Integration of Knowledge Graph Embedding into Topic Modeling with Hierarchical Dirichlet Process

Dingcheng Li, Siamak Zamani Dadaneh, Jingyuan Zhang, Ping Li
Cognitive Computing Lab (CCL)
Baidu Research USA
10900 NE Eighth St, Bellevue, WA 98004, USA
1195 Bordeaux Dr, Sunnyvale, CA 94089, USA
lidingcheng@baidu.com, zamanys4@gmail.com
zhangjingyuan03@baidu.com, liping11@baidu.com

Abstract

Leveraging domain knowledge is an effective strategy for enhancing the quality of inferred low-dimensional representations of documents by topic models. In this paper, we develop *topic modeling with knowledge graph embedding* (TMKGE), a Bayesian nonparametric model to employ knowledge graph (KG) embedding in the context of topic modeling, for extracting more coherent topics. Specifically, we build a hierarchical Dirichlet process (HDP) based model to flexibly borrow information from KG to improve the interpretability of topics. An efficient online variational inference method based on a stick-breaking construction of HDP is developed for TMKGE, making TMKGE suitable for large document corpora and KGs. Experiments on three public datasets illustrate the superior performance of TMKGE in terms of topic coherence and document classification accuracy, compared to state-of-the-art topic modeling methods.

1 Introduction

Topic models, such as Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2017) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), play significant roles in helping machines interpret text documents. Topic models consider documents as a bag of words. Given the word information, topic models formulate documents as mixtures of latent topics, where these topics are generated via the multinomial distributions over words. Bayesian methods are utilized to extract topical structures from the document-word frequency representations of the text corpus. Without supervision, however, it is found that the topics generated from these models are often not interpretable (Chang et al., 2009; Mimno et al., 2011). In recent studies, incorporating knowledge of different forms as a supervision has become a powerful strategy for discovering meaningful topics (Andrzejewski et al., 2009).

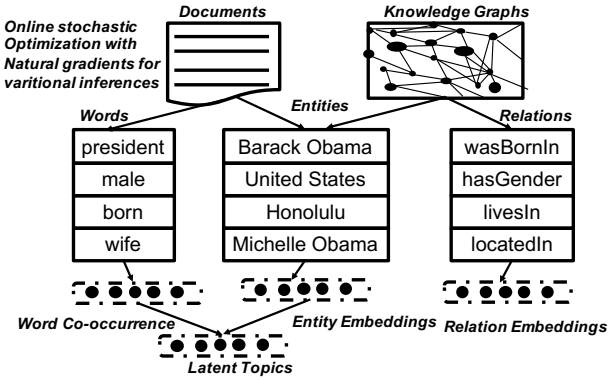

Figure 1: An overview of the proposed TMKGE framework. Entities are shared by both documents and knowledge graphs. Entity embeddings generated by *TransE* a knowledge graph embedding package are passed into TMKGE to generate hidden topics.

Most conventional approaches take prior domain knowledge into account to improve the topic coherence (Andrzejewski et al., 2009; Andrzejewski and Zhu, 2009; Hu et al., 2014; Jagarlamudi et al., 2012; Doshi-Velez et al., 2015). One commonly used domain knowledge is based on word correlations (Andrzejewski et al., 2009; Chen et al., 2013; Chen and Liu, 2014). For example, *must-links* and *cannot-links* among words are generated by domain experts to help topic modeling (Andrzejewski et al., 2009). Another useful form of knowledge for topic discoveries is based on word semantics (Andrzejewski and Zhu, 2009; Chemudugunta et al., 2008; Hu et al., 2014; Jagarlamudi et al., 2012; Doshi-Velez et al., 2015). In particular, word embedding (Pennington et al., 2014; Goldberg and Levy, 2014), in which bag of words are transformed into vector representations so that contexts are embedded into those word vectors, are used as semantic regularities to enhance topic models (Nguyen et al., 2015; Li et al., 2016; Das et al., 2015; Batmanghelich et al., 2016).

Proceedings of NAACL-HLT 2019, pages 940–950
Minneapolis, Minnesota, June 2 - June 7, 2019.

Knowledge graph (KG) embedding (Bordes et al., 2013) learns a low-dimensional continuous vector space for entities and relations to preserve the inherent structure of KGs. Yao et al. (2017) proposes KGE-LDA to incorporate embeddings of KGs into topic models to extract better topic representations for documents and shows promising performance. However, KGE-LDA forces words and entities to have identical latent representations, which is a rather restrictive assumption that prevents the topic model from recovering correct underlying latent structures of the data, especially in scenarios where only partial KGs are available.

This paper develops *topic modeling with knowledge graph embedding* (TMKGE), a hierarchical Dirichlet process (HDP) based model to extract more coherent topics by taking advantage of the KG structure. Unlike KGE-LDA, the proposed TMKGE allows for more flexible sharing of information between words and entities, by using a multinomial distribution to model the words and a multivariate Gaussian mixture to model the entities. With this approach, we introduce two proportional vectors, one for words and one for entities. In contrast, KGE-LDA only uses one, shared by both words and entities. Similar to HDP, TMKGE includes a collection of Dirichlet processes (DPs) at both corpus and document levels. The atoms of corpus-level DP form the base measure for document levels DPs of words and entities. Therefore, the atoms of corpus-level DP can represent word topics, entity mixture components, or both of them. Figure 1 provides an overview of TMKGE, where two sources of inputs, bag of words and KG embedding, extracted from corpus and KGs respectively, are passed into TMKGE.

As a nonparametric model, TMKGE does not assume a fix number of topics or entity mixture components as constraints. Instead, it learns the number of topics and entity mixture components automatically from the data. Furthermore, an efficient online variational inference algorithm is developed, based on Sethuraman's stick-breaking construction of HDP (Sethuraman, 1994). We in fact construct stick-breaking inference in a mini-batch fashion (Wang et al., 2011; Bleier, 2013), to derive a more efficient and scalable coordinate-accent variational inference for TMKGE.

