A Semantic Scan Statistic for Novel Disease Outbreak Detection

Kenton W. Murray [1]

Committee:
Daniel B. Neill (Adviser)
Chris Dyer
Roni Rosenfeld

August 16, 2013

[1]This work was funded in part by National Science Foundation Grants: IIS-0953330, IIS-0916345, IIS-0911032

A Master's Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Language Technologies
Language Technologies Institute
Carnegie Mellon University
July 1st, 2013

Acknowledgements

I would like to thank my committee for all the time and effort they put in advising this thesis. There were many long meetings and in-depth discussions over minute details and models that helped in getting this project to where it is. I very much appreciate all of the help.

I would also like to thank the many different students and faculty throughout the Language Technologies Institute whose helpful discussions and late night conversations helped shape and frame my thinking over the past two years. I appreciate all the help this provided and the intellectual growth it inspired.

Finally, I'd like to give a shout out to the members of the Event and Pattern Detection Lab. Thanks for the brainstorming and awesome presentations over the past two years.

Abstract

Anomalous pattern detection is a popular subfield in computer science aimed at detecting anomalous items and groupings of items in a dataset using methods from machine learning, data mining, and statistics. For anomaly detection tasks consisting of geospatially and temporally labeled data, spatial scan statistics have been successfully applied to numerous spatiotemporal data mining and pattern detection problems such as predicting crime waves or outbreaks of diseases [12, 7, 14, 15]. However, spatial scan statistics are limited by the ability to only scan over a structured set of data streams. When spatiotemporal data sets contain unstructured free text, spatial scan statistics require preprocessing data into structured categories. Manual labeling and annotating text can be time consuming or infeasible, while automatic classification methods that assign text field into a pre-defined set of event types can obscure the occurrence of novel events - such as a disease outbreak with a previously unseen pattern of symptoms - potentially drowning out the signal of the exact outliers the method is attempting to detect.

In this thesis, we propose the Semantic Scan Statistic, which integrates spatial scanning with unsupervised topic modeling to enable timely and accurate detection of novel disease outbreaks. We discuss some of the inherent challenges of working with free text data in an anomalous pattern detection framework, and we present some novel approaches to the problem using topic models by focusing on specifically adapting topic modeling algorithms to enable anomaly detection. We evaluate our approach using two years of free-text Emergency Department chief complaint data from Allegheny Country, PA, demonstrating the efficacy of the Semantic Scan Statistic and the benefits of incorporating unstructured text for spatial event detection. Using semi-synthetic disease outbreaks, a common evaluation method of the disease surveillance field, we show the ability to detect outbreaks of diseases is over 25% faster than current state-of-the-art methods that do not use textual information.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the world becomes increasingly digitized, more and more information is being stored as text in unstructured formats. This has a wide impact across many fields and disciplines within computer science, opening up many new and interesting research challenges. One of the most interesting aspects of this change is that datasets are getting increasingly varied and often exhibit mutliple different types of datasources and types. For instance, textual documents may also contain meta-data about time or location. On the other hand, structured datasets and databases are increasingly incorporating unstructured data sources.

Language Technologies are playing an increasingly important role as things become more digitized. Continuously, there are evermore areas that the field can impact and there are large potentials for huge contributions in data mining. In addition to information retrieval, extraction, and question answering; technologies and methods developed in machine learning, natural language processing, statistics, and data mining are impacting areas not traditionally thought of as language technologies. Areas that were once considered completely disparate from computer science are frequently intersecting the field. For instance, disease surveillance and public health have become evermore influenced by computational methods as records, and broad healthcare information have become digitized, while computers have broadly entered the healthcare profession. Not only do computers make it easier for health care professionals to accomplish their jobs and provide better quality care to patients, but new datasets are becoming availble to help improve data mining tasks outside of the industry.

Anomalous pattern detection is a popular and vibrant area of research

within Computer Science. Using Machine Learning, Data Mining, and Statistical Methods, the goal is to detect patterns of anomalous data in various data sources. In particular, one popular area is detecting anomalous subsets of a dataset - differentiating itself from outlier detection or simple anomaly detection. Yet as popular as the anomalous detection field is, significantly less work has been done on detecting anomalies in Natural Langauge Datasets - particularly spatial-temporal tagged free text. These increasingly availble datasets pose interesting challenges not normally encountered in normal natural language processing.

In this thesis, we discuss the utilization of unstructured, free text in an anomalous pattern detection framework. We focus on two main subsets of data mining and machine learning, topic modeling and spatial scan statistics. Topic models, which attempt to find latent, thematic structures in corpora of documents have become a very popular field of machine learning. Frequently, they are evaluated using metrics such as perplexity on held-out data. Less often, their value is measured through downstream tasks. Often, it is difficult to explicitly say how "good" a topic model is, and papers often present K-best lists of words in each learned topic as partial evidence of a well performing method. These lists, though useful for a reader to understand topic models, are not useful for evaluation between methods. By applying topic modeling algorithms to a new domain, we are able to show their efficacy and robustness through extrinsic evaluation - which is frequently not seen in the literature.

In additon to working with topic modeling, we make improvements to spatial scan statistics by removing the hardfast requirement to need structured data. Spatial scan statistics are a suite of methods that look at geospatial regions and find anomalous patterns. Most often, they operate by looking at counts of individual data streams, necessitating assingments to these data streams. We demonstrate the ability to incorporate unstructured, free text into our datastreams and increase detection power of spatial scan statistics.

We evaluate our methods using disease surveillance techniques. In particular, we test our algorithms by attempting to detect semi-synthetic outbreaks of diseases in real world hospital data. There is a lot of interest in this field as it poses both real world applications that are very tangible, but also presents a challenging research area which is just beginning to be investigated. We demonstrate the ability to work with potential outbreaks that have never been seen before, and to deal with data that is unstructured and lacks expert, manual labels.

In this thesis, we introduce a new suite of methods, the Semantic Scan

Statistic, that can incorporate topic models into spatial scan statistics. We discuss the challenges inherent with these methods and make improvements in both anomalous pattern detection and topic modeling. We begin with an in-depth discussion of these methods, then introduce the semantic scan statistic. We continue with a thorough discussion of our dataset and the challenges posed by a specific technical domain from a language technologies' perspective. Finally, we evaluate our methods using real world hospital data and demonstrate the efficacy of our methods.

Chapter 2

Background

Both topic modeling and spatial scan statistics are popular methods within machine learning and data mining that attempt to learn interesting structures from datasets. Topic models, frequently used in language technology applications, attempt to discover hidden mixtures of topics that describe a corpus of unstructured data. They model distributions of words using mixture models. Looking only at observed word counts in unstructured datasets, topic models try and find latent structures that explain the data. Spatial scan statistics aim to discern anomalous subsets and patterns within spatially located temporal data. Their objective is to determine if portions of a dataset cannot be explained by an baseline, underlying process and therefore may be potentially interesting. While extensions to topic models have attempted to incorporate time and spatial information into the models, no work has been done to detect anomalous, spatial-temporal regions. Likewise, spatial scan statistics have been extended to a variety of datasets, but have not been able to deal with unstructured datasets.

2.1 Spatial Scan Statistics

The spatial scan statistic was first presented by [10, 11] and is a powerful method for spatial surveillance problems. It detects clusters of anomalous data that are unexplainable by a baseline process. An extension of scan statistics, which attempt to determine if a point process is random, spatial scan statistics generalize to multiple dimensions such as spatial areas, multidimensional point processes, and varying sizes of the scanning window. They

are frequently used by the public health community for detecting spatial clusters of diseases such as breast cancer [12], lukemia [7], and West Nile [14]. Yet, they have been broadly applied to larger public interest datasets and other spatial-temporal, structured data sources such as crime detection [15].

Kulldorff's, [10], method initially looked at subsets of a spatial area using circles. A geographic area was examined exhautively using circles of varying radii with different centers. Records inside of the circles were compared to baselines to determine if they were anomalous. From there, spatial scan statistics have been expanded to look at a variety of other spatial areas such as rectangles [21], ellipses [13], and irregular shaped regions [5, 23, 24].

Spatial scan statistics monitor a set of spatial locations, s_i, each of which has a given observed count, c_i, and an expected count, b_i. They scan over regions, S, consisting of subsets of s_i's and maximize a likelihood ratio statistic. For a given region, S, an alternative hypothesis $H_1(S)$ is calculated which represents how interesting a cluster is in region S. This is compared to a null hypothesis, H_0, representing no anomalous clusters. A likelihood ratio $F(S)$ is calculated for a given region S, and is the ratio of the data likelihoods under the alternative and null hypothesis: In the most common, frequentist, hypothesis testing approach, this is given by:

$$F(S) = \log \frac{Pr(Data|H_1(S))}{Pr(Data|H_0)} \tag{2.1}$$

This is a relatively, simple, general algorithm that asks if the data can be explained by hypotheses. Alternative hypotheses are generated for each spatial region and compared to a baseline, null hypothesis. The null hypothesis is a baseline score for the region S, assuming that there is nothing anomalous going on. Frequently, this is simply an average expected count calculated from historic data for that region.

From this relatively basic principle, much work has been done on expanding spatial scan statistics in a variety of ways. From parameterizing spatial scan statistics, to determining different methods for looking at spatial regions, there is a large corpus of work in the field. Further extensions to spatial scan statistics use other scoring functions and to incorporate prior knowledge. The expectation-based poisson scan statistic focuses on regions with higher than expected counts as opposed to higher counts inside the region versus outside [22].

Many popular scan statistics are interested in models that can be parameterized. Parametric scan statistics such as [10] [22] [19], [16] assume a parametric model, such as Gaussian or Poisson distributed counts and maximize the log-likelihood ratio statistic $F(S)$ over all regions S.

Using the properties of the formulation of a particular scan statistic, various scoring functions can be defined compliant to equation 2.1. For instance, the expectation-based Poisson statistic log-likelinood ratio can be derived as [22]:

$$F(S) = \begin{cases} C \log \frac{C}{B} + B - C; & C > B \\ 0; & C \leq B \end{cases} \tag{2.2}$$

Where C is the aggregate counts for a region in a given time interval and B is the aggregate baseline. The counts are assumed to have been drawn from a Poisson distribution as individual counts are modeled through the use of a Poisson point process.

Likewise, Kulldorff's original statistic can be defined as:

$$F(S) = \begin{cases} C \log \frac{C}{B} + (C_{\text{all}} - C) \log \frac{C_{\text{all}} - C}{B_{\text{all}} - B} - C_{\text{all}} \log \frac{C_{\text{all}}}{B_{\text{all}}}; & \frac{C}{B} > \frac{C_{\text{all}}}{B_{\text{all}}} \\ 0; & \text{otherwise} \end{cases} \tag{2.3}$$

with C_{all} and B_{all} representing the total aggregate counts and baselines of all spatial locations [18].

In addition to the numerous different spatial scan statistics and many extensions from Kulldorff's original method, lots of work has been done on the theory of scan statistics. In particular, [18], provides a more theoretical treatment of spatial scan statistics over subset regions. It provides a proof of a class of score functions $F(S)$ that satisfy a property called 'linear time subset scanning' or LTSS. This property allows for extremely efficient unconstrained optimization over all subsets of the data. The framework requires ordering records according to a priority function, and searches over groups consisting of the top k higest priority records, requiring only a linear, rather than an exponential, number of subsets to be evaluated. [18] proves that this property, applied in a spatial setting, describes many commonly used methods including [10] [22] [17] and exponential [8].

2.2 Topic Modeling

Topic modeling algorithms are a popular set of methods for dealing with unstructured data and free text. In general, topic modeling algorithms attempt to fit a latent mixture of thematic topics to each individual document in a corpus. Each topic is a distribution over all of the words in a corpus and a document is represented as a mixture of these topics. Given a corpus of documents with only observed words, topic modeling algorithms attempt to learn posterior distribution of topics. One of the most well-known and basic topic models, Latent Dirichlet Allocation, or LDA, was proposed by [1] in 2003 and quickly became a common algorithm for classifying free text into topics in an unsupervised fashion. LDA, and other topic modeling algorithms, were an improvement over many other text classification methods because of the inherent property of allowing multiple topics to exist within a document and for words to have a probabilistically assigned likelihood of being generated from a specific topic.

