Detecting and Characterizing Events

Allison J. B. Chaney
Princeton University
achaney@cs.princeton.edu

Hanna Wallach
Microsoft Research
wallach@microsoft.com

Matthew Connelly
Columbia University
mjc96@columbia.edu

David M. Blei
Columbia University
david.blei@columbia.edu

Abstract

Significant events are characterized by interactions between entities, such as, countries, organizations, or individuals, that deviate from typical interaction patterns. Analysts, including historians, political scientists, and journalists, commonly read large quantities of text to construct an accurate picture of when and where an event happened, who was involved, and in what ways. In this paper, 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 that distinguishes between topics that describe "business as usual" and topics that deviate from these patterns. To demonstrate this model, we analyze a corpus of over two million U.S. State Department cables from the 1970s. We provide an open-source implementation of an inference algorithm for the model and a pipeline for exploring its results.

1 Introduction

Foreign embassies of the United States government communicate with one another and with the U.S. State Department through diplomatic cables. The National Archive collects these cables in a corpus, which traces the (unclassified) diplomatic history of the United States. The corpus contains, for example, over two million cables sent between 1973 and 1978.

Most of these cables describe diplomatic "business as usual," such as arrangements for visiting officials, recovery of lost or stolen passports, or obtaining lists of names for meetings and conferences. For example, the embassies sent 8,635 cables during the week of April 21, 1975. Here is one, selected at random:

> Hoffman, UNESCO Secretariat, requested info from PermDel concerning an official invitation from the USG RE subject meeting scheduled 10–13 JUNE 1975, Madison, Wisconsin. Would appreciate info RE status of action to be taken in order to inform Secretariat. Hoffman communicating with Dr. John P. Klus RE list of persons to be invited.

But, hidden in the corpus are also cables about important diplomatic events—the cables and events that are most interesting to historians, political sceintists, and journalists. For example, during that same week, the U.S. was in the last moments of the Vietnam war and, on April 30, 1975, lost its hold on Saigon. This triggered the end of the war and a max exodus of refugees. Here is one of the cables about this event:

> GOA program to move Vietnamese Refugees to Australia is making little progress and probably will not cover more than 100-200 persons. Press comment on smallness of program has recognized difficulty of getting Vietnamese out of Saigon, but "Canberra Times" Apr 25 sharply critical of government's performance. [...] Labor government clearly hopes whole matter will somehow disappear.

Our goal in this paper is to develop a tool to help historians, political scientists, and journalists wade through corpora of documents to find potentially significant events and the primary sources around them. We present *Capsule*, a probabilistic model for detecting and characterizing important events, such as the

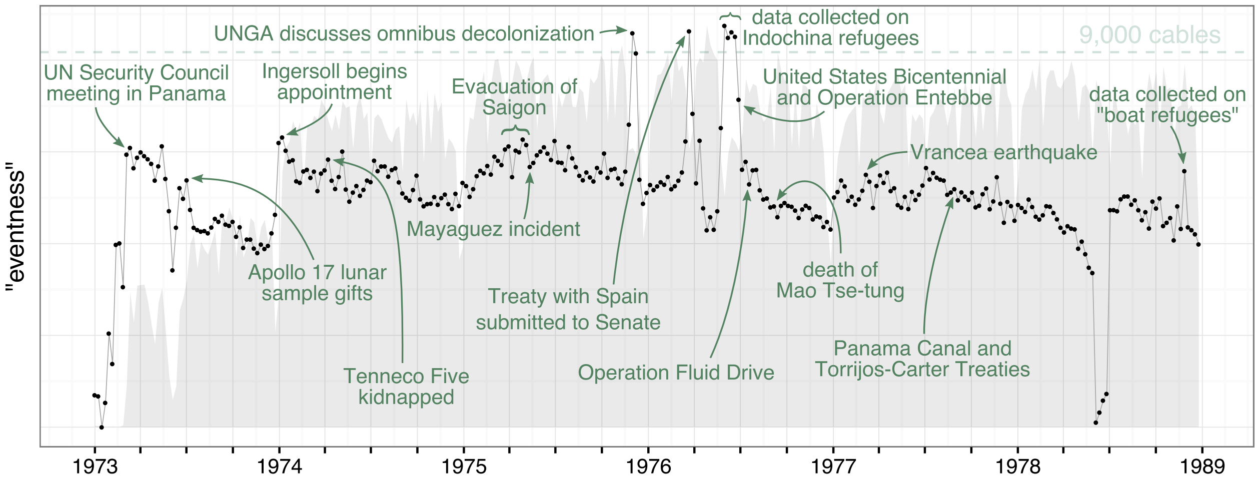

Figure 1: Measure of "eventness," or time interval impact on cable content (Eq. 3.3). Grey background indicates the number of cables sent over time. This comes from the model fit we discuss in Section 5. Capsule successful detects real-world events from National Archive diplomatic cables.

fall of Saigon, in large corpora of historical communication, such as diplomatic cables from the 1970s.

