal distribution

$$\mathbf{l}_d \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r). \tag{3}$$

Once the region and the location is generated, a topic z is selected dependent on both the latent region and the author of tweet:

$$P\left(z|\boldsymbol{\theta}^0, \boldsymbol{\theta}_u^{\text{user}}, \boldsymbol{\theta}_r^{\text{geo}}\right) = p\left(z|\boldsymbol{\theta}_j^0 + \boldsymbol{\theta}_{u,j}^{\text{user}} + \boldsymbol{\theta}_{r,j}^{\text{geo}}\right) \tag{4}$$

Here $\boldsymbol{\theta}^0$ is a global distribution over topics, $\boldsymbol{\theta}_u^{\text{user}}$ is a user-dependent distribution over topics and $\boldsymbol{\theta}_r^{\text{geo}}$ is a regional distribution over topics. The intuition is that the topic is heavily influenced where this tweet is written and user preferences. After generating the topic index z each word w in the tweet is generated by drawing from the aggregate distribution:

$$P\left(w|z, \boldsymbol{\phi}^0, \boldsymbol{\phi}_r^{\text{geo}}, \boldsymbol{\Pi}_z\right) = p\left(w|\boldsymbol{\phi}^0 + \boldsymbol{\phi}_r^{\text{geo}} + \boldsymbol{\Pi}_{z_d}\right). \tag{5}$$

In this case $\boldsymbol{\phi}^0$ parametrizes a global distribution over terms, $\boldsymbol{\phi}^{\text{geo}}$ describes the a region-dependence and $\boldsymbol{\Pi} \in \mathbb{R}^{\mathbb{K}\times\mathbb{V}}$ is a topic matrix where each row is a distribution over terms. With the above specification the generative story for a single tweet d can be expressed as follows:

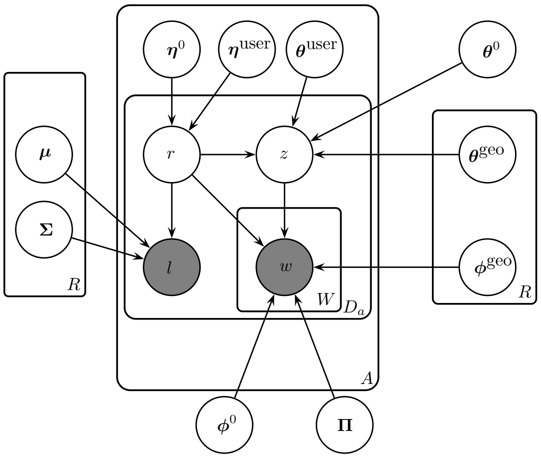

Figure 1: A graphical representation of our model

- Draw a latent region index

$$r_d \sim p(r_d|\boldsymbol{\eta}^0 + \boldsymbol{\eta}_u^{\text{user}})$$

- Draw a topic index

$$z_d \sim p(z_d|\boldsymbol{\theta}^0 + \boldsymbol{\theta}_u^{\text{user}} + \boldsymbol{\theta}_r^{\text{geo}})$$

- Draw a location

$$\mathbf{l}_d = \{l_0, l_1\} \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$$

- For each token w in $\mathbf{w}_d$ draw

$$w \sim p(w|\boldsymbol{\phi}^0 + \boldsymbol{\phi}_r^{\text{geo}}, \boldsymbol{\Pi}_{z_d})$$

This generative process applies to all tweets in the corpus. The graphical representation of the generation process is shown in Figure 1.

3.3 Sparse Modeling

As discussed in Section 3.1, the benefit of our approach is to learn discriminative features from data, rather than obtaining redundant ones in different components of the model. In order to achieve this goal, we also impose prior distributions over certain parts of our model. More specifically, for the following components in the model, we impose zero-mean Laplace distributions.

The rationale is that users in certain regions are likely to draw their words either from a location independent distribution or from a *small*, i.e. sparse corpus of additional terms which are more prevalent in a given location rather than globally. Likewise, we assume that topics consist of a background distribution of generic words plus a sparse set of additional words which are characteristic for the particular topic. Note that we do *not* require these words to be unique. That is, the word “jaguar” might for instance be more prevalent in the “animals” and in the “cars” topic. However, we do not expect it to be prevalent in a large number of topics beyond what a background language model would indicate.

