Detecting and Characterizing Events

Anonymous EMNLP submission

Abstract

Significant events are characterized by interactions between entities (e.g., countries, organizations, individuals) that deviate from typical interaction patterns. Investigators, such as historians, commonly read large quantities of text to construct an accurate picture of who, what, when, and where and event happened. In this work, we present the Capsule model for analyzing documents to identify and characterize events of potential significance. Specifically, we develop a model based on topic modeling to distinguish between topics that describe "business-as-usual" and topics that deviate from these patterns. To demonstrate this model, we analyze a corpus of over 2 million US State Department cables from the 1970s; we provide open-source implementations of an inference algorithm for the Capsule model and a pipeline to explore its results.

1 Introduction

Events are difficult to define; historians and political scientists read large quantities of text to construct an accurate picture of a single historical event. Events are interesting by definition: they are the hidden causes of anomalous observations. But they are also inherently abstract—we can observe that changes occur, but we cannot directly observe whether or not an event occurs.

Consider embassies sending diplomatic messages, such as shown in Figure 1. The Bangkok embassy, Hong Kong Embassy, and the State Department all have *typical concerns* about which they usually send messages. At date d, however, the message content changes for all three entities—again, we only observe the changes in message content, and do not observe the event directly. Our first goal is to determine *when* events happen, or identify these rare but pervasive deviations from the typical concerns.

Our second goal is to characterize *what* occurs. We rely on topic modeling (Blei, 2012) to summarize content and to characterize events.

We develop a Bayesian model that discovers the typical concerns of authors, identifies when events occur, and characterizes these events; we call this the *Capsule* model, as it encapsulates events.

We first review previous research related to event detect, summarization, and visualization. In Section 2, we describe the Capsule model and how to infer the latent parameters (the appendix provides further inference details). Section 3 provides an exploration of results on simulated and a real-world data set, and we conclude with a discussion in Section 4.

Related work. We first review previous work on automatic event detection and other related concepts.

In both univariate and multivariate settings, the goal is often the same: analysts want to predict whether or not a rare events will occur (Weiss and Hirsh, 1998; Das et al., 2008). Capsule, in contrast, is designed to help analysts explore and understand the original data: our goal is interpretability, not prediction.

A common goal is to identify clusters of documents; these approaches are used on news articles (Zhao et al., 2012; Zhao et al., 2007; Zhang et al., 2002; Li et al., 2005; Wang et al., 2007; Allan et al., 1998) and social media posts (VanDam, 2012; Lau et al., 2012; Jackoway et al., 2011; Sakaki et al., 2010; Reuter and Cimiano, 2012; Becker et al., 2010; Sayyadi et al., 2009). In the case of news articles, the task is to create new clusters as novel news stories appear—this does not help disentangle typical content from rare events of interest. Social media approaches identify rare events, but the methods are designed for short, noisy documents; they are not appropriate for larger documents that contain information about a variety of subjects.

Many existing methods use document terms as features, frequently weighted by tf-idf value (Fung et al., 2005; Kumaran and Allan, 2004; Brants et al., 2003; Das Sarma et al., 2011; Zhao et al., 2007; Zhao et al., 2012); here, events are bursts in groups of terms.

Topic models (Blei, 2012) reduce the dimensionality of text data; they have been used to help detect events mentioned in social media posts (Lau et al., 2012; Dou et al., 2012) and posts relevant to monitored events (VanDam, 2012). We rely on topic models to characterize both typical content and events, but grouped observations can also be summarized directly (Peng et al., 2007; Chakrabarti and Punera, 2011; Gao et al., 2012).

In addition to text data over time, author (Zhao et al., 2007), news outlet (Wang et al., 2007), and spatial information (Neill et al., 2005; Mathioudakis et al., 2010; Liu et al., 2011) can be used to augment event detection. Capsule uses author information in order to characterize typical concerns of authors.

Detecting and characterizing relationships (Schein et al., 2015; Linderman and Adams, 2014; Das Sarma et al., 2011) is related to event detection. When a message recipient is known, Capsule's author input can be replaced with a sender-receiver pair, but the model could be further tailored for interactions within networks.

2 The Capsule Model

In this section we develop the Capsule model. Capsule captures patterns in entity behavior and identifies time intervals in which many entities deviate from these patterns. The model relies on high dimensional entity behavior data over time, such as text messages being sent between entities. We first review topic models at a high level, and give the intuition on Capsule. Then, we formally specify our model and describe how we learn the hidden variables.

Background: Topic Models. Capsule relies on topic models to model text data. Topic models are algorithms for discovering the main themes in a large collection of documents; each document can then be summarized in terms of the global themes. More formally, a topic k is a probability distribution over the set of vocabulary words. Each document d is represented as a distribution over topics θ_d. Thus we can imagine that when we generate a document, we first pick which topics are relevant (and in what proportions). Under the LDA topic model (Blei et al., 2003), we know the number of words in each document. Then, for each word, we select a single topic from this distribution over topics, and finally select a vocabulary term from the corresponding topic's distribution over the vocabulary. Alternatively, we can cast topic modeling as factorization, such as in Poisson factorization (Gopalan et al., 2014), and draw a word count for each term in the vocabulary.

