Supplementary Material

to

Relating Romanized Comments to News Articles by Inferring Multi-*glyphic* Topical Correspondence

Abstract

This is supplementary material to our AAAI 2015 submission contains the following:

1 Complete Generative Model

The complete plate diagram for MCTM is shown in Figure 1, and the full generative story is described in Figure 2.

2 Collapsed-Blocked Gibbs Sampling

Notation. We use the following notation in the paper. $Dir()$, $Categ()$, $Beta()$, and $Bern()$ represent the Dirichlet, Categorical, Beta, Bernoulli distributions, while $SBP()$ represents a stick-breaking process. $Unif(z)$ is a distribution assigning uniform probability to each component of z. We use $v \sim P$ to say that we sample a value for variable v from distribution P. $[N]$ is the set $\{1, \ldots, N\}$. V_l is the vocabulary size of language l. K and J are the number of article and comment topics, respectively. z and y denote the topic assignments for article and comment word, respectively. θ is the set of topic vectors for an article. ϕ denotes the word distributions for article topics.

We want to **collapse** all real-valued variables and preserve only categorical variables so that we can converge faster, and also to avoid round-off errors. The main challenges that make the inference non-standard are:

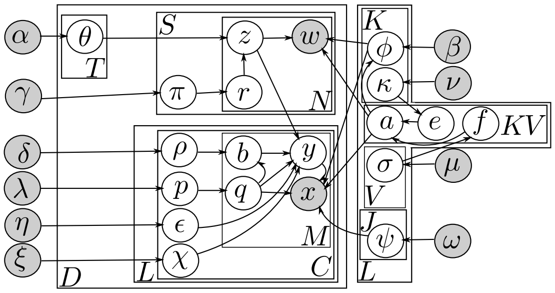

Figure 1: MCTM.

- (Topic sparsity parameter) $\kappa_{lk} \sim Beta(\nu)\ \forall l, k$.
- (Word sparsity parameter) $\sigma_{lv} \sim Beta(\mu)\ \forall l, v$.
- (Sparsity variables) $e_{lkv} \sim Bern(\kappa_{lk})$, $f_{lkv} \sim Bern(\sigma_{lv})$, and $a_{lkv} = e_{lkv} f_{lkv}, \forall l, k, v$.
- (Article topics) $\phi_{lk} \sim Dir(\beta a_{lk})\ \forall l, k$.
- (Comment topics) $\psi_{lj} \sim Dir(\omega)\ \forall l, j$.
- For each article w_d:
  - (Topic vectors) $\theta_{dt} \sim Dir(\alpha)\ \forall t$.
  - For each sentence w_{ds}:
    - $*$ (Topic vector distribution) $\pi_{ds} \sim SBP(\gamma)$.
    - $*$ For each word w_{dsn}:
      - · (Topic vector) $r_{dsn} \sim Categ(\pi_{ds})$.
      - · (Topic) $z_{dsn} \sim Categ(\theta_{dr_{dsn}})$.
      - · (Word) $w_{dsn} \sim Categ(\phi_{\ell z_{dsn}})$, where ℓ is the language of the article.
- For each comment x_{dlc} ($l = 1 \ldots L$):
  - (Topic source distribution) $p_{dlc} \sim Dir(\lambda)$.
  - (Segment distribution) $\rho_{dlc} \sim Dir(\delta)$.
  - (Corpus topic vector) $\epsilon_{dlc} \sim Dir(\eta)$.
  - (Comment topic vector) $\chi_{dlc} \sim Dir(\xi)$.
  - For each comment word x_{dlcm}:
    - $*$ (Topic source) $q_{dlcm} \sim Categ(p_{dlc})$.
    - $*$ (Article topic) If q_{dlcm}=1, $b_{dlcm} \sim Dir(\rho_{dlc})$, $y_{dlcm} \sim Unif(z_{db_{dlcm}})$, and $x_{dlcm} \sim Dir(\phi_{ly_{dlcm}} a_{ly_{dlcm}})$.
    - $*$ (Corpus topic) If q_{dlcm}=2, $y_{dlcm} \sim Dir(\epsilon_{dlc})$, and $x_{dlcm} \sim Dir(\phi_{ly_{dlcm}} a_{ly_{dlcm}})$.
    - $*$ (Comment topic) If q_{dlcm}=3, $y_{dlcm} \sim Dir(\chi_{dlc})$, and $x_{dlcm} \sim Dir(\psi_{ly_{dlcm}})$.

