A. Inference for CPGBN

Here we describe the derivation in detail for convolutional Poisson gamma belief network (CPGBN) with T hidden layers, expressed as

$$\begin{aligned}
&\boldsymbol{\theta}_j^{(T)} \sim \text{Gam}(\boldsymbol{r}, 1/c_j^{(T+1)}),\\
&...,\\
&\boldsymbol{\theta}_j^{(t)} \sim \text{Gam}(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}, 1/c_j^{(t+1)}),\\
&...,\\
&\boldsymbol{\theta}_j^{(1)} \sim \text{Gam}(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, 1/c_j^{(2)}),\\
&\boldsymbol{w}_{jk} = \boldsymbol{\pi}_{jk}\theta_{jk}^{(1)},\ \boldsymbol{\pi}_{jk} \sim \text{Dir}\big(\boldsymbol{\Phi}_{k:}^{(2)}\boldsymbol{\theta}_j^{(2)}/S_j\mathbf{1}_{S_j}\big),\\
&\boldsymbol{M}_j \sim \text{Pois}\big(\textstyle\sum_{k=1}^{K^{(1)}} \boldsymbol{D}_k * \boldsymbol{w}_{jk}\big).
\end{aligned} \tag{9}$$

Note using the relationship between the gamma and Dirichlet distributions ($e.g.$, Lemma IV.3 of Zhou & Carin (2012)), the elements of $\boldsymbol{w}_{jk}$ in the first hidden layer can be equivalently generated as

$$w_{jks} \sim \text{Gam}\big(\boldsymbol{\Phi}_{k:}^{(2)}\boldsymbol{\theta}_j^{(2)}/S_j, 1/c_j^{(2)}\big),\ s = 1, \dots, S_j. \tag{10}$$

Note the random variable $\theta_{jk}^{(1)}$, which pools the random weights of all words in document j, follows

$$\theta_{jk}^{(1)} = \textstyle\sum_{s=1}^{S_j} w_{jks} \sim \text{Gam}(\boldsymbol{\Phi}_{k:}^{(2)}\boldsymbol{\theta}_j^{(2)}, 1/c_j^{(2)}). \tag{11}$$

As described in Section 3.1, we have

$$\boldsymbol{m}_{jk\cdot\cdot} \sim \text{Pois}(\boldsymbol{w}_{jk}), \tag{12}$$

$$((\boldsymbol{d}'_{jk1\cdot}, \dots, \boldsymbol{d}'_{jkV\cdot})' \,|\, m_{jk\cdot\cdot}) \sim \text{Multi}(m_{jk\cdot\cdot}; \boldsymbol{D}_k(:)) \tag{13}$$

leading to the following conditional posteriors:

$$\begin{aligned}
&(\boldsymbol{w}_{jk} \,|\, -) \sim \text{Gam}(\boldsymbol{m}_{jk\cdot\cdot} + r_k, 1/(1 + c_j^{(2)})).\\
&(\boldsymbol{D}_k(:) \,|\, -) \sim \text{Dir}((\boldsymbol{d}'_{\cdot k1\cdot}, \dots, \boldsymbol{d}'_{\cdot kV\cdot})' + \eta\mathbf{1}_{|V|F}).
\end{aligned} \tag{14}$$

Since $\boldsymbol{w}_{jk} = \boldsymbol{\pi}_{jk}\theta_{jk}^{(1)}$, from (12) we have

$$m_{jk\cdot\cdot s} \sim \text{Pois}(\pi_{jks}\theta_{jk}^{(1)}), \tag{15}$$

$$m_{jk\cdots} \sim \text{Pois}(\theta_{jk}^{(1)}). \tag{16}$$

Since $\sum_{s=1}^{S_j} \pi_{jks} = 1$ by construction, we have

$$(\boldsymbol{m}_{jk\cdot\cdot} \,|\, m_{jk\cdots}) \sim \text{Multi}(m_{jk\cdots}; \boldsymbol{\pi}_{jk}), \tag{17}$$

and hence the following conditional posteriors:

$$(\theta_{jk}^{(1)} \,|\, -) \sim \text{Gam}(m_{jk\cdots} + \boldsymbol{\Phi}_{k:}^{(2)}\boldsymbol{\theta}_j^{(2)}, 1/(1 + c_j^{(2)})), \tag{18}$$