Summary of contributions: TMKGE is a Bayesian nonparametric model to extract more coherent topics by taking advantage of knowledge graph structures. We introduce two proportional vectors for more flexible sharing of information between words and entities. We derive an efficient and scalable parameter estimation algorithm via online variational inference. Finally, we empirically demonstrate the effectiveness of TMKGE in topic discovering and document classification.

2 Background and Related Work

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a popular probabilistic model that learns latent topics from documents and words, by using Dirichlet priors to regularize the topic distributions. The generated topics from LDA models, however, are often not interpretable (Chang et al., 2009; Mimno et al., 2011), in part because LDA models are unsupervised without using prior knowledge or external resources.

In recent years, prior knowledge are leveraged to guide the process of topic modeling (Andrzejewski and Zhu, 2009; Hu et al., 2014; Jagarlamudi et al., 2012; Doshi-Velez et al., 2015). For example, the deep forest LDA (DF-LDA) model (Andrzejewski et al., 2009) is proposed to incorporate *must-links* and *cannot-links* among words into topic modeling. One weakness of the DF-LDA model is that the link information is domain-dependent. Later, general knowledge based LDA is introduced to leverage must-links from multiple domains (Chen et al., 2013). More recently, MetaLDA (Zhao et al., 2017) proposes to improve topic modeling by incorporating diverse meta information as priors for both document hyperparameter α and word hyperparameter β.

Besides the word correlations, word semantics are also utilized as one type of useful knowledge for topic modeling (Chemudugunta et al., 2008; Hu et al., 2014; Jagarlamudi et al., 2012). Word embeddings, as a low-dimensional continuous vectors of words (Mikolov et al., 2013; Bengio et al., 2003; Pennington et al., 2014) are regarded to be an efficient representations of word semantics. Latent Feature Topic Modeling (LFTM) is proposed to use pre-trained word embeddings in topic modeling (Nguyen et al., 2015). It incorporates the embedding of a word and its topics into the traditional multinomial distribution over words as the probability function of topic modeling. TopicVec extends LFTM by combining a word and its local contextual words together into the conventional multinomial distribution over words. It also

learns embedding representations for topics (Li et al., 2016). Gaussian-LDA goes further to improve topic modeling (Das et al., 2015) by taking into considerations the continuous nature of word embeddings. Shi et al. (2017) constructs a more unified framework, STE (skip-gram topic embedding) to address the problem of polysemy. Li et al. (2019) proposes a unified framework TMSA (Topic Modeling and Sparse Autoencoder) to improve topic discovery and word embedding simultaneously via a mutual learning mechanism.

Hu et al. (2016) proposes topic-based embeddings for learning from large knowledge graphs (KGE). KGE learns low-dimensional continuous vector space for both entities and relations to preserve the inherent structure of knowledge graphs. A Bayesian method is introduced by considering the embeddings of entities and relations as topics. Later, Yao et al. (2017) proposes knowledge graph embedding LDA (KGE-LDA) to encode entity embeddings learned from knowledge graphs into LDA and show that knowledge graph embeddings boost topic discoveries. Inspired by this work, we explore to utilize entity embeddings to encode prior knowledge for topic modeling.

3 Method

This section presents the TMKGE model and an efficient online variational inference for learning its parameters. We first provide a review of hierarchical Dirichlet process (HDP) (Teh et al., 2005).

3.1 Preliminaries of HDP

Dirichlet process (DP) (MacEachern and Müller, 1998) $G \sim \text{DP}(\gamma_0, G_0)$, with a base measure G_0 and a concentration parameter $\gamma_0 > 0$, is the distribution of a random probability measure G over a measurable space $(\Omega, \mathcal{B})$, such that for any measurable disjoint partition $(A_1, ..., A_Q)$ of Ω,

$$(G(A_1), ..., G(A_Q)) \sim \text{Dir}(\gamma_0 G_0(A_1), ..., \gamma_0 G_0(A_Q))$$

where "Dir" denotes a Dirichlet distribution.

Hierarchical Dirichlet process (HDP) (Teh et al., 2005), introduced for dealing with multiple (D) groups of data, is a distribution over a set of random probability measures over $(\Omega, \mathcal{B})$: one probability measure $G_d \sim \text{DP}(\alpha_0, G_0)$ for each group $d \in \{1, 2, ..., D\}$, and a global probability measure $G_0 \sim \text{DP}(\gamma_0, H)$ with a base measure H.

Stick-breaking construction Teh et al. (2005) shows that the draws from G_0 and G_d can be expressed as weighted sums of point masses:

$$G_0 = \sum_{k=0}^{\infty} \beta_k \delta_{\phi_k}, \qquad G_d = \sum_{k=0}^{\infty} \pi_{dk} \delta_{\phi_k}.$$

A more convenient stick-breaking construction, especially for deriving closed-form variational inference (Wang et al., 2011), is Sethuraman (1994)'s construction, which proceeds as follows. First, the global-level DP draw is represented as

$$\beta'_k \sim \text{Beta}(1, \gamma_0), \qquad \beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell),$$

Note that the distribution for $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{\infty}$ is also commonly written as $\boldsymbol{\beta} \sim \text{GEM}(\gamma_0)$ (Pitman, 2002). Subsequently, the group-level draws are constructed as

$$\begin{aligned} \psi_{dt} \sim & G_0, \qquad \pi'_{dt} = \text{Beta}(1, \alpha_0), \\ \pi_{dt} = & \pi'_{dt} \prod_{\ell=1}^{t-1} (1 - \pi'_{d\ell}), \qquad G_d = \sum_{t=1}^{\infty} \pi_{dt} \delta_{\psi_{dt}}. \end{aligned} \tag{1}$$

Alternatively, the group-level atoms $\{\psi_{dt}\}_{t=1}^{\infty}$ can be represented as $\psi_{dt} = \phi_{c_{dt}}$, where the auxiliary indicator variables c_{dt} are independently drawn from a multinomial Mult($\boldsymbol{\beta}$).