Figure 2: Generative story for the final model.

- Coupling between κ and σ—we solve this by introducing auxiliary variables e and f.
- Sampling r (due to the SBP prior)—we use the equivalence of the SBP to the Generalized Dirichlet (GD) distribution (Connor and Mosimann 1969) and derive the sampling update.
- Interdependence between q, b, y, and between a, e, f—we handle this issue by doing **blocked** Gibbs sampling.

The joint distribution is

$$p(\theta, \pi, r, z, w, \rho, p, \epsilon, \chi, q, b, y, x, \phi, a, e, f, \sigma, \kappa, \psi \\ |\alpha, \gamma, \delta, \eta, \xi, \lambda, \beta, \mu, \nu, \omega)$$

$$= p(\theta|\alpha)p(\pi|\gamma)p(r|\pi)p(z|r,\theta)p(w|z,\phi,a)p(p|\lambda)p(q|p)$$
$$p(\rho|\delta)p(b|q,\rho)p(\epsilon|\eta)p(\chi|\xi)p(y|q,b,z,\epsilon,\chi)p(x|q,y,\phi,a,\psi)$$
$$p(\sigma|\mu)p(\kappa|\nu)p(a|,e,f)p(e|\kappa)p(f|\sigma)p(\phi|a,\beta)p(\psi|\omega)$$

We integrate out $\theta, \pi, \rho, p, \epsilon, \chi, \phi, \sigma, \kappa, \psi$ and sample $r, z, w, q, b, y, x, a, e, f$, as discussed below. We follow the indexing notation for variables, with a small change in order to simplify the presentation: We imagine that the article is part of a tuple of L articles, so that:
$w = \{\{\{\{w_{dlsn}\}_{n=1}^{N_{dls}}\}_{s=1}^{S_{dl}}\}_{l=1}^{L}\}_{d=1}^{D}$, and
$x = \{\{\{\{\{x_{dll'cm}\}_{m=1}^{M_{dll'c}}\}_{c=1}^{C_{dll'}}\}_{l'=1}^{L}\}_{l=1}^{L}\}_{d=1}^{D}$.
Note that, if ℓ is the language of article d, $w_{dl} = \varnothing, \forall l \neq \ell$.

2.1 Sampling distributions

In the following, we give the forms of the conditional distributions used for sampling. We denote the resulting integrals as functions $F_{\text{[index]}}(\text{[variable]})$ for ease of exposition. The exact forms of these functions are given in the appendix.

Sampling r. Using the equivalence of the SBP to the GD (Ishwaran and James 2001), we can directly write the density of π as

$$p(\pi_{d\ell s}|\gamma) = \prod_{t=1}^{T_d-1} \frac{1}{B(\gamma_t)} \pi_{d\ell st}^{\gamma_{t1}-1} \left(1 - \textstyle\sum_{i=1}^{t} \pi_{d\ell si}\right)^{\gamma_{t2}-\gamma_{(t+1)1}-\gamma_{(t+1)2}}$$

where $B()$ is the Beta function and $\gamma_{T_d 1} + \gamma_{T_d 2} \triangleq 1$. Using this, we get the conditional distribution

$$p(r_{d\ell sn}|\text{others}) \propto \int_{\theta} p(\theta|\alpha)p(z|r,\theta) \int_{\pi} p(\pi|\gamma)p(r|\pi)$$
$$\propto F_{d\ell sn}(\alpha)F_{d\ell sn}(\gamma)$$

Sampling z. The conditional distribution is

$$p(z_{d\ell sn}|\text{others}) \propto F_{d\ell sn}(\alpha) \int_{\epsilon} p(\epsilon|\eta)p(y|q,b,z,\epsilon)$$
$$\int_{\phi,\psi} p(w|z,\phi,a)p(x|q,y,\phi,a,\psi)p(\phi|a,\beta)p(\psi|\omega)$$
$$= F_{d\ell sn}(\alpha)F_{d\ell sn}(\beta)F_{d\ell sn}(y).$$