LDA models a corpus, C, consisting of D documents, each with a potentially different number of words N, coming from the entire corpus' vocabulary V. The model assumes a generative process for a corpus where each document, d, has a mixture of topics, represented as a multinomial, θ, which is drawn from a Dirichlet. Each, word in the document has an individual topic assignment drawn from the multinomial θ, and then the word (w_{dn}) is drawn from a distribution over the vocabulary V conditioned on the topic selected (z_{dn}). The graphical model can be seen in figure 2.1. The overall probability of a corpus is given by:

$$p(C|\alpha,\beta) = \prod_{d=1}^{D} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta))d\theta_d \tag{2.4}$$

[1]

Given an observed corpus, the goal is to learn the posterior distribution of topics given the observed words in the corpus:

$$p(\theta,\vec{z}|\vec{w},\alpha,\beta) = \frac{p(\theta,\vec{z},\vec{w}|\alpha,\beta)}{p(\vec{w}|\alpha,\beta)} \tag{2.5}$$

LDA was first proposed using a variational inference algorithm to determine the set of topics that existed in a corpus based off the frequency of

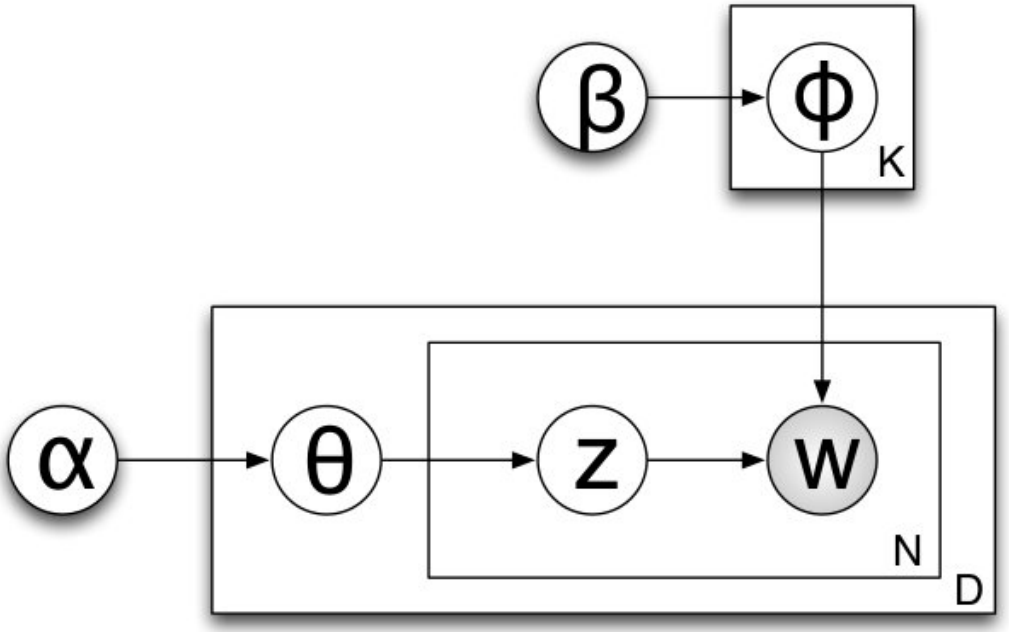

Figure 2.1: Latent Dirichlet Allocation as proposed by [1]. A corpus, C, consists of D documents of variable number of words, N. Each document has a set of thematic topics, θ that is a multinomial drawn from a Dirichlet distribution parameterized by α. Each of the K topics is a multinomial distribution over words drawn from another dirichlet distribution parameterized by β.

words in the observed documents. [6] propose a Gibbs Sampling methodology as an alternative to the variational inference method proposed in the original LDA paper. This has the benefit of easily being able to calculate the topic mixtures of all the documents, θ, and the distribution over the vocabulary for each topic ϕ, at any points during inference. This allows for easier extendability to more complex models.

2.2.1 Time Variant Topic Models

Attempting to modify LDA to account for topic shift over time is an active area of research. In many different settings, there are corpora that span large periods of time where a fixed, unchanging set of topics is not a realis-

tic assumption. Human language is constantly changing so limiting language modeling algorithms to be time invariant is frequently an unrealistic assumption. Numerous methods have been proposed to expand language models to incorporate additional information such as time, and the same is true for topic models. Many methods based upon LDA and other simple topic modeling algorithms have been proposed that extend topic models to allow topics to change.

In particular, [2], proposed Dynamic Topic Models which allow the Dirichlet hyperparameters, α and β, to vary over time using the Markov assumption with Gaussian Noise. t discrete time steps are chosen in advance. Topics "smoothly" evolve from the previous topics at time $t-1$. The authors evaluate their method on a corpus of the journal, *Science*, demonstrating the efficacy by predicting topics in future documents. The model is limited by its reliance on discrete time steps, the size of which needs to be fixed initially, and the impacts that different fixed values can have on the entire model.

Continuous time dynamic topic models, cDTM, extended dynamic topic models to remove the discrete time assumption by using Brownian motion [25]. The resolution of time steps is no longer required to be governed by predefined length time stemps, but is merely constrained by the time stamps of the documents in the corpus. In both the continuous and noncontinuous dynamic topic models, the Dirichlet hyperparameters, α and β, evolve over time, allowing the same topics to evolve. This differs from our method, presented later, as we do not let topics evolve, but rather allow new topics to form.

In contrast to the previous two methods, [26] propose a method where the topics are fixed but the topics' relative occurrences and correlations change over time. In other words, for a predefined number of topics, each topic has a fixed distribution of the likelihoods of words having been generated by that topic that is time invariant. In this method, the authors introduce a new hyperparameter, ψ, which describes a beta distribution that models time. The timestamp of a document, an observed variable, is drawn both from the topic distribution of the document, θ, and this beta distribution. The Gibbs sampling algorithm for this method is modified from the standard LDA Gibbs sampler so that in addition to drawing a topic variable, z_{di}, and the actual observed word, w_{di}, for every word in every document, a new timestep parameter, t_{di}, is drawn from Beta($\psi_{z_{di}}$) as well. The collapsed sampler approximates the model by drawing a timestep for each word in the document by using the latent topic variable z. Containing parallels to

our method as there are fixed topics and allowances for mixtures of topics to change, this method is also different from ours as it also does not allow for new topics, chich we believe to be a key aspect in an anomaly detection framework.

[3] extends basic topic models to utilize a temporal ordering of topics by learning a random topic initialization at t_0 and then allowing each time step to be based upon the previous slice in a manner similar to [2], but with an added parameter λ that governs how much the noise affects the change in topics between time slices. The model is reliant on user-specified parameters, which can be sensitive to tuning and require more work to determine good fixed values.

The multiscale dynamic topic model allows ϕ to vary over time slices by keeping track of sets of the empirical distributions of words on various time scales [9]. In other words, in addition to learning a standard LDA-based topic model at a given time slice, the model also keeps track of the counts of words for different time scales. The Dirichlet hyperparameter for the topics is adapted using the weighted sum of the empirical distributions over the different time scales. For this algorithm, instead of using variational inference or Gibbs sampling, the authors introduce a stochastic EM algorithm that sequentially updates the model only using newly observed data.

[27] modify basic LDA by allowing topic mixtures, θ, to vary over timesteps according to a Markov assumption, but keeping ϕ constant over time. Though similar to [2] this method does not allow α nor β to vary, merely θ. This method attempts to deal with continuous data streams which are assumed to be evenly sampled. This method attempts to solve the problem posed by allowing exchangability of timesteps in [2]. The generative story for this model is the same as LDA, being initialized at t_0 in the normal fashion, but then at each time step, θ_t is sampled from a multinomial distribution with expectation θ_{t-1}.

2.2.2 Nonparametric Topic Modeling

The most basic topic models such as LDA, are parametric; there are a fixed, predefined, number of topics. Unfortunately, parametric models have the downside that knowing the "correct" number of topics in a corpus is often fraught with trial and error and is frequently incorrect. Much work has been done on expanding these methods to nonparametric topic models, or models where the number of topics is also learned from the data. Most frequently,

Dirichlet Processes are used to determine the number of topics. Numerous representations of the Dirichlet Process exist, such as the Polya Urn Scheme, Stick Breaking, and Chinese Restaurant Processes. These are ways of viewing the problem as generalizations of clustering-like models to distributions over countably infinite number of parameters.

One way to view the dirichlet process is through the Chinese Restaurant Process metaphor. In this metaphor, customers come into a restaurant with an infinite number of tables. A customer chooses a table proportionally to its popularity.

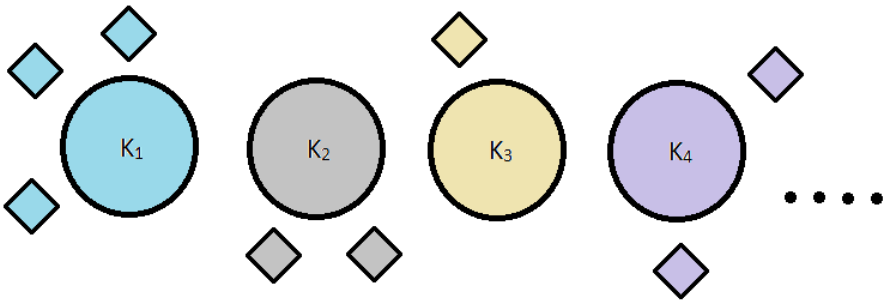

Figure 2.2: The Chinese Restaurant Process Metaphor. Here we have 4 occupied tables with 8 customers. The probability of assignment to tables K_1, K_2, K_3, and K_4 are $\frac{3}{8+\alpha}$,$\frac{2}{8+\alpha}$,$\frac{1}{8+\alpha}$, and $\frac{2}{8+\alpha}$ respectively. The probability of a new table is $\frac{\alpha}{8+\alpha}$.

As a table gets more popular, the probability of a new customer sitting at that table increases. This is often referred to as the rich-get-richer dynamic, and it prevents a nonparametric model from continuously increasing the number of parameters. More formally, the process can be viewed as:

$$Pr(z_i = k) = \begin{cases} \frac{N_k}{\sum_j N_j + \alpha} & \text{for an existing table} \\ \frac{\alpha}{\sum_j N_j + \alpha} & \text{for a new table} \end{cases} \tag{2.6}$$

At any given step, there is always the probability of selecting a new table dependent upon α. It will never be non-existent, but will decrease as the sample size increases.

Chapter 3

Methods

The **Semantic Scan Statistic**[1] melds scan statistics with topic modeling to incorporate unstructured data into an anomalous pattern detection framework. The key insight is that unstructured data, such as free text, has natural clusters within a collection that are indicative of similarity and thus representative of a datastream from a spatial scan statistic vantage point. From a generative model perspective, there is a hidden, latent set of themes pervasive in a collection of individual records. Each individual record has a mixture of some portion of these themes. When dealing with a standard corpus of natural language documents, these themes, or topics, represent the general ideas being discussed in each individual document. The semantic scan statistic views these as topics from an event detection perspective where each topic represents a potential datastream of interest. It is a standard, noisy-channel model where the true datastream has been obfuscated through the use of natural language such that it no longer exhibits an explicit, observable, expert label.

For instance, in the domain of disease surveillance, our notion of each individual topic is that of a class of similar diseases, which are modeled by the distribution over words given by patient descriptions. In other words, each disease presents symptoms. These symptoms are then expressed to a health care provider through human use of language and transcribed. Under the assumption that similar descriptions of symptoms are informative with regards to the diseases exhibited, useful information can be learned to structure our dataset using language models. The noisy channel model here is

[1]Much of this work is based upon a simpler version of the semantic scan statistic presented as a one page abstract in 2011 [28].

that a syndrome has gone through a noisy channel and instead of observing a disease, we observe words that describe symptoms.