$$(\boldsymbol{\pi}_{jk}|-) \sim \text{Dir}(\boldsymbol{m}_{jk\cdot\cdot}/m_{jk\cdots} + \boldsymbol{\Phi}_{k:}^{(2)}\boldsymbol{\theta}_j^{(2)}/S_j), \tag{19}$$

$$(c_j^{(2)}|-) \sim \text{Gam}(\textstyle\sum_{k=1}^{K^{(1)}} \boldsymbol{\Phi}_{k:}^{(2)}\boldsymbol{\theta}_j^{(2)} + a_0, 1/(\sum_{k=1}^{K^{(1)}} \theta_{jk}^{(1)} + b_0)). \tag{20}$$

The derivation for the parameters of layer $t \in \{2, ..., T\}$ is the same as that of gamma belief network (GBN) (Zhou et al., 2016), omitted here for brevity.

B. Sensitivity to Filter Width

To investigate the effect of the filter width of the convolutional kernel, we have evaluated the performance of CPFA ($i.e.$, CPGBN with a single hidden layer) on the SUBJ dataset with a variety of filter widths (unsupervised feature extraction + linear SVM for classification). We use the same CPFA code but vary its setting of the filter width. Averaging over five independent runs, the accuracy for filter wdith 1, 2, 3, 4, 5, 6, and 7 are 74.9 ± 0.9, 77.3 ± 0.4, 77.5 ± 0.5, 77.8 ± 0.4, 77.6 ± 0.5, 78.0 ± 0.4, and 77.5 ± 0.4, respectively. Note when the filter width reduces to 1, CPFA reduces to PFA ($i.e.$, no convolution). These results suggest the performance of CPFA has low sensitivity to the filter width. While setting the filter width as three may not be the optimal choice, it is a common practice for existing text CNNs (Kim, 2014; Johnson & Zhang, 2015a).

C. Hierarchical Visualization

Distinct from word-level topics learned by traditional topic models (Deerwester et al., 1990; Papadimitriou et al., 2000; Lee & Seung, 2001; Blei et al., 2003; Hinton & Salakhutdinov, 2009; Zhou et al., 2012), we propose novel phrase-level topics preserving word order as shown in Table. 3, where each phrase-level topic is often combined with several frequently co-occurred short phrases. To explore the connections between phrase-level topics of different layers learned by CPGBN, we follow Zhou et al. (2016) to construct trees to understand the general and specific aspects of the corpus. More specifically, we construct trees learned from TREC dataset, with the network structure set as $[K^{(1)}, K^{(2)}, K^{(3)}] = [200, 100, 50]$. We pick a node at the top layer as the root of a tree and grow the tree downward by drawing a line from node k at layer t to the top M relevant nodes k' at layer $t - 1$.

As shown in Fig. 5, we select the top 3 relevant nodes at the second layer linked to the selected root node, and the top 2 relevant nodes at the third layer linked to the selected nodes at the second layer. Considering the TREC corpus only consists of questions (questions about abbreviation, entity, description, human, location, or numeric), most of the topics learned by CPGBN are focused on short phrases on asking specific questions, as shown in Table. 3. Following the branches of the tree in Fig. 5, the root node covers very general question types on "how many, how long, what, when, why," and it is clear that the topics become more and more specific when moving along the tree from the top to bottom, where the shallow topics of the first layer tend to focus on a single question type, $e.g.$, the $183th$ bottom-layer node queries "how many" and the $88th$ one queries "how long."