We have

$$\begin{aligned} \boldsymbol{\eta}_r^0 &\sim \mathcal{L}(0, \omega^0) \quad & \boldsymbol{\eta}_{u,r}^{\text{user}} &\sim \mathcal{L}(0, \omega_u) & & \\ \boldsymbol{\theta}_z^{\text{geo}} &\sim \mathcal{L}(0, \lambda_l) \quad & \boldsymbol{\theta}_{u,z}^{\text{user}} &\sim \mathcal{L}(0, \lambda_u) \quad & \boldsymbol{\theta}_{r,z}^{\text{geo}} &\sim \mathcal{L}(0, \lambda_r) \\ \boldsymbol{\phi}_v^0 &\sim \mathcal{L}(0, \psi^0) \quad & \boldsymbol{\phi}_{r,v}^{\text{geo}} &\sim \mathcal{L}(0, \psi_l) & & \\ \mathbf{\Pi}_{z,v} &\sim \mathcal{L}(0, \psi_t) & & & & \end{aligned}$$

where $\mathcal{L}(\mu, b)$ is a Laplace distribution with mean μ and scale parameter b. A zero-mean Laplace prior has the same effect as placing an L_1 regularizer on these components, resulting in a sparse solution to the model. Here, a sparse modeling approach does not only encourage more discriminative features to be learned, but also leads to a more efficient learning algorithm, which will be introduced below. We use ISTA [2] algorithm to do sparse optimization in our work.

Note that besides Laplace distributions used in this paper, other distributions could be employed, too. For instance using a normal distribution as prior on all elements amounts to a latent Gaussian process induced by the parameters.

3.4 Inference Algorithm

Before we proceed with the inference algorithm, we introduce the following shorthands to simplify our notation:

$$\begin{aligned} P(z_d = k | \boldsymbol{\theta}^0, \boldsymbol{\theta}_u^{\text{user}}, \boldsymbol{\theta}_r^{\text{geo}}) &= \boldsymbol{\alpha}_{u,r,k} \\ P(w = v | z_d, \boldsymbol{\phi}^0, \boldsymbol{\phi}_r^{\text{geo}}, \mathbf{\Pi}) &= \boldsymbol{\beta}_{r,z,v} \\ P(r = t \,| \boldsymbol{\eta}^0, \boldsymbol{\eta}_u^{\text{user}}) &= \boldsymbol{\rho}_{u,t} \end{aligned}$$

We treat topic assignments z and latent region assignments r as latent variables and all other variables as model parameters. A mixture between EM and a Monte Carlo sampler is utilized to effectively learn all parameters for the model along the lines of [14]. In the E-step, we sample latent region assignments and topic assignments by fixing all other parameters by Gibbs sampling. In the M-step, we optimize model parameters by fixing all latent region assignments and topic assignments. We iterate this until convergence.

More specifically, in the E-step, we iteratively draw latent region assignments and topic assignments for all tweets. For each tweet, a latent region r is firstly drawn from the following distribution, conditioned on the old topic assignments:

$$r \sim P(\mathbf{l}_d | \boldsymbol{\mu}_j, \mathbf{\Sigma}_j) \times \boldsymbol{\rho}_{u,j} \times \boldsymbol{\alpha}_{u,j,k} \times \prod_{i=1}^{N_d} \boldsymbol{\beta}_{j,k,v} \tag{6}$$

where $P(\mathbf{l}_d | \boldsymbol{\mu}_j, \mathbf{\Sigma}_j)$ is the pdf function for a multivariate normal distribution and k is the old topic assignment. After r is sampled, we sample the topic assignment z for the same tweet, conditioned on the newly sampled r:

$$z \sim \boldsymbol{\alpha}_{u,r,k} \times \prod_{i=1}^{N_d} \boldsymbol{\beta}_{r,z,v} \tag{7}$$

where r is the new region index. In the M-step, we maximize the log likelihood of the model with respect to model parameters by fixing all region and topic assignments obtained in the E-step. For geographical modeling, the maximum likelihood estimation (MLE) of parameters can be obtained in closed form:

$$\boldsymbol{\mu}_j = \bar{N}_j \quad = \frac{1}{\#(d,j)} \sum_{d=1}^{D} \mathbb{I}(r_d = j) \mathbf{l}_d \tag{8}$$

$$\mathbf{\Sigma}_j = S_j \quad = \frac{1}{\#(d,j) - 1} \sum_{d=1}^{D} (\mathbf{l}_d - \boldsymbol{\mu}_j)^T (\mathbf{l}_d - \boldsymbol{\mu}_j) \tag{9}$$

where $\#(d,j)$ is the number of tweets assigned to region j. Indeed, $\boldsymbol{\mu}_j$ is set to the sample mean and $\mathbf{\Sigma}_j$ is set to the sample variance. For other parameters, unfortunately, no closed-form solutions exist. Therefore, we adopt gradient-based optimization methods to maximize the likelihood. Let L be the likelihood of the model. The gradients of model parameters can be obtained as follows. For $\boldsymbol{\eta}^0$ and $\boldsymbol{\eta}^{\text{user}}$, we have:

$$\begin{aligned} \partial \boldsymbol{\eta}_t^0(L) &= \sum_{u=1}^{U} d(u,t) - \sum_{u=1}^{U} d(u) \boldsymbol{\rho}_{u,t} \\ \partial \boldsymbol{\eta}_{u,t}^{\text{user}}(L) &= d(u,t) - d(u) \boldsymbol{\rho}_{u,t} \end{aligned} \tag{10}$$

