Incorporating Knowledge Graph Embeddings into Topic Modeling

Liang Yao, Yin Zhang,* Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang, Qinfei Chen
College of Computer Science and Technology
Zhejiang University, Hangzhou, China
{yaoliang, yinzh, wbg, shrineshine, iamzhangrui, zhyy, chenqinfei}@zju.edu.cn
*Corresponding Author

Abstract

Probabilistic topic models could be used to extract low-dimension topics from document collections. However, such models without any human knowledge often produce topics that are not interpretable. In recent years, a number of knowledge-based topic models have been proposed, but they could not process fact-oriented triple knowledge in knowledge graphs. Knowledge graph embeddings, on the other hand, automatically capture relations between entities in knowledge graphs. In this paper, we propose a novel knowledge-based topic model by incorporating knowledge graph embeddings into topic modeling. By combining latent Dirichlet allocation, a widely used topic model with knowledge encoded by entity vectors, we improve the semantic coherence significantly and capture a better representation of a document in the topic space. Our evaluation results will demonstrate the effectiveness of our method.

Introduction

Probabilistic topic models such as PLSA and latent Dirichlet allocation (LDA) (Hofmann 1999; Blei, Ng, and Jordan 2003) are widely used for text modeling and analysis. However, these unsupervised models without any human knowledge often result in topics that are difficult to interpret. In other words, they could not produce semantically coherent concepts (Chang et al. 2009; Mimno et al. 2011).

To overcome the drawback of interpretability in topic model, especially in LDA, some previous works incorporate prior domain knowledge in different forms into topic model. Although these efforts incorporate knowledge in many ways, they could not process knowledge in the form of fact-oriented triples in knowledge graphs, which is the main knowledge form for machines (Wang et al. 2014c).

Recently a new research direction called knowledge graph embedding has gained much attention (Bordes et al. 2011; 2013; Wang et al. 2014b; Guo et al. 2015). It aims at embedding components of a knowledge graph like WordNet and Freebase into continuous vectors, so as to simplify the knowledge representation while preserving the inherent structure of the original knowledge graph.

In this work, we propose a new knowledge-based topic model, called *Knowledge Graph Embedding LDA* (KGE-LDA), which combines topic model and knowledge graph embeddings. The proposed method explicitly models document-level word co-occurrence in a corpus with knowledge encoded by entities' vectors automatically learned from a knowledge graph in a unified model, which could extract more coherent topics and better representation of a document in the topic space.

The contributions of the paper are threefold:

- It proposes a novel knowledge-based topic model based on multi-relational knowledge graphs.
- It provides a Gibbs sampling inference method which could handle the knowledge encoded by knowledge graph embeddings properly.
- Experimental results on three widely used datasets demonstrate that our method outperforms several state-of-the-art knowledge-based topic models and entity topic models on two tasks.

Related Work

Knowledge-based Topic Models

To overcome the drawback of interpretability in topic model, especially in LDA, some previous works incorporated prior domain knowledge into topic model in different forms.

Word Correlation Knowledge The DF-LDA (Dirichlet Forest LDA) model in (Andrzejewski, Zhu, and Craven 2009) could incorporate knowledge in the form of must-links and cannot-links input by users. A must-link states that two words should share the same topic, while a cannot-link indicates two words should not be in the same topic. (Newman, Bonilla, and Buntine 2011) put forward two Bayesian regularization methods to improve topic coherence. Both methods exploited additional word co-occurrence data to improve the interpretability of learned topics.

Lately, General Knowledge based LDA (GK-LDA) (Chen et al. 2013) was put forward. GK-LDA could use must-link knowledge from multiple domains. (Xie, Yang, and Xing 2015) incorporated word correlation into LDA by building a Markov Random Field regularization. To learn word correlation knowledge automatically, AMC (topic modeling with Automatically generated Must-links and Cannot-links) was proposed (Chen and Liu 2014). AMC could learn must-link or cannot-link knowledge automatically from multiple domains to improve topic modeling in each domain.

Word Semantic Category Knowledge (Andrzejewski and Zhu 2009) proposed topic-in-set knowledge which restricts topic assignment of words to a subset of topics. Similarly, (Chemudugunta et al. 2008) proposed Concept-topic model to assign topics of words by utilizing human-defined concepts with a hierarchical structure. (Hu, Boyd-Graber, and Satinoff 2011) presented a framework that allows users to iteratively refine the topics by adding constraints that enforce a set of words must appear together in the same topic. (Jagarlamudi, Daumé III, and Udupa 2012) proposed to guide topic modeling by setting a set of seed words that users believe could represent certain topics. Recently, (Doshi-Velez, Wallace, and Adams 2015) proposed a method for achieving interpretability by exploiting controlled structured vocabularies in which words are organized into tree-structured hierarchies.