Sampling q, b, y. If $b_{d\ell\ell' cm} = s$ for some $s \in [S_{d\ell}]$, then $p(q_{d\ell\ell' cm}|\text{others})$ is a delta function at $q_{d\ell\ell' cm} = 1$, so that $q_{d\ell\ell' cm}$ never changes after the initial assignment. Due to this problem caused by the causal relationship between these variables, we sample them jointly, leading to a blocked Gibbs sampling update:

$$p(q_{d\ell\ell' cm}, b_{d\ell\ell' cm}, y_{d\ell\ell' cm}|\text{others}) \propto \int_{p} p(p|\lambda)p(q|p)$$
$$\int_{\rho} p(\rho|\delta)p(b|q,\rho) \int_{\epsilon,\chi} p(\epsilon|\eta)p(\chi|\xi)p(y|q,b,z,\epsilon,\chi)$$
$$\int_{\phi,\psi} p(w|z,\phi,a)p(x|q,y,\phi,a,\psi)p(\phi|a,\beta)p(\psi|\omega)$$
$$= F_{d\ell\ell' cm}(\lambda)\left(F_{d\ell\ell' cm}(\delta)F_{d\ell\ell' cm}(z)F_{d\ell\ell' cm}(\beta)\right)^{q_{d\ell\ell' cm1}}$$
$$\left(F_{d\ell\ell' cm}(\eta)F_{d\ell\ell' cm}(\beta)\right)^{q_{d\ell\ell' cm2}}\left(F_{d\ell\ell' cm}(\xi)F_{d\ell\ell' cm}(\omega)\right)^{q_{d\ell\ell' cm3}}$$

Sampling a, e, f. Similar to the situation discussed in Section 2.1, we find causal relationships here as well, and sample them jointly, giving a blocked Gibbs sampling update:

$$p(a_{\ell kv}, e_{\ell kv}, f_{\ell kv}|\mathbb{N}_{\ell kv} = \mathbb{M}_{12\ell kv} = 0, \text{others})$$
$$\propto p(a|e,f) \int_{\phi,\psi} p(w|z,\phi,a)p(x|q,y,\phi,a,\psi)p(\phi|a,\beta)p(\psi|\omega)$$
$$\int_{\sigma} p(\sigma|\mu)p(f|\sigma) \int_{\kappa} p(\kappa|\nu)p(e|\kappa)$$
$$= F_{\ell kv}(\beta)F_{\ell kv}(\nu)F_{\ell kv}(\mu)$$

Note that if $\mathbb{N}_{\ell kv} > 0$ or $\mathbb{M}_{12\ell kv} > 0$, we are forced to set $a_{\ell kv} = 1 = e_{\ell kv} = f_{\ell kv}$.

Foreign (*Rohin*) words used in English comments
desh logo hota log yeh tha chor sirf kiya mai ...
Foreign (English) words used in *Rohin* comments
and on of shame for india no this govt we ...
Compliments (English and *Rokan*)
English article nice writing good reading write great ...
Rokan lekhana chennagi baravanige good moodi sogasagi...
[article good writing good "turning out" nicely]
Commenter names (*Rokan*)
lalitha siddabasavaiah satya anupama almane lekhana ...

Table 1: Comment topics: interesting word classes automatically discovered by our model.

3 Qualitative evaluation of topics

We eyeballed the topics to verify semantic coherence, and found it to be satisfactory. The comment topics discovered were especially interesting, since we found that they could be categorized into different classes such as "compliments", "foreign words", "commenter names", etc. (Table 1). We also automatically learned an **"expletives"** topic. This can be useful for automatically learning a list of offensive terms, for use in flagging objectionable comments.

4 Classification Algorithms

Algorithm for Comment Category Detection Algorithm for MCTM. For the MCTM-DSGN, MCTM-DSGNP and MCTM-DSGNPC models, the algorithm for comment classification is as follows:
Input: article w_{dl}, comment $x_{dll'c}$, thresholds: $\tau_{MinCosSim}$

1. Let q_{max} =dominant (most frequent) topic source in $b_{dll'c}$.
2. If $q_{max} = 1$, $x_{dll'c}$ is topical.
3. Else If $q_{max} = 2$, $x_{dll'c}$ is corpus-topical.
4. Else If $q_{max} = 3$, $x_{dll'c}$ is comment-topical or robotic.
5. If $x_{dll'c}$ is topical, and if $\{s : Sim(w_{dls}, x_{dll'c}) > \tau_{MinCosSim}\} \neq \emptyset$, then $x_{dll'c}$ is specific.