The semantic scan statistic is a versitile and robust set of methods that is applicable to any geo-spatially tagged dataset containing unstructured data. It is not necessarily limited to only language, but applies to any spatialtemporal data with a latent, thematic unstructured data compenent. In this thesis, we present the semantic scan statistic for disease surveillance using written, free-text, medical records, but it is a more general set of methods for anomaly and event detection for unstructured data.

In this chapter, we discuss the semantic scan statistic and various ways of incorporating language models, and in particular, topic models, into a spatial scan statistic framework. We begin by discussing how these methods can be combined. We then discuss assigning records into datastreams using previously learned language models. We continue by talking about how changing the size of our corpora and modifiying simple topic models can be used for anomalous pattern detection. We conclude the chapter by exploring additional extensions to our basic models through nonparametric methods and modifications to preprocessing our data.

3.1 Topic Modeling in the Semantic Scan Statistic

As discussed previously, spatial scan statistics rely on time series analysis over streams of labeled classes of data in specific spatial regions. When dealing with structured datastreams, with easily defined counts and assignments to classes, this is a well-defined problem where significant prior work has already been done. In an unstructured context, such as text, the problem is less defined as it is non-trivial to declare what are individual data streams, and how free text maps to streams. The semantic scan statistic, uses unsupervised topic modeling to naively learn a set of topics over a dataset, then classifies individual free text items to a topic. Each topic is treated as an individual datastream, which allows for the application of standard spatial scan statistic methodologies.

After learning a topic model with K different topics for a corpus, each individual document is assigned a single topic most representative of it. The choice of assigning a single topic to a document, even though topic models are

mixture models, is dependent on the domain of the dataset and the potential anomalies of interest. For example, in a disease surveillance setting, individual disease cases taken from Emergency Department Chief Complaints in the dataset are by definition, individual diseases. Thus, from a generative story, potentially all the free text describing a case in a document was generated by the same class of diseases, represented as a topic. Many terms in the vocabulary describing a particular disease may also describe completely unrelated diseases, so the ability to model them initially using a mixture model will allow for greater flexibility during inference. There are other event detection domains where only having a single topic does not fit into the generative story, so language modeling methods used in the semantic scan statistic are inherently mixture models for greater generality. In this work, a single topic is chosen for each document due to the disease surveillance domain used in the evaluation, but broader applications of the method may use mixtures of topics. [20] have shown how spatial scan statistics can be used for multiple datastreams.

Having classified each individual document to the most likely topic, a standard spatial scan statistic is used. For every subset location, s_i counts are calculated based upon the assignments of documents to topics. The total number of datastreams is K, as each topic represents a datastream. In our experiments, we use the Expectation-Based Poisson scan statistic to calculate expected count and observed counts, then compare scores $F(S)$. The basic overview of the algorithm can be seen in Algorithm 1.

Algorithm 1: General Framework for Incorporating Unstructured Data into Spatial Scan Statistics

```
Learn a Language Model for Domain;
Assign labels to Free Text based upon Language Model;
Treat each label class as a datastream;
for each time window, t do
    for each subset of S, s_i do
        for each datastream do
            Calculate: F(S) using Language Model;
Return Most Anomalous F(S);
```

3.1.1 Subsets of the Corpus

Choosing a corpus of documents to train topic models on for spatial scan statistics is a non-evident task. As datasets are constantly changing with the addition of new documents in a spatialtemporal setting, looking at the full set of documents available may not be the most informative corpus. In this work, we look at the impacts of three different sets of data for a corpus. The main dataset to consider are "Recent Documents", which is a period of data leading up to the current time window in the scan. This is needed for the spatial scan statistic to adequately calculate a baseline for the null hypothesis. Generally, this is too large of a corpus to effectively model the interesting signals of potential interest since we are attempting to detect anomalous patterns occurring in close temporal proximity with the current time window.

Detection power may potentially decrease for any language model trained on a large corpus of data with only a small percentage of the data containing the actual anomalous information. The relative occurance of terms indicative of anomalies to the overall size of our corpus decreases with increased amounts of data, so the signal in the data may be harder to detect. We thus, also look at training on a subet of just "Current Documents". The scope of this set is dependent on the task at hand, so it is reliant on expert, domain knowledge to choose the proper scale. For instance, in a disease surveillance setting, there has been much research done on modeling disease outbreaks, how they progress over time, and the characteristic lengths of time they evolve over. From expert knowledge of the normal progression of these periods, an adequate timeslice to train the language model on can be selected. Note that while spatialtemporal scan statistics scan over both regions and time, the semantic scan statistic fits a topic model using the total number of documents in the entire spatial area for a given time, as subsets are potentially too sparse to learn a useful model.

In addition to using only documents temporally close to the current time window, we also consider a purely historic set of the data, which we will refer to as "Historic Documents". The intuition behind this is to prevent overfitting on recent data and gather information about larger trends for a more informed topic model.

In this paper we show experiments learning topic models on various combinations of "Historic Documents" and "Current Documents". "Recent Documents" is the dataset over which the method scans to calculate both base-

lines and potential subsets of interest. No additional language modeling is done on our "Recent Documents", but rather all documents in it are classified using a previously learned topic model from some combination of "Historic" and "Current Documents". Preliminary experiments conducted using language models trained on "Recent Documents" had less detection power. In all cases, "Current Documents" are a strict subset of "Recent Documents" and both are completely disjoint from the "Historic Documents", which are much older.

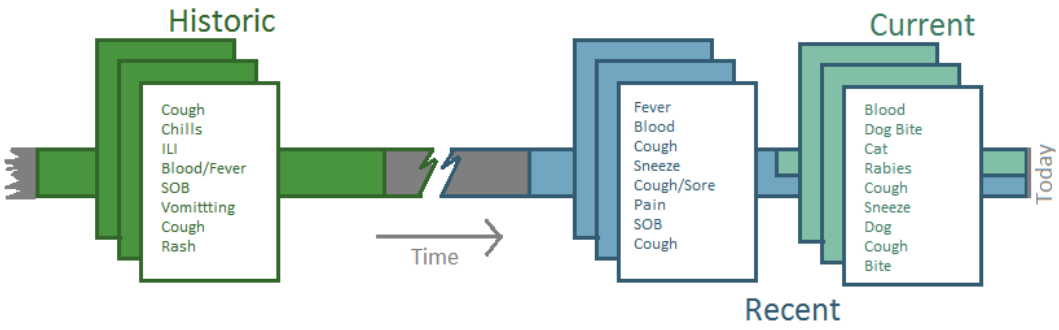

Figure 3.1: Different Corpora used for Different Topic Models in the Semantic Scan Statistic.

For the first portion of the thesis, we present the semantic scan statistic using three different topic modeling derived algorithms and evaluate the relative detection power across them. The algorithms differ primarily in how they mix historic and current data in the topic models and the sets of topics inferred from the corpora.

3.1.2 Incremental Topic Modeling

A challenging aspect when dealing temporally with text is that it often encompasses previously unseen data, meaning language models in general will not necessarily represent the new data well as some word types may never have been observed. In terms of topic modeling, new documents may not be represented well by existing topics (distributions over words), and thus pertinent information about the document may not be modeled as the likelihoods of words occuring may be incorrect. When incorporating topic models into spatial scan statistics, inferior models limit the detection power as signal is lost in the noise of improper classifications.

We therefore investigate three different topic modeling algorithms to see which one has the highest detection power for previously unseen data. The first method, the Static Topics Method, learns an initial set of topics solely from the "Historic Documents" and never varies. The other two methods incrementally modify topics to better fit new information and data while simultaneously attempting to prevent overfitting by using historical data.

3.2 Online Assignment

Online inference for new documents in topic models is a non-trivial problem and in our problem formulation, common pitfalls can have a more profound impact as interesting signals can be lost due to bad assignments. From an anomalous pattern detection vantage point, frequently, the most interesting aspect of a dataset can have a very low probability of occuring normally. Dimensionality reduction techniques such as LDA can drown out interesting aspects of infrequent terms. Due to random initializations, low term frequencies, and short document lengths, there is much more likely chance of initial random assignments of words to a topic to have a profound impact.Common methods such as resampling over the dataset will not necessarily perform well given the relative frequency of potential areas of interest.

We spent a lot of effort looking at ways to conduct online assignments. In a document, given a distribution over words for a topic, ϕ, and topic mixtures in a document, θ, transitioning from a mixture model where a document exhibits multiple topics, to one where a document is only assigned one is not evident. Common methods, such as summing the probabilities for a given topic for all words and taking the maximum did not perform well. Though one word may be heavily biased towards a specific topic, it may not be the most important word in a document. Thus, its high confidence of assigning to a specific topic can negatively impact performance. Likewise, multiplying probabilities may have a similar impact. There are potential issues with rare words significantly impacting the assignment due to very small probabilities.

From an anomaly detection perspective, we are interested in finding the most anomalous portion, so all words in a document matter, and the highest likelihood ones may actually matter less. In order to try and increase the impact of all words in a document, we also experimented with rerunning a Gibbs sampler over all documents again and figuring out which topic mixture was the greatest. Interestingly, this hurt our performance when evaluating a

downstream task of number of days to detect a disease outbreak.

Eventually, we settled upon this E-M inspired method to compute the assignments of topics given in algorithm 2.

Algorithm 2: Online Document Assignment

```
initialization;
θ_1 = ... = θ_K = 1/K;
while not Converged do
    for each word w_i do
        for each topic k do
            compute := Pr(z_i = k) ∝ φ_ik θ_k;
        Normalize: Σ_k Pr(z_i = k) = 1;
    Recompute θ;
    for each topic k do
        compute := θ_k ∝ α + Σ_i Pr(z_i = k);
    Normalize: Σ_k θ_k = 1;
Assign entire document to maximum θ_k ;
```

This assignment is done after we have learned ϕ for a topic model, values that represent the probability of a word given a topic. Learning these values took place during inference, and at this point, we do not allow these to vary at all. From here, it is a deterministic topic assignment for each document using algorithm 2. The classification of these documents may or may not be for the same set of documents we learned our topic model on. In other words, given a language model over a vocabulary for multiple different classes (topics), we assign documents by looking at the probabilities of a word in each class, for every word in the document, for any set of documents given (new or previously seen). In this algorithm, ϕ_{ik} represents the probability of a word indexed by i in topic k. θ_k is the mixture of topics in the document, but this is not learned during inference of our topic model, rather it is uniformly initialized for every document. As our ϕ values do not vary during assignment and since we uniformly initialize our topic mixtures, this is a deterministic algorithm that guarentees identical documents get assigned to the same topic.

3.3 Incremental and Static Semantic Scan Statistics

3.3.1 Static Topics Method

The first variant of topic modeling explored in the Semantic Scan Statistic is the Static Topics Method. In this method, a set of topics, K, are learned initially using a corpus of previously available training data, which we refer to as "Historic Documents". This method is a standard LDA topic model where posterior topic distributions are learned from a fixed, historic corpus. The model never changes during the scanning portion. Every document, d, in any subset of the spatial region S that is scored is classified using the online algorithm. Words that are not seen in training data were merely ignored when classifying cases. This method serves as our simplest baseline to demonstrate the efficacy of topic modeling when using scan statistics.

This method has some issues when encountering previously unseen data. In the case of a novel outbreak, it is conceivable that the vocabulary used was not in the training corpus. This has the potential for new, novel documents containing out-of-vocabulary (OOV's) to be less likely to be classified into meaningful topics, making any interesting signal undetectable. Additionally, there is a high likelihood that words in a new document will exist in the historic corpus, but not co-occur, making it less likely that they are assigned to the same topic. To the best of our knowledge, no previous work has been done on incorporating naive textual information into scan statistics. Therefore, even though we did not expect high detection power for the Static Topics Method, we use this as a baseline and propose two other topic modeling algorithms with better detection rates.

3.3.2 Dynamic Topics Method

The next method we call the Dynamic Topics Method, which retrains the set of topics frequently. Looking at a 14-day moving window up to most recent day, topics are simply learned using those limited cases, which we refer to as "Current Documents". No previous data is included in this method. It is very similar to [2], where the Markov assumption is assumption is broken and there is no dependence on the previous set of topics. In our method, at each timeslice, a completely new set of topics is learned. These topics are then used to classify cases in the scanning region, denoted "Recent Documents",

to one of the topics using the online classification algorithm. Remember that our "Current Documents" are a subset of our "Recent Documents" as can be seen by refering back to figure 3.1. The cases in the spatial region that are being considered are a superset of the data used to train this topic model.

