Inter and Intra Topic Structure Learning with Word Embeddings

He Zhao [1] **Lan Du** [1] **Wray Buntine** [1] **Mingyaun Zhou** [2]

Abstract

One important task of topic modeling for text analysis is interpretability. By discovering structured topics one is able to yield improved interpretability as well as modeling accuracy. In this paper, we propose a novel topic model with a deep structure that explores both inter-topic and intra-topic structures informed by word embeddings. Specifically, our model discovers inter topic structures in the form of topic hierarchies and discovers intra topic structures in the form of sub-topics, each of which is informed by word embeddings and captures a fine-grained thematic aspect of a normal topic. Extensive experiments demonstrate that our model achieves the state-of-the-art performance in terms of perplexity, document classification, and topic quality. Moreover, with topic hierarchies and sub-topics, the topics discovered in our model are more interpretable, providing an illuminating means to understand text data.

1. Introduction

Significant research effort has been devoted to developing advanced text analysis technologies. Probabilistic topic models such as Latent Dirichlet Allocation (LDA), are popular approaches for this task, which discover latent topics from text collections. One preferred property of probabilistic topic models is interpretability: one can explain that a document is composed of topics and a topic is described by words. Although widely used, most variations of standard vanilla topic models (e.g., LDA) assume topics are independent and there are no structures among them. This limits those models' ability to explore any hierarchical thematic structures. Therefore, it is interesting to develop a model that is capable of exploring topic structures and yields not only improved modeling accuracy but also better interpretability.

Table 1. Example local topics with top 10 words

1	journal science biology research journals international cell psychology scientific bioinformatics
2	fitness piano guitar swimming violin weightlifting lessons training swim weight
3	taylor prince swift william jovi bon woman gala pill jon
4	san auto theft grand andreas mobile gta game rockstar december

One popular direction to explore topic structure is using the hierarchical/deep representation of text data, such as the nested hierarchical Dirichlet process (nHDP) (Paisley et al., 2015), Deep Poisson Factor Analysis (DPFA) (Gan et al., 2015), and Gamma Belief Network (GBN) (Zhou et al., 2016; Cong et al., 2017). In general, these models assume that topics in the higher layers of a hierarchy are more general/abstract than those in the lower layers. Therefore, by revealing hierarchical correlations between topics, topic hierarchies provide an intuitive way to understand text data.

In addition to topic hierarchies, we are also interested in analyzing the fine-grained thematic structure within each individual topic. As we know, in conventional models, topics are discovered locally from the word co-occurrences in a corpus. So we refer those topics as *local topics*. Due to the limitation of the context of a target corpus, some local topics may be hard to interpret because of the following two effects: (1) They can mix the words which co-occur locally in the target corpus but are less semantically related in general; (2) Local topics can be dominated by specialized words, which are less interpretable without extra knowledge. For example, we show four example topics of our experiments in Table 1, where we can see: Topic 1 is composed of the words from both the "scientific publication" and "biology" aspects; Topic 2 is a mixture of "sports" and "music"; Topics 3 and 4 are very specific topics about "singer" and "video game" respectively. We humans are able to understand those local topics in the above way because we are equipped with the global semantics of the words, making us go beyond the local context of the target corpus. Therefore, we are motivated to propose a model which is able to automatically analyze the fine-grained thematic structures of local topics, further improving the interpretability of topic modeling.

[1]Faculty of Information Technology, Monash University, Australia [2]McCombs School of Business, University of Texas at Austin. Correspondence to: Lan Du <lan.du@monash.edu>, Mingyuan Zhou <mingyuan.zhou@mccombs.utexas.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Fortunately, word embeddings such as GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013), and FastText (Bojanowski et al., 2017) can be used as an accessible source of global semantic information for topic models. Learned from large corpora, word embeddings encode the semantics of words with their locations in a space, where more related words are closer to each other. For example in Topic 1, according to the distances of word embeddings, the words "biology, cell, psychology, bioinformatics" should be in one cluster and "journal, science, research, international, scientific" should be in the other. Therefore, if a topic model can leverage the information in word embeddings, it may discover the fine-grained thematic structures of local topics. Furthermore, it has been demonstrated that conventional topic models suffer from data sparsity, resulting in a large performance degradation on some shorter internet-generated documents like tweets, product reviews, and news headlines (Zuo et al., 2016; Zhao et al., 2017c). In this case, word embeddings can also serve as complementary information to alleviate the sparsity issue in topic models.

In this paper, we propose a novel deep structured topic model, named the **W**ord **E**mbeddings **D**eep **T**opic **M**odel, WEDTM[1], which improves the interpretability of topic models by discovering topic hierarchies (i.e., *inter topic structure*) and fine-grained interpretations of local topics (i.e., *intra topic structure*). Specifically, the proposed model adapts a multi-layer Gamma Belief Network which generates deep representations of topics as well as documents. Moreover, WEDTM is able to split a local topic into a set of *sub-topics*, each of which captures one fine-grained thematic aspect of the local topic, in a way that each sub-topic is informed by word embeddings. WEDTM has the following key properties: **(1)** Better interpretability with topic hierarchies and sub-topics informed by word embeddings. **(2)** The state-of-the-art perplexity, document classification, and topic coherence performance, especially for sparse text data. **(3)** A straightforward Gibbs sampling algorithm facilitated by fully local conjugacy under data augmentation.

