Detecting and Characterizing Events: Appendices

Anonymous EMNLP submission

A Inference

In this appendix, we describe the details of the inference algorithm for Capsule. Source code for this algorithm is available at is available at https://github.com/?????/capsule.

Conditional on a collection of observed documents, our goal is to estimate the posterior values of the hidden parameters, according to the Capsule model. Recall that our data is observed as word counts w_d for document d, with corresponding author and time interval information—a_d and i_d, respectively. The latent parameters of the model include global interval strengths ψ, interval descriptions π, entity concerns ϕ, and topics β,[1] as well as document-specific entity concerns θ and interval relevancy parameters ϵ.

As for many Bayesian models, the exact posterior for Capsule is not tractable to compute; we must instead approximate it. Thus, we develop an approximate inference algorithm for Capsule based on variational methods (Wainwright and Jordan, 2008).

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

$$\mathcal{L}(q) = \mathbb{E}_{q(\psi,\pi,\phi,\beta,\theta,\epsilon)}\big[\log p(w,\psi,\pi,\phi,\beta,\theta,\epsilon) - \log q(\psi,\pi,\phi,\beta,\theta,\epsilon)\big]. \quad (4)$$

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

$$q(\psi,\pi,\phi,\beta,\theta,\epsilon) = \prod_{t=1}^{T}\Big[q(\pi_t \mid \lambda_t^{\pi})q(\psi_t \mid \lambda_t^{\psi})\Big] \prod_{n=1}^{N}\Big[q(\phi_{n,0} \mid \lambda_{n,0}^{\phi})\, q(\beta_0^{(n)} \mid \lambda_{n,0}^{\beta})\Big] \prod_{k=1}^{K}\left[q(\beta_k \mid \lambda_k^{\beta}) \prod_{n=1}^{N} q(\phi_{n,k} \mid \lambda_{n,k}^{\phi})\right] \prod_{d=1}^{D}\left[\prod_{k=1}^{K} q(\theta_{d,k} \mid \lambda_{d,k}^{\theta}) \prod_{t=1}^{T} q(\epsilon_{d,t} \mid \lambda_{d,t}^{\epsilon})\right] \quad (5)$$

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

The expectations under q, which are needed to maximize the ELBO, have closed form analytic updates–we update each parameter in turn, following standard coordinate ascent variational inference techniques, as the Capsule model is specified with the required conjugate relationships that make this approach possible (Ghahramani and Beal, 2001).

To obtain simple updates, we first rely on auxiliary latent variables z. These variables, when marginalized out, leave the original model intact. The Poisson distribution has an additive property; specifically if $w \sim \text{Poisson}(a+b)$, then $w = z_1 + z_2$, where $z_1 \sim \text{Poisson}(z_1)$ and $z_2 \sim \text{Poisson}(z_2)$. We apply

[1] Note that for brevity we include entity-specific topics β_0 within β and their corresponding entity strength parameters ϕ_0 within ϕ and per-document entity relevancy parameters θ_0 within θ.

this decomposition to the word count distribution in Eq. 1 and define Poisson variables for each component of the word count:

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

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

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

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

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

$$\pi_t \mid \mathbf{W}, \psi, \phi, \beta, \theta, \epsilon, z \sim \text{Dirichlet}_V\left(\alpha_\pi + \sum_{d=1}^{D} \langle z^{\mathcal{T}}_{d,1,t}, \cdots, z^{\mathcal{T}}_{d,V,t}\rangle\right) \quad (6)$$

$$\beta_k \mid \mathbf{W}, \psi, \pi, \phi, \theta, \epsilon, z \sim \text{Dirichlet}_V\left(\alpha_\beta + \sum_{d=1}^{D} \langle z^{\mathcal{K}}_{d,1,k}, \cdots, z^{\mathcal{K}}_{d,V,k}\rangle\right) \quad (7)$$

$$\psi_t \mid \mathbf{W}, \pi, \phi, \beta, \theta, \epsilon, z \sim \text{Gamma}\left(s_\psi + |D(t)|s_\epsilon, r_\psi + \sum_{d\in D(t)} \epsilon_{d,t}\right) \quad (8)$$

$$\phi_{i,k} \mid \mathbf{W}, \psi, \pi, \beta, \theta, \epsilon, z \sim \text{Gamma}\left(s_\phi + |D(i)|s_\theta, r_\phi + \sum_{d\in D(i)} \theta_{d,k}\right) \quad (9)$$

$$\theta_{d,k} \mid \mathbf{W}, \psi, \pi, \phi\beta, \epsilon, z \sim \text{Gamma}\left(s_\theta + \sum_{v=1}^{V} z^{\mathcal{K}}_{d,v,k}, \phi_{a_d,k} + \sum_{v=1}^{V} \beta_{k,v}\right) \quad (10)$$

$$\epsilon_{d,t} \mid \mathbf{W}, \psi, \pi, \phi, \beta, \theta, z \sim \text{Gamma}\left(s_\epsilon + \sum_{v=1}^{V} z^{\mathcal{T}}_{d,v,t}, \psi_t + f(i_d, t)\sum_{v=1}^{V} \pi_{t,v}\right) \quad (11)$$