where $d(u,t)$ is the number of tweets produced by user u are assigned to the region t and $d(u)$ is the total number of tweets generated by user u. For the global topic distribution $\boldsymbol{\theta}^0$, user topic distributions $\boldsymbol{\theta}^{\text{user}}$ and regional topic distributions $\boldsymbol{\theta}^{\text{geo}}$, we have:

$$\partial \boldsymbol{\theta}_k^0(L) = \sum_{u=1}^{U} d(u,k) - \sum_{u=1}^{U} \sum_{t=1}^{R} d(u,t) \boldsymbol{\alpha}_{u,t,k} \tag{11}$$

$$\partial \boldsymbol{\theta}_{u,k}^{\text{user}}(L) = d(u,k) - \sum_{t=1}^{R} d(u,t) \boldsymbol{\alpha}_{u,t,k} \tag{12}$$

$$\partial \boldsymbol{\theta}_{t,k}^{\text{geo}}(L) = \sum_{u=1}^{U} d(u,t,k) - \sum_{u=1}^{U} d(u,t) \boldsymbol{\alpha}_{u,t,k} \tag{13}$$

where $d(u,k)$ is the number of tweets produced by user u assigned to the topic k and $d(u,t,k)$ is the number of tweets written by the user u in the region t assigned to the topic k. For the global language model $\boldsymbol{\phi}^0$, regional language models $\boldsymbol{\phi}^{\text{geo}}$ and topical language models $\mathbf{\Pi}$, we have:

$$\partial \boldsymbol{\phi}_v^0(L) = \sum_{t=1}^{R} n(t,v) - \sum_{t=1}^{R} \sum_{k=1}^{K} n(t,k) \boldsymbol{\beta}_{t,k,v} \tag{14}$$

$$\partial \boldsymbol{\phi}_{t,v}^{\text{geo}}(L) = n(t,v) - \sum_{k=1}^{K} n(t,k) \boldsymbol{\beta}_{t,k,v} \tag{15}$$

$$\partial \mathbf{\Pi}_{k,v}(L) = \sum_{t=1}^{R} n(t,k,v) - \sum_{t=1}^{R} n(t,k) \boldsymbol{\beta}_{t,k,v} \tag{16}$$

where $n(d,v)$ is the number of times term v appearing in tweet d, $n(t,k)$ is the number of terms associated to the topic k in region t, $n(t,v)$ is the number of times term v appearing region t, $n(t,k,v)$ is the number of terms v assigned to the topic k appearing in the region t. These gradients have an intuitive interpretation as the difference of the true counts and their expected counts.

3.5 Geograpical Location Modeling

In the previous section, we use a point estimate of regional means and covariance matrices in each M-step based on samples obtained in the E-step. However, this process is not very stable since only one sample of regional assignments for each tweet is taken into account. One way to reduce this instability would be to draw multiple samples per tweet and to use a set of samples for estimation purposes. However, this would introduce an inner loop in the E-step for each tweet, thus significantly increasing sampling time.

Instead, we apply a Bayesian treatment to mean vectors and covariance matrices and do not estimate them explicitly in M-step. The standard practice in multivariate normal distribution is to endow them with a set of conjugate parameters, that is, with a Gauss-Wishart prior. This is computationally expensive.

A cheaper (and equally reliable) approach is to place a non-informative Jeffrey's prior over the values of the mean parameters, that is

$$\boldsymbol{\mu} \sim \text{Unif}.$$

and a Jeffrey's distribution over the values of the covariance matrices to penalize large covariance matrices:

$$P(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(3/2)}.$$

The same treatment is also used in [1, 8]. By imposing these prior distributions, we can effectively integrate out $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, resulting in a collapsed Gibbs sampler for locations, similar to [9]. More specifically, we sample r from the following distribution:

$$r \sim T\left(\bar{N}_r, S_r \frac{(n+1)}{n(n-2)}, n-2\right) \boldsymbol{\rho}_{u,j} \boldsymbol{\alpha}_{u,j,k} \prod_{i=1}^{N_d} \boldsymbol{\beta}_{j,k,v} \qquad (17)$$

Here $T(a, b, n)$ is a multivariate Student-T distribution with the location as a, the scale matrix as b and n degree of freedom. Here, $\bar{N}_r$ and S_r are sample mean and sample respectively, as defined in (8). Sampling r does not require us to re-estimate the values of mean and covariance matrix in the M-step and hence reduce the computation cost of the inference algorithm.