Other Knowledge Forms (Andrzejewski et al. 2011) extended the topic-in-set knowledge by incorporating general knowledge specified by first-order logic. Lately, Probase-LDA (Yao et al. 2015), a method that combines topic model and a probabilistic knowledge base was put forward. The method can model text content with the consideration of probabilistic knowledge base for detecting better topics. (Du et al. 2015) linked LDA with constraints derived from document relative similarities. (Hu et al. 2016) firstly combined statistical topic representation with structural entity taxonomy, which provides a useful scheme to accurately induce grounded semantics.

Some recent works used word embeddings (Mikolov et al. 2013) to encode semantic regularities. (Nguyen et al. 2015) improved topic models by incorporating latent feature vector representations of words trained on very large corpora. (Das, Zaheer, and Dyer 2015) replaced LDA's parameterization of "topics" as categorical distributions over words with multivariate Gaussian distributions on the word embedding space, in which the semantic similarity is measured by Euclidean distance of word vectors. More recently, (Batmanghelich et al. 2016) proposed to use the von Mises-Fisher distribution to model the cosine distance between word vectors in a non-parametric topic model. These methods inspired us to integrate prior knowledge in the form of knowledge graph embedding into topic modeling.

Although above mentioned knowledge-based topic models utilized knowledge in many ways, they failed to handle knowledge in large scale fact-oriented triple knowledge graphs. In this work, we focus on this form of knowledge which is extensively used.

Knowledge Graph Embedding

A typical knowledge graph usually depicts knowledge as multi-relational data and represents knowledge as triple facts (*head entity*, relation, *tail entity*), which demonstrate the relation between two entities.

Knowledge graph embedding aims at embedding entities and relationships of knowledge graphs in vector spaces (Bordes et al. 2011; 2013; Wang et al. 2014b; 2014a; Guo et al. 2015; Xie et al. 2016). A knowledge graph is embedded into a low-dimensional continuous vector space while certain properties of the graph are preserved. Generally, each entity is treated as a point in the vector space and each relation is viewed as an operation over entity embeddings. For example, TransE (Bordes et al. 2013) interpreted a relation as a translation from the head entity to the tail entity. The embedding vectors are usually obtained by minimizing a global loss function regarding all entities and relations so that each entity vector captures both global and local structural patterns of the original knowledge graph. Thus, we can utilize entity embeddings to encode prior knowledge for topic modeling.

Knowledge Graph Embedding LDA

In this section, we present the KGE-LDA model, the Gibbs sampling inference and parameter learning method.

We incorporate entity embeddings into topic modeling by extending two classical entity topic models conditionally-independent LDA (CI-LDA) (Newman, Chemudugunta, and Smyth 2006) and correspondence LDA (Corr-LDA) (Blei and Jordan 2003). The two models can handle words and entities in the same topic space, but they only consider named entities recognized in text. To utilize triples in knowledge graphs, it's straightforward for us to use entity embeddings which encode knowledge graph structure instead of entities only.

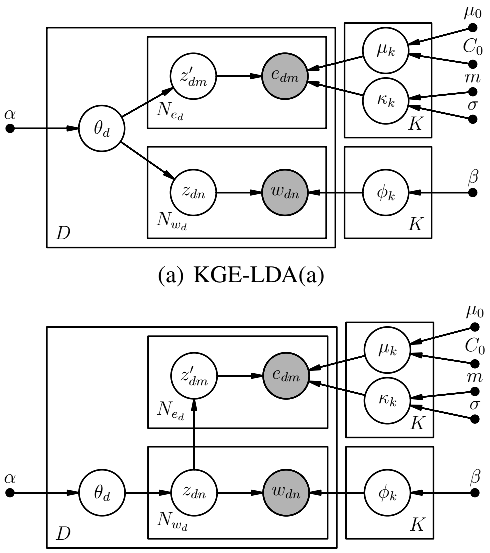
Since cosine distance is typically used to measure similarity between entity embeddings (Ji et al. 2015; He et al. 2015; Yang et al. 2015) and some knowledge graph embeddings lie on a unit sphere (ℓ^2 norm equals to 1) (Bordes et al. 2011; 2013; 2014; Yang et al. 2015; Garcia-Duran et al. 2016), we use the von Mises-Fisher (vMF) distribution (Mardia and Jupp 2009; Gopal and Yang 2014) to model them. The vMF distribution is widely used to model such directional data, which has also been employed by (Batmanghelich et al. 2016). Moreover, we found the inference is much more efficient using vMF distribution instead of multivariate Gaussian distribution in our preliminary experiment, which has also been shown in (Batmanghelich et al. 2016). vMF is a distribution that defines a probability density over points on a unit-sphere. The probability density function of the vMF distribution is