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

$$\omega_{d,v} \propto \langle \theta_{d,1}\beta_{1,v}, \cdots, \theta_{d,K}\beta_{K,v}, f(i_d, 1)\epsilon_{d,1}\pi_{1,v}, \cdots, f(i_d, T)\epsilon_{d,T}\pi_{T,v}\rangle. \quad (12)$$

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

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

References

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

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

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

Algorithm 1: Variational Inference for Capsule

Input: word counts w
Output: approximate posterior of latent parameters $(\psi, \pi, \phi, \beta, \theta, \epsilon)$ in terms of variational parameters $\lambda = \{\lambda^{\psi}, \lambda^{\pi}, \lambda^{\phi}, \lambda^{\beta}, \lambda^{\theta}, \lambda^{\epsilon}\}$
Initialize $\mathbb{E}[\beta]$ to slightly random around uniform
Initialize $\mathbb{E}[\psi]$, $\mathbb{E}[\pi]$, $\mathbb{E}[\psi]$, $\mathbb{E}[\theta]$, $\mathbb{E}[\epsilon]$ to uniform
for *iteration* $m = 1 : M$ **do**
  set λ^{ψ}, λ^{π}, λ^{ϕ}, λ^{β}, λ^{θ}, λ^{ϵ} to respective priors, excluding $\lambda^{\theta,rate}$ and $\lambda^{\epsilon,rate}$, which are set to 0
  update $\lambda^{\theta,rate} \mathrel{+}= \sum_V \mathbb{E}[\beta_v]$
  for *each document* $d = 1 : D$ **do**
    for *each term* $v \in V(d)^2$ **do**
      set $(K+T)$-vector $\omega_{d,v}$ using $\mathbb{E}[\pi]$, $\mathbb{E}[\theta]$, and $\mathbb{E}[\epsilon]$, as shown in Eq. 12
      set $(K+T)$-vector $\mathbb{E}[z_{d,v}] = w_{d,v} * \omega_{d,v}$
      update $\lambda_d^{\theta,shape} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 10)
      update $\lambda_d^{\epsilon,shape} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 11)
      update $\lambda_v^{\beta} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ (Eq. 7)
      update $\lambda_v^{\pi} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{T}}]$ (Eq. 6)
    end
    update $\lambda_d^{\theta,rate} \mathrel{+}= \mathbb{E}[\phi_{a_d}]$ (Eq. 10)
    update $\lambda_d^{\epsilon,rate} \mathrel{+}= \mathbb{E}[\psi]$ (Eq. 11)
    set $\mathbb{E}[\theta_d] = \lambda_d^{\theta,shape} / \lambda_d^{\theta,rate}$
    set $\mathbb{E}[\epsilon_d] = \lambda_d^{\epsilon,shape} / \lambda_d^{\epsilon,rate}$
    update $\lambda_{a_d}^{\phi,shape} \mathrel{+}= s_{\theta}$ (Eq. 9)
    update $\lambda_t^{\psi,shape} \mathrel{+}= s_{\epsilon} \forall t : f(i_d, t) \neq 0$ (Eq. 8)
    update $\lambda_{a_d}^{\phi,rate} \mathrel{+}= \theta_d$ (Eq. 9)
    update $\lambda^{\psi,rate} \mathrel{+}= \epsilon_d$ (Eq. 8)
  end
  set $\mathbb{E}[\phi] = \lambda^{\phi,shape} / \lambda^{\phi,rate}$
  set $\mathbb{E}[\beta_k] = \lambda^{\beta_{k,v}} / \sum_v \lambda^{\beta_k} \forall k$
  set $\mathbb{E}[\psi] = \lambda^{\psi,shape} / \lambda^{\psi,rate}$
  set $\mathbb{E}[\pi_t] = \lambda^{\pi_{t,v}} / \sum_v \lambda^{\pi_t} \forall t$
end
return λ