Figure 1 illustrates Capsule's analysis of two million cables from the National Archives' corpus. The y-axis represents "eventness," a loose measure of how strongly a week's cables deviate from typical diplomatic "business as usual" to discuss some matter that is common to many embassies. (We describe this measure of "eventness" in detail in section 3.)

The figure shows that Capsule detects many of the important moments during this five-year span, including the Air France hijacking and Israeli rescue operation "Operation Entebbe" (June 27–July 4, 1976), and the fall of Saigon (April 30, 1975). It also identifies other moments, such as the U.S. sharing lunar rocks with other countries (March 21, 1973) and the death of Mao Tse-tung (Sept. 9, 1976). Broadly speaking, Capsule gives a picture of the diplomatic history of these five years; it identifies and characterizes moments and source material that might be of interest to a historian.

The intuition behind Capsule is this: Embassies write cables throughout the year, usually describing typical diplomatic business, such as visits from government officials. Sometimes, however, important events occur, such as the fall of Saigon, that pull embassies away from their typical activities and lead them to write cables that discuss these events and their consequences. Capsule therefore operationalizes an "event" as a moment in history when multiple embassies deviate from their usual topics of discussion and each embassy deviates in the same way.

Capsule embeds this intuition into a Bayesian model that uses latent variables to encode what "business as usual" means for each embassy, to characterize the events of each week, and to identify the cables that discuss those events. Given a corpus of cables, the corresponding posterior distribution of the latent variables provides a filter for the cables that isolates important moments in diplomatic history. Figure 1 depicts the mean of this posterior distribution.

We present the Capsule model in section 3, providing both a formal model specification and guidance on how to use the model to detect and characterize real-world events. In section 4, we validate Capsule using simulated data, and in section 5, we use it to analyze over two million U.S. State Department cables. Although we describe Capsule in the context of diplomatic cables, it is suitable for exploring any corpus with the same underlying structure: text (or other discrete multivariate data) generated over time by known entities. This includes email, consumer behavior, social media posts, and opinion articles.

2 Related Work

We first review previous work on automatic event detection and other related concepts, to contextualize our approach in general and Capsule in particular.

In both univariate and multivariate settings, analysts often want to predict whether or not rare events will occur (Weiss and Hirsh, 1998; Das et al., 2008). In contrast, Capsule is intended to help analysts explore and understand their data; our goal is human interpretability rather than prediction or forecasting.

Events can be construed as either anomalies—temporary deviations from usual behavior—or "changepoints" that mark persistent shifts in usual behavior (Guralnik and Srivastava, 1999; Adams and MacKay, 2007). We focus on events as anomalies.

Event detection in the context of 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) usually means identifying clusters of documents. For news, the goal is to create new clusters as novel stories appear, without distinguishing between typical content and rare events; for social media, the goal is to identify rare events, but the resultant methods are intended for short documents, and are not appropriate for longer documents that contain information about a variety of subjects.

Many existing methods for detecting events from text focus on individual vocabulary terms, often weighted by tf-idf values (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). We characterize events by bursts in groups of terms.

Although groups of terms can be summarized directly (Peng et al., 2007; Chakrabarti and Punera, 2011; Gao et al., 2012), topic models (Blei, 2012) provide a way to automatically identify groups of related terms and reduce the dimensionality of text data. Researchers have previously used topic models to detect events mentioned in social media posts (Lau et al., 2012; Dou et al., 2012) and to find posts relevant to particular, monitored events (VanDam, 2012). Capsule uses topics to characterize both typical diplomatic content and potentially significant events.

In addition to modeling text over time, researchers have also used spatial information (Neill et al., 2005; Mathioudakis et al., 2010; Liu et al., 2011) and information about authors (Zhao et al., 2007) and news outlets (Wang et al., 2007) to enhance event detection. We rely on author information to characterize diplomatic "business as usual" for each embassy.

Event detection is closely related to detecting and characterizing relationships between entities (Schein et al., 2015; Linderman and Adams, 2014; Das Sarma et al., 2011). Capsule can trivially use sender–receiver pairs instead of authors, and the model specification can be tailored to reflect network structure.