Teh et al. (2008) also proposes a collapsed inference method as an alternative of stick-breaking inference. However, following Fox et al. (2011), we stick to the uncollapsed HDP model considering our truncated Dirichlet process has more computational efficiency and is simple to implement.

3.2 The TMKGE Model

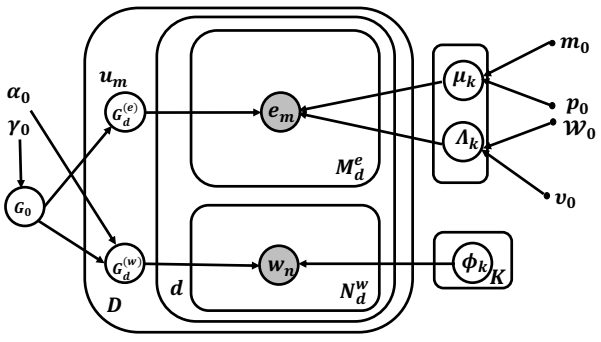

Figure 2: Graphical representation of the TMKGE framework. There are two components, the lower of which is the one for words and the upper of which is the one for entities. Both components share the Dirichlet process as priors. Since entities are represented with knowledge graph embeddings, therefore, each entity is generated with Gaussian priors while the one for words is still generated with Dirichlet priors.

Figure 2 is the graphical representation of TMKGE. Let D denote the number of documents in the corpus, where each document $d \in \{1, 2, ..., D\}$ contains $N_d^{(w)}$ words and $N_d^{(e)}$ entities. Throughout this work, superscripts (w) and (e) indicate word and entity related parameters, respectively. In each document d, the n-th word is represented by w_{dn}, where each word belongs to a vocabulary of size V, i.e., $w_{dn} \in \{1, 2, ..., V\}$. Furthermore, the P-dimensional embedding of the m-th entity is $\boldsymbol{e}_{dm}$, where the total number of unique entities in the corpus is E. We assume that entity embeddings are obtained from the "complete" knowledge graph, and hence they contain information independent of the corpus. In this paper, we use TransE (Bordes et al., 2013), a simple and effective tool for knowledge encoding, to calculate the embeddings of entities extracted from the documents. We should mention that we remove the normalization step of TransE and thus the output vectors ($\boldsymbol{e}_{dm}$) do not have unit ℓ_2 norm.

TMKGE builds upon HDP for joint modeling of word topics and entity mixtures. At the corpus level, word topics and entity mixtures correspond to atoms of a Dirichlet process $G_0 \sim \text{DP}(\gamma_0, H)$. At the document level, word topics and entity mixture components are atoms of independent DPs, with shared base measure G_0. Mathematically, for document d, we have

$$G_d^{(w)} \sim \text{DP}(\alpha_0, G_0), \qquad G_d^{(e)} \sim \text{DP}(\alpha_0, G_0),$$

where $G_d^{(w)}$ and $G_d^{(e)}$ are word and entity related DPs. Sethuraman's construction in (1) yields

$$G_d^{(w)} = \sum_{t=1}^{\infty} \pi_{dt}^{(w)} \delta_{\psi_{dt}^{(w)}}, \qquad G_d^{(e)} = \sum_{t=1}^{\infty} \pi_{dt}^{(e)} \delta_{\psi_{dt}^{(e)}}. \quad (2)$$

These DPs are then used to assign words and entities to topics and mixture components, respectively. In document d, let $z_{dn}^{(w)}$ denote the topic assigned to the n-th word, and $z_{dm}^{(e)}$ denote the mixture component assigned to the m-th entity. Using the mixing proportions of DPs in (2), we have

$$p(z_{dn}^{(w)} = t) = \pi_{dt}^{(w)}, \qquad p(z_{dn}^{(e)} = t) = \pi_{dt}^{(e)}.$$

For simplicity, we use index t to denote both word and entity related atoms, although they can correspond to different atoms of the global DPs.

The mixing proportions of corpus-level DP are used to map the document atoms to the shared global atoms. More precisely, we introduce the word and entity atoms mapping auxiliary variables $\boldsymbol{c}_d^{(w)} = \{c_{dt}^{(w)}\}_{t=1}^{\infty}$ and $\boldsymbol{c}_d^{(e)} = \{c_{dt}^{(e)}\}_{t=1}^{\infty}$. The mapping probabilities then can be expressed as

$$p(c_{dt}^{(w)} = k) = \beta_k, \qquad p(c_{dt}^{(e)} = k) = \beta_k.$$

TMKGE allows flexible sharing of information between knowledge graphs and documents. This is an important advantage, as in practice only partial relational information are available, and thus strictly forcing the topics and entity mixtures to share components may lead to reducing the power of model to correctly recover the latent structure of the data. Furthermore, the nonparametric nature of the model enables the automatic discovery of number of atoms for both words and entities, at document and corpus levels.

