Who, What, When, Where, and Why? A Computational Approach to Understanding Historical Events Using State Department Cables

Allison J.B. Chaney, Hanna Wallach, David M. Blei

October 6, 2015

> *We can do nothing but scrutinize historical events themselves if we want to discover what they are.*
>
> – Dean W.R. Matthews, *What is an Historical Event?*

Abstract

We develop computational methods for analyzing historical documents to identify events of potential historical significance. Significant events are characterized by interactions between entities (e.g., countries, organizations, individuals) that deviate from typical interaction patterns. When studying historical events, historians and political scientists commonly read large quantities of text to construct an accurate picture of who, what, when, and where—a necessary precursor to answering the more nuanced question, "Why?" Our methods help historians identify possible events from the texts of historical communication. Specifically, we build on topic modeling to distinguish between topics that describe "business-as-usual" and topics that deviate from these patterns, where deviations are also indicated by particular entities interacting during particular periods of time. To demonstrate our methods, we analyze a corpus of 2 million State Department cables from 1973 to 1977.

1 Introduction

For this work, we characterize an event in two ways: *when* it occurs and *what* occurs. We seek to learn these representations from observed data, specifically, a corpus of 2 million State Department cables from 1973 to 1977. Each cable is sent by an entity, such as a person or department, on a specific day, and containing a message in text format.

To learn *when* events occur, we can consider the timestamps on a collection of cables. To simplify our model, we can assume that we know roughly how long events last, and only need to discover the starting point for each event. To understand *what* occurs, we can summarize the cable message content with a topic model such as LDA (Blei et al., 2003) and model event content in that same space.

2 A Generative Model of Events

Our model is a generative process—we make assumptions about how the data came to be and describe these assumptions in terms of probability distributions. Given our model and observed data, the task is then to reverse the generative process to find the hidden quantities that (retrospectively) generated the data.

Consider an entity like the Bangkok American embassy. We can imagine that there is a stream of cables being sent by this embassy—some might be sent to the US State Department, others to another American embassy like Hong Kong, and perhaps a few are sent to individuals by name. Each of these cables has an associated send date, and we can represent the content of the cable with a topic model; we call these cable descriptions θ, which is a matrix of D cables (or documents) by K topics.[1]

An entity will usually talk about certain topics and with certain frequency. The Bangkok embassy, for instance, sent and average of 22 cables per day in the 1970s, and was concerned with topics regarding southeast Asia more generally. We can describe the general interests of entities in the same topic space we use to describe individual cables and we will call these per-entity interests ϕ.

Now imagine that at a particular time, an event occurs, such as the capture of Saigon during the Vietnam war. We do not directly observe that events occurs, but each event can again be described in the same topic space used to describe individual cables. Whether or not an event occurs at a particular time step is represented by ϵ_t and the content of the event (or its topical representation) is called π_t.

When an event occurs, both the frequency of cables being sent and the cable content changes. The Bangkok embassy sent 31 cables the day following the capture of Saigon (a 36% increase over the average), and the majority of these cables are about Vietnam war refugees. Thus we imagine that an entity's stream of cables is controlled by what it usually talks about (and how often) as well as the higher level stream of unobserved events. The influence of an event does not last indefinitely, however, so we model the decay of its magnitude with some function f.

When we analyze the cables with this model setup, we disentangle cables that represent "business as usual" from those that reflect the higher-order event stream. Consequently, we infer what that stream is, i.e., when something happened and what happened.

Recall that the key hidden values are event descriptions π ("what"), event occurrences ϵ ("when"), and entity interests ϕ. Since entities are tied to individuals and places, we can use them to describe "who" is involved and "where" and event occurs after fitting our model.

These hidden parameters interact with each other in the following formal generative process.

- For each day i with date a_i:
  - Generate event occurrence/strength $\epsilon \sim \text{Poisson}(\eta_\epsilon)$, where η_ϵ is a fixed, non-negative hyperparameter for the mean event strength.
  - Generate the day/event's description in terms of each topic k: $\pi_{ik} \sim \text{Gamma}(\alpha_0, \beta_0)$, where α_0 and β_0 are fixed hyperparameters.
- Draw the entity's base topics: $\phi_{0k} \sim \text{Gamma}(\alpha, \beta)$ (eventually for each entity, but for now, just limit data to only one entity)

[1] This allows us to represent the cable in terms of about 100 topics rather than in terms of hundreds of thousands of vocabulary words. We can discover these topics with LDA and treat them as fixed observations going forward.

- For each cable j on date c_j:
  - Set cable topic parameter: $\phi_{jk} = \phi_{0k} + \sum_i f(a_i, c_j)\pi_{ik}\epsilon_i$, where f is defined below.
  - Draw cable topic: $\theta_{jk} \sim \text{Gamma}(\beta_c \phi_{jk}, \beta_c)$.

We define the event decay function to be

$$f(a,c) = \begin{cases} 1 - \frac{c-a}{d}, & \text{if } a \le c < a + d \\ 0, & \text{otherwise,} \end{cases}$$

where d is the time distance (in days) after event a at which point the event is no longer relevant.

2.1 Inference

Posterior inference is the central computational problem. We want to learn the hidden values describe above (event descriptions π, event occurrences ϵ, and entity descriptions ϕ) from our observed data. We construct a black box variational inference algorithm following Ranganath et al. (2014) to determine the values of these latent parameters.

3 Discussion

Once we have determined the values of the hidden parameters in our model, we will have a discrete list of events in terms of when they occur and their topical content. We plan to compare these discovered events to a list of predetermined historical events to evaluate the effectiveness of our model.

Traditional topic models can describe documents, but they cannot identify when events occur—only a model like ours that explicitly models event occurrences and event content can attribute document observations to discrete events. Further, by modeling entities, we can distinguish between "business-as-usual" document content and content that is attached to particular events—we are also unable to capture this phenomenon with traditional topic models.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *JMLR*, 3:993–1022.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *AISTATS*.