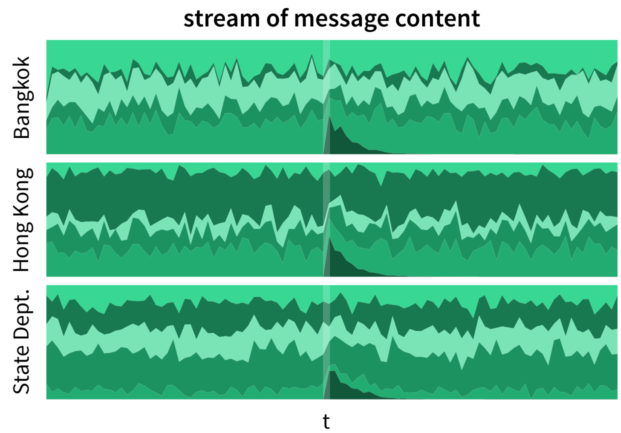

Figure 2: Cartoon intuition. The y-axis represents the stacked proportions of cables about various topics, while the x-axis represents time. The Bangkok embassy, Honk Kong embassy, and U.S. State Department all have typical diplomatic business, about which they usually send cables. When an event occurs during time interval t, the cables alter to cover the event before returning to "business as usual." Capsule discovers the embassies' typical concerns, as well as the timing and content of events.

Finally, there are connections between Capsule and recent work on Poisson processes. In particular, we can interpret Capsule as a collection of related discrete-time Poisson processes with random intensity measures. Further, marginalizing out the event strengths (described in section 3.1) reveals that the use of a vocabulary term by one embassy can "excite" the use of that term by another. This suggests a close relationship to Hawkes processes (Hawkes, 1971).

3 The Capsule Model

In this section, we present the Capsule model for detecting and characterizing significant diplomatic events. We first provide the intuition behind Capsule, and then formally specify the model. We also explain how to use Capsule to explore a corpus and how to learn the posterior distribution of the latent variables.

Consider an entity like the Bangkok embassy, as illustrated in figure 2. We can imagine that this entity sends a stream of diplomatic cables over time—some to the U.S. State Department, others to other

American embassies, such as the one in Hong Kong. Embassies usually write cables that describe typical diplomatic business. For example, the Bangkok embassy might write about topics regarding southeast Asia more generally. We can think of a topic as being a probability distribution over vocabulary terms.

Now imagine that an event, such as the capture of Saigon during the Vietnam war, occurs during a particular time interval t. We cannot directly observe the occurrence of this event, but we can observe the stream of cables and the event's impact on it. When the event occurs, multiple entities deviate from their usual topics of discussion simultaneously, before returning to their usual behavior, as depicted in figure 2. For example, the day after the capture of Saigon, the majority of the diplomatic cables written by the Bangkok embassy and several other entities were about Vietnam war refugees. If we think of the event as another probability distribution over vocabulary terms, then each entity's stream of cables reflects its typical concerns, as well as any significant events.

3.1 Model Specification

We now define the Capsule model. Our data come from *entities* (e.g., embassies) who send *messages* (e.g., diplomatic cables) over *time*; specifically, we observe the number of times n_{dv} that each vocabulary term v occurs in each message d. Each message is associated with an author entity a_d and a time interval t_d within which that message was sent.

We model each message with a bank of Poisson distributions—one for each vocabulary term:

$$n_{dv} \sim \text{Poisson}\left(\lambda_{dv}\right). \tag{1}$$

The rate λ_{dv} blends the different influences on message content. Specifically, it blends three types of *topics*, intended to capture "business-as-usual" discussion and content related to significant events.

We operationalize each topic as a specialized probability distribution over vocabulary terms (the set of unique words in the corpus of messages), as is common in topic models (Blei et al., 2003; Canny, 2004; Gopalan et al., 2014)—i.e., each term is associated with each topic, but with a different probability.

Each message blends 1) general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ about diplomacy (e.g., terms about diplomats, terms about communication), 2) an entity topic $\boldsymbol{\eta}_{a_d}$ specific to the author of that message (e.g., terms about Asia),[1] and 3) event topics $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T$ that are specific to the events in recent time intervals (e.g., terms about a coup, terms about the death of a dignitary).

Topic Type	Top Terms
General	visit, hotel, schedule, arrival
Entity	soviet, moscow, ussr, agreement
Event	saigon, evacuation, vietnam, help

Table 1: The highest-probability vocabulary terms for examples of the three types of topics (general, entity, and event). These examples come from the analysis that we describe section 5.

Examples of these three types of topics are in table 1. The general topic relates to planning travel, the entity topic captures words related to the U.S.S.R., and the event topic captures words related to the evacuation of Saigon toward the end of the Vietnam War.

