Detecting and Characterizing Events

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

Hanna Wallach
Microsoft Research
wallach@microsoft.com

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

Matthew Connelly
Columbia University
mjc96@columbia.edu

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 an 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

Foreign embassies of the United States government communicate with each other and with the U.S. State Department through cabled message. The National Archive collects these documents in a running corpus, which traces the (unclassified) diplomatic history of the United States. It has collected, for example, about two million cables sent between 1973 and 1978.

Typically, a cable from this collection describes 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 of primary interest to historians. During that same week, the United States was in the last moments of the Vietnam war and, on April 30, 1975, lost its hold on Saigon. This resulted in the end of the Vietnam War and a max exodus of refugees from the country. One of the cables around this event is

> 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 method to help historians and political scientists wade through their collections, such as the 1970s cables, to find potentially important events, such as the fall of Saigon,

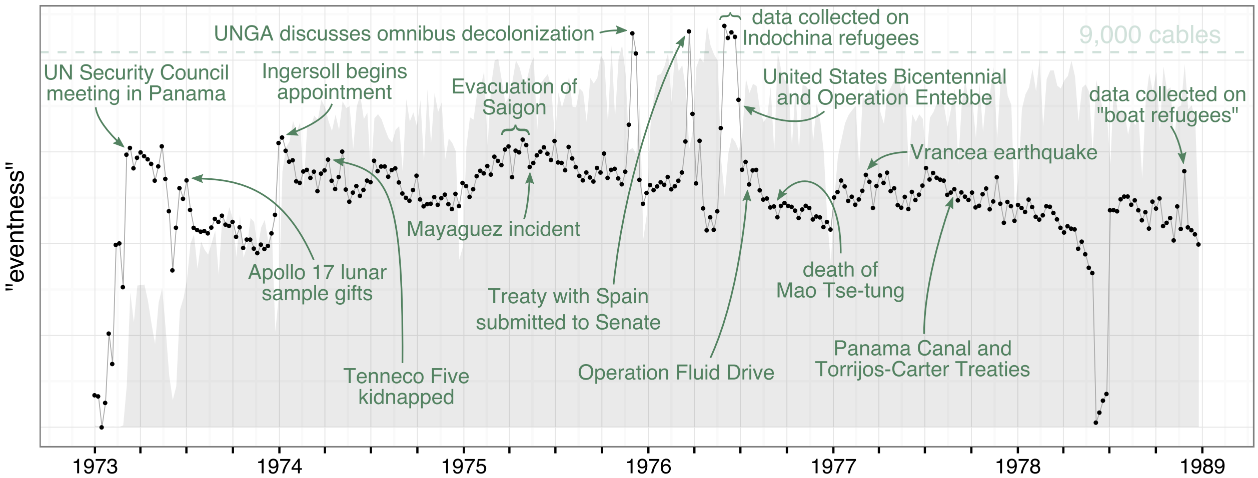

Figure 1: Measure of "eventness," or time interval impact on cable content (Eq. 2). Grey background indicates the number of cables sent over time. This comes from the model fit we discuss in Section 3. Capsule successful detects real-world events from National Archive diplomatic cables.

and the primary sources around them. We develop *Capsule*, a probabilistic model for detecting and characterizing important events in large collections of historical communication.

Figure 1 illustrates Capsule's analysis of the two million cables from the National Archives. The y-axis is "eventness", a loose measure how strongly a week's cables deviate from the usual diplomatic chatter to discuss a matter that is common to many embassies. (This is described in detail in Section 2.)

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 business such as the visiting of a government official. Sometimes, however, there is an important event, e.g., the fall of Saigon. When an event occurs, it pulls embassies away from their typical business to write cables that discuss what happened and its consequences. Thus Capsule effectively defines an "event" to be a moment in history when embassies deviate from what each usually discusses, and when each embassy deviates in the same way.

Capsule embeds this intuition into a Bayesian model. It uses hidden variables to encode what "typical business" means for each embassy, how to characterize the events of each week, and which cables discuss those events. Given a corpus, the corresponding posterior distribution provides a filter on the cables that isolates important moments in the diplomatic history. Figure 1 illustrates the mean of this posterior.

Capsule can be used to explore any corpora 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.