3.6 Implementation Notes

Several implementation notes warrant a detailed discussion here. Firstly, the bottleneck of sampling z is to evaluate many exponential functions as we expand Equation (7):

$$\frac{\exp\Big(\boldsymbol{\theta}^0_k + \boldsymbol{\theta}^{\text{user}}_{u,k} + \boldsymbol{\theta}^{\text{geo}}_{r,k}\Big)}{\sum_{i=1}^{K} \exp\Big(\boldsymbol{\theta}^0_i + \boldsymbol{\theta}^{\text{user}}_{u,i} + \boldsymbol{\theta}^{\text{geo}}_{r,i}\Big)} \prod_{i=1}^{N_d} \frac{\exp\Big(\boldsymbol{\phi}^0_{w_i} + \boldsymbol{\phi}^{\text{geo}}_{r,w_i} + \boldsymbol{\Pi}_{k,w_i}\Big)}{\sum_{j=1}^{V} \exp\Big(\boldsymbol{\phi}^0_j + \boldsymbol{\phi}^{\text{geo}}_{r,j} + \boldsymbol{\Pi}_{k,j}\Big)}$$

The key to speed up the sampling procedure here is to reduce the number of exponential functions to be evaluated. We rewrite the above equation as:

$$\begin{aligned} &\exp\Big[\boldsymbol{\theta}^0_k + \boldsymbol{\theta}^{\text{user}}_{u,k} + \boldsymbol{\theta}^{\text{geo}}_{r,k} + \sum_{i=1}^{N_d}\Big(\boldsymbol{\phi}^0_{w_i} + \boldsymbol{\phi}^{\text{geo}}_{r,w_i} + \boldsymbol{\Pi}_{k,w_i}\Big) \\ &- \log \sum_{i=1}^{K} \exp\Big(\boldsymbol{\theta}^0_i + \boldsymbol{\theta}^{\text{user}}_{u,i} + \boldsymbol{\theta}^{\text{geo}}_{r,i}\Big) \\ &- N_d \log \sum_{j=1}^{V} \exp\Big(\boldsymbol{\phi}^0_j + \boldsymbol{\phi}^{\text{geo}}_{r,j} + \boldsymbol{\Pi}_{k,j}\Big)\Big] \qquad (18) \end{aligned}$$

The logarithm of a sum of components can be efficiently computed as $\log \sum_i \exp(x_i) = m + \log[\sum_i \exp(x_i - m)]$ where m is the maximum element in x_i and can be cached since they are constant in the E-step. Therefore, we only need to calculate one exponential function for sampling z per tweet, which significantly reduces the computational cost.

The second technique to speed up the inference algorithm is to efficiently calculate gradients (14), (11), and (10). A naïve calculation would lead to a very inefficient implementation. Taking the gradients of $\boldsymbol{\Pi}$ as an example, the expanded form of gradients is as follows:

$$\sum_{t=1}^{R} n(t,k,v) - \sum_{t=1}^{R} n(t,k) \frac{\exp(\boldsymbol{\phi}^0_v + \phi^{\text{geo}}_{t,v} + \boldsymbol{\Pi}_{k,v})}{\sum_{i=1}^{V} \exp(\boldsymbol{\phi}^0_i + \boldsymbol{\phi}^{\text{geo}}_{t,i} + \boldsymbol{\Pi}_{k,i})}$$

where the second part of the gradients, which is the expected counts, requires the calculation for all the possible combinations of topics and latent regions. However, because of sparse modeling in Section (3.3), we can effectively calculate the second parts by utilizing the sparsity of the model as follows:

$$\begin{aligned} &n(k,v) - \exp(\boldsymbol{\phi}^0_v) \sum_{t=1}^{R} n(t,k) \frac{1}{C_{t,k}} \\ &- \sum_{t=1}^{R} n(t,k) \frac{1}{C_{t,k}} \exp(\boldsymbol{\phi}^0_v) \Big[\exp(\boldsymbol{\phi}^{\text{geo}}_{t,v}) - 1)\Big] \\ &- \sum_{t=1}^{R} n(t,k) \frac{1}{C_{t,k}} \exp(\boldsymbol{\phi}^0_v) \exp(\boldsymbol{\phi}^{\text{geo}}_{t,v}) \Big[\exp(\boldsymbol{\Pi}_{k,v}) - 1\Big] \end{aligned}$$

where $C_{t,k} = \sum_{i=1}^{V} \exp(\boldsymbol{\phi}^0_i + \boldsymbol{\phi}^{\text{geo}}_{t,i} + \boldsymbol{\Pi}_{k,i})$. The gradients are decomposed into three parts. The first part is a global term for all terms and therefore can be calculated once and cached. The second part only exists for those $\boldsymbol{\phi}^{\text{geo}}_{t,v}$ are not zero. Similarly, the third part is non-zero only when both $\boldsymbol{\phi}^{\text{geo}}_{t,v}$ and $\boldsymbol{\Pi}_{k,v}$ are not zero. Thus, if we employ a reasonable L_1 regularizer on both regional and topical language models, most of those elements would be driven to zero and therefore the second and third parts can be very efficiently calculated. Similar decomposition also works for other gradients.

The last but not the least important technique is how to initialize the model. Different initialization values of parameters can lead to significantly different results. Here, we use the following initialization steps. Again, taking language models as an example, we firstly initialize $\boldsymbol{\phi}^0$ as log frequencies of terms in the whole corpus and $\boldsymbol{\phi}^{\text{geo}}_r$ as log frequencies of terms in region r minus the same term in $\boldsymbol{\phi}^0$. Then, we initialize $\boldsymbol{\Pi}$ as all zero and optimize over $\boldsymbol{\Pi}$ by fixing $\boldsymbol{\phi}^0$ and $\boldsymbol{\phi}^{\text{geo}}$. Similar strategy can be also applied to $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ values. For latent regions, we initialize them by a K-Means algorithm.