2. Related Work

Deep/hierarchical topic models: Several approaches have been developed to learn hierarchical representations of documents and topics. The Pachinko Allocation model (PAM) (Li & McCallum, 2006) assumes the topic structure is modeled by a directed acyclic graph (DAG), which is document specific. nCRP (Blei et al., 2010) models topic hierarchies by introducing a tree structure prior constructed with multiple CRPs. Paisley et al. (2015); Kim et al. (2012); Ahmed et al. (2013) further extend nCRP by either softening its constraints or applying it to different problems respectively. Poisson Factor Analysis (PFA) (Zhou et al., 2012) is a nonnegative matrix factorization model with Poisson link, which is a popular alternative to LDA, for topic modeling. The details of the close relationships between PFA and LDA can be found in Zhou (2018). There are several deep extensions to PFA for documents, such as DPFA (Gan et al., 2015), DPFM (Henao et al., 2015), and GBN (Zhou et al., 2016). Among them, GBN factorizes the factor score matrix (topic weights of documents) in PFA with nonnegative gamma-distributed hidden units connected by the weights drawn from the Dirichlet distribution. From a modeling perspective, GBN is related to PAM, while GBN assumes there is a corpus-level topic hierarchy shared by all the documents. As reported by Cong et al. (2017), GBN outperforms other hierarchical models including nHDP, DPFA, and DPFM. Despite having the attractive properties, these deep models barely consider intra topic structures or the sparsity issue associated with internet-generated corpora.

Word embedding topic models: Recently, there is a growing interest in applying word embeddings to topic models, especially for sparse data. For example, WF-LDA (Petterson et al., 2010) extends LDA to model word features with the logistic-normal transform, where word embeddings are used as word features in Zhao et al. (2017b). LF-LDA (Nguyen et al., 2015) integrates word embeddings into LDA by replacing the topic-word Dirichlet multinomial component with a mixture of a Dirichlet multinomial component and a word embedding component. Due to the non-conjugacy in WF-LDA and LF-LDA, part of the inference has to be done by MAP optimization. Instead of generating tokens, Gaussian LDA (GLDA) (Das et al., 2015) directly generates word embeddings with the Gaussian distribution. The model proposed in Xun et al. (2017) further extends GLDA by modeling topic correlations. MetaLDA (Zhao et al., 2017c; 2018a) is a conjugate topic model that incorporates both document and word meta information. However, in MetaLDA, word embeddings have to be binarized, which can lose useful information. WEI-FTM (Zhao et al., 2017b) is a focused topic model where a topic focuses on a subset of words, informed by word embeddings. To our knowledge, topic hierarchies and sub-topics are not considered in most of the existing word embedding models.

3. The Proposed Model

Based on the PFA framework, WEDTM is a hierarchical model with two major components: one for discovering the inter topic hierarchies and the other for discovering intra topic structures (i.e., sub-topics) informed by word embeddings. The two components are connected by the bottom-layer topics, detailed as follows. Assume that each document j is presented as a word count vector $\boldsymbol{x}_j^{(1)} \in \mathbb{N}_0^V$, where V is the size of the vocabulary; the pre-trained L

[1] https://github.com/ethanhezhao/WEDTM

dimensional real-valued embeddings for each word $v \in \{1, \cdots, V\}$ are stored in a L-dimensional vector $\boldsymbol{f}_v \in \mathbb{R}^L$. Now we consider WEDTM with T hidden layers, where the t-th layer is with K_t topics and k_t is the index of each topic. In the bottom layer ($t = 1$), there are K_1 (local) topics, each of which is associated with S sub-topics. To assist clarity, we split the generative process of the model into three parts, shown as follows:

$$\text{Generating documents} \begin{cases} \boldsymbol{\theta}_j^{(1)} \sim \text{Gam}\left[\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, p_j^{(2)}/(1-p_j^{(2)})\right], \\ \boldsymbol{\phi}_{k_1}^{(1)} \sim \text{Dir}\left(\boldsymbol{\beta}_{k_1}\right), \\ \boldsymbol{x}_j^{(1)} \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right), \end{cases}$$

$$\text{Inter structure} \begin{cases} \boldsymbol{\theta}_j^{(T)} \sim \text{Gam}\left(\boldsymbol{r}, 1/c_j^{(T+1)}\right), \\ \cdots \\ \boldsymbol{\theta}_j^{(t)} \sim \text{Gam}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}, 1/c_j^{(t+1)}\right) (t < T), \\ \boldsymbol{\phi}_{k_t}^{(t)} \sim \text{Dir}\left(\eta_0 \mathbf{1}\right) (t > 1), \\ \cdots \end{cases}$$

$$\text{Intra structure} \begin{cases} \boldsymbol{w}_{k_1}^{<s>} \sim \mathcal{N}\left[\mathbf{0}, \text{diag}(1/\boldsymbol{\sigma}^{<s>})\right], \\ \alpha_{k_1}^{<s>} \sim \text{Gam}(\alpha_0^{<s>}/S, 1/c_0^{<s>}), \\ \beta_{vk_1}^{<s>} \sim \text{Gam}\left(\alpha_{k_1}^{<s>}, e^{\boldsymbol{f}_v^\top \boldsymbol{w}_{k_1}^{<s>}}\right), \\ \beta_{vk_1} := \sum_s^S \beta_{vk_1}^{<s>}, \end{cases}$$

where $^{(t)}$ is the index of the layer that a variable belongs to and $^{<s>}$ is the index of sub-topic s. To complete the model, we impose the following priors on the latent variables:

$$r_{k_T} \sim \text{Gam}(\gamma_0/K_T, 1/c_0),$$
$$\gamma_0 \sim \text{Gam}(a_0, 1/b_0), p_j^{(t)} \sim \text{Beta}(a_0, b_0),$$
$$c_j^{(t)} \sim \text{Gam}(e_0, 1/f_0), \alpha_0^{<s>} \sim \text{Gam}(e_0, 1/f_0),$$
$$c_0^{<s>} \sim \text{Gam}(e_0, 1/f_0), \sigma_l^{<s>} \sim \text{Gam}(a_0, 1/b_0).$$