This method is potentially prone to short-term fluctuations, increasing the possibility of overfitting the alternative hypothesis on clusters of data that are not actually anomalous. By allowing the topics to vary, we have compensated for the issues presented in the Static Topics Method, but this method does not encorporate any previously seen data. Since the distribution of topics are recalculated every day, emerging patterns have the potential to quickly appear in a single topic. Due to long term, systematic fluctuations in datasets, this may not be ideal behavior. For instance, in a disease surveillance setting, it is well known that certain infectious diseases are more prone during different time periods (flu season). By not using historic data, this knowledge may not be incorporated and overscored or over-fitted. An important point to note with topic modeling algorithms is that a topic is merely a distribution over words with no identifiability. Thus, when we learn a new set of topics everyday, there is no guarentee of consistency across the topics. There is a completely new model every day that may be completely unrelated to the previously learned models and not incorporate useful prior information. As data between consecutive days has a significant overlap, topics learned under succesive models may be similar, but due to random initializations, the numbering of topics will likely be different. In other words, Topic 3 in one model may be very similar to Topic 7 in another model, just the numerical identifications may differ.

3.3.3 Incremental Topics Method

In order to allow the topics to vary slightly over time and respond to previously unseen data without overfitting, the third method, the Incremental Topics Method, allows topics to be inferred from both historic and current data. Rather than resampling just the most recently seen data, as is the case with the Dynamic Topics Method, the Incremental Topic Method also takes into account the topics learned in the Static Topics Method. The goal of this is to detect novel outbreaks without being constrained by the current 14 day timeslice of data.

This algorithm consists of a two-step method. In the first, a set of K topics is learned from the "Historic Documents" using a purely LDA model.

Then, using a corpus of "Current Documents", a new set of K' topics is learned, but the total number of topics is $K + K'$ and the distributions in K are not allowed to change.

In this method, a set of static topics is learned from previously available training data, identically to the static topic method above. Then, a new corpus is used to calculate the posterior over topics, but some of the toics are fixed. Rather than just learning a new set of topics using Gibbs sampling for the dynamic topics, a modified Gibbs sampler runs so that the counts of words assigned to the existing static topicsis held constant, but assignments of words are allowed to vary for the new set of topics.

Though the incremental method is different from both the static and dynamic, it can easily be reduced to either. If K is set to 0, there are no previously learned topics and the model reduces to LDA. The corpus chosen is our "Current Documents" and the model is identical to the Dynamic Topics Method. Likewise, if K' is 0, no topics are learned using current data and the method reduces to LDA. In this sceneario, the corpus chosen is our "Historic Documents" and the model is identical to the Static Topics Method. Yet, when both K and K' are non-zero, the topic model is no longer LDA as it contains, fixed observable distributions of topics during portions of the inference on the posterior.

The intuition behind this method is that for emerging topics, we do not want the set of topics to vary as smoothly as in the dynamic method as that can be prone individual fluctuations in the data and lose the signal of anomalous events. Additionally, we did not want the set of topics to be as fixed as the static method as that will likely fail on previously unseen textual data. Instead, the goal is to maintian a slightly similar set of topics similar to the static topics where normal data can still be well classified but also allow a set of incremental topics to capture new patterns in a corpus.

This method is designed to be very different than other topic modeling methods with changing topics, as the goal of our method is to have emerging topics be different from existing ones and be useful for anomalous event detection. The majority of the other models with changing topics, as discussed in the previous chapter, attempt to smooth the transition of topics over time. Often using a Markov property assumption, these models do not want too much change in the topics from one timeslice to another as that can impact the intuition of what the set of topics means. On the other hand, we also do not want too much change in all of our topics over time, but we would like emerging topics to be readily identifiable.

3.3.4 Gibbs Sampler

Exact inference on LDA is intractable. A variety of methods have been proposed to estimate the posterior distribution of the model. The initial paper proposing LDA used a variational inference approach, but Gibbs Samplers are also very common. Due to the nature of our incremental topics and the extension from LDA, we decided to use a collapesed Gibbs sampler based on [6].

The equation for the sampler given in [6] is:

$$P(z_{i,j} = k|\vec{z}_{\neg i,j}, \vec{w}) \propto \left(\frac{n_{\neg i,j}^{(w_i)} + \beta}{n_{i,j}^{(.)} + V\beta}\right)\left(\frac{n_{\neg i,j}^{(d_i)} + \alpha}{n_{\neg i,.}^{(d_i)} + K\beta}\right) \tag{3.1}$$

In Equation 3.1, $\vec{z}$ is a vector of topic assignments for all words in all documents, indexed by i and j respectively, with k indexing topics. α and β are fixed hyperparameters describing the Dirichlet distributions. V is the size of the vocabulary and K is the number of topics. n is an integer count of the superscript (w_i number of occurances of the word indexed by i in the corpus, and d_i being the occurances in a single document), with subscripts of the form, $\neg i, j$, being values to be excluded from the count.

The implementation is:

Algorithm 3: Gibbs Sampler for Latent Dirichlet Allocation from [6]

```
initialization;
for each document d_j do
    for each word w_i do
        Randomly Assign z_i ∈ [1, K];
while not Converged do
    for each document d_j do
        for each word w_i do
            Remove current assignment, z_{i,j};
            for each topic k do
                Compute Pr(z_{i,j}) using 3.1 ;
            Sample a topic ;
            Add current assignment, z_{i,j}^{new};
```

For the dynamic and static methods mentioned above, the posterior of the topic models can be learned simply by using this sampler. The only

difference between the two is the underlying corpora. For the incremental method, the sampler changes significantly. Initially, the model learns a posterior distribution over a set of topics, K, using the "Historical Documents". Another posterior is learned for the K' topics on the corpus of "Current Documents". Finally, a modified sampler is run over all the topics, $K + K'$, using only the corpus "Current Documents". Only the K' topics are allowed to vary, keeping the historical information from the original topics, but allowing that information to influence new ones.

Generally, seeding a Monte Carlo Markov Chain Sampler is done randomly, and over time, the samples approach the true posterior. In our work here, we follow this common practice for the static and dynamic methods, but do not initialize the incremental this way. In the first step, the portion where we learn a set of dynamic and a set of static topics, we randomly initialize the sampler. After learning the posteriors for both models, we use the information given as the initialization of the incremental method.

The sampler has changed in the following ways:

Algorithm 4: Updated Gibbs Sampler for Incremental Topics

```
initialization;
for each document d_j do
    for each word w_i do
        Assign z_i probabilistically ∈ [1, K] using posterior predictive
        distribution of φ from earlier model;
while not Converged do
    for each document d_j do
        for each word w_i do
            Remove current assignment, z_{i,j};
            for each topic k do
                Compute Pr(z_{i,j}) using 3.1 ;
            if current_assignment, z_{i,j} ∈ static_topics then
                Re-Add current assignment;
            Sample a topic from new distribution ;
            if new_assignment, z^{new}_{i,j} ∉ static_topics then
                Add new_assignment, z^{new}_{i,j};
```

Note that the counts for the incremental topics may change, but not for the static set of topics as we readd the count back to the initial topic. We

may then also increase a count for an incremental topic if our sample draws it, meaning that the total number of counts may vary. This is due to the fact of how a topic is defined. The posterior predictive distribution of ϕ, which is defined as:

$$\hat{\phi}_k^{(w)} = \frac{n_k^{(w)} + \beta}{n_k^{(\cdot)} + V\beta} \tag{3.2}$$

is completely dependent on the counts regarding the number of times words have been assigned to a specific topic. Thus, changing those counts for the static topics during our sampling would allow the topic to vary. Note that this does mean from the sampler's perspective, the number of words in our corpora changes - going up or down as words are assigned to the incremental topics. Again, the intuition behind this is that we are attempting to detect novel, emerging patterns while using historic data.

3.4 Nonparametric Topic Models

One of the biggest challenges when dealing with any clustering problem is determining what is the correct number of clusters. As discussed above, the goal of our work was to detect clusters of anomalous diseases. In particular, we are attempting to cluster anomalous cases into a topic and determine that the topic is anomalous. The goal of our incremental method was to have a set of topics that would be more peaked and biased towards anomalous cases or an emerging trend. By using nonparameteric methods, we are hoping a similar thing happens - the potential to create a new topic when there is something anomalous. It is approaching the same problem as before, just from a different direction.

3.4.1 Document Level Nonparametric Topic Model

The first nonparametric model that we tried aimed to preserve a document level mixing of topics. One of the most beneficial aspcets of Latent Dirichlet Allocation is that each document is represented as a mixture over topics. The same set of topics is shared across documents - meaning that different documents may exhibit some combinations of the same topics, just in different proportion. Our first nonparametric model attempted to keep the property where every document has a different mixture of the same set of topics.

$$P(z_{i,j} = k|\vec{z}_{\neg i,j}, \vec{w}) \propto \begin{cases} \frac{n_{\neg i,j}^{(w_i)}+\beta}{n_{i,j}^{(.)}+V\beta} \frac{n_{\neg i,j}^{(d_i)}+\alpha}{n_{\neg i,.}^{(d_i)}+K\beta} & \text{for all existing topics} \\ \frac{n^{(w_i)}}{\sum n_i} \frac{\alpha}{n_{\neg i,.}^{(d_i)}+K\beta} & \text{for creating a new topic} \end{cases} \tag{3.3}$$

Note that the equation for the standard, parametric, LDA Gibbs Sampler is

$$P(z_{i,j} = k|\vec{z}_{\neg i,j}, \vec{w}) \propto \hat{\phi}_k^{(w)} \hat{\theta}_k^{(d)} \tag{3.4}$$

simply by removing the current assignments.

In our nonparametric equation 3.3, we do not have a distribution over terms, $\hat{\phi}$ (or the similar equation remvoving the current count), for a new topic. As it was not entirely obvious what the new topic's distribution may look like, we used the empirical distribution of the entire corpus as $\hat{\phi}_0$ initially.

$$\hat{\phi}_0 = \frac{n^{(w_i)}}{\sum_i n^{(w_i)}} \tag{3.5}$$

Unfortunately, this model had a high propensity to keep creating new topics. Through our implementation, we had set up an upper bound on the number of topics equivalent to the number of topics in our parametric models. Thus, the probability of creating a new topic when that threshold was reached was 0. This caused our model to exhibit one of the standard behaviors of a Dirichlet Process - a rich get richer. Yet in our formulation, almost all of the probability mass was assigned to one topic, making for a very uninteresting model with low detection power for anomalous events. As this was not producing any useful results, we tried modifying the topic assignment equations slightly. In an attempt to spread the probability mass across topics more evenly, we changed the method to randomly assign a topic once the upper threshold for number of topics had been reached - if the choice had been to create a new topic. In practice, this had the somewhat expected result of merely randomly assigning topics without any interesting properties.

In order to be consistent with other methods, we constrained the model to an upper number of topics. This model turns out to be equivalent to the model in [4], but instead of ∞ topics, we set an upper threshold to K.

Inference on this model was intractable as well. Therefore, we extended the Gibbs sampler from [6] once again, but modified our calculation of θ to use the Chinese Restaurant Process.

Unfortunately, the propensity to create a new topic was very high as many of our short documents did not share common words. Having set an upper limit of topics equivalent to our parametric LDA models, the majority of cases were assigned into one topic. Experiments to change this by reverting back to a normal, parametric Gibbs sampler after the upper limit of topics had been reached did not perform well either, neither did random initializations.

3.4.2 Corpus Level Nonparametric Topic Model

After the previous nonparameteric method did not work, we looked at using a global mixture of topics. Instead of a document level θ, we had a corpus level one. This way when a new topic was created it would be easier for a new topic to also exhibit words in other documents. Unfortunately, this has the result of requiring either one topic assignment per document (Naive-Bayes) or requiring all documents to be of length one. We decided to use documents of length one to demonstrate that unigrams have some detection power, but topic models are better.