We present the model in Section 2, providing both a formal model specification and guidance on how to use its posterior to detect and characterize real-worlds events. In Section 3, we evaluate Capsule and explore its results on a collection of U.S. State Department cables and on simulated data.

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 that analysts want to predict whether or not 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.

Events can also be construed as "change points" to mark when typical observations shift semi-permanently from one value to another (Guralnik and Srivastava, 1999; Adams and MacKay, 2007). Both varieties of events are important, but we focus on temporary shifts away from normal.

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, usually 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 the 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 can use a sender-receiver pair in place of an author, but the model could be further tailored for network interactions.

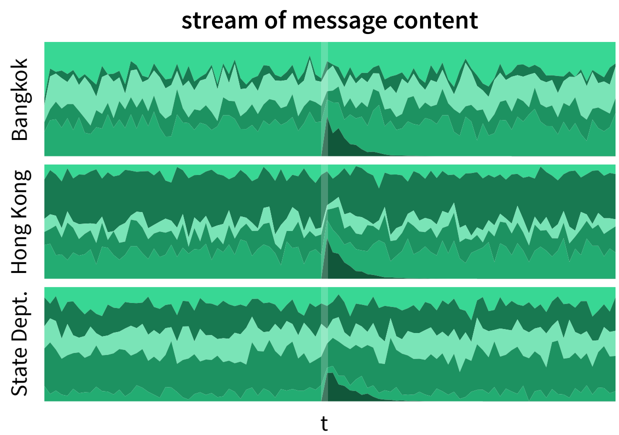

Figure 2: Cartoon intuition of Capsule; the y axis is the stacked proportion of messages about various subjects during a given time interval. The Bangkok embassy, Honk Kong embassy, and State Department all have typical concerns about which they usually send messages. When an events occurs at time t, the stream of message content alters to include the event, then fades back to "business as usual." Capsule discovers entities' typical concerns as well as the event occurrence and content.

2 The Capsule Model

In this section we develop the Capsule model for detecting and characterizing events. Capsule relies on text data sent between entities over time, and builds on topics models. We first give the intuition on Capsule, then formally specify the model. We also describe how to explore a corpus using Capsule and how we learn its hidden variables.

Consider an entity like the Bangkok American embassy, shown in Figure 2. 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 will 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—significant events impact multiple parties

topic type	top terms
general	visit, hotel, schedule, arrival
entity	soviet, moscow, ussr, agreement
event	saigon, evacuation, vietnam, help

Table 1: Top vocabulary terms for examples of each of the three topic varieties; these three types of topics blend to form the distribution of each message. They come from the model fit we discuss in Section 3.

simultaneously. The day following the capture of Saigon, for instance, the majority of the diplomatic cables sent by the Bangkok embassy and several other entities 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 now define Capsule in detail. Our data are *entities* sending *messages* over *time*. The observed variables are $w_{d,v}$, the number of times term v occurs in message d. The message is associated with an entity (or author) a_d and a time (or date) interval i_d.

We model each message with a bank of Poisson distributions, one per term in the vocabulary, $w_{d,v} \sim \text{Poisson}(\lambda_{d,v})$. The rate $\lambda_{d,v}$ blends the different influences on the content of the message, which are defined in terms of different types of *topics*. A topic, as in typical topic modeling (Blei et al., 2003; Canny, 2004; Gopalan et al., 2014), is a distribution over terms.

Specifically, the message blends general topics about diplomacy (e.g., diplomats, communication) β_k, an entity topic that is specific to the author of the message (e.g., terms about France) η_{a_d},[1] and an event topic that is specific to the events of relevant recent weeks (e.g., terms about an international crisis) γ_t. Notice how messages share these topics in different configurations: all messages share the general topics; messages from the same entity share the entity topics; and messages from the same interval share the event topics.

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.

[1]These entity-specific topics are similar to background topics (Paul and Dredze, 2012).

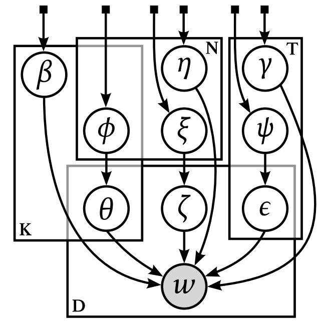

Figure 3: The graphical model for Capsule. Observed words w depend on general topics β, entity-specific topics η, and event topics γ, as well as document representations θ, ξ, and ϵ. Variables ϕ and ζ represent entity concerns (with general topics and entity-specific topics, respectively) and ψ represents the event strength of a given time interval. Hyper-parameters are indicated by black squares, but not labeled for visual simplicity.