The messages share the three types of topics in different ways: all messages share the general topics, messages written by a single entity share an entity topic, and messages in the same time interval use the event topics in similar ways. Each message blends its corresponding topics with a set of message-specific strengths. As a result, each message captures a different mix of general diplomacy discussion, entity-specific terms, and recent events. Specifically, the Poisson rate for vocabulary term v in message d is

$$\lambda_{dv} = \sum_{k=1}^{K} \theta_{dk}\beta_{kv} + \zeta_d \eta_{a_d v} + \sum_{t=1}^{T} f(t_d, t)\,\epsilon_{dt}\gamma_{tv}, \tag{2}$$

where θ_{dk} is message d's strength for general topic k, ζ_d is message d's strength for a_d's entity topic, and ϵ_{dt} is message d's strength for event topic t. The function $f(\cdot)$ ensures that the events influences decay over time. As described in appendix B, we

We found that an exponential decay function, as in equation (8), works well in practice.

As described in Appendix B, we assessed the sensitivity of o ur model to different settings of event duration τ three different decay f unctions f: exponential, linear, and step functions. We found that fitting C apsule with an exponential decay function,

[1] The entity-specific topics play a similar role to the background topics introduced by Paul and Dredze (2012).

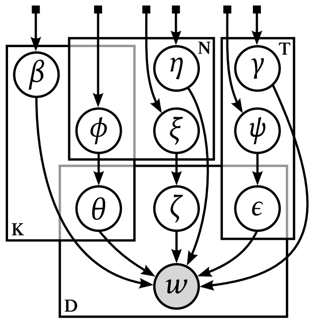

Figure 3: Graphical model for Capsule. Observed term counts depend on general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, entity topics $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_A$, and event topics $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T$, as well as message-specific strengths $\boldsymbol{\theta}_d$, ζ_d, and $\boldsymbol{\epsilon}_d$. Variables $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A$ and $\xi_1, \ldots, \xi_A$ represent entity-specific strengths, while $\psi_1, \ldots, \psi_T$ allow time intervals to be more or less "eventful." Black squares denote hyperparameters (unlabeled for visual simplicity).

or

$$f(i_d, t) = \begin{cases} 0, & \text{if } t \leq i_d < t + \tau \\ \exp\left\{\frac{-(i_d - t)}{\tau/5}\right\}, & \text{otherwise,} \end{cases} \tag{3}$$

provided the best performance and the most interpretable results.

We place hierarchical gamma priors over the message-specific strengths, introducing entity-specific strengths $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A$ and $\xi_1, \ldots, \xi_A$ that allow different entities to focus on different topics and event strengths $\psi_1, \ldots, \psi_T$ that allow different time intervals to be more or less "eventful." We place Dirichlet priors over the topics. The graphical model is in figure 3 and the generative process is in figure 4.

Given a corpus of messages, learning the posterior distribution of the latent variables uncovers the three types of topics, the message- and entity-specific strengths, and the event strengths. In section 3.3, we explain how an analyst can use the event strengths as a filter that isolates potentially significant messages.

3.2 Learning the Posterior Distribution

In order to use Capsule to to explore a corpus of messages, we must first learn the posterior distribution of the latent variables—the general topics, the entity topics, the event topics, the message- and entity-specific strengths, and the event strengths—conditioned on the observed term counts. As for many Bayesian models, this posterior distribution is not tractable to

- for $k = 1, \ldots, K$,
  - draw general topic $\boldsymbol{\beta}_k \sim \text{Dirichlet}_V(\alpha, \ldots, \alpha)$
  - for each entity $a = 1, \ldots, A$,
    - draw entity-specific strength $\phi_{ak} \sim \text{Gamma}(s, r)$
- for each entity $a = 1, \ldots, A$,
  - draw entity topic $\boldsymbol{\eta}_a \sim \text{Dirichlet}_V(\alpha, \ldots, \alpha)$
  - draw entity-specific strength $\xi_a \sim \text{Gamma}(s, r)$
- for each time interval $t = 1, \ldots, T$,
  - draw event topic $\boldsymbol{\gamma}_t \sim \text{Dirichlet}_V(\alpha, \ldots, \alpha)$
  - draw event strength $\psi_t \sim \text{Gamma}(s, r)$
- for each message $d = 1, \ldots, D$, sent during time interval t_d by author entity a_d,
  - for each general topic $k = 1, \ldots, K$,
    - draw message-specific strength $\theta_{dk} \sim \text{Gamma}(s, \phi_{a_d k})$
  - draw message-specific strength $\zeta_d \sim \text{Gamma}(s, \xi_{a_d})$
  - for each time interval $t = 1, \ldots, T$,
    - draw message-specific strength $\epsilon_{dt} \sim \text{Gamma}(s, \psi_t)$
  - for each vocabulary term $v = 1, \ldots, V$,
    - set $\lambda_{dv} = \sum_{k=1}^{K} \theta_{dk}\beta_{kv} + \zeta_d \eta_{a_d v} + \sum_{t=1}^{T} f(t_d, t)\, \epsilon_{dt}\gamma_{tv}$
    - draw term counts $n_{d,v} \sim \text{Poisson}(\lambda_{dv})$