We defined $Sim(w_{dls}, x_{dll'c})$ as the cosine similarity between the topic vectors correponding to segment and the comment. The topic vector for w_{dls} is the vector $\boldsymbol{u} \in \mathbb{R}^K$ such that $\boldsymbol{u}_k$ = the number of times topic k occurs in z_{dls}. The topic vector for $x_{dll'c}$ is the vector $\boldsymbol{v} \in \mathbb{R}^K$ such that $\boldsymbol{v}_k$ = the number of times article topic[1] k occurs in $y_{dll'c}$.

[1] We ignore occurrences of comment topics.

Thus, $Sim(w_{dls}, x_{dll'c}) = \frac{\boldsymbol{u}^T \boldsymbol{v}}{||\boldsymbol{u}|| \, ||\boldsymbol{u}||}$. In our experiments, we set $\tau_{MinCosSim} = .1$.

Algorithm for MCTM-D. For the MCTM-D model, the algorithm for comment classification is as follows:

Input: article w_{dl}, comment $x_{dll'c}$, thresholds: $\tau_{MinCosSim}$, $\tau_{FracCorpCommRob}$

1. Sort the comments x_{dl} in decreasing order of $Sim(w_{dl}, x_{dll'c'})$.
2. Take the top $\tau_{FracCorpCommRob}$ fraction of the comments and form the set $\mathcal{T}$. The remaining comments for the set $\mathcal{N}$.
3. If $x_{dll'c} \in \mathcal{T}$, $x_{dll'c}$ is topical.
4. Else If $x_{dll'c} \in \mathcal{N}$, $x_{dll'c}$ is corpus-topical or comment-topical or robotic.
5. If $x_{dll'c}$ is topical, and if $\{s : Sim(w_{dls}, x_{dll'c}) > \tau_{MinCosSim}\} \neq \emptyset$, then $x_{dll'c}$ is specific.

In our experiments, we set $\tau_{MinCosSim} = .1$, $\tau_{FracCorpCommRob} = .3$.

References

Connor, R. J., and Mosimann, J. E. 1969. Concepts of independence for proportions with a generalization of the dirichlet distribution. *JASA*.

Ishwaran, H., and James, L. F. 2001. Gibbs sampling methods for stick-breaking priors. *JASA* 96(453).

Appendix

Counting notation. We use the following counting notation in the equations below. $\mathbb{N}$ is used for counts on the articles, $\mathbb{M}$ for counts on the comments, $\mathbb{V}$ for counts of words in a topic, and $\mathbb{K}$ for count of topics that contain a word. Indexes use the convention followed in the paper, e.g. d for articles, l for article language, etc. h is an index over topic sources. $\mathbb{N}_{dls>t}$ is the #times a topic was generated using a topic vector greater than t.

Sampling equations. The exact forms for the integrals given in Section 2.1 are as follows.

$$F_{d\ell sn}(\alpha) = \frac{\alpha_{z_{d\ell sn}} + \mathbb{N}^{-d\ell sn}_{dr_{d\ell sn} z_{d\ell sn}}}{\sum_k \alpha_k + \mathbb{N}^{-d\ell sn}_{dr_{d\ell sn} k}}$$

$$F_{d\ell sn}(\gamma) = \frac{\gamma_{r_{d\ell sn} 1} + \mathbb{N}^{-d\ell sn}_{d\ell s r_{d\ell sn}}}{\gamma_{r_{d\ell sn} 1} + \mathbb{N}^{-d\ell sn}_{d\ell s r_{d\ell sn}} + \gamma_{r_{d\ell sn} 2} + \mathbb{N}^{-d\ell sn}_{d\ell s > r_{d\ell sn}}}$$
$$\prod_{t=1}^{r_{d\ell sn}-1} \frac{\gamma_{t2} + \mathbb{N}^{-d\ell sn}_{d\ell s>t}}{\gamma_{t1} + \mathbb{N}^{-d\ell sn}_{d\ell st} + \gamma_{t2} + \mathbb{N}^{-d\ell sn}_{d\ell s>t}}$$

$$F_{d\ell sn}(\beta) = \frac{\beta_{w_{d\ell sn}} + \mathbb{N}^{-d\ell sn}_{\ell z_{d\ell sn} w_{d\ell sn}} + \mathbb{M}_{12\ell z_{d\ell sn} w_{d\ell sn}}}{\sum_{v \in a_{\ell z_{d\ell sn}}} \beta_v + \mathbb{N}^{-d\ell sn}_{z_{d\ell sn} v} + \mathbb{M}_{12\ell z_{d\ell sn} v}}$$