4. EXPERIMENTS

In this section, we demonstrate the effectiveness of our model on real-world datasets. We compare our model with several state-of-the-art models. Our dataset is a sample of the Twitter Firehose stream[1], issued to Yahoo!. In Twitter, two types of location information are associated to tweets: 1) geographical locations and 2) Twitter Places[2]. For geographical locations, each tweet is associated to a real-valued latitude and longitude vector. For Twitter Places, we convert them into real-valued latitudes and longitudes. After doing this, we remove all tweets without locations. We also preprocess all the remaining tweets by detecting whether

[1] https://dev.twitter.com/docs/streaming-api/methods

[2] http://blog.twitter.com/2010/06/twitter-places-more-context-for-your.html

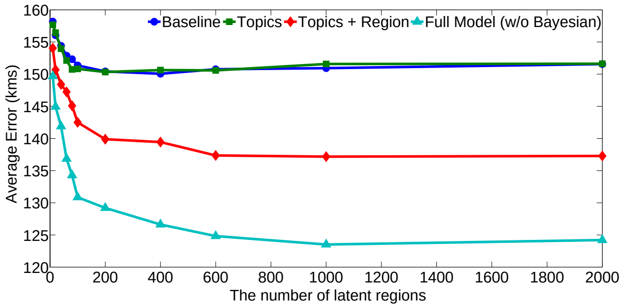

Figure 2: The comparison of location prediction on Yahoo! dataset. The X-axis is the number of latent regions and Y-axis is the average Euclidean distance in kilometers (kms) between predicted locations and true locations.

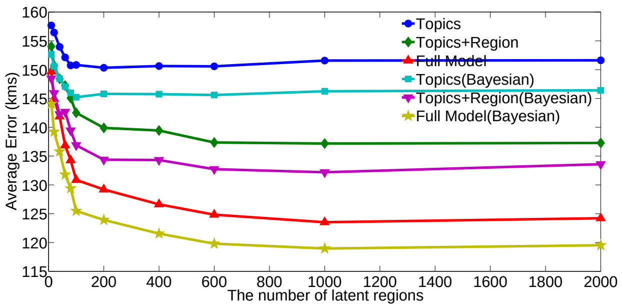

Figure 3: The comparison of non-Bayesian models and Bayesian models on the task of location prediction on Yahoo! dataset. The X-axis is the number of latent regions and Y-axis is the average Euclidean distance in kilometers (kms) between predicted locations and true locations.

a tweet is in English. This step is done by a dictionary based method. We randomly sample 10,000 users from the dataset, with their full set of tweets between January 2011 and May 2011, resulting 573,203 distinct tweets. The size of the dataset is significantly larger than the ones used in some similar studies (e.g, [7, 17]).

4.1 Location Prediction

In addition to demonstrating that our model can discover interesting topics and users' geographical patterns, we also wish to show that our model can be used in a quantitative fashion. Here, we focus on the task of location prediction for tweets. Differing from the work done by Eisenstein et al. [7] where their aim is to predict the location for a user and the way they defined the location of a user may not be very appropriate (the first location shown in their dataset), our goal is to predict the location for each new tweet, based on the words used in the tweet and its authors' information. Based on our statistics, only 1% $\sim$ 2% of tweets have either geographical locations (including Twitter Places) explicitly attached, meaning that we cannot easily locate a majority of tweets. However, it has been shown (e.g., [5, 4]) that geographical locations can be used to predict users' behaviors and uncover users' interests and therefore it is potentially invaluable for many perspectives, such as behavior targeting and online advertisements. In addition to our dataset, we also apply our model to an open source dataset[3], denoted as CMU dataset, and compare the best reported results.

Evaluation Metric: For each new tweet, we predict its location as $\hat{\mathbf{l}}_d$. We calculate the Euclidean distance between predicted value and the true location and average them over the whole test set $\frac{1}{N}\sum \text{Dis}(\hat{\mathbf{l}}_d, \mathbf{l}_d)$ where $\text{Dis}(a, b)$ is the Euclidean distance function and N is the total number of tweets in the test set.

Baselines: The following methods are used as baselines in our dataset to compare with the full model proposed in Section (3).

- Yin et al. [17]: Their method is essentially to have a global set of topics shared across all latent regions. There is no regional language models in the model. Besides, no user level preferences are learned in the model. The prediction is done by two steps: 1) choosing the region index that can maximize the test tweet likelihood, and 2) use the mean location of the region as the predicted location. We re-implemented their method in our work. This method is denoted as Baseline.

- Our model without $\boldsymbol{\phi}^{\text{geo}}$, $\boldsymbol{\eta}^{\text{user}}$ and $\boldsymbol{\theta}^{\text{user}}$: This is essentially very similar to Baseline. The only difference is that Baseline is under PLSA formalism and our model is in SAGE formalism. We denote this method as Topics.