3.5 Character N-Grams

The dataset, as described more in detail in Chapter 4, is very noisy and full of misspellings. This poses an interesting challenge from a language modeling perspective. Misspellings change the distribution of words in a corpus by increasing the number of potential out-of-vocabulary (OOV) terms in a test set.

Normally, in a language technologies perspective, when dealing with sparseness, data smoothing techniques are applied. These methods help deal with infrequent terms and make models more robust. One of the challenges with the problem of anomaly detection is that very rare occurences may actually be of interest. It is beneficial to have expected baselines for even rare terms that could potentially just be misspellings.

In order to smooth our dataset, without loosing information, we looked at N-Gram character subsets. This has an interesting property of constraining the total vocabulary size. For instance, choosing an N-gram of size 3 has 27^3 total possibilities in the vocabulary assuming only the English alphabet without capitalization or numbers while also allowing whitespace to map to a single character. The aim is that small misspellings will have a smaller

impact on the overall results. If the word "cough" is spelled "couugh" the goal would be to have the extra "u" matter less as "cou" and "ugh" both exist.

Unfortunately, one of the downsides of splitting a word into character n-grams is that there is not as nice of a generative story. LDA, and other topic models, have a nice property that a generative story describes in detail how a specific word was created under the model. For instance, in LDA, the generative story is that each document has a distribution over topics chosen, θ, and from this distribution, each word draws a topic assignment variable z. Based upon this topic assignment, a word is drawn from the corresponding topic, ϕ, which is a distribution over all words in the vocabulary. Thus, there is a simplified world view of how documents in a corpus are generated - by choosing a set of topics for a document and then choosing words probabilistically according to the topics represented in a document. By splitting words into groupings of characters, the nice language-motivated model breaks down. Instead, the generative story is that for each document in a corpus, a set of topics is chosen. Then, for each n-gram observed, a topic is chosen for that substring and the n-gram is generated conditioned on that topic. There is less of a compelling generative story here, and more of a non-theoretically supported tweaking, yet it does not impact the semantic scan statistic as long as we do not attempt to generate new data from our topic model. We briefly discuss how this impacts the semantic scan statistic in our evaluations.

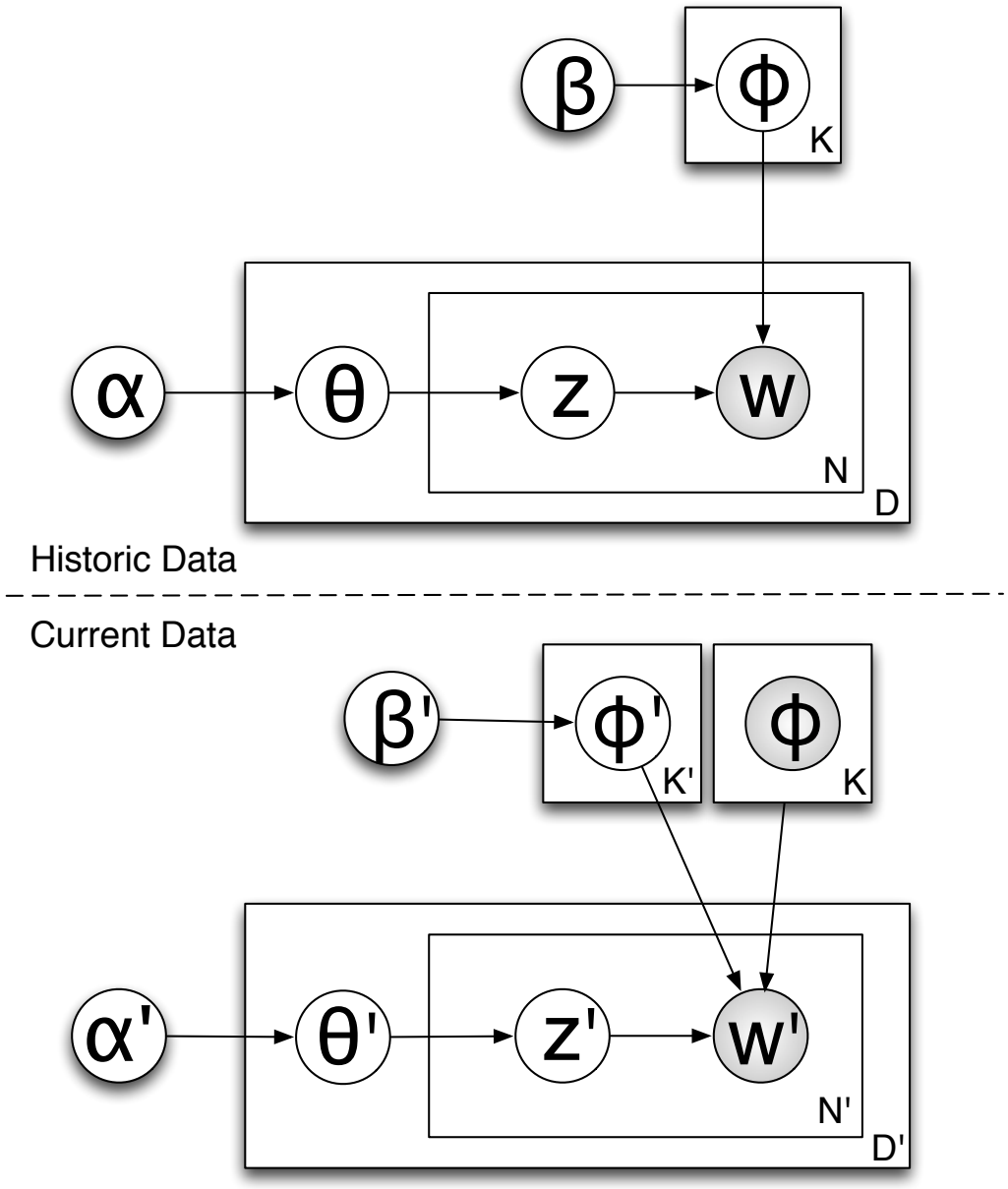

Figure 3.2: Semantic Scan Statistic Topic Model. The static topics method is the top portion of the figure where a set of K topics is learned only on historical data. The dynamic topics method is the bottom portion of the figure where K' topics are learned using current data and $K = 0$. The incremental topics method is a two-step process when K topics are learned using historical data and K' topics are learned on current data. After initialization learning both K and K', ϕ_i is fixed $\forall k \in K$ and are an observed set of varaibles in the resulting model's inference.

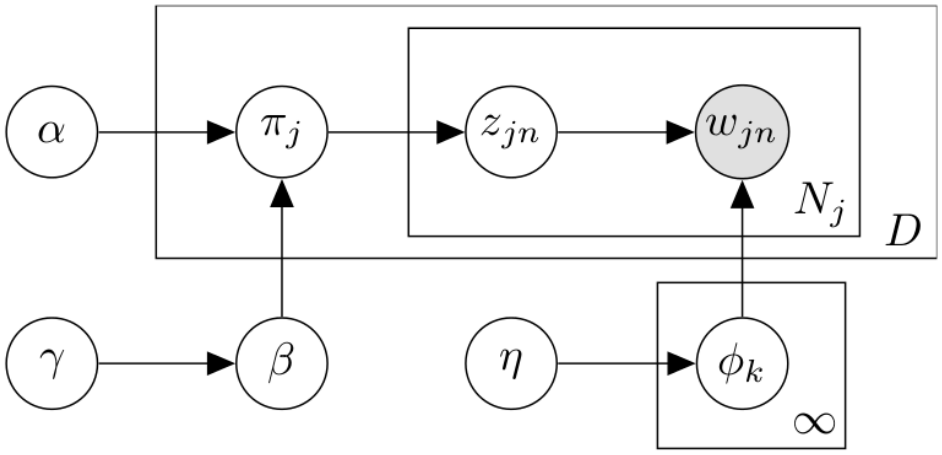

Figure 3.3: Nonparametric LDA [4]

Chapter 4

Dataset

In this thesis, we worked with a spatiotemporal dataset comprised of hospital records. Obtained from a networked healthcare provider in Western Pennsylvania, it is a useful dataset for disease surveillance - a common application of anomalous pattern detection. The combination of a unique domain, very noisy data, geospatailly annotated data, and the goal of detecting anomalous patterns, necessitates an in-depth discussion of our dataset and its properties. In particular, the dataset contains both a date and geographic location for every record, along with some expert labels. In addition, each record contains noisy free text, with little grammatical structure, frequent misspellings, and minimal standardization across records.

The dataset comprises hospital data from the University of Pittsburgh Medical Center (UPMC) health care provider network in Allegheny County, Pennsylvania. We use the dataset to evaluate the semantic scan statistic. Using semi-synthetic injects, a common practice in the disease surveillance literature, we model outbreaks of diseases in spatial regions. We evaluate the semantic scan statistic's ability to detect when an outbreak is occurring.

In this chapter, we discuss the dataset, the methods for using this in an anomaly detection framework, and discuss the interesting language technology challenges inherent with working in this domain. We begin with a discussion about the dataset itself, and in particular, the partial structure given it by the medical community. We continue by explaining how semi-synthetic injects are created for evaluation and the relationship they have to the disease surveillance community. Finally, we conclude this chapter by discussing the structure of the language in the corpus, including descriptive statistics and unique aspects of interest to a language technologies domain.

Table 4.1: Example Case Formats

Date	Location	Adm_Reason	ICD-9	Prodrome
01.01.2004	15213	COUGH AND NAUSEA	789	9
02.03.2004	15232	BLEEDING	556.3	6
08.17.2004	15217	BLEEDINGPERPATIENT	444.3	3
01.23.2005	15216	ILI	789	9
07.04.2005	15232	ETOH	421	6
12.13.2005	15235	VERY CONGESTED	487.1	9

4.1 Hospital Data

We evaluated our methods using hospital emergency department (ED) data collected from ten Allegheny County, Pennsylvania hospitals from January 1, 2004 to December 31, 2005. The dataset contains every emergency department visit to one of those ten hospitals after being anonymized to remove any personally identifiable information. In total, the data consisted of ∼340K records of individual ED visits, each of which contained five attributes: Admission Date, Home Zipcode, Chief Complaint, International Classification of Diseases-9th Edition (ICD-9) code, and prodrome. They appear similar to the records given in Table 4.1. The first three attributes are populated upon a patient's admittance to the ED, and are directly used by our semantic scan methods; the last two attributes are generally not populated until the patient's discharge, and we use these attributes for evaluation and comparison purposes only. The closest methods to our own utilize the last two attributes, which can be populated days after a patient is first admitted and after the other fields are entered into a Electronic Health Record's database.

The chief complaints field of the dataset (also referred to as "Admission Reason" and "Adm_Reason"), is a free-text field recorded by a triage nurse upon a patient's admittance to the ED. Chief complaints are generally short (a few words or a phrase, such as "pain in rt arm"), have little grammatical structure, and are very noisy (with frequent misspellings, inconsistent use of terms and abbreviations, etc.). They pose an interesting challenge from a language technologies perspective as they contain significant amounts of information in a small number of words, and additionally, frequently lack structure commonly used in many tasks.

The International Classification of Diseases-9th Edition Codes, or ICD-9 Codes, are standardized codes used to manually classify diseases and ail-

ments into specific groups. They are primarily used for billing insurance companies in the United States, but also serve a beneficial, unintended, secondary purpose of applying some minimal structure to our dataset. There are hundreds of these ICD-9 codes, many of which are not represented in our corpus. UPMC has implemented an internal system for their hospital network that maps many of the codes into broader syndromic categories referred to as "Prodromes". These are hard assignments, reliant on experts manually defining which ICD-9 codes map to which prodromes. The system defines 8 different prodromes. 7 of the 8 prodromes contain enough records to be of interest in our evaluation. The final one related to skin ailments that were less likely to be diagnosed through an Emergency Department and were too sparse to use for any meaningful evaluation.