Each message blends its corresponding topics with different strengths, which are drawn per message. Each message represents a different mix of the events of recent weeks, entity-specific items, and general diplomacy.

Putting this together, the Poisson rate for term v in document d is

$$\lambda_{d,v} = \theta_d^\top \beta_v + \zeta_d \eta_{a_d,v} + \sum_{t=1}^{T} f(i_d, t) \epsilon_{d,t} \gamma_{t,v}, \tag{1}$$

where θ_d corresponds to strength of general diplomacy, ζ_d corresponds to strength of entity-specific concerns, and ϵ_d corresponds to strength of events; 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 find that exponential decay, as in Equation (5), performs well.

We place gamma priors on the topic strengths and Dirichlet priors on the topics. The distributions of general and entity topic strengths are defined hierarchically by entity, capturing the different topics that each entity tends to discuss. The prior on the entity strength is also defined hierarchically; different weeks are more or less "eventful." The graphical model is shown in Figure 3 and the generative process is in Figure 4.

- for each time step $t = 1{:}T$,
  - draw interval description over vocabulary (event topic) $\gamma_t \sim \text{Dirichlet}_V(\alpha)$
  - draw interval strength $\psi_t \sim \text{Gamma}(s_\psi, r_\psi)$
- for each entity $n = 1{:}N$,
  - draw entity-specific topics over vocabulary $\eta_n \sim \text{Dirichlet}_V(\alpha)$
  - draw entity-specific topic strength $\xi_n \sim \text{Gamma}(s_\xi, r_\xi)$
- 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 $\zeta_d \sim \text{Gamma}(s_\zeta, \xi_{a_d})$
  - 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 $\epsilon_{d,t} \sim \text{Gamma}(s_\epsilon, \psi_t)$
  - for each vocabulary term $v = 1{:}V$,
    - set $\lambda_{d,v} = \theta_d^\top \beta_v + \zeta_d \eta_{a_d} + \sum_{t=1}^{T} f(i_d, t)\epsilon_{d,t}\gamma_{t,v}$
    - draw word counts $w_{d,v} \sim \text{Poisson}\left(\lambda_{d,v}\right)$

Figure 4: The generative process for Capsule.

Given a collection of messages, posterior inference uncovers the different types of topics and how each message exhibits them. We will see below, how inferences about the event strengths enable us to filter the corpus to find important messages.

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 strength prior reveals that word use from one entity can "excite" word use in another, which suggests a close relationship to Hawkes processes (Hawkes, 1971).

Detecting and characterizing events. Once we estimate the posterior distribution of the Capsule parameters, described in the following section, we can use the expectations of the latent parameters to explore the original data. To detect events, we consider the proportion of the document about event j, and take a weighted average of these proportions:

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

where $\varepsilon_{d,t} = f(i_d, t)\mathbb{E}[\epsilon_{d,t}]$. This measure of "eventness" provides an estimate of the proportion of words that are related to a real-world event in that interval. Figure 1 shows events detected with this metric.

Given an identified event, we can characterize it in terms of its top terms under γ, but we can also use weighted event relevancy parameters $\varepsilon_{d,t}$ to sort documents; Section 3 explores relevant documents for events found in the National Archive diplomatic cables data. In addition to detecting and characterizing events, Capsule can be used to explore entity concerns and the general themes in a given collection.

To make Capsule more accessible, we developed an open source tool for visualizing its results.[2] Our tool creates a navigator of the documents and latent parameters, allowing users to explore events, entities, topics, and the original documents. Figure 5 shows several screenshots of this browsing interface.