- Our model without $\boldsymbol{\eta}^{\text{user}}$ and $\boldsymbol{\theta}^{\text{user}}$: This variation of our model can learn regional language models while user preferences are still missing here. We denote this method as Topics + Region.

For the comparison on the CMU dataset, we compare with:

- Eisenstein et al. [7]: The model is to learn a base topic matrix that can be shared across all latent regions and a different topic matrix as the regional variation for each latent region. No user level preferences are learned in the model. The best reported results are used in the experiments.

- Eisenstein et al. [6]: The original SAGE paper. The best reported results are used in the experiments.

- Wing and Baldridge [16]: Their method is essentially to learn regional language models per explicit regions. The best reported results are used in the experiments.

For our model, the prediction is conducted in two steps. Firstly, a region index that can maximize the likelihood of test tweet is chosen. Next, the mean location of the corresponding region is used as the predicted location.For Bayesian treatment of geographical modeling discussed in Section (3.5), the mean vectors are estimated after the whole inference algorithm finishes.

Experimental Results: Firstly, we show the basic comparison between our model and other baselines discussed above on the Yahoo! dataset. The results are shown

[3] http://www.ark.cs.cmu.edu/GeoText/

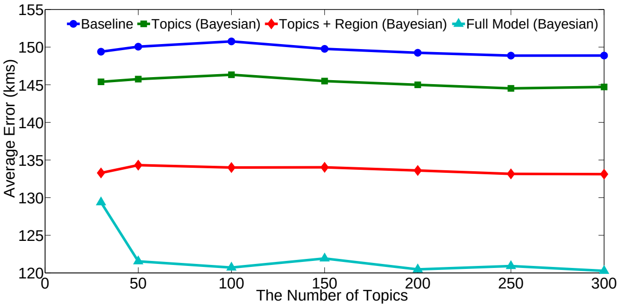

Figure 4: The comparison of models with different number of topics by fixing the number of latent regions (as 400) on `Yahoo!` dataset. The X-axis is the number of topics and Y-axis is the average Euclidean distance in kilometers (kms) between predicted locations and true locations.

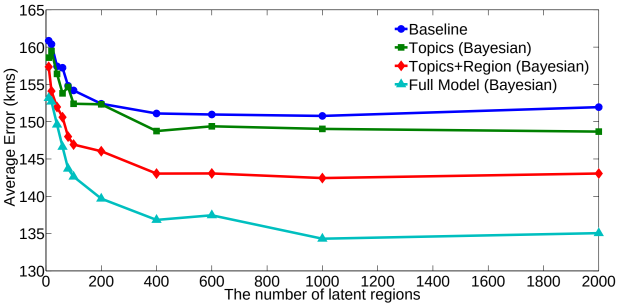

Figure 5: The comparison of models with randomly selected users on `Yahoo!` dataset. The X-axis is the number of latent regions and Y-axis is the average Euclidean distance in kilometers (kms) between predicted locations and true locations.

in Figure (2). In this experiment, we fix the number of topics to 50 for all models. For all models, we adopt a five-fold cross validation setting. The numbers reported here are averaged across different folds. One major impression is that the average error decreases as the number of latent regions increases, although it becomes flat after 500 latent regions. This makes sense because we predict the locations based on the mean locations of latent regions. Therefore, the more regions the model has, the more flexible the prediction would be. As we discussed above, `Topics` method is very similar to `Baseline` method and therefore, not very surprisingly, the performance of these two models is approximately the same. For `Topics + Region` model, the performance is significantly better over `Baseline` model and `Topics` model. The main reason might be that regional language models learn special terms for different regions and therefore these terms become discriminative when we perform location predictions. Moreover, our sparse modeling approach also contributes to learned discriminative terms in regional language models. By incorporating user regional preferences ($\boldsymbol{\eta}^{\text{user}}$), our full model performs the best on the `Yahoo!` dataset. This partially validates that users might have stable mobility patterns in their usage of micro-blogging environments and therefore we can learn this pattern through their historical content. Indeed, Cho et al. [5] found that users who frequently use location sharing services demonstrate surprisingly stable patterns and they successfully used a two-component Gaussian mixture model to predict users' locations in the future. Note that the full model used in this experiment is the one **without** Bayesian geographical modeling that is discussed in Section (3.5).

The next set of experiments is to show whether the Bayesian treatment of geographical modeling can lead to additional improvements of predication performance. As we previously discussed, non-Bayesian modeling in locations may lead to unstable results. The experimental setting follows the one used above and results are shown in Figure (3). Two observations can be made from the figure. Firstly, all models with Bayesian modeling lead to significantly improvements over their non-Bayesian counterparts. The second observation is that, although Bayesian modeling can improve the performance, major improvements still comes from whether certain components are "on" or "off". In short, Bayesian modeling in locations enjoys better predictive performance and a more efficient inference algorithm, as discussed in previous sections.