Each atom of corpus DP (G_0) corresponds to a set of parameters for both words and entities. Atom k contains topic-word Dirichlet distribution $\boldsymbol{\phi}_k = (\phi_{k1}, ..., \phi_{kV})^T$, and entity Gaussian mixture parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}$. Given $\boldsymbol{\phi}_k$ and topic assignment variables, the generative process for n-th word of document d is

$$\begin{aligned} z_{dn}^{(w)} &\sim \text{Mult}(\boldsymbol{\pi}_d^{(w)}), \\ (w_{dn} | z_{dn}^{(w)} = t, c_{dt}^{(w)} = k, \boldsymbol{\phi}_k) &\sim \text{Mult}(\boldsymbol{\phi}_k). \end{aligned}$$

In a similar fashion, the generative process of m-th entity of document d is

$$\begin{aligned} z_{dm}^{(e)} &\sim \text{Mult}(\boldsymbol{\pi}_d^{(e)}), \\ (\boldsymbol{e}_{dm} | z_{dm}^{(e)} = t, c_{dt}^{(e)} = k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &\sim \text{N}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}), \end{aligned}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ are the mean vector and precision matrix of multivariate Gaussian distribution.

Furthermore, we impose conjugate priors on both word and entity components parameters as:

$$\begin{aligned} \boldsymbol{\phi}_k &\sim \text{Dir}(\eta, ..., \eta), \qquad \boldsymbol{\mu}_k \sim \text{N}\big(\boldsymbol{m}_0, (\rho_0 \boldsymbol{\Lambda}_k)^{-1}\big), \\ \boldsymbol{\Lambda}_k &\sim \text{Wishart}(\nu_0, \boldsymbol{W}_0). \end{aligned}$$

3.3 Online Variational Inference

In this section, inspired by (Wang et al., 2011), we propose an online variational inference algorithm for efficient learning of TMKGE model parameters. We use a fully factorized variational distribution based on stick-breaking construction, and perform online mean-field variational inference.

In addition to topic parameters $\boldsymbol{\phi}_k$ and entity mixture parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}$, other parameters of interest are corpus-level stick proportions $\boldsymbol{\beta}' = \{\beta'_k\}_{k=1}^{\infty}$, document-level stick proportions for words $\boldsymbol{\pi}_d'^{(w)} = \{\pi_{dt}'^{(w)}\}_{t=1}^{\infty}$ and entities $\boldsymbol{\pi}_d'^{(e)} = \{\pi_{dt}'^{(e)}\}_{t=1}^{\infty}$, topic assignments for words

$\boldsymbol{z}_d^{(w)} = \{z_{dn}^{(w)}\}_{n=1}^{N_d^{(w)}}$, mixture assignments for entities $\boldsymbol{z}_d^{(e)} = \{z_{dm}^{(e)}\}_{m=1}^{N_d^{(e)}}$, and mapping variables $\boldsymbol{c}_d^{(w)}$ and $\boldsymbol{c}_d^{(e)}$. Denote $\Theta^{(w)}$ and $\Theta^{(e)}$ respectively the word and entity related parameters. Then the variational distribution factorizes as

$$q(\boldsymbol{\beta}', \Theta^{(w)}, \Theta^{(e)}) = q(\boldsymbol{\beta}')q(\Theta^{(w)})q(\Theta^{(e)}).$$

For corpus-level stick proportions, we assume a Beta distribution:

$$q(\boldsymbol{\beta}') = \prod_{k=1}^{K-1} \text{Beta}(\beta'_k|u_k, v_k),$$

where the number of global atoms is truncated at K, thereby $q(\beta'_K = 1) = 1$. For the word related parameters $\Theta^{(w)}$, we have

$$\begin{aligned}
q(\Theta^{(w)}) =& q(\boldsymbol{c}^{(w)})q(\boldsymbol{z}^{(w)})q(\boldsymbol{\pi}'^{(w)})q(\boldsymbol{\phi}),\\
q(\boldsymbol{c}^{(w)}) =& \prod_{d=1}^{D}\prod_{t=1}^{T-1} \text{Mult}(\boldsymbol{\varphi}_{dt}^{(w)}),\\
q(\boldsymbol{z}^{(w)}) =& \prod_{d=1}^{D}\prod_{n=1}^{N_d^{(w)}} \text{Mult}(\boldsymbol{\varsigma}_{dt}^{(w)}),\\
q(\boldsymbol{\pi}'^{(w)}) =& \prod_{d=1}^{D}\prod_{t=1}^{T-1} \text{Beta}(\pi'_{dt}|a_{dt}^{(w)}, b_{dt}^{(w)}),\\
q(\boldsymbol{\phi}) =& \prod_{k=1}^{K} \text{Dir}(\boldsymbol{\lambda}_k).
\end{aligned}$$

The variational distributions for entity related parameters have a similar form to the above distributions, except the Gaussian mixture parameters, which are expressed as follows:

$$q(\boldsymbol{\mu}_k) = \text{N}\big(\boldsymbol{m}_k, (\rho_k\boldsymbol{\Lambda}_k)^{-1}\big), \quad q(\boldsymbol{\Lambda}_k) = \text{Wishart}(\nu_k, \boldsymbol{W}_k).$$

In standard variational inference theory, the evidence lower bound (ELBO), which is the lower bound to the marginal log likelihood of the observed data, is maximized to find the best variational approximation to the true intractable posterior. Given the modeling framework of TMKGE, the ELBO can be written as