Figure 4: Generative process for Capsule. We use s and r to denote top-level (i.e., fixed) shape and rate hyperparameters; they can be set to different values for different variables.

compute; approximating it is therefore our central statistical and computational problem. We introduce an approximate inference algorithm for Capsule, based on variational methods (Jordan et al., 1999),[2], which we outline in appendix A.[3] This algorithm produces a fitted variational distribution which be can then be used as a proxy for the true posterior distribution.

[2] Source code: https://github.com/ajbc/capsule.

[3] Appendices are in the supplemental material.

3.3 Detecting and Characterizing Events

Having approximated the posterior distribution of the latent variables, we can use the mean of this distribution to explore the data. Specifically, we can explore "business-as-usual" content using the posterior expected values of the general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ and the entity topics $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_A$, and we can detect and characterize significant events using the posterior expected values of the event strengths and event topics.

$$m_j = \frac{1}{\sum_d f(i_d, j)} \sum_d \frac{\varepsilon_{d,j}}{\zeta_d + \sum_t \varepsilon_{d,t} + \sum_k \mathbb{E}[\theta_{d,k}]},$$

where $\varepsilon_{d,t} = f(i_d, t)\mathbb{E}[\epsilon_{d,t}]$. TODO... This measure of "eventness" estimates the fraction of term occurrences that are related to the event in time interval t. Figure 1 illustrates real-world events detected using this measure.

We can characterize an event t by selecting the highest-probability vocabulary terms from $\mathbb{E}[\boldsymbol{\gamma}_t]$. We can also identify the most strongly-associated messages by computing $f(t_d, t)\ \mathbb{E}[\epsilon_{dt}]$ for each message d and ordering the messages accordingly. In section 5, we explore the cables associated with significant events in the National Archives' corpus of diplomatic cables. To make Capsule more accessible for historians, political scientists, and journalists, we have released an open-source tool for visualizing its results.[4] This tool allows analysts to browse a corpus of messages and the mean of the corresponding posterior distribution, including general topics, entity topics, and event topics. Figure 5 contains several screenshots of the tool's browsing interface.

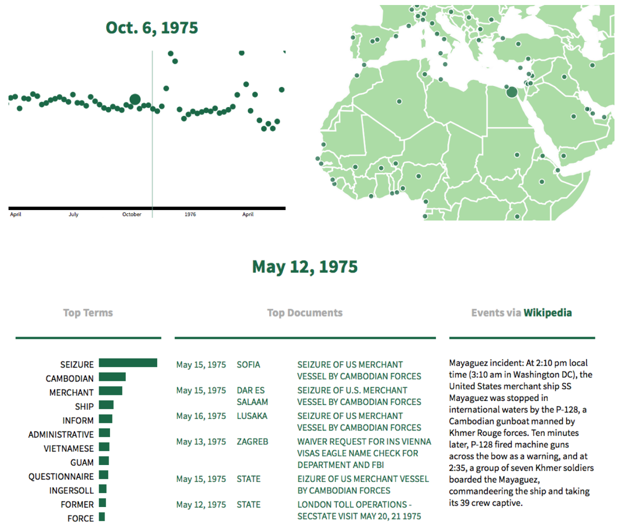

Figure 5: Screenshots of the Capsule visualization tool used to explore U.S. State Department cables. Top left: events over time (similar to Figure 1). Top right: entities located on a map. Bottom: summary of the week of May 12, 1975, including top vocabulary terms, relevant cables, and text from Wikipedia.

4 Model Validation with Simulated Data

Before using Capsule to explore a corpus of real messages (described in section 5), we provide a quantitative validation of the model using simulated data.

We used the generative process in figure 4 to create ten data sets, each with 100 time intervals, ten general topics, ten entities, and roughly 20,000 messages. We then used these data sets to compare Capsule's event detection performance to that of four baseline methods. We also compared the methods' abilities to identify the most relevant messages for each event.

4.1 Detecting Events

For each data set, we ordered the time intervals from most to least eventful, using the "eventness" measure described in section 3.3 and the simulated values of the latent variables. We then treated these ranked lists of time intervals as "ground truth" and assessed how well each method was able to recover them.