We first take a look at the bottom layer of the model, i.e., the process of generating the documents, which follows a PFA framework. In this part, WEDTM models the word counts $\boldsymbol{x}_j^{(1)}$ in a document by a Poisson (Pois) distribution and factorizes the Poisson parameters into a product of the factor loadings $\boldsymbol{\Phi}^{(1)} \in \mathbb{R}_+^{V \times K_1}$ and hidden units $\boldsymbol{\theta}_j^{(1)}$. $\boldsymbol{\theta}_j^{(1)}$ is the first-layer latent representation (unnormalized topic weights) of document j, each element of which is drawn from a gamma (Gam) distribution[2]. The k_1-th column of $\boldsymbol{\Phi}^{(1)}$, $\boldsymbol{\phi}_{k_1}^{(1)} \in \mathbb{R}_+^V$ is the word distribution of topic k_1, drawn from a Dirichlet (Dir) distribution. We then explain the component for discovering inter topic hierarchies, which is similar to the structure of GBN (Zhou et al., 2016). Specifically, the shape parameter of $\boldsymbol{\theta}_j^{(1)}$ is factorized into $\boldsymbol{\theta}_j^{(2)}$ and $\boldsymbol{\Phi}^{(2)} \in \mathbb{R}_+^{K_1 \times K_2}$, where $\boldsymbol{\theta}_j^{(2)}$ is the second-layer latent representation of document j and $\boldsymbol{\phi}_{k_2}^{(2)} \in \mathbb{R}_+^{K_1}$ models the correlations between topic k_2 and all the first-layer topics. Note that strictly speaking, k_2 is not a "real" topic as it is not a distribution over words. But it can be interpreted with words by $\boldsymbol{\Phi}^{(1)}\boldsymbol{\phi}_{k_2}^{(2)}$. By repeating this construction, we are able to build a deep structure to discover topic hierarchies.

Now we explain how sub-topics are discovered for the bottom-layer topics with the help of word embeddings. First of all, WEDTM applies individual asymmetric Dirichlet parameters $\boldsymbol{\beta}_{k_1} \in \mathbb{R}_+^V$ for each bottom-layer (local) topic $\boldsymbol{\phi}_{k_1}^{(1)}$. We further construct $\beta_{vk_1} = \sum_s^S \beta_{vk_1}^{<s>}$, where $\beta_{vk_1}^{<s>}$ models how strongly word v is associated with sub-topic s in local topic k_1. For each sub-topic s, we introduce an L-dimensional sub-topic embedding: $\boldsymbol{w}_{k_1}^{<s>} \in \mathbb{R}^L$. As $\beta_{vk_1}^{<s>}$ is gamma distributed, its scale parameter is constructed by the dot product of the embeddings of sub-topic s and word v through the exponential function.

The basic idea of our model is summarized as follows:

1. In terms of sub-topics, we assume each (local, bottom-layer) topic is associated with several sub-topics, in a way that the sub-topics contribute to the prior of the local topic via a sum model (Zhou, 2016). Therefore, if a word dominates in one or more sub-topics, it is likely that the word will still dominate in the local topic. With this construction, a sub-topic is expected to capture one fine-grained thematic aspect of the local topic and each sub-topic can be directly interpreted with words via $\boldsymbol{\beta}_{k_1}^{<s>} \in \mathbb{R}_+^V$.

2. To leverage word embeddings to inform the learning of sub-topics, we introduce the sub-topic embedding for each of them, $\boldsymbol{w}_{k_1}^{<s>}$, which directly interacts with the word embeddings. Therefore, sub-topic embeddings are learned with both the local context of the target corpus and the global information of word embeddings. According to our model construction, the probability density function of β_{vk_1} is the convolution of S covariance-dependent gamma distributions (Zhou, 2016). Therefore, if the sub-topic embeddings of s and word embeddings of v are close, the dot product of them will be large, giving a large expectation of $\beta_{vk_1}^{<s>}$. The large expectation means that v has a large weight in sub-topic s of k. Finally, $\beta_{vk_1}^{<s>}$ further contributes to the local topic's prior β_{vk_1}, informing $\phi_{vk_1}^{(1)}$ of the local topic.

3. It is also noteworthy the special case of WEDTM, where $S = 1$, meaning that there are no sub-topics and each local topic k_1 is associated with one topic embedding vector $\boldsymbol{w}_{k_1}$. Consequently, in WEDTM, there are three latent variables capturing the weights between the words and local topic k_1: $e^{\mathbf{F}^\top \boldsymbol{w}_{k_1}}$ ($\mathbf{F} \in \mathbb{R}^{L \times V}$ is

[2]The first and second parameters of the gamma distribution are the shape and scale respectively.

the embeddings of all the words), $\boldsymbol{\beta}_{k_1}$, and $\boldsymbol{\phi}_{k_1}^{(1)}$, each of which is a vector over words. It is interesting to analyze the connections and differences of them. $e^{\mathbf{F}^\top \boldsymbol{w}_{k_1}}$ is the prior of $\boldsymbol{\beta}_{k_1}$, while $\boldsymbol{\beta}_{k_1}$ is the prior of $\boldsymbol{\phi}_{k_1}^{(1)}$. So $e^{\mathbf{F}^\top \boldsymbol{w}_{k_1}}$ is the closest one to the word embeddings, i.e., the global semantic information, while $\boldsymbol{\phi}_{k_1}^{(1)}$ is the closest one to the data, i.e., the local document context of the target corpus. Therefore, unlike conventional topic models with $\boldsymbol{\phi}_{k_1}^{(1)}$ only, the three variables of WEDTM give three different views to the same topic, from global to local, respectively. We qualitatively show this interesting comparison in Section 5.4.

4. The last but not least, word embeddings in WEDTM can be viewed to serve as the prior/complementary information to assist the learning of the whole model, which is important especially for sparse data.

4. Inference

Unlike many other word embeddings topic models, the fully local conjugacy of WEDTM facilitates the derivation of an effective Gibbs sampling algorithm. As the sampling for the latent variables in the process of generating documents and modeling inter topic structure are similar to GBN, the details can be found in Zhou et al. (2016). Here we focus on the sampling of the latent variables for modeling intra topic structure.

Assume that sampled by Eq. (28) in Appendix B of Zhou et al. (2016), the latent count for the bottom-layer local topics are $x_{vjk_1}^{(1)}$, which counts how many words v in document j are allocated with local topic k_1.