All previous experiments are the ones with fixed topics and different latent regions. Here we show how the predictive performance varies for different number of topics. The basic setting remains the same as the previous two sets of experiments and the results are shown in Figure (4). The main observation is that the performance does not change too much as the number of topics varies. As we mentioned before, all these models make predictions based on the mean vectors of latent regions. Therefore, a fixed number of regions will limit the predictive power of these models and hence the performance is sort of bounded in a range. In other words, enlarging the number of topics does not give models the flexibility to learn regions well.

Another interesting experiment is not to randomly sample tweets but randomly sample users. In this setting, all users in the test set are never shown in the training set and therefore we do not have sufficient user preference data. This setting might be more realistic in Twitter because the majority of users never use geo-related features and hence it is highly likely that some users will adopt this feature in the future. In order to effectively predict locations, we use the following strategy to learn a "prior" distribution for users. Taking $\boldsymbol{\eta}^{\text{user}}$ as an example, since the test user is not in our training set, we optimize over $\boldsymbol{\eta}^{\text{user}}$ by fixing all other parameters on the fly. Therefore, the obtained values for this user is essentially the prior regional distribution for this user, without any tweets observed. After having this prior distribution, we can effectively predict locations as usual. We do this optimization for users on the fly for all other user-related parameters. The results are shown in Figure (5). The main observation from the figure is that the performance from all models is significantly worse than the experiments with randomly selected tweets. This partially validates that all these models suffer from certain difficulties for "new" users and "new" content. However, the relative improvement of performance remains the same as previous experiments, suggesting that our model can learn reasonable prior distributions for users, in order to achieve better predictive performance.

For the `CMU` dataset, we download their dataset and run our model on it. Note that previous models (e.g., [7, 16]) are designed to predict the locations for users. In our case, we can do finer grained predictions on tweet level. To make fair

# of latent regions	[[7]]	[[16]]	[[6]]	`Topics`	`Topics + Region`	Full Model
10	494	479	501	540.60	481.58	449.45
20	494	479	501	522.18	446.03	420.83
40	494	479	501	513.06	414.95	395.13
60	494	479	501	507.37	410.09	380.04
80	494	479	501	499.42	408.38	374.01
100	494	479	501	498.94	407.78	372.99

Table 2: Comparison of models on `CMU` dataset. All numbers are Kilometers. For [7, 16, 6], the median number reported in the paper is used. We do not re-run their models and only report numbers from corresponding papers.

comparisons, two strategies can be applied here: 1) obtain the predicted location for each tweet and take the mean locations over them and 2) obtain the dominant region index for tweets by the same user and use the mean value for it as the prediction. In our experiments, we have tried both strategies and found no significant difference between them. Therefore, we only report the results from the first strategy. The results are shown in Table (2). Firstly, we see that our full model outperforms all previous models significantly. In addition, as the number of latent regions increases, the predictive performance increases, which also validates the results in our `Yahoo!` dataset. Here is some analysis why our model outperforms others. For [7] and [6], they used a topic-variation matrix per region, which might be too expensive to be applied over a large number of regions while the authors in those papers found that their model peaks at around 50 regions and 10 topics and the predictive performance deteriorates otherwise for excessive number of parameters, resulting in over-fitting. In our case, we use global topics and background topics to factor out common words. In addition, we use two signals: regional topic distribution and regional word unigrams. For [7, 6], their model has a single location for all tweets per user. On the contrary, our model assumes that each user has a distribution over regions and each tweet is associated with a region, thus we can accommodate user movements. Also, their models used a two-stage training which does not enable the language model to influence how many regions are needed. However, we use a joint training procedure for both regions and topics and we re-sample the user regions in our training phase where their models assume that regions assignments are given at the first place.

4.2 Qualitative Study

In this section, we take one run of our full model on `Yahoo!` dataset as an example to demonstrate what kinds of topics can be obtained. Firstly, we show some samples of regional language models. As we see in the previous section, these language models play a vital role in location predictions. Since in our model, regions are latent variables and do not correspond to cities or regions in the real-world. It might be difficult to demonstrate topics. Here, we assign the mean vectors of latent regions to nearest existing cities and manually pick 5 cities as an example, shown in Table (3). Terms are the ones with largest magnitudes in $\boldsymbol{\phi}^{\text{geo}}$. It is very interesting to see that most top ranked terms are actually the name of these locations. Remember that our method is fully unsupervised. In addition, we can see that top ranked terms in different regions vary significantly. Another interesting observation is that users tend to tweet with their locations when they are in airports. This can be seen in region "United States->California->San Francisco" and "United Kingdom->England->London". In addition to geographical language models, we also show some examples from the global topic matrix $\mathbf{\Pi}$. These language models are designed so that broader topics will be captured here. The examples are shown in Table (4). Again, these topics are manually picked and the "title" of these topics is assigned by the authors of the paper since these topics are learned without any explicit labels. We can see that these topics are relatively broad, compared to regional language models and widely discussed across regions. Some topics might have captured recent unrest in the Middle East.