Of the remaining 7 prodromes, six were well-defined syndromic categories where much previous work has been done on detecting anomalous patterns and outbreaks of diseases. Defined by public health experts, these prodromes contained ICD-9 cases related to gastrointestinal, fever, respiratory, hemorrhagic, botulinic, and encephalitic diseases.

The final prodrome was referred to as "unmapped". These were diseases that either did not fit into one of the other prodromes, or had not been defined as mapping to one by experts. These "unmapped" ICD-9 codes represented 75% of all ED visits for our data. This highlights the challenges encountered when expert labeling is required and motivates working with the unstructured aspects of records.

4.2 Experimental Subsets of the Corpus

From the two years of data, we constrained ourselves to creating semi-synthetic outbreaks only using ICD-9 codes that had at least 10 cases. This left our evaluation to a total of 556 ICD-9 codes. Of these, 54 were mapped to one of the six, expertly defined prodromes, while the remaining 502 ICD-9 codes were assigned to the "unmapped" prodrome that represents the bulk of our dataset. All our models look at all cases when learning a language model, we just do not use the rarer cases when simulating an outbreak due to data sparsity.

For our evaluation of the semantic scan statistic, we compared the performance of our methods on the 54 "mapped" ICD-9 codes and a random sample of 108 "unmapped" ICD-9 codes. As a baseline, we also included the

performance of a standard spatial scan method which scans over all prodrome categories (including "unmapped"). Note that this "prodrome" method uses the additional information in the prodrome field, which is not present at admission into the emergency department.

The "unmapped" ICD-9 codes were randomly partitioned into two sets of 54 codes each; for the former, our evaluation was performed identically to the mapped ICD-9 codes. The goal of this is to demonstrate that without expert labels, the semantic scan statistic is able to detect anomalous patterns of disease outbreaks. This is a novel contribution to the disease surveillance field, and the broader anomalous pattern detection field in general. Evaluating on these "unmapped" cases is very useful for two main reasons. First off, it allows emergency departments to begin disease surveillance as a patient is admitted to the Emergency Department instead of at discharge. Secondly, with approximately three-quarters of the dataset lacking a specific mapping, this enables a more robust disease surveillance program that is not reliant on expert labels.

The remaining 54 ICD-9 codes were denoted as "novel unmapped" ICD-9 codes, and used to approximate detection performance for previously unseen outbreak types. A major benefit of the semantic scan statistic, and using unstructured data in general, is that it allows for robust anomaly detection models even when there is the potential for completely new and previously unseen data to constitute the anomlous patterns. Take for example, the word "swine". It does not occur at all in any of the hundreds of thousand of cases in 2004 and 2005, yet the word "flu" does. Four years later, a major outbreak of swine flu occurs. The ability to detect completely novel outbreaks would be potentially very beneficial to the disease surveillance community. This is much broader than just disease surveillance and can be applicable in any anomalous pattern detection task.

To create novel outbreaks, we used a "leave-one-out" test in which all occurrences of the given ICD-9 code were removed from the background data before evaluating detection performance. We simulated novelty by completely removing entire diseases from our corpus, yet still simulating outbreaks using them. This evaluation was used to compare the detection power of our methods on disease outbreaks with novel, unexpected, and previously unseen symptom patterns. We removed the cases of specific ICD-9 codes, including their chief complaints, but did not remove all occurrences of the vocabulary terms they contained. For instance, many different ICD-9 codes contain words related to "flu" and "cough". We still had these terms in

our dataset after removing an individual ICD-9 code. The aim being that these would more accurately reflect how a novel outbreak would occur in real life. The 54 cases randomly selected had a wide variety in degree of novelty ranging from the few occurrences of the vocabulary of their chief complaints remaining in the dataset, to many cases remaining.

4.3 Semi-Synthetic Injects

We now briefly discuss how we used the dataset to approximate disease outbreaks. 100 outbreaks were created using the semi-synthetic inject method. Semi-synthetic injects sample from existing data, replicating cases according to various distributions. In our work, we used an inject scheme where injects grow linearly with an expected change each day of a given inject delta. We used an inject delta of two, meaning that the expected value for the number of injects in an outbreak would be two higher than the previous day. The total length of an outbreak lasted 10 days. For each of our 162 ICD-9 Codes (54 Mapped, 54 Unmapped, and 54 Novel), 100 code specific outbreaks were created by randomly sampling existing cases for the specific ICD-9 code and inserting those ICD-9, Prodrome, and Chief Complaints fields into the outbreaks. The only field differing across diseases in this sampling is the ICD-9 corresponding to the Chief Complaint. The same 100 outbreaks were used to generate experiments for all 162 ICD-9 codes.

To evaluate, we compared the existing prodrome method with the three different scan statistics based upon topic modeling. For each of the 100 cases, we compared the 7 prodromes, and the 108 ICD-9 codes' detection powers.

Minimal preprocessing was done to the chief complaint datafield so as not to dampen the noise in the dataset. All words were converted to lowercase. Punctuation was removed. Slashes, ampersands, plus symbols, and other punctuation indicative of two words was removed and the token was separated into two words.

In many natural language processing applications, stop word lists and stemming algorithms are used, but we did not do this. These methods effectively smooth the distribution over the vocabulary and reduce its size. This is desirable in many free text applications such as information retrieval, question answering, etc., but was not desirable here. As we are trying to detect anomalous patterns, smoothing noise in the data can negatively impact detection power and drown out the interesting signal.

4.4 Corpus Statistics

Working with a very unique domain and technical corpus has the potential to impact the distribution of terms and possibly make the corpus have different properties than standard human communication. In particular, we were interested if the properties frequently observed in human language would apply here as medical ailments may have widely varying distributions of occuring. For instance, if one out ten cases is flu related and another one out of ten relate to broken bones, the short chief complaints in our dataset had the potential to not follow a power law distribution. So, we looked at the distribution of our corpus to see how common language laws applied. In general, they did, and the frequency of terms obeyed a power law distribution. The most common word in our dataset was "pain", and it represented about approximately 11% of the total terms in all of the 2004 and 2005 data. This is relatively the same as is observed in other domains, just with different terms being the most frequent than in other areas.

4.4.1 Misspelling

One of the major challenges with working with this dataset is the frequency of misspellings within it. This is to be expected given the nature of how it was created. Healthcare providers are focused on patient interactions and treating potentially life threatening ailments, rather than caring about data entry. As long as the information entered can convey the relevant information to other healthcare providers, there is no additional incentive to clean up and fix typos or spelling errors.

There were a couple of common errors in the dataset. Many times, words were concatenated, frequently combining three or four words at a time without a single space or breaking character. Other providers would just enter the number of the ICD-9 code without adding any free text. Also prevalent were other common spelling errors observed in most settings when working with text.

One prime example is the word "vomiting". Even after preprocessing steps to attempt and normalize textual fields, there were still over 200 different spelllings of this particular word. Table 4.2 contains examples of some of the misspellings. You can see errors from simply transposing two characters, to repeating a character too many times, to combining multiple words, to adding additional random characters.

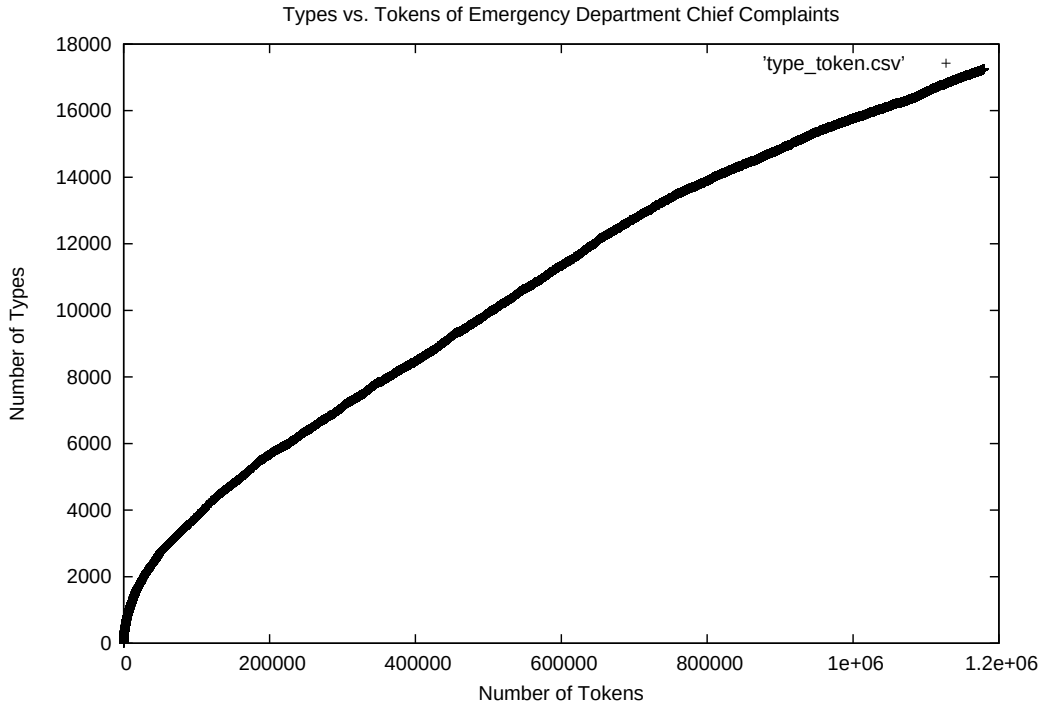

Figure 4.1: Type-Token Curve for 2004 and 2005 Emergency Departments in Allegheny County UPMC Hospitals

An interesting aspect of our goal is that we are attempting to detect anomalous records, which means common methods to smooth our dataset may not be ideal. Methods aimed at fixing errors and misspellings have the potential to actually obfuscate interesting information, in essence, smoothing out the anomalous signals. As our corpus still adhered to power law distributions like most human language, most of our methods did not try and correct spelling errors and instead tried to focus simply on getting a good language model.

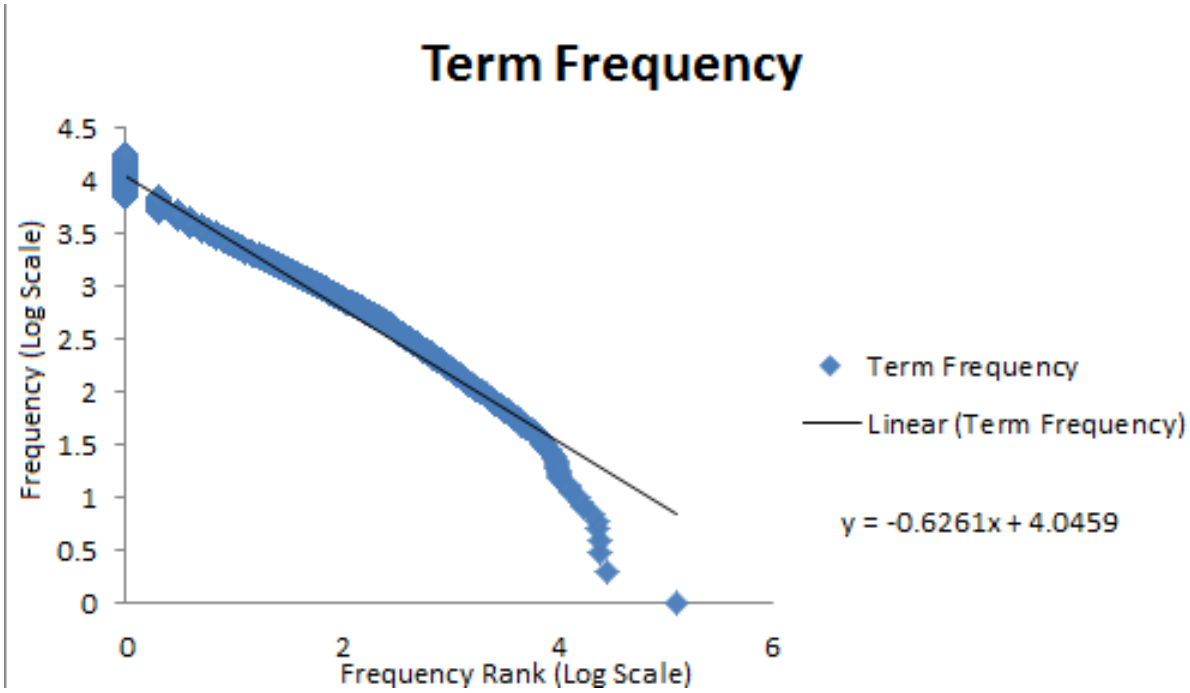

Figure 4.2: Frequency of Terms vs. Relative Frequency Ranking in Log-Log Scale for 2004 and 2005 Emergency Departments in Allegheny County UPMC Hospitals. The corpus obeys a standard power law distribution, even in this very specific domain.