Sample $\beta_{vk_1}^{<s>}$. We first sample:

$$\left(h_{vk_1}^{<1>}, \cdots, h_{vk_1}^{<S>}\right) \sim \text{Mult}\left(h_{vk_1}, \frac{\beta_{vk_1}^{<1>}}{\beta_{vk_1}}, \cdots, \frac{\beta_{vk_1}^{<S>}}{\beta_{vk_1}}\right), \tag{1}$$

where $h_{vk_1} \sim \text{CRT}\left(x_{v\cdot k_1}^{(1)}, \beta_{vk_1}\right)$ (Zhou & Carin, 2015; Zhao et al., 2017a), and $x_{v\cdot k_1}^{(1)} := \sum_j x_{vjk_1}^{(1)}$ [3]. Then:

$$\beta_{vk_1}^{<s>} \sim \frac{\text{Gam}(\alpha_{k_1}^{<s>} + h_{vk_1}^{<s>}, 1)}{e^{-\pi_{vk_1}^{<s>}} + \log \frac{1}{q_{k_1}}}, \tag{2}$$

where $q_{k_1} \sim \text{Beta}(\beta_{\cdot k_1}, x_{\cdot\cdot k_1}^{(1)})$ (Zhao et al., 2018b) and we define $\pi_{vk_1}^{<s>} := \boldsymbol{f}_v^\top \boldsymbol{w}_{k_1}^{<s>}$.

[3] We hereafter use $\cdot$ of a dimension to denote the sum over that dimension.

Sample $\alpha_k^{<s>}$. We first sample $g_{vk_1}^{<s>} \sim \text{CRT}\left(h_{vk_1}^{<s>}, \alpha_{k_1}^{<s>}\right)$, then:

$$\alpha_{k_1}^{<s>} \sim \frac{\text{Gam}(\alpha_0^{<s>}/S + g_{\cdot k_1}^{<s>}, 1)}{c_0^{<s>} + \log\left(1 + e^{\pi_{vk_1}^{<s>}} \log \frac{1}{q_{k_1}}\right)}. \tag{3}$$

It is noteworthy that the hierarchical construction on $\alpha_{k_1}^{<s>}$ is closely related to the gamma-negative binomial process and can be considered as a (truncated) gamma process (Zhou & Carin, 2015; Zhou, 2016) with an intrinsic shrinkage mechanism on S. It means that the model is able to automatically learn the number of effective sub-topics.

Sample $\boldsymbol{w}_{k_1}^{<s>}$.

$$\begin{aligned}
\boldsymbol{w}_{k_1}^{<s>} &\sim \mathcal{N}(\boldsymbol{\mu}_{k_1}^{<s>}, \boldsymbol{\Sigma}_{k_1}^{<s>}), \\
\boldsymbol{\mu}_{k_1}^{<s>} &= \\
\boldsymbol{\Sigma}_{k_1}^{<s>} &\left[\sum_v^V \left(\frac{h_{vk_1}^{<s>} - \alpha_{k_1}^{<s>}}{2} - \omega_{vk_1}^{<s>} \log\log \frac{1}{q_{k_1}}\right) \boldsymbol{f}_v\right], \\
\boldsymbol{\Sigma}_{k_1}^{<s>} &= \left[\text{diag}(1/\boldsymbol{\sigma}^{<s>}) + \sum_v^V \omega_{vk_1}^{<s>} \boldsymbol{f}_v (\boldsymbol{f}_v)^\top\right]^{-1},
\end{aligned} \tag{4}$$

where $\omega_{vk_1}^{<s>} \sim \text{PG}\left(h_{vk_1}^{<s>} + \alpha_{k_1}^{<s>}, \pi_{vk_1}^{<s>} + \log\log \frac{1}{q_{k_1}}\right)$ and PG denotes the Pólya gamma distribution (Polson et al., 2013). To sample from PG, we use an accurate and efficient approximate sampler in Zhou (2016).

Omitted derivations, details, and the overall algorithm are in the supplementary materials.

5. Experiments

We evaluate the proposed WEDTM by comparing it with several recent advances including deep topic models and word embedding topic models. The experiments were conducted on four real-world datasets including both regular and sparse texts. We report perplexity, document classification accuracy, and topic coherence scores. We also qualitatively analyze the topic hierarchies and sub-topics.

5.1. Experimental Settings

In the experiments, we used a regular text dataset (20NG) and three sparse text datasets (WS, TMN, Twitter), the details of which are as follows: **1. 20NG**, 20 Newsgroup, consists of 18,774 articles with 20 categories. Following Zhou et al. (2016), we used the 2000 most frequent terms after removing stopwords. The average document length is 76. **2. WS**, Web Snippets, contains 12,237 web search snippets with 8 categories, used by Li et al. (2016); Zhao et al. (2017c;b). The vocabulary contains 10,052 tokens and there are 15 words in one snippet on average. **3. TMN**, Tag My