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 goal is to compute the posterior values of the hidden parameters—general topics β, entity topics η, event topics γ, entity concerns ϕ (for general topics) and ξ (for their own topic), overall event strengths ψ, and document-specific strengths for general topics θ, entity topics ζ, and event topics 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 (Jordan et

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

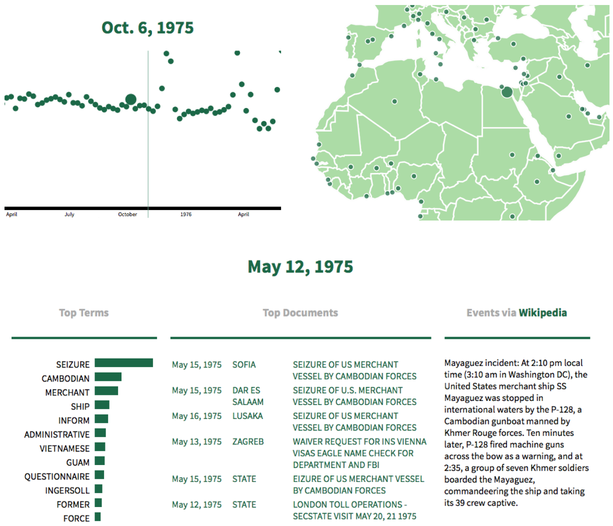

Figure 5: Screenshots of Capsule visualization of US State Department cables. Top-left: events over time (similar to Figure 1). Right-top: entities shown on a map. Bottom: time interval summary, including top terms, relevant documents, and text scraped from Wikipedia.

al., 1999),[3] which is detailed in Appendix A.[4] 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 simulated data and a collection of over 2 million U.S. State Department diplomatic cables from the 1970s.

3.1 Results on Simulated Data

Prior to exploring Capsule results on data of historical interest, we provide a quantitative assessment of the model on simulated data.

We generated ten data sets, each with 100 time steps, 10 general topics, and 100 entities. Each simulation contained about 20,000 documents and followed the generative process assumed by Capsule, as shown in Figure 4.

To evaluate event detection, we created a ranked list of all time intervals and computed the overlap between a method and the simulated ground at every threshold; this generates an curve under which we can compute the area and normalized based on ideal performance—we refer to this metric as event detection AUC:

$$\frac{\sum_{i=1}^{T} |\text{Truth}_i \cap \text{Model}_i|}{\sum_{i=1}^{T} i}, \tag{2}$$

where Model_i is a set of the top i most eventful intervals according to the model, and Truth_i is the known set of the top i most eventful intervals. As the data is simulated, we can order all intervals by their known "eventness"—this metric captures how well the model recovers the true ordering.

The most successful of the baseline methods for event detection were based on absolute error in word count relative to the mean. This can be computed for all words:

$$\sum_{v=1}^{V} \left[\sum_{d=1}^{D} \text{abs} \left(w_{d,v} - \frac{1}{|D|} \sum_{d=1}^{D} w_{d,v} \right) \right], \tag{3}$$

and can also be weighted by tf-idf,

$$\sum_{v=1}^{V} \text{tf-idf}(v) \left[\sum_{d=1}^{D} \text{abs} \left(w_{d,v} - \frac{1}{|D|} \sum_{d=1}^{D} w_{d,v} \right) \right]. \tag{4}$$

We also considered metrics that computed deviations on the entity and document level, but the simplest overall metrics performed best.

Figure 6 shows that Capsule[5] outperforms these approaches for event detection. We also consider an "event only" model—this is a model that only uses the interval-related subset of Capsule's parameters; comparing to this shows that is it important to model "business as usual" for improved event detection. LDA based approaches like average deviation from mean in topic space ((Dou et al., 2012)) do not perform well for event detection as deviations in topic space are too coarse to provide a meaningful signal.

Once events have been identified, our next task is to identify relevant documents; to evaluate this,

[3] Source code is available at `https://github.com/ajbc/capsule`.

[4] Appendices are located in the supplemental materials document.

[5] The model was set with the same number of topics $K = 10$ and exponential decay f used to simulate the data. More details on the decay function surround its formal definition in Equation (5).