$$F_{d\ell sn}(y) = \left(\frac{\mathbb{N}^{-d\ell sn}_{d\ell s z_{d\ell sn}} + 1}{\mathbb{N}^{-d\ell sn}_{d\ell s z_{d\ell sn}}} \right)^{\mathbb{M}_{d\ell s z_{d\ell sn}}}$$

$$F_{d\ell\ell' cm}(\lambda) = \frac{\lambda_{q_{d\ell\ell' cm}} + \mathbb{M}^{-d\ell\ell' cm}_{d\ell\ell' c q_{d\ell\ell' cm}}}{\sum_h \lambda_h + \mathbb{M}^{-d\ell\ell' cm}_{d\ell\ell' ch}}$$

$$F_{d\ell\ell' cm}(\delta) = \frac{\delta_{b_{d\ell\ell' cm}} + \mathbb{M}^{-d\ell\ell' cm}_{d\ell\ell' c b_{d\ell\ell' cm}}}{\sum_s \delta_s + \mathbb{M}^{-d\ell\ell' cm}_{d\ell\ell' cs}}$$

$$F_{d\ell\ell' cm}(z) = \frac{\mathbb{N}_{d\ell b_{d\ell\ell' cm} y_{d\ell\ell' cm}}}{N_{d\ell b_{d\ell\ell' cm}}}$$

$$F_{d\ell\ell' cm}(\beta) = \frac{\beta_{x_{d\ell\ell' cm}} + \mathbb{N}_{\ell' y_{d\ell\ell' cm} x_{d\ell\ell' cm}} + \mathbb{M}^{-d\ell\ell' cm}_{12\ell' y_{d\ell\ell' cm} x_{d\ell\ell' cm}}}{\sum_{v \in a_{\ell' y_{d\ell\ell' cm}}} \beta_v + \mathbb{N}_{\ell' y_{d\ell\ell' cm} v} + \mathbb{M}^{-d\ell\ell' cm}_{12\ell' y_{d\ell\ell' cm} v}}$$

$$F_{d\ell\ell' cm}(\eta) = \frac{\eta_{y_{d\ell\ell' cm}} + \mathbb{M}^{-d\ell\ell' cm}_{d\ell\ell' c y_{d\ell\ell' cm}}}{\sum_k \eta_k + \mathbb{M}^{-d\ell\ell' cm}_{d\ell\ell' ck}}$$

$$F_{d\ell\ell' cm}(\xi) = \frac{\xi_{y_{d\ell\ell' cm}} + \mathbb{M}^{-d\ell\ell' cm}_{3d\ell\ell' c y_{d\ell\ell' cm}}}{\sum_j \xi_j + \mathbb{M}^{-d\ell\ell' cm}_{3d\ell\ell' cj}}$$

$$F_{d\ell\ell' cm}(\omega) = \frac{\omega_{x_{d\ell\ell' cm}} + \mathbb{M}^{-d\ell\ell' cm}_{3\ell' y_{d\ell\ell' cm} x_{d\ell\ell' cm}}}{\sum_v \omega_v + \mathbb{M}^{-d\ell\ell' cm}_{3\ell' y_{d\ell\ell' cm} v}}$$

$$F_{\ell kv}(\beta) = \frac{\Gamma(\sum_{v' \in a_{\ell k}, v' \neq v} \beta_{v'} + \beta_v a_{\ell kv})}{\Gamma(\sum_{v' \in a_{\ell k}, v' \neq v} \beta_{v'} + \mathbb{N}_{\ell kv'} + \mathbb{M}_{12\ell kv'} + \beta_v a_{\ell kv})}$$

$$F_{\ell kv}(\nu) = (\nu_1 + \mathbb{V}^{-\ell kv}_{\ell k})^{e_{\ell kv}} (\nu_2 + V_\ell - \mathbb{V}^{-\ell kv}_{\ell k} - 1)^{1 - e_{\ell kv}}$$

$$F_{\ell kv}(\mu) = (\mu_1 + \mathbb{K}^{-\ell kv}_{\ell v})^{f_{\ell kv}} (\mu_2 + K - \mathbb{K}^{-\ell kv}_{\ell v} - 1)^{1 - f_{\ell kv}}$$