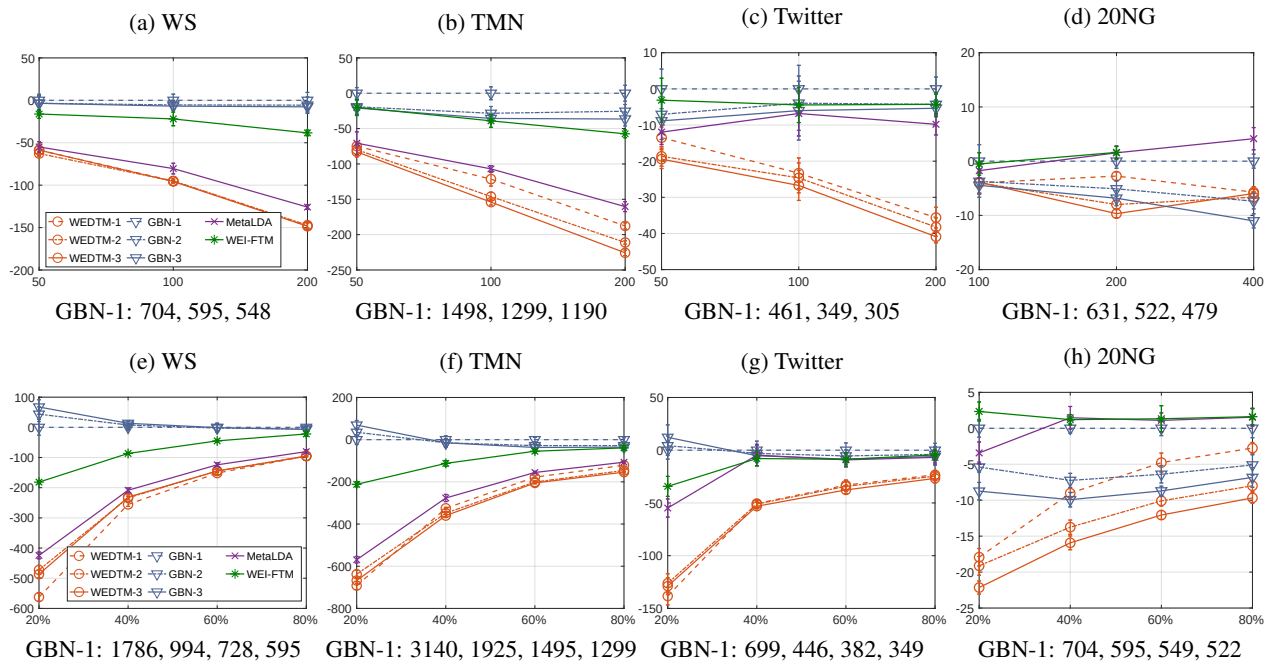

Figure 1. (a)-(d): Relative per-heldout-word perplexity[6] (the lower the better) with the varied K_1 and fixed proportion (80%) of training words of each document. (e)-(h): Relative per-heldout-word perplexity[6] with the varied proportion of training words of each document and fixed K_1 (100 on WS, TMN, and Twitter; 200 for 20NG). The error bars indicate the standard deviations of 5 random trials. The number attached to WEDTM and GBN indicates the number of layers (i.e., T) used.

News, consists of 32,597 RSS news snippets from Tag My News with 7 categories, used by Nguyen et al. (2015); Zhao et al. (2017c;b). Each snippet contains a title and a short description. There are 13,370 tokens in the vocabulary and the average length of a snippet is 18. **4. Twitter**, was extracted in 2011 and 2012 microblog tracks at Text REtrieval Conference (TREC)[4] and preprocessed in Yin & Wang (2014). It has 11,109 tweets in total. The vocabulary size is 6,344 and a tweet contains 21 words on average.

We compared WEDTM with: **1. GBN** (Zhou et al., 2016), the state-of-the-art deep topic model. **2. MetaLDA** (Zhao et al., 2017c; 2018a), the state-of-the-art topic model with binary meta information about document and/or word. Word embeddings need to be binarized before used in the model. **3. WEI-FTM** (Zhao et al., 2017b), the state-of-the-art focused topic model that incorporates real-valued word embeddings.

It is noteworthy that GBN was reported (Cong et al., 2017) to have better performance than other deep (hierarchical) topic models such as nHDP (Paisley et al., 2015), DPFA (Gan et al., 2015), and DPFM (Henao et al., 2015). MetaLDA and WEI-FTM were reported to perform better than other word embedding topic models including WF-LDA (Petterson et al., 2010) and GPUDMM (Li et al., 2016) as well as short text topic models like PTM (Zuo et al., 2016). Therefore, we considered the three above competitors to WEDTM.

Originally MetaLDA (when no document meta information is provided) and WEI-FTM follow the LDA framework, where the topic distribution for document j is $\boldsymbol{\theta_j} \sim \text{Dir}(\alpha_0 \mathbf{1})$ and α_0 is a hyperparameter (usually set to 0.1). For a fair comparison, we replaced this part with the PFA framework with the gamma-negative binomial process (Zhou & Carin, 2015), which is equivalent to GBN when $T = 1$ and closely related to the hierarchical Dirichlet Process LDA (HDPLDA) (Teh et al., 2012).

For all the models, we used 50-dimensional GloVe word embeddings pre-trained on Wikipedia[5]. Except for MetaLDA, where we followed the paper to binarise the word embeddings, the other three models used the original real-valued embeddings. The hyperparameter settings we used for WEDTM and GBN are $a_0 = b_0 = 0.01, e_0 = f_0 = 1.0, \eta_0 = 0.05$. For MetaLDA and WEI-FTM, we collected 1000 MCMC samples after 1000 burnins; for GBN and WEDTM, we collected 1000 for $T = 1$ and 500 for $T > 1$ MCMC samples after 1000 for $T = 1$ and 500 for $T > 1$ burnins, to estimate the posterior mean. Due to the shrinkage effect of WEDTM on S, discussed in Section 4, we set $S = 5$ which is large enough for all the topics.

[4] http://trec.nist.gov/data/microblog.html

[5] https://nlp.stanford.edu/projects/glove/

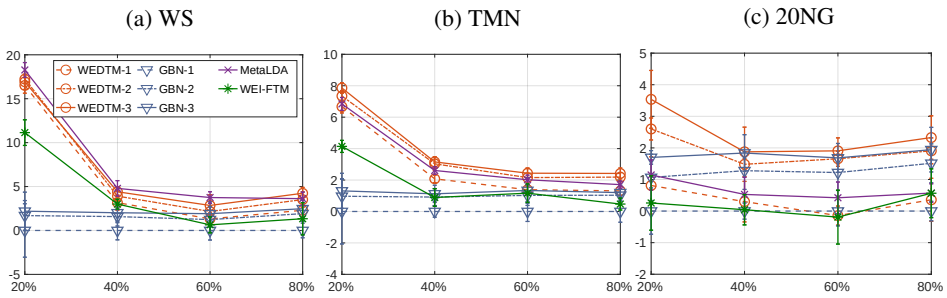

Figure 2. Relative document classification accuracy[6] (%) on WS, TMN, and 20NG with the varied proportion of training words of each training document. The results with $K_1 = 100$ on WS and TMN, $K_1 = 200$ on 20NG are reported.