For Capsule itself, we used our approximate inference algorithm to obtain a fitted variational distribution for each simulated data set. We then ordered the time intervals using our "eventness" measure and the posterior expected values of the latent variables.

For our first baseline, we constructed an "event-only" version of Capsule by dropping the first and second terms in equation (2). We used this baseline to test whether modeling "business as usual" discussion makes it easier to detect significant events. We obtained a fitted variational distribution for this model using a variant of our approximate inference algorithm, and then ordered the time intervals using our "eventness" measure, modified appropriately, and the posterior expected values of the latent variables.

For our second baseline, we drew inspiration from

[4]Source code: https://github.com/ajbc/capsule-viz; demo: http://www.princeton.edu/~achaney/capsule/.

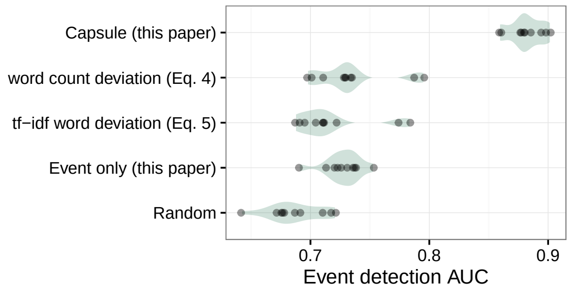

Figure 6: Event detection performance using ten simulated data sets. Each dot represents the performance (equation (5)) of a single method on a single data set; the shaded green area summarizes the distribution of performance for a single method.

previous work on event detection in the context of news articles, and focused on each time interval's deviation in term counts from the average. Specifically, we ordered the time intervals $t = 1, \ldots, T$ for each simulated data set according to this measure:

$$m_t = \sum_{v=1}^{V} \sum_{\substack{d=1 \\ t_d = t}}^{D} \left| n_{dv} - \frac{1}{D} \sum_{d=1}^{D} n_{dv} \right|. \quad (4)$$

We added tf-idf term weights for our third baseline:

$$m_t = \sum_{v=1}^{V} \text{tf-idf}\,(v) \sum_{\substack{d=1 \\ t_d = t}}^{D} \left| n_{dv} - \frac{1}{D} \sum_{d=1}^{D} n_{dv} \right|. \quad (5)$$

Finally, we randomly ordered the time intervals for each data set to serve as a straw-man baseline.

We also experimented with baselines that involved term-count deviations on the entity level and topic-usage deviations on the document level (Dou et al., 2012), but found that they were not competitive.

For each data set, we compared each method's ranked list of time intervals to the corresponding "ground-truth" list of time intervals, by dividing the sum of the lists' actual set overlap at each rank by the sum of their maximum set overlap at each rank:

$$\frac{\sum_{r=1}^{T} |S_r^{\text{truth}} \cap S_r^{\text{method}}|}{\sum_{r=1}^{T} r}, \quad (6)$$

where S_r^{truth} is a set of the top r time intervals according to the "ground-truth" list and S_r^{method} is a set of the top r time intervals according to the method.

Figure 6 shows that Capsule outperforms all four baseline methods. These results serve as a sanity check for the model and its implementation.

4.2 Identifying Relevant Messages

For each data set, we created a list of the most relevant messages for each time interval t by computing $f(t_d, t)\, \epsilon_{dt}$ for each message d (using the simulated values of ϵ_{dt}) and ordering the messages accordingly. We then treated these ranked lists of messages as "ground truth" and assessed how well Capsule and the baseline methods were able to recover them.

For Capsule, we used our approximate inference algorithm to obtain a fitted variational distribution for each simulated data set, and then, for each time interval, ordered the messages according to $f(t_d, t)\, \mathbb{E}[\epsilon_{dt}]$. For our second and third baselines, we ordered the messages according per-document versions of equation (3) and equation (4)—i.e.,

$$m_{dt} = \sum_{v=1}^{V} \left| n_{dv} - \frac{1}{D} \sum_{d=1}^{D} n_{dv} \right| \quad (7)$$

and

$$m_{dt} = \sum_{v=1}^{V} \text{tf-idf}\,(v) \left| n_{dv} - \frac{1}{D} \sum_{d=1}^{D} n_{dv} \right|. \quad (8)$$

For each data set, we compared each method's ranked list of documents for each time interval to the corresponding "ground-truth" list, by computing precision at 10 documents. The average precision for Capsule was was 0.44, while the average precision for the "event-only" version of the model was 0.09. The other baselines recovered zero relevant messages.