The Capsule Model. Topic models are often applied to provide a structure for an otherwise unstructured collection of documents. Documents, however, are often accompanied by metadata, such as the date written or author attribution; this information is not exploited by traditional topic models. The Capsule model uses both author and date information to identify and characterize events that influence the content of the collection.

Consider an entity like the Bangkok American embassy, shown in Figure 1. We can imagine that there is a stream of messages (or *diplomatic cables*) being sent by this embassy—some might be sent to the US State Department, others to another American embassy like Hong Kong. An entity will usually talk about certain topics; the Bangkok embassy, for instance, is concerned with topics regarding southeast Asia more generally.

Now imagine that at a particular time t, an event occurs, such as the capture of Saigon during the Vietnam war. We do not directly observe that events occur, but we do observe the message stream. Using this stream, each event be described as a distribution over the vocabulary, similar to how topics are distributions over these same terms. When an event occurs, the message content changes for multiple entities. The day following the capture of Saigon, the

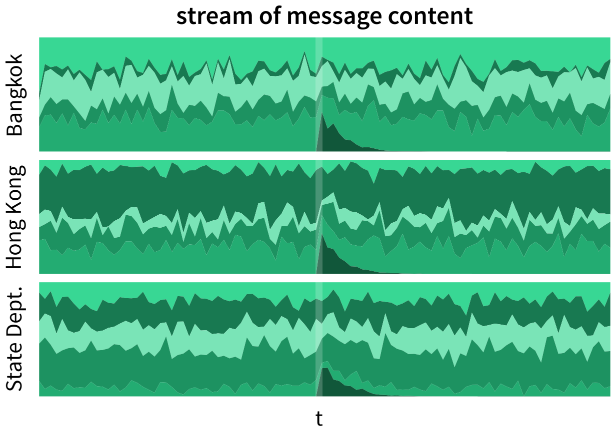

Figure 1: Cartoon intuition of Capsule. The Bangkok embassy, Honk Kong embassy, and State Department all have typical concerns about which they usually send messages. When an events occurs at time t, the stream of message content alters to include the event, then fades back to "business as usual." Capsule discovers both entities' typical concerns and the event locations and content.

majority of the diplomatic cables sent by the Bangkok embassy were about Vietnam war refugees. Thus we imagine that an entity's stream of messages is controlled by what it usually talks about as well as the higher level stream of unobserved events.

Model Specification. We formally describe Capsule. The observed data are word counts $w_{d,v}$ for document d and vocabulary term v; each document d also has an author (or entity) a_d and a time (or date) interval i_d associated with it.

The hidden variables of this model are general topics of conversation β, authors' typical concerns ϕ, event descriptions π, event strengths ψ, and document-specific topics θ and event relevancy ϵ.

As in topic modeling, we represent the general topics of conversation with a $K \times V$ matrix β, where K is a low dimensional number of topics that we wish to capture, and V is the size of our vocabulary; each row β_k is normalized such that it represents the probability of seeing vocabulary word v when discussing topic k. As a generative process, we draw these general topics from a Dirichlet distribution, or $\beta_k \sim \text{Dirichlet}_V(\alpha_\beta)$.

In addition to using these general topics to represent entity concerns, each entity n has its own exclusive topic $\beta_0^{(n)}$, which can be appended as a bias row to the general topics β. These entity-specific topics are drawn from a Dirichlet, just as the general topics.

The concerns of author n are represented with ϕ_n, a $(K+1)$-dimensional topic vector, where each element is drawn from a gamma distribution, or $\phi_{n,k} \sim \text{Gamma}(s_\phi, r_\phi)$,[1] and the first element of the concern vector $\phi_{n,0}$ describes how much the entity n relies on its exclusive topic $\beta_0^{(n)}$

Similar to topic modeling, we represent the contents of each document in topic space; each document d has a $(K+1)$-dimensional latent parameter θ_d to describe the particular contents of that document. Unlike traditional topic models, each document d's topics depend on the concerns of the author a_d; each document topic $\theta_{d,k}$ is drawn from a gamma distribution parameterized by the corresponding author concerns $\phi_{a_d,k}$: $\theta_{d,k} \sim \text{Gamma}(s_\theta, \phi_{a_d,k})$.

To represent events, we consider discrete intervals of time. Each interval t has a corresponding interval strength ψ_t and description π_t. Event strengths are a single value for each interval t, and are drawn from a gamma distribution: $\psi_{n,k} \sim \text{Gamma}(s_\psi, r_\psi)$. These strengths indicate how important the interval is in determining message content. Interval descriptions are similar to topics: each description is a V-dimensional vector drawn from a Dirichlet distribution over the vocabulary terms, or $\pi_k \sim \text{Dirichlet}_V(\alpha_\pi)$.