4.4.2 Stop Words

In many language technology tasks, removing stop words can help increase performance by removing words that reveal very little information. The combination of both the task in this work, and the dataset, made determining useful stop word lists much more of a challenge. Common stop word lists are not applicable in this domain. For instance, "a", is frequently used as an abbreviation for parts of many medical ailments, such as "a. fib" for "atrial fibrillation". Removing what could be a non-informative word in a normal dataset has the potential to remove important information in this domain.

Again, with the aim of detecting anomalous patterns in our dataset and without knowing where an interesting signal may be, we decided not to use stop words lists. Though potentially hurting performance of our model by including common words, it also makes are model more robust in extensions

Table 4.2: Examples of Various Spellings of the Word "Vomiting" in the UPMC Dataset

VOMITTTING	VOMIYTTING
VOMMIT	VOMMITING
VOMMITNG	VOMMITTING
VOMMITTONG	VOMNITING
COUGHINGVOMITINGDIAHRREA	VOMOTING
VOMOTTING	VOMTIING
VOMTING	VOMTITING
VOMTTING	VOMITINH
VOMITINHG	VOMITINIG
VOMITITNG	VOMITNG
VOMITNGSEVERE	VOMITNING
VOMITTED	VOMITTIG
VOMITTIN	VOMITTING

to other domains and datasets. If this method was used in an area that is multilingual, no additional work would be needed to adapt this to a non-English corpus.

Chapter 5

Experiments

Our experiments demonstrate the success of using topic models in an anomalous pattern detection framework. In general, unstructured, free text has the ability to improve detection power for anomalous pattern detection in a disease surveillance task. We show that the semantic scan statistic performs well across a variety of different types of outbreaks and across different types of topic models.

Our evaluation demonstrates the efficacy of topic models on an extrinsic task, providing proof in the abilities of topic models to improve downstream performance tasks. Throughout this chapter, we discuss our results using the three main topic models described earlier: Static, Dynamic, and Incremental. The static topic model is trained only on one full year of historic data. It does not change. The dynamic topic model is updated from scratch every day and is trained on the previous two weeks of data. The incremental topic model, as described earlier in figure 3.2, uses the same full year of historic data as the static method and is updated in the second step with the previous two weeks of data like the dynamic model. All three of our models use 25 topics. The incremental method trains using 20 topics that are fixed and then add 5 topics that vary when exposed to the most recent data. In addition to these three models, we also briefly discuss smaller scale evaluations of extensions to these models, such as nonparametric, annealing schemes, and changing data preprocessing. Throughout, we demonstrate the efficacy of topic models and the use of unstructured, free text in anomalous pattern detection.

We evaluated our experiments using semi-synthetic injects as described in section 4.3. With 100 simulated outbreaks, and 162 different ICD-9 codes, comparing multiple methods was very computationally expensive. The re-

sults presented here took approximately one month to run using over 20 cores continuously. We create three different outbreak evaluation sets. The first, we refer to as Labeled. It consists of the 54 ICD-9 codes that have been manually classified by humans into the broad Prodrome categories. The second, we call Unlabeled and it is a sample of 54 cases that were under the broad Prodrome "Unmapped". The final set of outbreaks is our Novel set. This is another subsample of 54 "Unmapped" ICD-9 cases, but we have removed these cases from the underlying, non-outbreak distribution.

In general, we can see that topic models can learn a thematic, latent structure of our dataset. Looking at a few topics learned by our models, we can see that similar words are clustered in the same topic. Just by inspection of some of the topics in table 5.1, we see that topic models are able to learn interesting structure.

5.1 Labeled Outbreaks

As a baseline to the state-of-the-art methods, the first set of experiments we ran was on the set of labeled outbreaks. Labeled outbreaks are ICD-9 codes that map to specific prodromes in our dataset. These are cases where experts have annotated the data with a known ICD-9 code which gives structure to the records. As mentioned before, much previous work has focused on anomalous pattern detection on datasets similar to this. Our goal with these experiments was to show that we are still able to detect anomalous patterns in these records without using the expert labels.

As you can see in both table 5.2 and figure 5.1, the semantic scan statistic performs worse than the prodome method for all topic models. This is to be expected as the prodrome method uses expertly labeled data. To reiterate the point mentioned earlier, these experiments were intended to show that we are able to perform at a level only slightly worse than the state-of-the-art without needing labeled data. Additionally, the comparison between the prodrome method and the methods that use the chief complaints, is not entirely fair as those fields would be updated at different points in an actual healthcare setting. The prodrome method is reliant on ICD-9 codes that are typically assigned at discharge whereas chief complaints are entered upon patient admittance. In reality, the detection difference would be much less significant. Nonetheless, the semantic scan static is able to perform well here.

injury	bleeding	detox	fistula
head	bleed	pt	bee
laceration	rectal	per	sting
inj	nose	evaluation	foot
fall	syncope	depression	possible
arm	vaginal	fall	depression
left	v	seizure	swollen
eye	d	suicidal	aspiration
lac	infection	complications	tracheoesophageal
hand	n	psychiatric	esophageal
to	from	alcohol	clotted
right	gi	crack	stings
facial	urinary	cocaine	fistual
wrist		pysch	tracheal
lip			
rt			
leg			
chin			
forehead			
finger			
face			
knee			
mouth			

Table 5.1: Example topics learned from emergency department data. Note how similar words, including misspellings, are often highly likely in the same topic such as “fistual” and “fistula” or “sting” and “stings”.

5.2 Unlabeled Outbreaks

The second set of outbreaks we evaluated on were the unlabeled cases. These were ICD-9 codes that did not have an expert label. As such, the performance of the prodrome method was much degraded and the semantic scan statistic handily beat out the prodrome method regardless of the type of topic model used in it. This is an important result for the semantic scan statistic as it shows the ability to perform well in anomalous pattern detection settings without needing any labels. Lack of manually labeled data is increasingly indicative of our evermore digitizing world and there will be fewer datasets

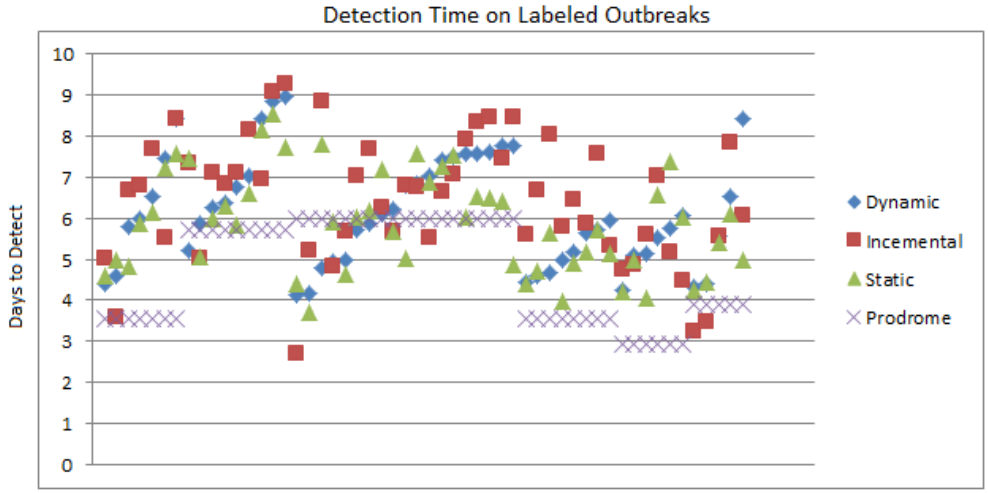

Figure 5.1: Number of days needed to detect outbreak for Static, Dynamic, and Incremental Methods. Also included is the Prodrome Method which is the current state-of-the-art in the literature. The Prodrome Method performs the best as it uses labels from humans to categorize cases. Note that the prodrome method is invariant to the ICD-9 case injected, only relying on the mapping to the broder prodrome.

with expert labels such as the ones observed in the labeled outbreaks in section 5.1.

All of our topic models performed similarily on these outbreaks, with the dynamic method slightly outperforming the other two methods. We can conclusively say that the use of textual data can be used for disease surveillance. The average results for the three methods compared to the prodrome method can be seen in table 5.3. Comparing the prodrome method here to the results in table 5.2 show exactly how reliant this method is on labeled data.

In addition to the three standard topic models we evaluated on all outbreaks, we examined a few extensions to our models using some of the outbreaks. We attempted to look at the impact of misspellings within the dataset to see if normalization could improve our results. We also investigated whether annealing update steps could potentially get our sampler to a better optimum. Neither method was overwhelmingly better, but we

Table 5.2: Detection Time in Days for Labeled Outbreaks

	Dynamic Method	Incremental Method	Static Method	Prodrome Method
Prodrome 1	6.171	6.252	5.881	3.55
Prodrome 2	7.083	7.430	6.848	5.7
Prodrome 3	6.358	6.704	6.108	5.99
Prodrome 5	5.153	6.414	4.955	3.56
Prodrome 6	5.307	5.318	5.535	2.92
Prodrome 7	5.842	5.236	5.030	3.88

Table 5.3: Average Performance of Dynamic, Incremental, and Static Methods Compared to the Prodrome Method on the Unlabeled Outbreaks

Method	Average Days to Detect	Average Percent Detected
Prodrome	8.700	41.0%
Dynamic	5.472	93.1%
Incremental	5.718	91.5%
Static	5.490	91.1%

Table 5.4: Unlabeled Outbreak Detection Power for Dynamic Methods

Method	Days to Detect	% Detected
Full Words	5.472	93.1%
3-Grams	5.56	91.9%

discuss the n-gram subset briefly now since the results were close.

5.2.1 Character N-Gram

The character n-gram method aimed to deal with the noisy portions of our data set and correct for misspellings. We experimented using our Dynamic method on this unlabeled outbreak dataset with a character N-gram of size 3. This method performed well. It was approximately the same detection power as looking at the entire words with our Dynamic Method. As can be seen in table 5.4, our average days to detect and percent detected are similar, though looking at entire words performed slightly better.

Interestingly, the outbreak cases where our methods performed well were not correllated at all. This can be seen in figure 5.2 and implies that there likely is some structure that can be found using words, but that in some cases

n-grams can help deal with the noise. Since the generative story was not as interesting and there were no major improvements in detection power, we did not investigate this line of inquiry any further on any other outbreaks.

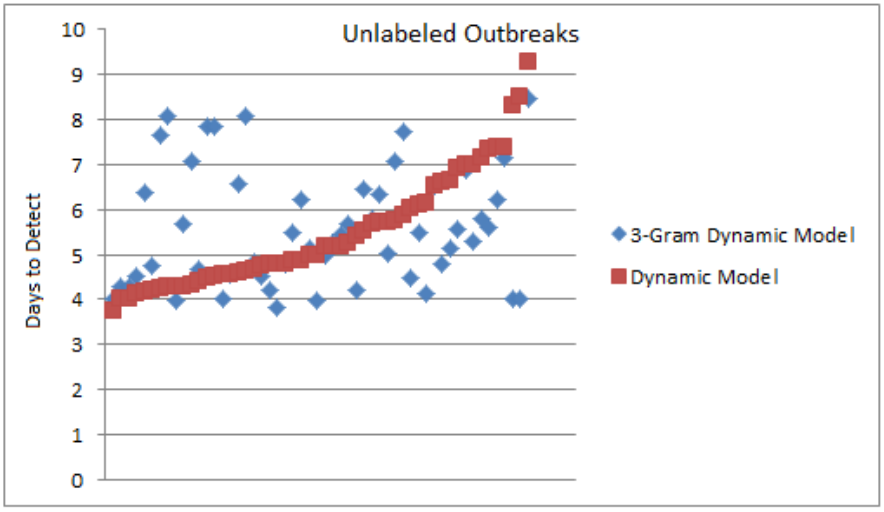

Figure 5.2: Days to Detect an Unlabeled Outbreak using 3-gram Characters and Full Words

5.3 Novel Outbreaks

Our set of novel outbreaks were generated when we removed all occurrences of an individual ICD-9 code from the dataset and only used those cases to simulate an outbreak. One of the potential benefits of using language modeling is that it should be able to still detect anomalous disease patterns even when it has not seen anything similar to it before. This is a major advantage over previous methods, as structured information was required. Often, this was in the form of manually labeled data which would be unlikely to be available for novel disease patterns.