5 Exploratory Analysis

Capsule is intended to help analysts explore and understand their data. In this section, we demonstrate its capabilities by analyzing a corpus of over two million U.S. State Department cables from the 1970s.

5.1 Data

The National Archive collects diplomatic cables sent between the U.S. State Department and its foreign embassies. We obtained a subset of this corpus from the Central Foreign Policy Files at the National Archives, via the History Lab at Columbia University.[5] The subset contains over two million cables sent between 1973 and 1978. In addition to the text

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

of the cables, each message is labeled with its author (e.g., the U.S. State Department, a particular embassy, or a named individual), its recipients (often several), and the date the cable was sent. We used a vocabulary of 6,293 terms and omitted cables with fewer than three terms, resulting in 2,021,852 cables sent between 22,961 entities. We used weekly time intervals, as few cables were sent on the weekends.

5.2 Model Settings

We ran our approximate inference algorithm for Capsule to obtain a fitted variational distribution. We used $K = 100$ general topics, the exponential decay function in equation (8) with $\tau = 4$, and top-level hyperparameters $s = r = 0.3$. With these settings, a single iteration of the algorithm took about an hour.[6]

5.3 Quantitative Results

Similar to the validation on simulated data discussed in section 4, we can validate Capsule on this real-world data. Here, we focus on event detection and held-out data likelihood.

The History Lab at Columbia University provided us with a list of thirty-nine real-world events during that took place between 1973 and 1978. These events are present in at least one of six reputable collections of historic events, such as the Office of the Historian's Milestones in the History of U.S. Foreign Relations.[7]

We ran Capsule and baseline comparison methods to recover these events, and used the nDCG metric to evaluate the methods. The nDCG metric is discounted cumulative gain,

$$\text{DCG} = \sum_{j=1}^{T} \frac{\mathbf{1}[\text{interval at rank } j \text{ in known events}]}{\log j}, \tag{9}$$

divided by the ideal DCG value, or

$$\text{nDCG} = \frac{\text{DCG}}{\text{ideal DCG}}. \tag{10}$$

As shown in Table 2, Capsule outperforms the baselines.

[6]Each iteration of our algorithm considers all messages. Modifying it to stochastically sample the data would reduce the time required to obtain an equivalent fitted variational distribution.

[7]https://history.state.gov/milestones/1969-1976

Model	LL, 10 iter.	LL, final
Full Capsule	-1.62e7	-1.52e7
Entity Topics Only	-1.64e7	–
General Topics Only	-1.71e7	-1.53e7
Event Only	-1.79e7	–

Table 3: Log likelihood (LL) computed on validation data at 10 iterations and at convergence—the event only and entity only models are small enough that they converge with very few iterations. The full Capsule model achieves the lowest log likelihood in both cases.

Additionally, we computed held-out validation data likelihood on the model and each of its component parts; Table 3 shows that the full Capsule model captures the data better than any of its component parts individually.

5.4 Exploration

Having validated that Capsule can detect real-world events, we now turn to our primary goal—using Capsule to explore and understand a corpus of messages.

Figure 1 shows the "eventness" measure described in section 3.3 over time. High values—which are often anomalous peaks—correspond to real-world events. One of the tallest peaks occurs during the week of December 1, 1975, when the United Nations General Assembly (UNGA) discussed omnibus decolonization. As described in section 3.3, we can characterize this event by computing $f(t_d, t)\, \mathbb{E}[\epsilon_{dt}]$ for each message d and then sorting the messages accordingly. Table 4 lists the top-ranked cables.

Another notable event was the seizure of the S.S. Mayaguez, an American merchant vessel, during May, 1975, at the end of the Vietnam War. The top-ranked cables for this event are in table 5. We can examine the individual cables to confirm their relevancy and learn more about the event. For example, here is the most relevant cable, according to Capsule:

> In absence of MFA Chief of Eighth Department Avramov, I informed American desk officer Yankov of circumstances surrounding seizure and recovery of merchant ship Mayaguez and its crew. Yankov promised to inform the Foreign Minister of US statement today (May 15). Batjer

Method	nDCG
Capsule	0.693
Average tf-idf weighted word count deviation	0.652
Average unweighted word count deviation	0.642
Single term maximum tf-idf weighted deviation	0.561
Random (10k ave)	0.557
Single term maximum unweighted deviation	0.555

Table 2: Evaluation of Capsule and comparison baselines on a collection of 39 real-world events. Capsule performs best.