Just as we describe each document d in terms of relevant topics with the θ_d parameters, we also describe the relevancy of each time interval with the ϵ_d parameters. These interval relevancy parameters are drawn from gamma distributions and depend on the overall strength ψ of the corresponding interval; for interval t and document d (written at time i_d), we have $\epsilon_{d,t} \sim \text{Gamma}(s_\epsilon, \psi_{i_d,t})$.

Conditional on the hidden variables and the author and time metadata, Capsule is a model of how document word counts came to be. For document d and vocabulary term v, we generate the word counts form a Poisson distribution parameterized by the documents topics θ_d and relevant events ϵ, as well as

[1]We use the shape-rate parameterization for all Gamma distributions.

global topic β and event descriptions π:

$$w_{d,v} \sim \text{Poisson}\left(\theta_d^\top \beta_v^{(a_d)} + \sum_{t=1}^{T} f(i_d, t)\epsilon_{d,t}\pi_{t,v}\right), \quad (1)$$

where f is some function of decay. This function is important because events should not remain at their full strength indefinitely, but should decay over time. In our experiments, we consider step functions, linear decay, and exponential decay. Figure 2 gives the full generative process for Capsule.

- for each time step $t = 1{:}T$,
  - draw interval description over vocabulary $\pi_t \sim \text{Dirichlet}_V(\alpha)$
  - draw interval strength $\psi_t \sim \text{Gamma}(s_\psi, r_\psi)$
- for each entity $n = 1{:}N$,
  - draw entity-specific topics over vocabulary $\beta_0^{(n)} \sim \text{Dirichlet}_V(\alpha)$
  - draw entity-specific topic strength $\phi_{n,0} \sim \text{Gamma}(s_\phi, r_\phi)$
- for each topic $k = 1{:}K$,
  - draw general topic distribution over vocabulary $\beta_k \sim \text{Dirichlet}_V(\alpha)$
  - for each entity $n = 1{:}N$,
    - draw general entity concern $\phi_{n,k} \sim \text{Gamma}(s_\phi, r_\phi)$
- for each document $d = 1{:}D$ sent at time i_d by author a_d,
  - draw local entity concern $\theta_{d,0} \sim \text{Gamma}(s_\theta, \phi_{a_d,0})$
  - for each topic $k = 1{:}K$,
    - draw local entity concern $\theta_{d,k} \sim \text{Gamma}(s_\theta, \phi_{a_d,k})$
  - for each time $t = 1{:}T$,
    - draw local interval relevancy $\epsilon_{d,t} \sim \text{Gamma}(s_\epsilon, \psi_{i_d,t})$
  - for each vocabulary term $v = 1{:}V$,
    - draw word counts $w_{d,v} \sim \text{Poisson}\left(\theta_d^\top \beta_v^{(a_d)} + \sum_{t=1}^{T} f(i_d, t)\epsilon_{d,t}\pi_{t,v}\right)$

Figure 2: The generative process for Capsule.

Learning the hidden variables. In order to use the Capsule model to explore the observed documents, we must compute the posterior distribution. Conditional on the observed word counts w, our goals to to compute the posterior values of the hidden parameters—global interval strengths ψ, interval descriptions π, entity concerns ϕ, and topics β, as well as document-specific entity concerns θ and interval relevancy parameters ϵ.

As for many Bayesian models, the exact posterior for Capsule is not tractable to compute; approximating it is our central statistical and computational problem. We develop an approximate inference algorithm for Capsule based on variational methods (Wainwright and Jordan, 2008).[2]

Variational inference approaches the problem of posterior inference by minimizing the KL divergence from an approximating distribution q to the true posterior p. This is equivalent to maximizing the ELBO,

$$\mathcal{L}(q) = \mathbb{E}_{q(\psi,\pi,\phi,\beta,\theta,\epsilon)}\big[\log p(w, \psi, \pi, \phi, \beta, \theta, \epsilon) - \log q(\psi, \pi, \phi, \beta, \theta, \epsilon)\big]. \quad (2)$$

We define the approximating distribution q using the mean field assumption:

$$q(\psi, \pi, \phi, \beta, \theta, \epsilon) = \prod_{t=1}^{T}\left[q(\pi_t \mid \lambda_t^\pi) q(\psi_t \mid \lambda_t^\psi)\right]$$
$$\prod_{n=1}^{N}\left[q(\phi_{n,0} \mid \lambda_{n,0}^\phi)\, q(\beta_0^{(n)} \mid \lambda_{n,0}^\beta)\right]$$
$$\prod_{k=1}^{K}\left[q(\beta_k \mid \lambda_k^\beta) \prod_{n=1}^{N} q(\phi_{n,k} \mid \lambda_{n,k}^\phi)\right]$$
$$\prod_{d=1}^{D}\left[\prod_{k=1}^{K} q(\theta_{d,k} \mid \lambda_{d,k}^\theta) \prod_{t=1}^{T} q(\epsilon_{d,t} \mid \lambda_{d,t}^\epsilon)\right] \quad (3)$$

The variational distributions $q(\pi)$ and $q(\beta)$ are both Dirichlet-distributed with free variational parameters λ^π and λ^β, respectively. Similarly, $q(\psi)$, $q(\phi)$, $q(\theta)$ and $q(\epsilon)$ are all gamma-distributed with corresponding free variational parameters λ^ψ, λ^ϕ, λ^θ, and λ^ϵ.