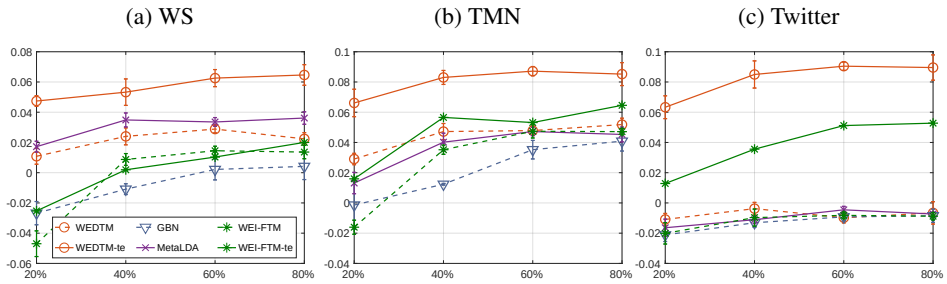

Figure 3. Topic coherence (NPMI, the higher the better) on WS, TMN, and Twitter with the varied proportion of training words of each document. The results with $K_1 = 100$ are reported. For WEDTM and WEI-FTM, the top words of a topic are generated by ranking the word distribution and the dot product of topic and word embeddings (denoted "-te").

5.2. Perplexity

Perplexity is a measure that is widely used (Wallach et al., 2009) to evaluate the modeling accuracy of topic models. Here we randomly chose a certain proportion of the word tokens in each document as training and used the remaining ones to calculate per-heldout-word perplexity. Figure 1 shows the relative perplexity[6] results of all the models on all the datasets, where we varied the number of bottom-layer topics as well as the proportion of training words. The proposed WEDTM performs significantly better than the others, especially on sparse data. There are several interesting remarks of the results: (1) The perplexity advantage of WEDTM over GBN becomes obvious when the corpus becomes sparse (e.g., WS/TMN/Twitter V.S. 20NG and 20% V.S. 80% training words). It shows that using word embeddings as the prior information benefits the model. (2) In general, increasing the depth of the model leads to better perplexity. However, when the data are too sparse (e.g. WS with 20% training words), the single-layer WEDTM and GBN perform better than their multi-layer counterparts. (3) Although MetaLDA and WEI-FTM leverage word embeddings as well, the proposed WEDTM outperforms them significantly. Perhaps the way that WEDTM incorporates word embeddings is more effective.

5.3. Document Classification

We consider the multi-class classification task for predicting the categories for test documents to evaluate the quality of the latent document representation (unnormalized topic weights) extracted by these models.[7] In this experiment, following Zhou et al. (2016), we ran the topic models on the training documents and trained a L_2 regularized logistic regression using the LIBLINEAR package (Fan et al., 2008) with the latent representation $\boldsymbol{\theta}_j^{(1)}$ as features. After training, we used the trained topic models to extract the latent representations of the test documents and the trained logistic regression to predict the categories. For all the datasets, we randomly selected 80% documents for training and used the remaining 20% for testing. Figure 2 shows the relative document classification accuracy[6] results for all the models. It can be observed that with word embeddings, WEDTM outperforms GBN significantly, the best on TMN and 20NG, and the second-best on WS. Again, we see a similar phe-

[6]We subtracted the score of GBN with only one layer (GBN-1) from the score of each model. The lines plot the differences. So GBN-1 is the horizontal line on "0". The absolute score of GBN-1 is given below each figure.

[7]The results of Twitter are not reported because each document of it is associated with multiple categories.

nomenon: word embeddings help more on the sparser data and increasing the network depth improves the accuracy.

5.4. Topic Coherence

Topic coherence is another popular evaluation of topic models (Zuo et al., 2016; Zhao et al., 2017b;b). It measures the semantic coherence in the most significant words (top words) in a topic. Here we used the Normalized Pointwise Mutual Information (NPMI) (Aletras & Stevenson, 2013; Lau et al., 2014) to calculate topic coherence score of the top 10 words of each topic and report the average score of all the topics.[8]

To compare with the other models, in this experiment, we set $S = 1$ for WEDTM. Recall that in WEDTM, from global to local, there are three ways to interpret a topic. Here we evaluate NPMI for two of them: $e^{\mathbf{F}^\top \boldsymbol{w}_{k_1}}$ and $\boldsymbol{\phi}_{k_1}^{(1)}$. Figure 3 shows the NPMI scores for all the models on WS, TMN, and Twitter. It is not surprising to see that the top words generated by $e^{\mathbf{F}^\top \boldsymbol{w}_{k_1}}$ in WEDTM always gain the highest NPMI scores, meaning that the topics are more coherent. This is because the topic embeddings in WEDTM directly interact with word embeddings. Moreover, if we just compare the topics generated by $\boldsymbol{\phi}_{k_1}^{(1)}$, WEDTM also gives more coherent topics than the other models. This demonstrates that the proposed model is able to discover more interpretable topics.

5.5. Qualitative Analysis

As one of the most appealing properties of WEDTM is its interpretability, we conducted the extensive qualitative evaluation of the quality of the topics discovered by WEDTM, including topic embeddings, sub-topics, and topic hierarchies. More qualitative analysis including topic hierarchy visualization and synthetic document generation is shown in the supplementary materials.

Demonstration of topic embeddings: We demonstrate that WEDTM discovers more coherent topics by comparing with those of GBN in Table 2. Here we set $S = 1$ as well. This demonstration further explains the numerical results in Figure 3. It is also interesting to compare the local interpretation ($\boldsymbol{\phi}_k^{(1)}$) and global interpretation (topic embeddings) of the same topic in WEDTM. For example, in the fifth set, the local interpretation (5.b) is about "networks and security," while the global interpretation (5.c) generalizes it with more general words related to "communications." We can also observe that although the local interpretation of WEDTM is not as close to word embeddings as the global interpretation, as informed by the global interpretation, the local interpretation of WEDTM's topics is still considerably more coherent than those in GBN.