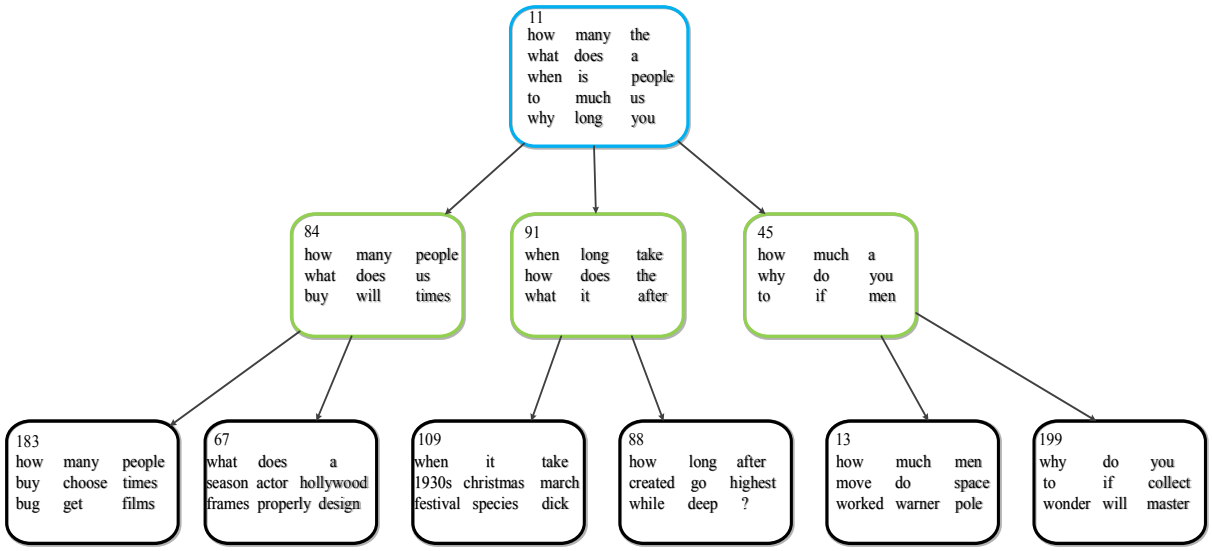

Figure 4. The $[6, 3, 1]$ phrase-level tree that includes all the lower-layer nodes (directly or indirectly) linked to the $11th$ node of the top layer, taken from the full $[200, 100, 50]$ network inferred by CPGBN on TREC dataset.

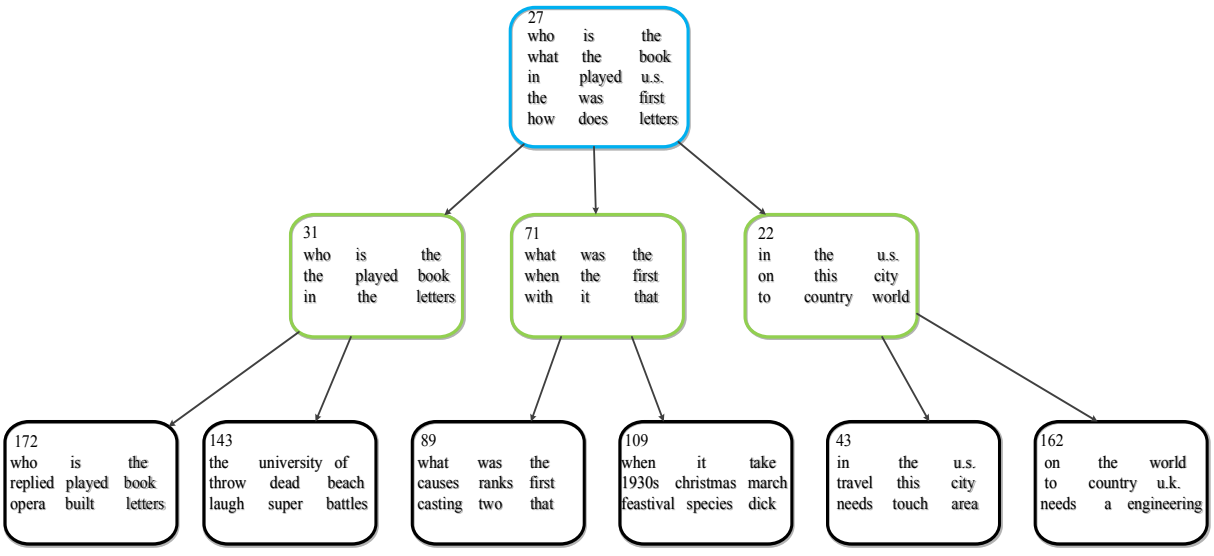

Figure 5. The $[6, 3, 1]$ phrase-level tree that include all the lower-layer nodes (directly or indirectly) linked to the $27th$ node of the top layer, taken from the full $[200, 100, 50]$ network inferred by CPGBN on TREC dataset.