The expectations under q, which are needed to maximize the ELBO, have closed form analytic updates, as detailed in Appendix A. We update each parameter in turn, following standard coordinate ascent variational inference techniques. Full details

[2]Source code will be released on github upon publication.

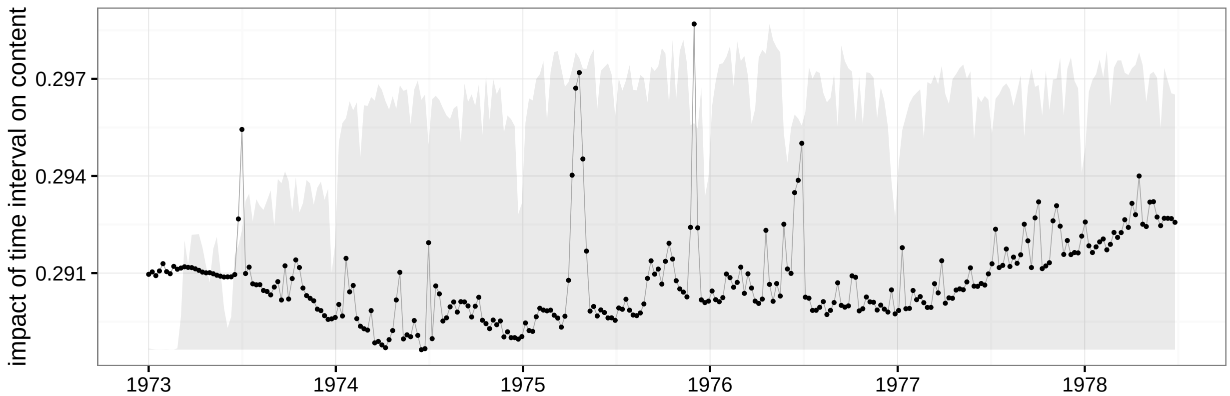

Figure 3: Measure of time interval impact on cable content: $\psi_t \frac{1}{|D_t|} \sum_{d \in D_t} \epsilon_{d,t}$, where D_t is the set of all cables sent in interval t. The grey background indicates the number of cables sent over time.

on our inference algorithm can be found in the appendix. This algorithm produces a fitted variational distribution which can then be used as a proxy for the true posterior, allowing us to explore a collection of documents with Capsule.

3 Evaluation

In this section we explore the performance of Capsule on a collection of US State Department cables. These cables were sent between 1973 and 1978 and obtained from the History Lab at Columbia,[3] which received them from the Central Foreign Policy Files at the National Archives. In addition to the text of the cables themselves, each document is supplemented with information about who sent the cable (e.g., the State Department, the U.S. Embassy in Saigon, or an individual by name), who received the cable (often multiple entities), and the date the cable was sent. To test our model, we used a vocabulary of size 6,293 and omitted cables with fewer than three terms, resulting in a collection of 2,139,324 messages sent between 27,134 entities. We selected a weekly duration for the time intervals, as few cables were sent on the weekends.

We fit Capsule with 100 topics and using an exponential decay with mean lifetime of 3—this indicates that most intervals would no longer be relevant after about 3 weeks. To detect when an event occurs, we multiply the average event relevancy ϵ for all documents in a given interval together with interval strength ψ, or $\psi_t \frac{1}{|D_t|} \sum_{d \in D_t} \epsilon_{d,t}$, where D_t is the set of all cables sent in interval t.

Figure 3 shows this measure over the duration of the data set. The highest time intervals, ones in which we declare events to be detected, include the tallest peak the week of December 1, 1975, just prior to the Indonesian invasion of East Timor, which Began December 7, 1975. The second tallest peak occurs the week of April 21, 1975, just prior to the fall of Saigon on April 30, 1975. For any given week, we can sort the documents by their interval relevancy parameters ϵ; Tables 1 and 2 show the top cables for these two events, which reflect the real-world events those weeks.

Other event peaks include the week of July 2, 1973; the top three words under event its description π are *bicentennial*, *hijack*, and *mercenary*. Top cables under event relevancy ϵ surround the bicentennial celebration of United States (July 4, 1973) and the Air France hijacking incident that began on June 27: Israeli operatives rescued hostages from this incident on July 4th.

Another peak occurs the week of April 17, 1978 surrounding a UN special session on disarmament; the top three words under event its description π are *SSOD* (acronym for "special session on disarmament", *disarmament*, and *ICS* (likely an acronym for "incident command system").