Demonstration of sub-topics: In Figure 4, We show the sub-topics discovered by WEDTM for the topics used as examples at the beginning of the paper (Table 1). It can be observed that the intra topic structures with sub-topics clearly help to explain the local topics. For example, WEDTM successfully splits Topic 1 into sub-topics related to "journal" and "biology," and Topic 2 into "music" and "sports". Moreover, with the help of word embeddings, WEDTM discovers general sub-topics for specific topics. For example, Topic 3 and 4 are more interpretable with the sub-topics of "singer" and "game" respectively. The experiment also empirically demonstrates the shrinkage mechanism of the model: for most topics, the effective sub-topics are less than the maximum number $S = 5$.

Demonstration of topic hierarchies: Figure 5 shows an example that jointly demonstrates the inter and intra structures of WEDTM. The tree is a cluster of topics related to "health," where the topic hierarchies are discovered by ranking $\{\boldsymbol{\Phi}^{(t)}\}_t$, the leaf nodes are the topics in the bottom layer, and each bottom-layer topic is associated with a set of sub-topics. In WEDTM, the inter topic structures are revealed in the form of topic hierarchies while the intra topic structures are revealed in the form of sub-topics. Combining the two kinds of topic structures in this way gives a better view of the target corpus, which may further benefit other text analysis tasks.

6. Conclusion

In this paper, we have proposed WEDTM, a deep topic model that leverages word embeddings to discover inter topic structures with topic hierarchies and intra topic structures with sub-topics. Moreover, with the introduction to sub-topic embeddings, each sub-topic can be informed by the global information in word embeddings, so as to discover a fine-grained thematic aspect of a local topic. With topic embeddings, WEDTM provides different views to a topic, from global to local, which further improves the interpretability of the model. As a fully conjugate model, the inference of WEDTM can be done by a straightforward Gibbs sampling algorithm. Extensive experiments have shown that WEDTM achieves the state-of-the-art performance on perplexity, document classification, and topic quality. In addition, with topic hierarchies, sub-topics, and topic embeddings, the model can discover more interpretable structured topics, which helps to get better understandings of text data. Given the local conjugacy, it is possible to derive more scalable inference algorithms for WEDTM, such as stochastic variational inference and stochastic gradient MCMC, which is a good subject for future work.

[8] We used the Palmetto package with a large Wikipedia dump to compute NPMI (http://palmetto.aksw.org).

Table 2. Top 10 words of five sets of example topics on the WS dataset. Each set contains the top words of 3 topics: topic 'a' is generated by $\phi_k^{(1)}$ in GBN-3; topic 'b' is generated by $\phi_k^{(1)}$ in WEDTM-3; topic 'c' is generated by $e^{\mathbf{F}^\top \boldsymbol{w}_{k_1}}$ in WEDTM-3. Topic 'a' and 'b' are matched by the Hellinger distance of $\phi_k^{(1)}$. Topic 'b' and 'c' are different ways of interpreting one topic in WEDTM.

Topic	Index	Top 10 words	NPMI
1	a	engine car buying home diesel selling fuel automobile violin jet	-0.055
	b	engine motor diesel fuel gasoline jet electric engines gas technology	0.202
	c	engine diesel engines gasoline steam electric fuel propulsion motors combustion	0.224
2	a	party labor democratic political socialist movement union social news australian	0.168
	b	party political communist democratic socialist labor republican parties conservative leader	0.188
	c	party democratic communist labour liberal socialist conservative opposition elections republican	0.219
3	a	cancer lung tobacco intelligence artificial information health symptoms smoking treatment	-0.006
	b	cancer lung tobacco information health smoking treatment gov research symptoms	0.050
	c	cancer breast diabetes pulmonary cancers patients asthma cardiovascular cholesterol obesity	0.050
4	a	oscar academy awards swimming award winners swim oscars nominations picture	0.020
	b	art awards oscar academy gallery museum surrealism sculpture picasso arts	0.076
	c	paintings awards award art museum gallery sculpture painting picasso portrait	0.087
5	a	security computer network nuclear weapons networking spam virus spyware national	0.059
	b	security network wireless access networks spam spyware networking national computer	0.061
	c	wireless internet networks devices phone broadband users network wi-fi providers	0.143

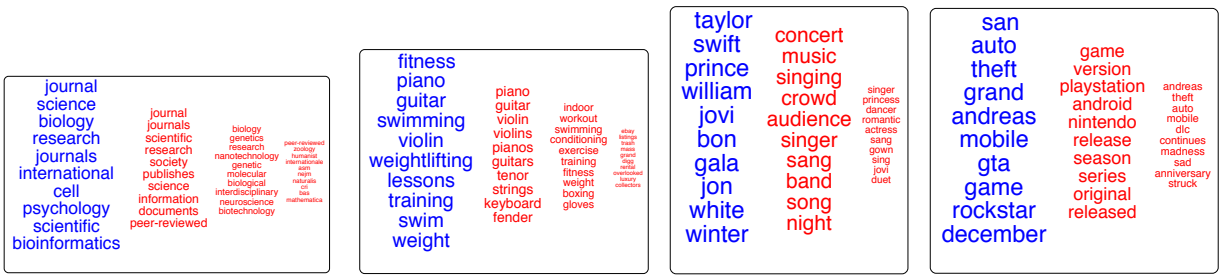

Figure 4. The sub-topics (red) of the example topics (blue). Larger font size indicates larger weight ($\sum_v^V \beta_{vk}^{<s>}$) of a sub-topic to the local topic. We set $S = 5$ and trimmed off the sub-topics with extreme small weights.