Given these novel outbreaks, the semantic scan statistic was able to effectively detect outbreaks ranging from very novel (many terms non-existent in the outbreak-free dataset) to not at all anomalous (all terms in the outbreak vocabulary existed in other ICD-9 cases). Incorporating unstructured data

Table 5.5: Average Days to Detect for Dynamic, Static, and Incremental Methods on Novel Outbreaks

Dynamic	Incremental	Static
5.436	5.998	5.981

allows spatial scan statistics to detect new anomalous patterns that were not predefined by a human.

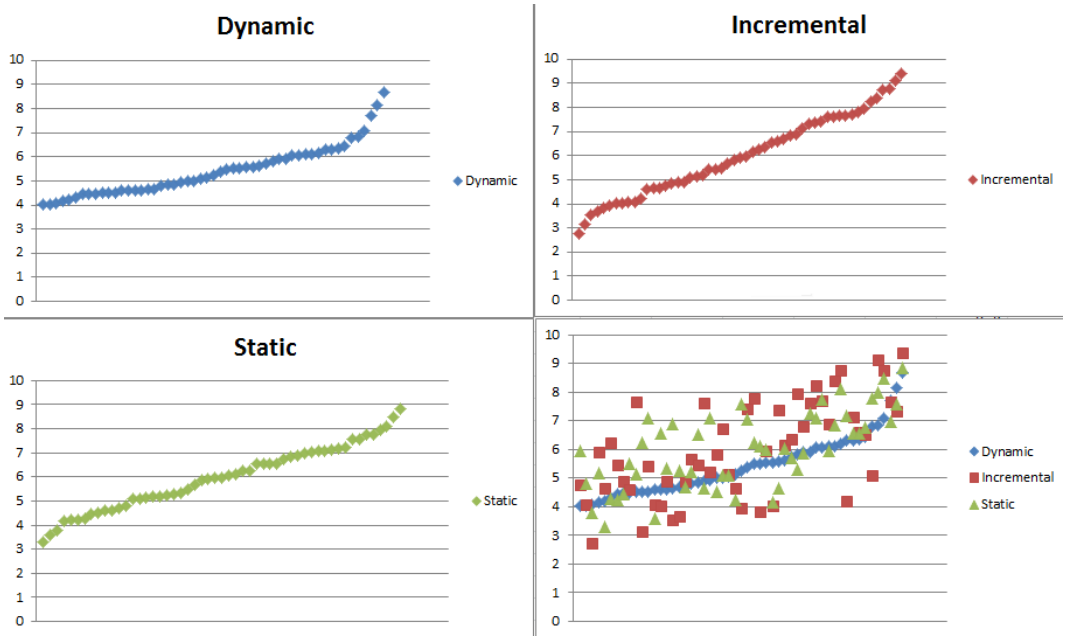

Figure 5.3: Number of days needed to detect outbreak for Static, Dynamic, and Incremental Methods on Novel Outbreaks. The bottom right plot shows how the different methods performed on the same ICD-9 Case (as opposed to sorted in the other 3 plots which are sorted by performance). Note the wide variance between the methods.

As we were removing cases from the original, non-outbreak dataset, the results of the prodrome method should improve slightly. Fewer cases improve detection power for that method. Preliminary results on a development dataset showed consistent improvements, but very minor. The topic modeling methods of the semantic scan statistic are sufficiently better so it was not worthwhile to rerun the prodome method on the test dataset.

Days to Detect	% Detected
7.82	56.1%

Table 5.6: 25 Nonparametric Topics' Detection Power for 54 "Unmapped" ICD-9

5.3.1 Novelty

One of the interesting aspects of this work is the ability to detect novel disease outbreaks where words may not occur outside of an outbreak. We looked at the impact of these terms on detection power and how the percentage of an inject's vocabulary affects it. Interestingly, the "novelty" of an outbreak had little correlation to the detection performance of our methods. In some cases, we were able to detect very novel outbreaks well, and in other cases we were unable to.

5.4 Nonparametric Models

We also ran a few experiments using our nonparametric model formulation. Unsurprisingly, the nonparametric model we evaluated did not perform very well on our dataset. This was expected since our model did not use co-occurrences of terms in a document. Essentially, we were looking at how unigrams would fare in detecting outbreaks, but allowing for a variety of classes. The goal was for a new topic to be created if a word was sufficiently different than the background empirical distribution already observed. There was some modest success here (detection power is still better than a lack of labels for the prodrome case), but the major takeaway is that parametric topic models are effective in a spatial scan statistic setting.

5.5 Hellinger Scores

To gauge the accuracy of our model, we were interested in how closely a learned set of topics' approximated the injected outbreak's distribution. In order to do so, we decided to model the difference between the distribution of the topic chosen for any given outbreak and the empirical distribution of the outbreak.

Method (Experimental Corpus)	Percentage Correct Topic Identified as Most Anomalous	Hellinger Distance (Most Anomalous Topic vs. Outbreak Distribution)
Dynamic (Novel Corpus)	72.2%	0.721
Incremental (Novel Corpus)	46.9%	0.796
Static (Novel Corpus)	44.8%	0.800
Dynamic (Unlabeled Corpus)	72.5%	0.720
Incremental (Unlabeled Corpus)	52.0%	0.810
Static (Unlabeled Corpus)	51.6%	0.814
Dynamic (Labeled Corpus)	64.3%	0.724
Incremental (Labeled Corpus)	46.1%	0.789
Static (Labeled Corpus)	53.2%	0.774

Table 5.7: Hellinger Scores and Percentage Correct Topic Chosen as Most Anomalous (Day 6 of Outbreak)

Given two discrete probability distributions, $P = (p_1 \dots p_k)$ and $Q = (q_1 \dots q_k)$, the Hellinger distance is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{5.1}$$

The Hellinger distance was preferable in this work over other similarity measurements for a couple of reasons. Unlike the more popular KL-Divergence, the Hellinger distance is symmetric. Thus $H(P,Q) = H(Q,P)$ whereas $KL(P,Q) \neq KL(Q,P)$. This is beneficial for a couple of reasons such as comparing topics with exchangeability may complicate calculations. Yet, the nicest property of the Hellinger distance is that it is bounded between 0 and 1. This was desirable due to the fact that we are interested in comparing two very different distributions. We would like to obtain a measure of how close a topic (distribution over words) is to the empirical distribution of an inject. By definition, a topic is a distribution over all unique terms, V, in our corpus. The total number of unique terms in an inject is a small percentage of the total vocabulary size, so using an unbounded divergence measure such as Kullback-Leibler has the potential of diverging to infinity.

As we can see in table 5.7, our Hellinger distance scores are not that close. This is likely due to the fact that a large percentage of our vocabulary has zero probability in our injects. The more interesting number is the number

of times we correctly identified the topic in our model closest to our inject distribution. Since all of our methods had 25 topics, randomly selecting would give us a value of 4%. We are performing significantly better than random. Again, this is a testament to the efficacy of utilizing topic models for anomalous pattern detection.

Chapter 6

Conclusion

In this thesis, we presented the Semantic Scan Statistic which brings language technologies and anomalous pattern detection together. We show that incorporating unstructured, free text into existing spatial scan statistic frameworks can increase detection power and reduce the need for manually labeled data. In particular, we have shown efficacy of using noisy, unstructured data in an anomalous pattern detection framework.

In addition to the contributions to anomalous pattern detection, we have also provided a new extrinsic evaluation of topic models that demonstrates their efficacy and power in a novel downstream task. We show that language modeling in general can be used for data mining techniques that have previously relied on structured data. Furthermore, we are opening broader areas of computer science to language technologies.

As our world gets more digitized, the needs for technologies that can parse large amounts of unstructured data and human language information will become increasingly important. Further developments in technology will likely increase the likelihood that future datasets will contain more varied datasources, such as spatiotemporal information. The need for technologies, and the core research behind them, like the Semantic Scan Statistic, will continue to grow into the future.

Bibliography

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, January 2003.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[3] L. Bolelli, S. Ertekin, and C. Lee Giles. Topic and trend detection in text collections using latent dirichlet allocation. *ECIR*, 2009.

[4] M. Bryant and E. B. Sudderth. Truly nonparametric online variational inferencefor hierarchical dirichlet processes. *NIPS*, 2012.

[5] L. Duczmal and R. Assuncao. A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, pages 269–286, 2004.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, pages 5228–5235, April 2004.

[7] U. Hjalmars, M. Kulldorff, G. Gustafsson, and N. Nagarwalla. Childhood lukemia in sweden: Using gis and a spatial scan statistic for cluster detection. *Statistics in Medicine*, pages 707–715, 1996.

[8] L. Huang, M. Kulldorff, and D. Gregorio. A spatial scan statistic for survival data. *Biometrics*, 63:109–118, 2007.

[9] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Online multiscale dynamic topic models. *KDD*, 2010.

[10] M. Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, pages 1481–1496, 1997.

[11] M. Kulldorff. Prospective time period geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, pages 61–72, 2001.

[12] M. Kulldorff, E. J. Feuer, B. A. Miller, and L. S. Freedman. Breast cancer clusters in the northeast united states: A geographic analysis. *American Journal of Epidemiology*, pages 161–170, 1997.

[13] M. Kulldorff, L. Huang, L. Picle, and L. Duczmal. An elliptic spatial scan statistic. *Statistics in Medicine*, pages 3929–3943, 2006.

[14] F. Mostashari, M. Kulldorff, J. J. Hartman, J.R. Miller, and V. Kulasekera. Dead bird clustering: A potential early warning system for west nile virus activity. *Emerging Infectious Diseases*, pages 641–646, 2003.

[15] T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, pages 223–229, 2010.

[16] D. B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 25, 2000.

[17] D. B. Neill. Detection of spatial and spatio-temporal clusters. *Tech. rep CMU-CS-06-142. Ph.D. thesis. Carnegie Mellon University, Department of Computer Science*, 2006.

[18] D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, pages 337–360, 2011.

[19] D. B. Neill, A. W. Moore, and G. F. Cooper. A bayesian spatial scan statistic. *In Y. Weiss, et al., eds. Advances in Neural Information Processing Systems*, 18:1003–1010, 2006.

[20] D. B. Neill, A. W. Moore, and G. F. Cooper. A multivariate bayesian scan statistic. *Advances in Disease Surveillance*, 2:60, 2007.

[21] D. B. Neill, A. W. Moore, and M. R. Sabhnani. Detecting elongated disease clusters. *Morbidity and Mortality Weekly Report*, 2005.

[22] D. B. Neill, A. W. Moore, M. R. Sabhnani, and K. Daniel. *Proceedings of the 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 218–227, 2005.

[23] G. P. Patil and C. Taille. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, pages 183–197, 2004.

[24] T. Tango and K. Takahashi. A flexibily shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 2005.

[25] C. Wang, D. Blei, and David Heckerman. Continuous time dynamic topic models. *Uncertainty in Artificial Intelligence*, 2008.

[26] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 424–433, 2006.

[27] X. Wei, J. Sun, and X. Wang. Dynamic mixture models for multiple time series. *IJCAI*, pages 2909 – 2914, 2007.

[28] Y.Liu and D. B. Neill. Detecting previously unseen outbreaks with novel symptom patterns. *Emerging Health Threats Journal: ISDS Conference Abstracts*, 2011.