Examples of general topics of conversation are shown in Table 3 and entity-exclusive topics are shown in Table 4; these show us how entity topics ab-

[3] http://history-lab.org

ϵ	date	entity	subject
0.1237	1975-12-03	STATE	PRESIDENT'S TALKING POINTS ON PORTUGUESE TIMOR
0.1210	1975-12-03	STATE	PRESIDENT'S TALKING POINTS ON PORTUGUESE TIMOR
0.1153	1975-12-04	STATE	TIMOR WE ARE REPEATING FYI A DAO MESSAGE
0.1126	1975-012-04	STATE	LEGAL PROBLEMS RELATING TO PORTUGUESE TIMOR
0.1053	1975-12-07	SECRETARY PEKING	US SUPPORT FOR TIMOR RESOLUTION
0.1021	1975-12-01	STATE	INVASION OF PORTUGUESE TIMOR

Table 1: Top documents for the time interval of week December 1, 1975, just prior to the Indonesian invasion of East Timor, which Began December 7, 1975.

ϵ	date	entity	subject
0.0908	1975-04-24	MANSFIELD, MIKE	ASSISTANCE IN EVACUATING FAMILY FROM SOUTH VIETNAM
0.0886	1975-04-24	RAILSBACK, TOM	ASSISTANCE IN EVACUATING FRIEND FROM SOUTH VIETNAM
0.0877	1975-04-24	MANSFIELD, MIKE	ASSISTANCE IN EVACUATING FAMILY FROM SOUTH VIETNAM
0.0863	1975-04-24	WILLIAMS, HARRISON	ASSISTANCE IN EVACUATING FAMILY FROM VIETNAM
0.0860	1975-04-24	KOCH, EDWARD	ASSISTANCE IN EVACUATING FAMILY FROM SOUTH VIETNAM
0.0858	1975-04-21	SCHWEIKER, RICHARD	SUPPORT IN EVACUATING FAMILY FROM VIETNAM
⋮	⋮	⋮	⋮
0.0812	1975-04-25	KETCHUM, WILLIAM	MOVEMENT OF SOUTH VIETNAMESE REFUGEES TO GUAM
0.0800	1975-04-21	SCOTT, HUGH	WHEREABOUTS OF MISSIONARIES IN VIETNAM

Table 2: Top documents for the time interval of week April 21, 1975, just prior to the fall of Saigon on April 30, 1975.

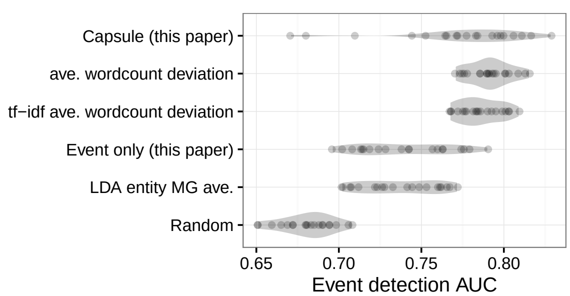

Figure 4: Event detection performance on twenty simulated datasets. Capsule is able to detect events as well as comparison methods, but its performance has higher variance.

sorb location-specific words, preventing these terms from overwhelming the general topics.

These exploratory results show that our model is successfully capturing when multiple entities are discussing the same subjects and that our model can be used to explore the underlying data by providing a structured scaffold from which to view the data.

We also considered held out log likelihood in evaluating the fitness of our model, as shown in Table 5. We see that the event structure is crucial to fitting the data well. We held out 5% of the data.

Finally, we simulated data according to our generative process in order to compare our method to baseline and existing approaches. To evaluate event detection, we created a ranked list of all time intervals and computed the overlap between a model and the simulated ground at every threshold; this generates an curve under which we cam compute the area and normalized based on ideal performance—we refer to this metric as event detection AUC. The most successful of the comparison methods for event detection was average absolute error in wordcount, both unweighted and weighted by tf-idf. Figure 4 shows that Capsule can outperform these approaches for event detection, but that it has higher variance in performance. The other comparison method in Figure 4 is based on LDA; we fit a multinomial Gaussian to the topic representation of all documents and then computed the average probability of seeing the topic distributions of documents in the time interval. Time intervals with the lowest probability were marked as most likely to have events. All other baselines performed close to random for event detection.

This method of fitting a multinomial Gaussian to LDA representations of documents also performed

top terms
OUTLOOK, REVIEW, HIRE, PERSONNEL, INVITE, PREPARE, NECESSARY
ARREST, INCIDENT, SECURITY, FAMILY, OFF, GUARD, DEATH, JAIL
LOCATE, HOME, SON, DEATH, PLEASE, CONTACT, FATHER, DEPARTMENT
REQUEST, REFUGEE, RESPONSE, SERVICE, SALE, ASYLUM, APPRECIATE
MARKET, REPORT, COPY, COMMERCIAL, FOOD, IMPORT, COMMERCE
FEAR, LEADERSHIP, BACK, ARM, ROLE, PLAY, THREATEN
HOTEL, TRAVEL, RESERVATION, VISIT, ARRANGE, SCHEDULE, STAY

Table 3: Top vocabulary terms for a selection of topics, according to topic distributions β_k.