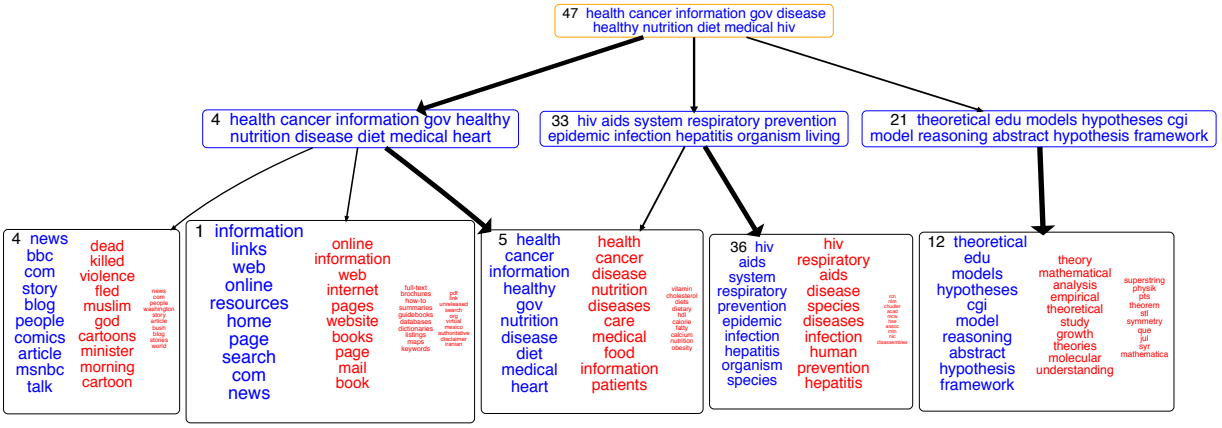

Figure 5. One example sub-tree of the topic hierarchy discovered by WEDTM on the WS dataset with $K_1 = 50$ and $S = 5$. The tree is generated in the same way to Zhou et al. (2016). A line from node k_t at layer t to node k_{t-1} at layer $t-1$ indicates that $\phi_{k_{t-1}k_t}^{(t)} > 1.5/K_{t-1}$ and its width indicates the value of $\phi_{k_{t-1}k_t}^{(t)}$ (i.e. topic correlation strength). The outside border of the text box is colored as orange, blue, or black if the node is at layer three, two, or one, respectively. For the leaf nodes, sub-topics are shown in the same way to Figure 4.

References

Ahmed, A., Hong, L., and Smola, A. Nested Chinese restaurant franchise process: Applications to user tracking and document modeling. In *ICML*, 2013.

Aletras, N. and Stevenson, M. Evaluating topic coherence using distributional semantics. In *Proc. of the 10th International Conference on Computational Semantics*, pp. 13–22, 2013.

Blei, D., Griffiths, T., and Jordan, M. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2): 7, 2010.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *TACL*, 5: 135–146, 2017.

Cong, Y., Chen, B., Liu, H., and Zhou, M. Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In *ICML*, pp. 864–873, 2017.

Das, R., Zaheer, M., and Dyer, C. Gaussian LDA for topic models with word embeddings. In *ACL*, pp. 795–804, 2015.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

Gan, Z., Chen, C., Henao, R., Carlson, D., and Carin, L. Scalable deep Poisson factor analysis for topic modeling. In *ICML*, pp. 1823–1832, 2015.

Henao, R., Gan, Z., Lu, J., and Carin, L. Deep Poisson factor modeling. In *NIPS*, pp. 2800–2808, 2015.

Kim, J. H., Kim, D., Kim, S., and Oh, A. Modeling topic hierarchies with the recursive Chinese restaurant process. In *CIKM*, pp. 783–792. ACM, 2012.

Lau, J. H., Newman, D., and Baldwin, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pp. 530–539, 2014.

Li, C., Wang, H., Zhang, Z., Sun, A., and Ma, Z. Topic modeling for short texts with auxiliary word embeddings. In *SIGIR*, pp. 165–174, 2016.

Li, W. and McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, pp. 577–584, 2006.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionally. In *NIPS*, pp. 3111–3119, 2013.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. Improving topic models with latent feature word representations. *TACL*, 3:299–313, 2015.

Paisley, J., Wang, C., Blei, D., and Jordan, M. Nested hierarchical Dirichlet processes. *TPAMI*, 37(2):256–270, 2015.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *EMNLP*, pp. 1532–1543, 2014.

Petterson, J., Buntine, W., Narayanamurthy, S. M., Caetano, T. S., and Smola, A. J. Word features for Latent Dirichlet Allocation. In *NIPS*, pp. 1921–1929, 2010.

Polson, N., Scott, J., and Windle, J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–1349, 2013.

Teh, Y. W., Jordan, M., Beal, M., and Blei, D. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2012.

Wallach, H. M., Mimno, D. M., and McCallum, A. Rethinking LDA: Why priors matter. In *NIPS*, pp. 1973–1981, 2009.

Xun, G., Li, Y., Zhao, W. X., Gao, J., and Zhang, A. A correlated topic model using word embeddings. In *IJCAI*, pp. 4207–4213, 2017.

Yin, J. and Wang, J. A Dirichlet multinomial mixture model-based approach for short text clustering. In *SIGKDD*, pp. 233–242. ACM, 2014.

Zhao, H., Du, L., and Buntine, W. Leveraging node attributes for incomplete relational data. In *ICML*, pp. 4072–4081, 2017a.

Zhao, H., Du, L., and Buntine, W. A word embeddings informed focused topic model. In *ACML*, pp. 423–438, 2017b.

Zhao, H., Du, L., Buntine, W., and Liu, G. MetaLDA: A topic model that efficiently incorporates meta information. In *ICDM*, pp. 635–644, 2017c.

Zhao, H., Du, L., Buntine, W., and Liu, G. Leveraging external information in topic modelling. *Knowledge and Information Systems*, pp. 1–33, 2018a.

Zhao, H., Rai, P., Du, L., and Buntine, W. Bayesian multi-label learning with sparse features and labels, and label co-occurrences. In *AISTATS*, pp. 1943–1951, 2018b.

Zhou, M. Softplus regressions and convex polytopes. *arXiv preprint arXiv:1608.06383*, 2016.

Zhou, M. Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis*, 2018.

Zhou, M. and Carin, L. Negative binomial process count and mixture modeling. *TPAMI*, 37(2):307–320, 2015.

Zhou, M., Hannah, L., Dunson, D. B., and Carin, L. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, pp. 1462–1471, 2012.

Zhou, M., Cong, Y., and Chen, B. Augmentable gamma belief networks. *JMLR*, 17(163):1–44, 2016.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., and Xiong, H. Topic modeling of short texts: A pseudo-document view. In *SIGKDD*, pp. 2105–2114, 2016.