$$\begin{aligned}
\mathcal{L}(q) = \sum_d \Bigg\{ & \mathbb{E}\Big[\log\Big(p(\boldsymbol{w}_d|\boldsymbol{c}_d^{(w)}, \boldsymbol{z}_d^{(w)}, \boldsymbol{\phi})p(\boldsymbol{c}_d^{(w)}|\boldsymbol{\beta}')\\
& \times p(\boldsymbol{z}_d^{(w)}|\boldsymbol{\pi}'^{(w)}_d)p(\boldsymbol{e}_d|\boldsymbol{c}_d^{(e)}, \boldsymbol{z}_d^{(e)}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\\
& \times p(\boldsymbol{c}_d^{(e)}|\boldsymbol{\beta}')p(\boldsymbol{z}_d^{(e)}|\boldsymbol{\pi}'^{(e)}_d)p(\boldsymbol{\pi}'^{(w)}_d|\alpha_0)p(\boldsymbol{\pi}'^{(e)}_d|\alpha_0)\Big)\Big]\\
& + H\big(q(\boldsymbol{c}_d^{(w)})\big) + H\big(q(\boldsymbol{z}_d^{(w)})\big) + H\big(q(\boldsymbol{\pi}'^{(w)}_d)\big)\\
& + H\big(q(\boldsymbol{c}_d^{(e)})\big) + H\big(q(\boldsymbol{z}_d^{(e)})\big) + H\big(q(\boldsymbol{\pi}'^{(e)}_d)\big)\Bigg\}\\
& + \mathbb{E}\Big[\log\Big(p(\boldsymbol{\beta}')p(\boldsymbol{\phi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\Big)\Big] + H\big(q(\boldsymbol{\beta}')\big)\\
& + H\big(q(\boldsymbol{\phi})\big) + H\big(q(\boldsymbol{\mu}, \boldsymbol{\Lambda})\big),
\end{aligned}$$

where $H(\cdot)$ is the entropy term for variational distribution. By taking derivatives of this lower bound with respect to each variational parameter, we derive the coordinate ascent update steps.

We develop an online variational inference for TMKGE, to process large datasets (Wang et al., 2011; Hoffman et al., 2010). Given the existing corpus-level parameters, first a document d is sampled and then its optimal document-level variational parameters are computed. For word related variational parameters, these updates include

$$\begin{aligned}
a_{dt}^{(w)} &= 1 + \sum_n \varsigma_{dnt}^{(w)},\\
b_{dt}^{(w)} &= \alpha_0 + \sum_n \sum_{s=t+1}^{T} \varsigma_{dns}^{(w)},\\
\varphi_{dtk}^{(w)} &\propto \exp\Big(\sum_n \varsigma_{dns}^{(w)}\mathbb{E}_q\big[\log p(w_{dn}|\boldsymbol{\phi}_k)\big]\mathbb{E}_q\big[\log\beta_k\big]\Big),\\
\varsigma_{dnt}^{(w)} &\propto \exp\Big(\sum_k \varphi_{dtk}^{(w)}\mathbb{E}_q\big[\log p(w_{dn}|\boldsymbol{\phi}_k)\big]\mathbb{E}_q\big[\log\pi_{dt}^{(w)}\big]\Big)
\end{aligned} \tag{3}$$

where expectations are with respect to variational distributions and have closed forms. For entity related variational parameters, similar updates can be derived, with the term $\mathbb{E}_q\big[\log p(\boldsymbol{e}_{dm}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\big]$ replacing $\mathbb{E}_q\big[\log p(w_{dn}|\boldsymbol{\phi}_k)\big]$. Following Wang et al. (2011), for the corpus-level variational parameters, we use the following gradients:

$$\begin{aligned}
\partial\lambda_{kv} &= -\lambda_{kv} + \eta + D\sum_t \varphi_{dtk}^{(w)}\big(\sum_n \varsigma_{dnt}^{(w)} I[w_{dn} = v]\big),\\
\partial\boldsymbol{m}_k &= -\boldsymbol{m}_k + \frac{D\sum_{m,t}\varphi_{dtk}^{(e)}\varsigma_{dmt}^{(e)}\boldsymbol{e}_{dm} + \rho_0\boldsymbol{m}_0}{Dr_k + \rho_0},\\
\partial\rho_k &= -\rho_k + \rho_0 + Dr_k,\\
\partial\nu_k &= -\nu_k + \nu_0 + Dr_k,\\
\partial\boldsymbol{W}_k &= -\boldsymbol{W}_k + \Big(\boldsymbol{W}_0^{-1} + D\sum_{m,t}\varphi_{dtk}^{(e)}\varsigma_{dmt}^{(e)}\boldsymbol{e}_{dm}\boldsymbol{e}_{dm}^T\Big)^{-1},\\
\partial u_k &= -u_k + 1 + D\sum_t\big(\varphi_{dtk}^{(w)} + \varphi_{dtk}^{(e)}\big),\\
\partial v_k &= -v_k + \gamma_0 + D\sum_t\sum_{\ell=k+1}^{K}\big(\varphi_{dt\ell}^{(w)} + \varphi_{dt\ell}^{(e)}\big),
\end{aligned} \tag{4}$$

where r_k is defined as $\sum_{m,t}\varphi_{dtk}^{(e)}\varsigma_{dmt}^{(e)}$. The corpus-level parameters are then updated using these gradients (among them, the first, the fifth and the sixth are natural gradients while the other four are approximations from the posterior of Gaussian Wishart scale matrix W. It appears difficult to obtain natural gradients for those four.) and a learning rate parameter ϵ_t. For instance, for topic-words distribution parameters we have

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \epsilon_{t_0}\partial\boldsymbol{\lambda}. \tag{5}$$

			number of top words and PMI scores					
model	parameters	data source	5	10	15	20	25	30
TMKGE	K=300, T=20	20 Newsgroups	**20.8**	91.1	210.0	380.0	**602.0**	**876.0**
HDP	K=300, T=20		20.0	**91.6**	**212.6**	**384.1**	598.4	868.7
LDA	K=100		13.5	64.6	163.4	285.0	455.2	671.1
KGE-LDA	K=30		18.9	69.8	187.5	320.6	482.7	616.5
TMKGE	K=300, T=20	NIPS	16.6	**97.1**	160.3	**299.6**	**474.5**	**685.5**
HDP	K=300, T=20		**16.7**	66.8	157.2	280.2	444.0	643.1
LDA	K=100		13.9	67.6	**161.9**	297.0	471.2	681.1
KGE-LDA	K=30		14.3	97.2	163.4	285.3	453.3	645.4
TMKGE	K=300, T=20	Ohsumed	**21.6**	**123.3**	**237.3**	**407.7**	**624.2**	**895.5**
HDP	K=300, T=20		15.6	70.7	168.2	338.9	582.9	864.9
LDA	K=100		11.9	65.6	131.9	257.0	481.2	691.1
KGE-LDA	K=30		15.6	116.5	185.4	354.2	585.4	795.6

Table 1: Topic Coherence of all models on three datasets with different number of top words. A higher PMI score implies a more coherent topic. Improvements of TMKGE over other methods are significant.

