Grounding Topic Models with Knowledge Bases: Supplementary Material

A Inference via Collapsed Gibbs Sampling

Here we describe the inference algorithm for LGSA based on collapsed Gibbs Sampling.

Given a document corpus $\mathcal{D}$, the informative priors over entities $\{\boldsymbol{p}^\eta, \boldsymbol{p}^\zeta, \lambda^\eta, \lambda^\zeta\}$, and the hyperparameters $\{\alpha, \beta\}$, LGSA specifies the following full posterior distribution:

$$\begin{aligned} &p(\boldsymbol{\Lambda}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{r}, \boldsymbol{e}, \boldsymbol{y} | \mathcal{D}, \boldsymbol{p}^\eta, \boldsymbol{p}^\zeta, \lambda^\eta, \lambda^\zeta, \alpha, \beta) \propto \\ &\Big(p(\boldsymbol{\Lambda}|\beta) p(\boldsymbol{\eta}|\boldsymbol{p}^\eta, \lambda^\eta) p(\boldsymbol{\zeta}|\boldsymbol{p}^\zeta, \lambda^\zeta) p(\boldsymbol{\theta}|\alpha) p(\boldsymbol{z}|\boldsymbol{\theta}) \\ &p(\boldsymbol{r}, \boldsymbol{e}|\boldsymbol{z}, \boldsymbol{\Lambda}) p(\boldsymbol{m}_\mathcal{D}|\boldsymbol{e}, \boldsymbol{\eta}) p(\boldsymbol{y}|\boldsymbol{M}_D) p(\boldsymbol{w}_D|\boldsymbol{y}, \boldsymbol{e}, \boldsymbol{\zeta}) \Big). \end{aligned} \tag{A.1}$$

where $\boldsymbol{m}_\mathcal{D}$ and $\boldsymbol{w}_\mathcal{D}$ are the mentions and words in the document collections, respectively; $\boldsymbol{M}_\mathcal{D} = \{M_1, \ldots, M_D\}$ is the counts of mentions of the documents. The constant of proportionality is the marginal likelihood of the observed data.

The task of posterior inference for LGSA is to determine the probability distribution of the hidden variables given the observed mentions and words. However, exact inference is intractable due to the difficulty of calculating the normalizing constant in the above posterior distribution.

We use collapsed Gibbs Sampling, a well-established Markov chain Monte Carlo (MCMC) technique for approximate inference. In collapsed Gibbs Sampling, the distributions $\boldsymbol{\Phi} = \{\boldsymbol{\Lambda}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\theta}\}$ are first marginalized (collapsed), a Markov chain over the latent indicators $\{\boldsymbol{z}, \boldsymbol{r}, \boldsymbol{e}, \boldsymbol{y}\}$ is then constructed, whose stationary distribution is the posterior. We obtain samples of latent variables from the Markov chain. Point estimates for the collapsed distributions $\boldsymbol{\Phi}$ can then be computed given the samples, and predictive distributions are computed by averaging over multiple samples.

Sampling Procedure

Gibbs Sampler repeatedly samples each latent variable conditioned on the current states of other hidden variables and observations; a configuration of latent states of the system is then obtained. Next we provide the derivation of the sampling formulas (i.e. Eqs.(1-4) in the paper).

By marginalizing out $\boldsymbol{\Phi}$ in Eq.(A.1), we obtain:

$$\begin{aligned} &p(\boldsymbol{z}, \boldsymbol{r}, \boldsymbol{e}, \boldsymbol{y}|.) \\ &\propto p(\boldsymbol{z}|\alpha) P(\boldsymbol{r}, \boldsymbol{e}|\boldsymbol{z}, \beta) p(\boldsymbol{m}_\mathcal{D}|\boldsymbol{e}, \boldsymbol{p}^\eta, \lambda^\eta) p(\boldsymbol{y}|\boldsymbol{M}_\mathcal{D}) \\ &\cdot p(\boldsymbol{w}_\mathcal{D}|\boldsymbol{y}, \boldsymbol{e}, \boldsymbol{p}^\zeta, \lambda^\zeta) \\ &= \int p(\boldsymbol{\theta}|\alpha) P(\boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{\theta} \int p(\boldsymbol{\Lambda}|\beta) p(\boldsymbol{r}, \boldsymbol{e}|\boldsymbol{z}, \boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \\ &\cdot \int p(\boldsymbol{\eta}|\boldsymbol{p}^\eta, \lambda^\eta) p(\boldsymbol{m}_\mathcal{D}|\boldsymbol{e}, \boldsymbol{\eta}) d\boldsymbol{\eta} \cdot p(\boldsymbol{y}|\boldsymbol{M}_\mathcal{D}) \\ &\cdot \int p(\boldsymbol{\zeta}|\boldsymbol{p}^\zeta, \lambda^\zeta) p(\boldsymbol{w}_\mathcal{D}|\boldsymbol{y}, \boldsymbol{e}, \boldsymbol{\zeta}) d\boldsymbol{\zeta}. \end{aligned} \tag{A.2}$$

The conditional of z_{dj} can be computed as:

$$\begin{aligned} &p(z_{dj} = z | e_{dj} = e, \boldsymbol{r}_{-dj}, \boldsymbol{z}_{-dj}, .) \propto \\ &p(z_{dj} = z | \boldsymbol{z}_{-dj}, \alpha) p(e_{dj} = e | z_{dj} = z, \boldsymbol{r}_{-dj}, .) \end{aligned} \tag{A.3}$$

The first term of Eq.(A.3) is:

$$p(z_{dj} = z | \boldsymbol{z}_{-dj}, \alpha) = \frac{p(z_{dj} = z, \boldsymbol{z}_{-dj} | \alpha)}{p(\boldsymbol{z}_{-dj} | \alpha)}. \tag{A.4}$$

As we assume each z is generated from a multinomial distribution $\boldsymbol{\theta}$, and the hyperparameter for conjugate Dirichlet prior is α, we have:

$$\begin{aligned} p(\boldsymbol{z}|\alpha) &= \int P(\boldsymbol{\theta}|\alpha) P(\boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \prod_d \frac{\Gamma(K\alpha)}{\prod_z \Gamma(\alpha)} \prod_z \theta_{dz}^{\alpha - 1} \cdot \prod_d \prod_z \theta_{dz}^{n_d^{(z)}} d\boldsymbol{\theta} \\ &= \prod_d \frac{\Gamma(K\alpha)}{\prod_z \Gamma(\alpha)} \cdot \frac{\prod_z \Gamma(n_d^{(z)} + \alpha)}{\Gamma(n_d^{(\cdot)} + K\alpha)}, \end{aligned}$$

where $n_d^{(z)}$ is the number of times that topic z has been associated with a mention of document d. Marginal counts are represented with dots (i.e. $n_d^{(\cdot)}$ is obtained by marginalizing $n_d^{(z)}$ over z). Combining the above equation with Eq.(A.4) leads to:

$$\begin{aligned} \frac{p(z_{dj} = z, \boldsymbol{z}_{-dj}|\alpha)}{p(\boldsymbol{z}_{-dj}|\alpha)} &= \frac{\Gamma(n_d^{(z)} + \alpha) \Gamma(n_{d,-dj}^{(\cdot)} + K\alpha)}{\Gamma(n_{d,-dj}^{(z)} + \alpha) \Gamma(n_d^{(\cdot)} + K\alpha)} \\ &= \frac{n_{d,-dj}^{(z)} + \alpha}{n_{d,-dj}^{(\cdot)} + K\alpha}, \end{aligned} \tag{A.5}$$

where the count with subscript $-ij$ denotes a quantity with the current instance (i.e. mention m_{dj}) excluded. Here we use the identity $\Gamma(x+1) = x\Gamma(x)$.