entity	top terms
STATE	REQUEST, FOLLOW, EMBASSY, MEET, MAKE, STATE, DEPARTMENT
BANGKOK	BANGKOK, THAILAND, THAI, REFUGEE, EMBASSY, FOLLOW, REPORT
JERUSALEM	JERUSALEM, ISRAELI, BANK, REPORT, SAY, COMMENT, ONE
STOCKHOLM	SWEDISH, SWEDEN, TRADE, MEET, EMBASSY, FOLLOW, MAKE
CASABLANCA	CASABLANCA, MOROCCO, MOROCCAN, REQUEST, PLEASE, FOLLOW, NOTE
KAMPALA	UGANDAN, NAIROBI, AFRICAN, IMPERIALIST, VOICE, KENYA, MISSIONARY
NDJAMEAN	CHADIAN, CHAD, LAGOS, DROUGHT, INITIATION, AUSTERITY, GOC

Table 4: Top vocabulary terms for a selection of entities according to entity-exclusive topics $\beta_0^{(n)}$.

method	held out log likelihood
full Capsule model	-2.41e7
event only	**-1.82e7**
topics only	-2.69e7
entities only	-2.43e7

Table 5: Held-out data log likelihood.

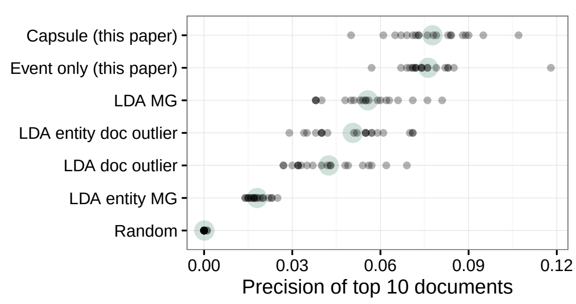

Figure 5: Precision of recovering the top ten most relevant documents, averaged over all time intervals. Capsule performs best, averaged over twenty simulations.

well for recovering relevant documents. This approach can be altered to fit a per-entity multinomial Gaussian, but this performs worse. Simply finding documents based on absolute deviation from the mean works well in LDA topic space (relative to overall mean or entity mean), but not over the full vocabulary. Word count deviations, which performed well for event detection, performed worse than random for document recovery. Both Capsule and its event-only partial model outperform all comparison methods in terms of document recovery. Figure 5 shows precision of recovering the top ten documents.

We also assessed the sensitivity of our model to various parameter settings. We simulated data sets with three different event relevancy decay functions f: exponential, linear, and step functions. Then, we fit Capsule to each of these datasets using every permutation of f and multiple settings for event decay duration. In all cases, we found that the model was not sensitive to decay shape or duration, but that exponential decay generally performed the best on all data types.

4 Discussion

We have presented Capsule, a Bayesian model that identifies when events occur, characterizes these events, and discovers the typical concerns of author entities. We have shown that Capsule outperforms deterministic baseline methods and explored its results on a real-world datasets. We anticipate that Capsule can be used by historians, political scientist, and others who wish to explore and investigate events in large text corpora.

A Inference

In this appendix, we describe the details of the variational inference algorithm for Capsule. This algorithm fits the parameters of the variational distribution q in Eq. 3 so that it is close in KL divergence to the posterior.

Recall that the variational distributions $q(\pi)$ and $q(\beta)$ are both Dirichlet-distributed with free variational parameters λ^{π} and λ^{β}, respectively. Similarly, the variational distributions $q(\psi)$, $q(\phi)$, $q(\theta)$ and $q(\epsilon)$ are all gamma-distributed with corresponding free variational parameters λ^{ψ}, λ^{ϕ}, λ^{θ}, and λ^{ϵ}. For these gamma-distributed variables, each free parameter λ has two components: shape s and rate r.

Minimizing the KL divergence between the true posterior p and the variational approximation q is equivalent to maximizing the ELBO (Eq. 2). We achieve this with closed form coordinate updates, as the Capsule model is specified with the required conjugate relationships that make this approach possible (Ghahramani and Beal, 2001).

To obtain simple updates, we first rely on auxiliary latent variables z. These variables, when marginalized out, leave the original model intact. The Poisson distribution has an additive property; specifically if $w \sim \text{Poisson}(a + b)$, then $w = z_1 + z_2$, where $z_1 \sim \text{Poisson}(z_1)$ and $z_2 \sim \text{Poisson}(z_2)$. We apply this decomposition to the word count distribution in Eq. 1 and define Poisson variables for each component of the word count:

$$z^{\mathcal{K}}_{d,v,k} \sim \text{Poisson}(\theta_{d,k}\beta_{k,v})$$

$$z^{\mathcal{T}}_{d,v,t} \sim \text{Poisson}\left(f(i_d,t)\epsilon_{d,t}\pi_{t,v}\right).$$