The rest of corpus-level variational parameters in (4) can be similarly updated. To ensure that the parameters converge to a stationary point, the learning rate satisfies (Hoffman et al., 2010; Sato, 2001) $\sum_{t_0=1}^{\infty} \epsilon_{t_0} = \infty$ and $\sum_{t_0=1}^{\infty} \epsilon_{t_0}^2 < \infty$.

Following Wang et al. (2011), we use $\epsilon_{t_0} = (\tau_0 + t_0)^{-\kappa}$, where $\kappa \in (0.5, 1]$ and $\tau_0 > 0$. To improve the stability of online variational inference, we use a mini-batch of documents to compute the natural gradients. That is, the contribution of the single document d in (4) is replaced by sum of contributions of documents in the mini-batch $\mathcal{S}$, and the factor D is replaced by $D/|\mathcal{S}|$. The overall scheme of online variational inference for TMKGE is shown in Algorithm 1.

Algorithm 1 Online variational inference for the proposed TMKGE framework.

Initialize corpus-level variational parameters.
while *Stopping criterion is not met* **do**
  Sample a random document d from the corpus.
  Update $\boldsymbol{a}_d^{(w)}$, $\boldsymbol{b}_d^{(w)}$, $\boldsymbol{\varphi}_d^{(w)}$ and $\boldsymbol{\varsigma}_d^{(w)}$ using (3).
  Update $\boldsymbol{a}_d^{(e)}$, $\boldsymbol{b}_d^{(e)}$, $\boldsymbol{\varphi}_d^{(e)}$ and $\boldsymbol{\varsigma}_d^{(e)}$ similar to (3).
  Compute the natural gradients using (4).
  Set $\epsilon_{t_0} = (\tau_0 + t_0)^{-\kappa}$ and $t_0 \leftarrow t_0 + 1$.
  Update all corpus-level parameters as (5).
end

4 Experiments

We evaluate TMKGE on two experimental tasks and compare its performance to those of LDA, HDP and KGE-LDA. For LDA and HDP, we use the online variational inference implementations. More precisely, we will evaluate our framework by the test whether it finds coherent and meaningful topics and the test whether it can achieve good performance in document classification.

We run our experiments on three popular datasets; 20 Newsgroups, NIPS and the Ohsumed corpus. The 20 Newsgroups dataset contains 18,846 documents evenly categorized into 20 different categories.

The NIPS dataset contains 1,740 papers from the NIPS conference. The Ohsumed corpus is from the MEDLINE database. We consider the 13,929 unique Cardiovascular diseases abstracts in the first 20,000 abstracts of the year 1996. Each document in the set has one or more associated categories from the 23 disease categories. The documents belonging to multiple categories are eliminated so that 7,400 documents belonging to only one category remain. The datasets are tokenized with Stanford CoreNLP (Manning et al., 2014). After standard pre-processing (such as removing stop words), there are 20,881 distinct words in the 20 Newsgroups dataset, 14,482 distinct words in the NIPS dataset and 8,446 distinct words in the Ohsumed dataset.

4.1 External knowledge source

The knowledge graph we employ is WordNet (Miller, 1995). WordNet is a large lexical knowledge graph. Entities in WordNet are synonyms which express distinct concepts. Relations in WordNet mainly involve conceptual-semantic and lexical relations. We use a subset of WordNet (WN18) introduced in Bordes et al. (2011) and employed in Yao et al. (2017) as well. WN18 contains 151,442 triplets with 40,943 entities and 18 relations. We link tokenized words to entities in WN18 via NLTK (Bird and Loper, 2004).

20 Newsgroups			NIPS			Ohsumed		
lord	tcp/ip	kuwait	distribution	network	tube	vietnam	hemagglutinin	shbg
God	drive	iraq	gaussian	learning	regression	veterans	anti-tumor	patients
elohim	system	kuwaiti	posterior	model	svs	mthfr	mthfr	globulin
jesus	computer	sabah	covariance	neural	support	income	tumor	testicular
subject	information	abdulla	ensemble	data	fraction	white	pbl	hormone
israel	space	gulf	matrix	figure	erros	proportion	antibody	levels
armenian	windows	amir	KL	information	vapnik	drinking	meh	group
christ	data	ahmed	divergence	units	algorithm	era	ab test	sex
john	message	sheikh	approximate	problem	smola	lifetime	verapamil	binding
group	software	saudi	algorithm	recognition	vector	interview	radioactivity	treatment
101.2	98.5	119.3	152.6	91.1	106.3	105.4	135.4	152.2
20 Newsgroups			NIPS			Ohsumed		
internet	drive	car	distribution	control	kernel	gene	cancer	treatment
mail	windows	cars	bayesian	trajectory	support	dna	tumor	therapy
email	dos	engine	gaussian	robot	xi	protein	survival	dose
list	card	oil	prior	controller	vector	region	tumors	drug
message	disk	miles	posterior	arm	margin	genetic	carcinoma	effects
address	mac	dealer	probability	model	examples	analysis	breast	placebo
fax	scsi	speed	variables	forward	set	mutation	stage	trial
network	memory	buy	markov	motor	kernels	sequence	malignant	oral
send	system	ford	distribution	trajectories	svm	molecular	chemotherapy	mg
e-mail	apple	drive	approximation	inverse	machines	mrna	primary	effective
89.2	84.4	63.6	154.8	86.2	88.3	149.9	107.7	106.5

Table 2: Example topics learned from three datasets by TMKGE with $K = 300$ and $T = 20$, and KGE-LDA with $K = 30$. The last row for each model is the topic coherence computed using the 4,776,093 Wikipedia documents as reference. Some medical short words: pbl = Peripheral blood leucocyte, meh = Mean erythrocyte hemoglobin.