Entertainments
lady bieber album music beats artist video listen itunes apple produced movies #bieber lol new songs
Sports
yankees match nba football giants wow win winner game weekend horse #nba
Politics
obama election middle east china uprising egypt russian tunisia #egypt afghanistan people eu

Table 4: Examples of $\mathbf{\Pi}$, global topic matrix. The terms are top ranked terms in each language model.

5. CONCLUSIONS

In this paper, we address the problem of modeling geographical topical patterns on Twitter by introducing a novel sparse generative model, which utilizes both statistical topic models and sparse coding techniques to provide a principled method for uncovering different language patterns and common interests shared across the world. Our approach is vital for applications such as behavior targeting, user profiling, content recommendation and topic tracking and the method can be easily extended in a number of ways. We show that interesting topics can be identified by the model and we demonstrate its effectiveness on the task of predicting locations of new messages and outperform non-trivial baselines. Main contributions of this work include a) a sparse additive model of content and locations that incorporate multiple facets of micro-blogging environments without switch variables, b) sparse coding techniques and Bayesian treatments are smoothly embedded in our modeling, resulting in an efficient and effective implementation and c) outperforms several state-of-the-art algorithms in the task of location predictions and demonstrate interesting patterns from real-world datasets. For future work, we wish to model human mobility explicitly by introducing user level regional

Location with Top Ranked Terms
United States->New York->Brooklyn
brooklyn ave flatbush avenue mta prospect 5th #brooklyn spotlight carroll bushwick museum broadway madison vanderbilt coney slope eastern subway new york pkwy #viernesnayobon #mets otsego greenwich starbucks
United States->California->San Francisco
sfo francisco san airport international millbrae terminal flight burlingame bart mateo boarding bayshore telecommute landed heading bay airlines united bound flying #sfo camino groupon caltrain moon tsa baggage california engineer valley
United States->Pennsylvania->Philadelphia
philadelphia #philadelphia phl #jobs market others #job street philly walnut septa chestnut the cherry sansom arch spruce citizens locust btw temple pennsylvania rittenhouse passyunk bitlyetq7a6 bookrenters pike international
United Kingdom->England->London
winds lhr hounslow terminal the cloudy mph ickenham bath heathrow temperature airport car only airways uxbridge sun splendid fair london british lounge tothers harmondsworth speedbird whens for stars day flight dominos navigation brunel
Australia->New South Wales->Sydney
sydney #sydney bondi george street mascot domestic syd surry station cnr platforms harbour darlinghurst qantas hoteloxford eddy haymarket terminal wales australia chalmers uts pitt #marketing junction darling centre #citijobs citigroup druitt

Table 3: Examples of $\phi^{\mathbf{geo}}$, geographical language models. The terms are top ranked terms in each language model.

components. In addition, temporal factors should also be considered for the task of location prediction.

Acknowledgements

We thank Marco Pennacchiotti and Ana-Maria Popescu for helpful discussions. In addition, we thank Twitter to provide the dataset and the permission for publication of this work. We also appreciate the reviews from anonymous reviewers. This material is based upon work supported in part by the National Science Foundation under Grant Numbers IIS-0545875 and IIS- 0803605.

6. REFERENCES

[1] A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *Proceedings of KDD 2009*, pages 39–48, New York, NY, USA. ACM.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, March 2009.

[3] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS 2006*, pages 241–248.

[4] Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring millions of footprints in location sharing services. In *ICWSM 2011*.

[5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of KDD 2011*, pages 1082–1090, New York, NY, USA. ACM.

[6] J. Eisenstein, A. Ahmed, and E. Xing. Sparse additive generative models of text. In *Proceedings of ICML 2011*, pages 1041–1048, New York, NY, USA, June. ACM.

[7] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP 2010*, pages 1277–1287, Stroudsburg, PA, USA. Association for Computational Linguistics.

[8] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis 2nd edition*. Chapman-Hall, 2003.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[10] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang. Equip tourists with knowledge mined from travelogues. In *Proceedings of WWW 2010*, pages 401–410, New York, NY, USA. ACM.

[11] T. Hofmann. Unsupervised learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 2001.

[12] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of WWW 2006*, pages 533–542, New York, NY, USA. ACM.

[13] S. Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *Proceedings of WSDM 2010*, pages 281–290, New York, NY, USA. ACM.

[14] H. M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of ICML 2006*, pages 977–984, New York, NY, USA. ACM.

[15] C. Wang, J. Wang, X. Xie, and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, GIR '07, pages 65–70, New York, NY, USA, 2007. ACM.

[16] B. P. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of ACL 2011*, pages 955–964, Stroudsburg, PA, USA. Association for Computational Linguistics.

[17] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proceedings of WWW 2011*, pages 247–256, New York, NY, USA. ACM.

[18] X. Zhu, D. M. Blei, and J. Lafferty. TagLDA: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, Madison, 2006.