The $\mathcal{K}$ and $\mathcal{T}$ superscripts indicate the contributions from entity concerns and events, respectively. Given these variables, the total word count is deterministic:

$$w_{d,v} = \sum_{k=1}^{K} z^{\mathcal{K}}_{d,v,k} + \sum_{t=1}^{T} z^{\mathcal{T}}_{d,v,t}.$$

Coordinate-ascent variational inference is derived from complete conditionals, i.e., the conditional distributions of each variable given the other variables and observations. These conditionals define both the form of each variational factor and their updates. The following are the complete conditional for each of the gamma- and Dirchlet-distributed latent parameters. The notation $D(i)$ is used for the set of documents sent by entity i; $D(t)$ is the set of documents sent impacted by events at time t (e.g., all documents after the event in the case of exponential decay).

$$\pi_t \mid \mathbf{W}, \psi, \phi, \beta, \theta, \epsilon, z \sim \text{Dirichlet}_V\left(\alpha_\pi + \sum_{d=1}^{D}\langle z^{\mathcal{T}}_{d,1,t}, \cdots, z^{\mathcal{T}}_{d,V,t}\rangle\right) \quad (4)$$

$$\beta_k \mid \mathbf{W}, \psi, \pi, \phi, \theta, \epsilon, z \sim \text{Dirichlet}_V\left(\alpha_\beta + \sum_{d=1}^{D}\langle z^{\mathcal{K}}_{d,1,k}, \cdots, z^{\mathcal{K}}_{d,V,k}\rangle\right) \quad (5)$$

$$\psi_t \mid \mathbf{W}, \pi, \phi, \beta, \theta, \epsilon, z \sim \text{Gamma}\left(s_\psi + |D(t)|s_\epsilon, r_\psi + \sum_{d\in D(t)}\epsilon_{d,t}\right) \quad (6)$$

$$\phi_{i,k} \mid \mathbf{W}, \psi, \pi, \beta, \theta, \epsilon, z \sim \text{Gamma}\left(s_\phi + |D(i)|s_\theta, r_\phi + \sum_{d\in D(i)}\theta_{d,k}\right) \quad (7)$$

$$\theta_{d,k} \mid \mathbf{W}, \psi, \pi, \phi\beta, \epsilon, z \sim \text{Gamma}\left(s_\theta + \sum_{v=1}^{V} z^{\mathcal{K}}_{d,v,k}, \phi_{a_d,k} + \sum_{v=1}^{V}\beta_{k,v}\right) \quad (8)$$

$$\epsilon_{d,t} \mid \mathbf{W}, \psi, \pi, \phi, \beta, \theta, z \sim \text{Gamma}\left(s_\epsilon + \sum_{v=1}^{V} z^{\mathcal{T}}_{d,v,t}, \psi_t + f(i_d,t)\sum_{v=1}^{V}\pi_{t,v}\right) \quad (9)$$

The complete conditional for the auxiliary variables has the form $z_{d,v} \mid \psi, \pi, \phi, \beta, \theta, \epsilon \sim \text{Mult}(w_{d,v}, \omega_{d,v})$, where

$$\omega_{d,v} \propto \langle \theta_{d,1}\beta_{1,v}, \cdots, \theta_{d,K}\beta_{K,v}, f(i_d, 1)\epsilon_{d,1}\pi_{1,v}, \cdots, f(i_d, T)\epsilon_{d,T}\pi_{T,v} \rangle. \quad (10)$$

Intuitively, these variables allocate the data to one of the entity concerns or events, and thus can be used to explore the data.

Given these conditionals, the algorithm sets each parameter to the expected conditional parameter under the variational distribution. The mean field assumption guarantees that this expectation will not involve the parameter being updated. Algorithm 1 shows our variational inference algorithm.