4.2 Model parameters

In the experiments, for each method, we report the results based on the hyperparameter settings that obtain the best performances. For TMKGE and HDP, we report the results for $K = 300$, $T = 20$ and $K = 100$, $T = 10$ cases. For LDA and KGE-LDA, respectively, we have $K = 100$ and $K = 30$. Throughout this work we fix the dimension of entity embedding as $P = 5$. For online variational inference, we run the algorithms for 1000 iterations, with mini-batch size of 100.

4.3 Topic Coherence

We assess the performance of the proposed TMKGE model based on topic coherence. Topic coherence has been shown to be more consistent with human judgment than other typical topic model metrics such as perplexity (Chang et al., 2009; Newman et al., 2010). We perform both quantitative and qualitative analysis of the topics discovered by TMKGE, and compare its performance to those of LDA, HDP and KGE-LDA.

4.3.1 Quantitative Analysis

We evaluate the coherence of discovered topics by the point-wise mutual information (PMI) Topic Coherence metric. The PMI Topic Coherence is implemented following Newman et al. (2010):

$$PMI(k) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

where k refers to a topic, N refers to the number of top words of k, $p(w_i)$ is the probability that w_i appears in a document, $p(w_i, w_j)$ is the probability that w_i and w_j co-occur in the same document. A higher PMI score implies a more coherent topic. Following KGE-LDA, 4,776,093 Wikipedia articles are employed for obtaining topic coherence scores. Different from Yao et al. (2017), which used a fixed value of N (the number of top words, e.g. $N = 5$ or $N = 10$), we vary N in a range from 5 to 30. (Lau and Baldwin, 2016) suggests that calculating topic coherence over several different cardinalities and averaging results in a substantially more stable evaluation.

Table 1 shows the average topic coherence for different methods and datasets. We can observe that for the three datasets, TMKGE achieves highest topic coherence in almost all top word sizes. In the few cases which TMKGE does not rank highest, there only exist subtle differences with the top performing result. This shows that knowledge graph embedding improves the coherence of discovered topics. Further, for the top 10 words, the topic coherence of all three datasets are higher than those obtained by KGE-LDA. This shows that topic modeling based on HDP for both entity embedding and words enjoys incomparable advantages over LDA-based modeling.

4.3.2 Qualitative Analysis

Table 2 shows example topics with their PMI scores learned from the three corpora by KGE-LDA and our TMKGE model. For comparison, we report similar topics to those listed in the KGE-LDA paper. It can be seen that TMKGE finds quite closely related words in a topic. For example, for the second column of 20 Newsgroups, topic words from both TMKGE and KGE-LDA are related to computers. However, it can be noted that words from TMKGE focus more on the core words of computer science. In contrast, words from the same topic in KGE-LDA seems to be closer to the brand, such as windows, mac or apple. In addition, topics found from TMKGE are more diverse than those found in KGE-LDA. For 20 Newsgroups, the three topics we list here refer to theology, computer science and middle east respectively while the three topics from KGE-LDA refer to internet, computer and car respectively. Both TMKGE and KGE-LDA discover probability-related and machine learning topics with different top words from NIPS dataset. Roughly speaking, KGE-LDA discovers gene-related, cancer-related and treatment-related topics from Ohsumed corpus. TMKGE discovers more diverse and more specific topics. For example, one topic TMKGE discovers is about Vietnamese veterans, cancer-related and sexual-disease topics. From the perspective of topic coherence, we can also see that TMKGE obtains higher PMI score in most of those topics. The whole trend is consistent with the average PMI score reported in the last section. Overall, TMKGE performs better than other topic models, including LDA, HDP and KGE-LDA in terms of average PMI and also in qualitative case studies.

4.4 Document Classification

We evaluate our proposed method through document classification, we follow the approach in (Li and McCallum, 2006) for document classification.

We have conducted a five-way classification on the `comp` subject of 20 Newsgroups dataset and on the top five most frequent labels of Ohsumed dataset (no labels for nips dataset), where each class of documents is divided into 75% training and 25% testing. For each class, the LDA, HDP and TMKGE models are trained on the training documents, and then the predictive likelihood for the test documents is calculated using the E-step in the variational inference procedure of LDA. A document is classified correctly if its corresponding model produces the highest likelihood.

class	LDA	HDP	KGE-LDA	TMKGE
20 Newsgroup				
pc	68.6	78.9	67.2	**78.9**
os	71.7	80.7	70.7	**82.3**
mac	82.0	**87.1**	68.1	86.5
windows.x	84.0	83.5	64.4	**84.9**
graphics	81.2	81.9	65.4	**83.0**
Ohsumed				
C04	50.6	73.0	59.1	**73.8**
C10	46.2	63.0	54.4	**64.9**
C14	51.5	44.6	33.2	**52.3**
C21	86.5	89.5	83.7	**89.7**
C23	68.2	81.9	75.3	**86.1**

Table 3: Document classification accuracy a five-way classification on the comp subject of 20 Newsgroups dataset and on the top five most frequent labels of ohsumed dataset (no labels for NIPS dataset).