$f(t_d, t)\mathbb{E}[\epsilon_{dt}]$	Date	Author Entity	Subject
4.60	1975-12-05	Canberra	30th UNGA: Item 23, Guam, Obmibus Decolonization and ...
4.26	1975-12-05	Mexico	30th UNGA-Item 23: Guam, Omnibus Decolonization and ...
4.21	1975-12-06	State	30th UNGA-Item 23: Guam, Omnibus Decolonization and ...
4.11	1975-12-03	Dakar	30th UNGA: Resolutions on American Samoa, Guam and ...
4.08	1975-12-04	Monrovia	30th UNGA: Item 23: Resolutions on decolonization and A...

Table 4: Top-ranked cables for the week of December 1, 1975, when the United Nations General Assembly discussed decolonization resolutions. Capsule accurately recovers cables related to this real-world event. Typos are intentionally copied from the data.

A third week of interest occurs in early July of 1976. On July 4th, the US celebrated its Bicentennial, but on the same day, Israeli forces completed a hostage rescue mission—an Air France flight from Tel Aviv had been hijacked and taken to Entebbe, Uganda. This event, like many events, is mostly discussed the week following the real-world event; relevant cables are shown in Appendix B, Table 6. The cable from Stockholm describing the "Ugandan role in Air France hijacking" begins with the following content, which reveals further information about the event.

> 1. We provided MFA Director of Political Affairs Leifland with Evidence of Ugandan assistance to hijackers contained in Ref A. After reading material, Leifland described it a "quite good", and said it would be helpful for meeting MFA has scheduled for early this morning to determine position GOS will take at July 8 UNSC consideration of Israeli Rescue Operation. ...

Capsule assumes that only one event occurs in each time interval—this example is a clear violation of this assumption, but it also demonstrates that the model successfully captures both events, even when they overlap.

In addition to events, Capsule can be used to explore the general themes of a corpus and entities' typical concerns. Examples of general topics of conversation are shown in Appendix B, Table 7 and entity-exclusive topics are shown in Appendix B, Table 8; these show us how entity topics absorb 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.

6 Conclusion

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 comparison methods and explored its results on a real-world datasets. We anticipate that Capsule can be used by historians, political scientists, and others who wish to investigate events in large text corpora.

References

Ryan Prescott Adams and David JC MacKay. 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

$f(t_d, t)\mathbb{E}[\epsilon_{dt}]$	**Date**	**Author Entity**	**Subject**
5.06	1975-05-15	Sofia	Seizure of US merchant vessel by Cambodian forces
5.05	1975-05-15	Dar es Salaam	Seizure of U.S. merchant vessel by Cambodian forces
4.92	1975-05-16	Lusaka	Seizure of US merchant vessel by Cambodian forces
4.61	1975-05-13	Zagreb	Waiver request for INS Vienna visas Eagle name check...
4.59	1975-05-15	State	eizure of US merchant Vessel by Cambodian forces

Table 5: Top-ranked cables for the week of May 12, 1975, when the S.S. Mayaguez, an American merchant vessel, was captured. Capsule accurately recovers cables related to this real-world event. Typos are intentionally copied from the data.

James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 291–300.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, January.

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 ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 330–337.

John Canny. 2004. Gap: a factor model for discrete data. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129.

Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 11:66–73.

Kaustav Das, Jeff Schneider, and Daniel B Neill. 2008. Anomaly pattern detection in categorical datasets. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 169–176.

Anish Das Sarma, Alpa Jain, and Cong Yu. 2011. Dynamic relationship and event discovery. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 207–216.

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 International Conference on Very Large Data Bases (VLDB)*, 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 International Conference on Information and Knowledge Management (CIKM)*, pages 1173–1182.

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 (NIPS)*, pages 3176–3184. Curran Associates, Inc.

Valery Guralnik and Jaideep Srivastava. 1999. Event detection from time series data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 33–42.

Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

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.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 297–304.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *Proceedings of the International Conference on Computational Linguistics (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 ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 106–113.

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 ACM SIGMM International Workshop on Social Media (WSM)*, pages 3–8.

Michael Mathioudakis, Nilesh Bansal, and Nick Koudas. 2010. Identifying, attributing and describing spatial bursts. *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, 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 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 218–227.

Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from twitter. *Health*, 11:16–6.

Wei Peng, Charles Perng, Tao Li, and Haixun Wang. 2007. Event summarization for system management. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1028–1032.

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 International World Wide Web Conference (WWW)*, pages 851–860.

Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *Proceedings of the International AAAI Conference on Web and Social Media (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 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1045–1054.

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

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 784–793. ACM.

Gary M Weiss and Haym Hirsh. 1998. Learning to predict rare events in event sequences. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 359–363.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–88.

Qiankun Zhao, Prasenjit Mitra, and Bi Chen. 2007. Temporal and information flow based event detection from social text streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 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.