Algorithm 1: Variational Inference for Capsule

Input: word counts w
Output: approximate posterior of latent parameters (ψ, π, ϕ, β, θ, ϵ) in terms of variational parameters $\lambda = \{\lambda^\psi, \lambda^\pi, \lambda^\phi, \lambda^\beta, \lambda^\theta, \lambda^\epsilon\}$
Initialize $\mathbb{E}[\beta]$ to slightly random around uniform
Initialize $\mathbb{E}[\psi]$, $\mathbb{E}[\pi]$, $\mathbb{E}[\psi]$, $\mathbb{E}[\theta]$, $\mathbb{E}[\epsilon]$ to uniform
for *iteration* $m = 1 : M$ **do**
  set λ^ψ, λ^π, λ^ϕ, λ^β, λ^θ, λ^ϵ to respective priors, excluding $\lambda^{\theta,rate}$ and $\lambda^{\epsilon,rate}$, which are set to 0
  update $\lambda^{\theta,rate} \mathrel{+}= \sum_V \mathbb{E}[\beta_v]$
  for *each document* $d = 1 : D$ **do**
    for *each term* $v \in V(d)$[4] **do**
      set $(K+T)$-vector $\omega_{d,v}$ using $\mathbb{E}[\pi]$, $\mathbb{E}[\theta]$, and $\mathbb{E}[\epsilon]$, as shown in Eq. 10
      set $(K+T)$-vector $\mathbb{E}[z_{d,v}] = w_{d,v} * \omega_{d,v}$
      update $\lambda_d^{\theta,shape} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 8)
      update $\lambda_d^{\epsilon,shape} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 9)
      update $\lambda_v^{\beta} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 5)
      update $\lambda_v^{\pi} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{T}}]$ (Eq. 4)
    end
    update $\lambda_d^{\theta,rate} \mathrel{+}= \mathbb{E}[\phi_{a_d}]$ (Eq. 8)
    update $\lambda_d^{\epsilon,rate} \mathrel{+}= \mathbb{E}[\psi]$ (Eq. 9)
    set $\mathbb{E}[\theta_d] = \lambda_d^{\theta,shape} / \lambda_d^{\theta,rate}$
    set $\mathbb{E}[\epsilon_d] = \lambda_d^{\epsilon,shape} / \lambda_d^{\epsilon,rate}$
    update $\lambda_{a_d}^{\phi,shape} \mathrel{+}= s_\theta$ (Eq. 7)
    update $\lambda_t^{\psi,shape} \mathrel{+}= s_\epsilon \forall t : f(i_d, t) \neq 0$ (Eq. 6)
    update $\lambda_{a_d}^{\phi,rate} \mathrel{+}= \theta_d$ (Eq. 7)
    update $\lambda^{\psi,rate} \mathrel{+}= \epsilon_d$ (Eq. 6)
  end
  set $\mathbb{E}[\phi] = \lambda^{\phi,shape} / \lambda^{\phi,rate}$
  set $\mathbb{E}[\beta_k] = \lambda^{\beta_{k,v}} / \sum_v \lambda^{\beta_k} \forall k$
  set $\mathbb{E}[\psi] = \lambda^{\psi,shape} / \lambda^{\psi,rate}$
  set $\mathbb{E}[\pi_t] = \lambda^{\pi_{t,v}} / \sum_v \lambda^{\pi_t} \forall t$
end
return λ

References

James Allan, Ron Papka, and Victor Lavrenko. 1998. Online new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 291–300. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3:993–1022, March.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 330–337. ACM.

Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. *ICWSM*, 11:66–73.

Kaustav Das, Jeff Schneider, and Daniel B Neill. 2008. Anomaly pattern detection in categorical datasets. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–176. ACM.

[4] $V(d)$ is the set of vocabulary indices for the collection of words in document d. We could also iterate over all V, but as zero word counts give $\mathbb{E}[z_{d,v}] = 0 \ \forall v \notin V(d)$, the two are equivalent.

Anish Das Sarma, Alpa Jain, and Cong Yu. 2011. Dynamic relationship and event discovery. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 207–216. ACM.

Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X Zhou. 2012. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 93–102. IEEE.

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, pages 181–192. VLDB Endowment.

Wei Gao, Peng Li, and Kareem Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1173–1182. ACM.

Zoubin Ghahramani and Matthew J Beal. 2001. Propagation algorithms for variational bayesian learning. *Advances in neural information processing systems*, pages 507–513.

Prem K Gopalan, Laurent Charlin, and David Blei. 2014. Content-based recommendations with poisson factorization. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3176–3184. Curran Associates, Inc.

Alan Jackoway, Hanan Samet, and Jagan Sankaranarayanan. 2011. Identification of live news events using twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 25–32. ACM.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304. ACM.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *COLING*, pages 1519–1534.

Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. 2005. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 106–113. ACM.

Scott W Linderman and Ryan P Adams. 2014. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*.

Xueliang Liu, Raphaël Troncy, and Benoit Huet. 2011. Using social media to identify events. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 3–8. ACM.

Michael Mathioudakis, Nilesh Bansal, and Nick Koudas. 2010. Identifying, attributing and describing spatial bursts. *Proceedings of the VLDB Endowment*, 3(1-2):1091–1102.

Daniel B Neill, Andrew W Moore, Maheshkumar Sabhnani, and Kenny Daniel. 2005. Detection of emerging space-time clusters. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 218–227. ACM.

Wei Peng, Charles Perng, Tao Li, and Haixun Wang. 2007. Event summarization for system management. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1028–1032. ACM.

Timo Reuter and Philipp Cimiano. 2012. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 22. ACM.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *ICWSM*.

Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. 2015. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054. ACM.

Courtland VanDam. 2012. A probabilistic topic modeling approach for event detection in social media. Master's thesis, Michigan State University.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January.

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 784–793. ACM.

Gary M Weiss and Haym Hirsh. 1998. Learning to predict rare events in event sequences. In *KDD*, pages 359–363.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In

Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 81–88. ACM.

Qiankun Zhao, Prasenjit Mitra, and Bi Chen. 2007. Temporal and information flow based event detection from social text streams. In *AAAI*, volume 7, pages 1501–1506.

Wayne Xin Zhao, Rishan Chen, Kai Fan, Hongfei Yan, and Xiaoming Li. 2012. A novel burst-based text representation model for scalable event detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 43–47. Association for Computational Linguistics.