Table 3 presents the average classification accuracy for TMKGE, HDP and LDA over five repeated simulations. The table includes the classification accuracy for KGE-LDA, where the learned topic proportions are used as features for SVM classifier. For the majority of document classes, TMKGE has the best classification accuracy, except for the class `mac`. As shown, the SVM classifier based on KGE-LDA has significantly worst performance. For more complete comparisons, we run experiments on all subjects of 20 Newsgroups and also report experimental results published in Shi et al. (2017) in Table 4. TMKGE achieves the best performance on all models.

Model	Acc (%)	Model	Acc (%)
BOW	79.7	STE-Diff	82.9
Skip-Gram	75.4	LDA	77.5
TWE	81.5	TMSA	83.5
PV	75.4	HDP	82.4
GPU-DMM	48.0	KGE-LDA	70.5
STE-Same	80.4	TMKGE	**88.8**

Table 4: Document classification: all subjects of 20 Newsgroups dataset for more complete comparisons. Clearly shown is the best performances of TMKGE

A few points can be observed from the superior performance of TMKGE. Firstly, it looks the addition of unnormalized knowledge graph embedding into TMKGE as a proportional vector to the word vector boosts the performance. Secondly, the selection of HDP over LDA plays an essential role. This can be indicated from the poor performance of KGE-LDA (which is even worse than BOW). More impressively, TMKGE achieves

even much better performances than STE-Diff, TWE and TMSA, all of which involve the integration of word embedding and topic modeling. Impressively, TMKGE shows its supremacy over the state of the art model, TMSA with high margins. This shows that the knowledge graph structure included into the entity embedding conveys more information than pure word embedding. Meanwhile, this also shows that the two proportional vectors generated with online HDP enables the flexible sharing of information between words and entities. Accordingly, more coherent topics are extracted and the classification result are boosted as well.

5 Conclusion

This paper presents TMKGE, a Bayesian nonparametric model based on hierarchical Dirichlet process for incorporation of entity embeddings from external knowledge graphs into topic modeling. The proposed method allows for flexible sharing of information between documents and knowledge graph. Specifically, TMKGE avoids forcing the words and entities to identical latent factors, thus making it a suitable framework for scenarios where only partial relational information are available. Furthermore, as a Bayesian nonparameteric model, TMKGE learns the number of word topics and entity mixture components automatically from the data. We have derived an efficient and scalable online variational inference for TMKGE.

Comprehensive experiments on three different datasets suggest that TMKGE significantly outperforms SOA methods in terms of both topic coherence and document classification accuracy.

References

David Andrzejewski and Xiaojin Zhu. 2009. Latent dirichlet allocation with topic-in-set knowledge. In *SemiSupLearn '09 Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48, Boulder, Colorado.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning (ICML)*, pages 25–32, Montreal, Canada.

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Arnim Bleier. 2013. Practical collapsed stochastic variational inference for the HDP. *CoRR*, abs/1312.0412.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2787–2795, Lake Tahoe, NV.

Antoine Bordes, Jason Weston, Ronan Collobert, Yoshua Bengio, et al. 2011. Learning structured embeddings of knowledge bases. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, San Francisco, CA.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 288–296, Vancouver, Canada.

Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, and Mark Steyvers. 2008. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *Proceedings of the 7th International Semantic Web Conference (ISWC)*, pages 229–244, Karlsruhe, Germany.

Zhiyuan Chen and Bing Liu. 2014. Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1116–1125, New York, NY.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM)*, pages 209–218, San Francisco, CA.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embed-

dings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL)*, pages 795–804, Beijing, China.

Finale Doshi-Velez, Byron C Wallace, and Ryan Adams. 2015. Graph-sparse LDA: A topic model with structured sparsity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2575–2581, Austin, TX.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. 2011. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 856–864, Vancouver, Canada.

Thomas Hofmann. 2017. Probabilistic latent semantic indexing. In *ACM SIGIR Forum*, pages 211–218. ACM.

Changwei Hu, Piyush Rai, and Lawrence Carin. 2016. Topic-based embeddings for learning from large knowledge graphs. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1133–1141, Cadiz, Spain.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine learning*, 95(3):423–469.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 204–213, Avignon, France.

Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 483–487, San Diego, CA.

Dingcheng Li, Jingyuan Zhang, and Ping Li. 2019. Tmsa: A mutual learning model for topic discovery and wordembedding. In *Proceedings of the SIAM conference on Data Mining (SDM)*, Calgary, Canada.

Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. 2016. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 666–675, Berlin, Germany.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, pages 577–584, Pittsburgh, PA.

Steven N MacEachern and Peter Müller. 1998. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics (ACL): system demonstrations*, pages 55–60, Baltimore, MD.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, Lake Tahoe, NV.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272, Edinburgh, UK.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 100–108, Los Angeles, LA.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Jim Pitman. 2002. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11(5):501–514.

Masa-Aki Sato. 2001. Online model selection based on the variational bayes. *Neural computation*, 13(7):1649–1681.

Jayaram Sethuraman. 1994. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun Ping Lai. 2017. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 375–384, Shinjuku, Tokyo.

Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2005. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems (NIPS)*, pages 1385–1392, Vancouver, Canada.

Yee W Teh, Kenichi Kurihara, and Max Welling. 2008. Collapsed variational inference for HDP. In *Advances in neural information processing systems (NIPS)*, pages 1481–1488, Vancouver, Canada.

Chong Wang, John Paisley, and David Blei. 2011. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 752–760, Fort Lauderdale, FL.

Liang Yao, Yin Zhang, Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang, and Qinfei Chen. 2017. Incorporating knowledge graph embeddings into topic modeling. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 3119–3126, San Francisco, CA.

He Zhao, Lan Du, Wray L. Buntine, and Gang Liu. 2017. Metalda: A topic model that efficiently incorporates meta information. In *Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM)*, pages 635–644, New Orleans, LA.