The second term of Eq.(A.3) is the probability of generating entity e conditioned on topic z, which requires summing over the probabilities of all paths in z that could have generated u:

$$p(e_{dj} = e|z_{dj} = z, \boldsymbol{r}_{-dj}, .) = \sum_{\boldsymbol{r}(e \in \boldsymbol{r})} p(\boldsymbol{r}|\boldsymbol{r}_{-dj}, z_{dj} = z, .). \quad \text{(A.6)}$$

The probability of a path $\boldsymbol{r}$ is the product of the topic-specific transition probabilities along the path from root c_0 to leaf $c_{|\boldsymbol{r}|-1}$ (i.e. entity e):

$$p(\boldsymbol{r}|\boldsymbol{r}_{-dj}, z_{dj} = z, .) = \prod_{h=0}^{|\boldsymbol{r}|-2} p(c_{h+1}|c_h, z_{dj} = z, \boldsymbol{r}_{-dj}.).$$

Here $p(c_{h+1}|c_h, z_{dj} = z, \boldsymbol{r}_{-dj}.)$ can be derived analogously to Eqs.(A.4-A.5), where the Dirichlet-Multinomial conjugates ensure the tractability of the integrals. We then obtain:

$$p(c_{h+1}|c_h, z_{dj} = z, \boldsymbol{r}_{-dj}.) = \frac{n^{(z),-dj}_{c_h,c_{h+1}} + \beta}{n^{(z),-dj}_{c_h,\cdot} + |C(c_h)|\beta}, \quad \text{(A.7)}$$

where $n^{(z),-dj}_{c_h,c_{h+1}}$ is the number of paths in topic z that go from c_h to c_{h+1}, with the path of mention m_{dj} excluded.

Finally, by combining Eqs.(A.3-A.7) we obtain the sampling formula for z_{dj} as Eq.(1-2) in the paper. Note that we omit the subscripts/superscripts $-ij$ in Eq.(1-2) to avoid cluttering of notation. Eq.(3) and Eq.(4) are derived in a similar manner.

Distribution Estimation

After a sufficient number of sampling iterations as described above, we obtain a set of samples. The unknown distributions can then be computed by integrating across the samples. Specifically, for any single sample we can estimate $\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\Lambda}, \boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ as:

$$\hat{\theta}_{dz} = \frac{n^{(z)}_d + \alpha}{n^{(\cdot)}_d + K\alpha},$$

$$\hat{\theta'}_{de} = \frac{n^{(e)}_d}{n^{(\cdot)}_d},$$

$$\hat{\Lambda}_{zcc'} = \frac{n^{(z)}_{c,c'} + \beta}{n^{(z)}_{c,\cdot} + |C(c)|\beta},$$

$$\hat{\eta}_{em} = \frac{n^{(m)}_e + \lambda^\eta p^\eta_{em}}{n^{(\cdot)}_e + \lambda^\eta},$$

$$\hat{\zeta}_{ew} = \frac{n^{(w)}_e + \lambda^\zeta p^\zeta_{ew}}{n^{(\cdot)}_e + \lambda^\zeta}.$$

Finally, based on the estimated $\hat{\boldsymbol{\Lambda}}$, we can compute the topic representations $(\boldsymbol{\phi}, \boldsymbol{\tau})$ by summing over all possible paths:

$$\hat{\phi}_{ze} = \sum_{\boldsymbol{r}(e \in \boldsymbol{r})} \prod_{h=0}^{|\boldsymbol{r}|-2} \hat{\Lambda}_{z c_h c_{h+1}},$$

$$\hat{\tau}_{zc} = \sum_{\boldsymbol{r}(e \in \boldsymbol{r})} \prod_{h=0}^{c_{h+1}=c} \hat{\Lambda}_{z c_h c_{h+1}}.$$

It is straightforward to see that $\sum_e \hat{\phi}_{ze} = 1$. That is, $\hat{\boldsymbol{\phi}}_z$ is a distribution over entities.

Inference on New Documents

Given a new-arriving document d, we can infer its topic distribution $\boldsymbol{\theta}_d$ and entity distribution $\boldsymbol{\theta}'_d$ to reveal its major themes and entities. The inference can be carried out using the Gibbs Sampling described above, but this time with the topic and entity statistics (i.e. $\boldsymbol{\Lambda}$, $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$) fixed.