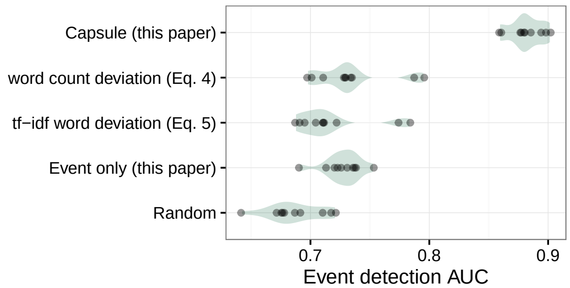

Figure 6: Event detection performance on ten simulated datasets; each dot is the performance on a single dataset, and the shaded green describes the distribution of the performances. Capsule detects events better than comparison methods.

we calculate precision of recovering the top N documents. Both Capsule and its event-only partial model outperform all comparison methods in terms of document recovery. For Capsule, average precision at 10 documents was 0.44; the event-only model had average precision of 0.09. LDA performed slightly worse than the event-only model, and the other comparison methods (similar to Equations 3 and 4) recovered zero relevant documents–equivalent to random.

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

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

provided the best performance and the most interpretable results.

3.2 Results on U.S. State Department Diplomatic Cables

As Capsule is intended to be used to explore large collections of documents, we must demonstrate its use in that context. This sections describes and explores the application of Capsule to a real-world collection of diplomatic messages.

Data. The National Archive collects communications between the U.S. Sate Department and its embassies. We obtained a collection of these diplomatic messages from the History Lab at Columbia,[6] which received them from the Central Foreign Policy Files at the National Archives. The communications in this data set were sent between 1973 and 1978.

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. We used a vocabulary of size 6,293 and omitted cables with fewer than three terms, resulting in a collection of 2,021,852 messages sent between 22,961 entities. We selected a weekly duration for the time intervals, as few cables were sent on the weekends.

Model Settings. We fit Capsule with $K = 100$ general topics and using the exponential decay f, shown in Equation (5), with event duration $\tau = 4$. With these settings on the cables data, fitting the model takes about one hour per iteration.[7]

Quantitative Results. The History Lab at Columbia provided a list of 39 real-world events in the time period covered by the cables data; they validated that these events were present in at least one of six reputable collections of events, such as the Office of the Historian list of milestones.[8]

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,

$$\text{DCG} = \sum_{j=1}^{T} \frac{\mathbf{1}[\text{interval at rank } j \text{ in known events}]}{\log j}, \tag{6}$$

divided by the ideal DCG value, or

$$\text{nDCG} = \frac{\text{DCG}}{\text{ideal DCG}}. \tag{7}$$

As shown in Table 2, Capsule outperforms the baselines.

Additionally, we can compute held-out validation data likelihood on the model and each of its component parts; Table 3 shows that the full Capsule model

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

[7] Our algorithm is batch–we consider each data point for every iteration. Modifying the algorithm to stochastically sample the data would reduce the time required to achieve an equivalent model fit.

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

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.

captures the data better than any of its component parts individually.

Model Exploration. The evaluations to this point are useful in validating that Capsule captures its intended constructs, but the objective of the model is not prediction; rather, it is to be used as a scaffold to explore large collections of documents. We now turn to exploring the cables data using Capsule.

We begin our exploration by detecting events using Capsule. With Section 2 as our metric of "eventness," we consider this metric over time, which is shown in Figure 1. Here, high values—often peaks—correspond to real-worlds events, several of which are labeled.

One of the tallest peak occurs the week of December 1, 1975, during which the United Nations General Assembly (UNGA) discussed omnibus decolonization. As discussed in Section 2, we sort documents by their weighted event relevancy parameters $f(i_d, t)\epsilon_{d,t}$ to find cables that reflect an event. Table 4 shows the top cables for this discussion. Capsule accurately identifies this real-world event and recovers relevant cables.

Another notable event was the seizure of the S.S. Mayaguez, an American merchant vessel, in May of 1975—at the end of the Vietnam War. The top documents for this week are shown in Table 5. We can inspect individual documents to confirm their relevancy and learn more about the events. For instance, the content of the most relevant document, according to Capsule, is as follows.

> 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

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

Model	LL at 10 iterations	LL at convergence
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.

$f * \epsilon$	Date	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 documents for the time interval of week December 1, 1975, when the UN discussed decolonization resolutions; Capsule recovers relevant documents related to this real-world event. Typos intentionally copied from original data.

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.

4 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*.

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*

$f * \epsilon$	**Date**	**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 documents for the week during which the S.S. Mayaguez was captured. Capsule identifies documents relevant to the real-world event. Typos intentionally copied from original data.

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.