$$f(x|\mu,\kappa) = C_l(\kappa)\exp(\kappa\mu^\top x); C_l(\kappa) = \frac{\kappa^{0.5l-1}}{(2\pi)^{0.5l} I_{0.5l-1}(\kappa)} \tag{1}$$

where $x \in \mathbb{R}^l$ lies on a $l-1$ dimensional sphere, i.e., $\|x\|_2 = 1$. μ is the mean parameter with $\|\mu\|_2 = 1$ and $\kappa > 0$ is the concentration parameter, the former defines the direction of the mean and the latter determines the spread of the probability mass around the mean. $I_\nu(a)$ is the modified Bessel function of the first kind at order ν and argument a. $\mu^\top x$ is the cosine similarity between x and mean μ, κ plays the role of the inverse of variance.

Representation and Generative Process

We name our model based on CI-LDA as KGE-LDA(a) and our model based on Corr-LDA as KGE-LDA(b). The graph-



(a) KGE-LDA(a)

(b) KGE-LDA(b)

Figure 1: The graphical representation of KGE-LDA.

ical representations of KGE-LDA(a) and KGE-LDA(b) are given in Figure 1. We illustrate the mathematical notations in Table 1.

Here we introduce the details of our model. Let D be the number of documents where each document d has N_{w_d} words and N_{e_d} entities linked to existing knowledge graphs by existing entity linking tools, w_{dn} is the n-th word in d and e_{dm} is the L-dimensional entity embedding of m-th entity in d, which is obtained by TransE in this work. We choose TransE because it is simple and effective, and it achieves the state-of-the-art performance in encoding knowledge. Moreover, entity vectors of TransE naturally have unit ℓ^2 norm which need not to be post-processed. z_{dn} and z'_{dm} are latent topic assignments for w_{dn} and e_{dm} respectively. Let K be the number of topics, ϕ_k is the V-dimensional topic-word multinomial for topic $k \in 1 \dots K$, where V is the vocabulary size and θ_d is the K-dimensional document-topic multinomial for d. Since our entities are continuous vectors on a unit-sphere, we characterize each entity topic $k \in 1 \dots K$ as a vMF distribution with parameters (μ_k, κ_k). α and β are hyperparameters of the Dirichlet priors on θ_d and ϕ_k respectively. μ_0, C_0 are the hyperparameters of the prior vMF distribution of μ_k. m and σ are the mean and standard deviation of the prior logNormal distribution of κ_k.

The generative process of KGE-LDA(a) is given as:

1. For each document d draw $\theta_d \sim \text{Dir}(\alpha)$.
2. For each topic k in $1 \dots K$:
   (a) Draw $\phi_k \sim \text{Dir}(\beta)$.
   (b) Draw $\mu_k \sim \text{vMF}(\mu_0, C_0)$.
   (c) Draw $\kappa_k \sim \text{logNormal}(m, \sigma^2)$.
3. For each of the N_{w_d} words in document d:
   (a) Draw a topic $z_{dn} \sim \text{Mult}(\theta_d)$.
   (b) Draw a word $w_{dn} \sim \text{Mult}(\phi_{z_{dn}})$.

Symbol	Description
D	The number of documents.
K	The number of topics.
V	The number of unique words.
L	The dimension of entity embeddings.
N_{w_d}	The number of words in document d.
N_{e_d}	The number of entities in document d.
w_{dn}	The n-th word in document d.
e_{dm}	The embedding of the m-th entity in document d.
z_{dn}	The latent topic assignment for w_{dn}.
z'_{dm}	The latent topic assignment for e_{dm}.
ϕ_k	The topic-word multinomial for topic k.
θ_d	The document-topic multinomial for document d.
(μ_k, κ_k)	The vMF distribution parameters of entity topic k.
α	Hyperparameter of the Dirichlet prior on θ_d.
β	Hyperparameter of the Dirichlet prior on ϕ_k.
(μ_0, C_0)	Hyperparameter of the prior vMF distribution.
(m, σ)	Hyperparameter of the prior log-normal distribution.

Table 1: Mathematical notations.

4. For each of the N_{e_d} entities in document d:
   (a) Draw a topic $z'_{dm} \sim \text{Mult}(\theta_d)$.
   (b) Draw entity embedding $e_{dm} \sim \text{vMF}(\mu_{z'_{dm}}, \kappa_{z'_{dm}})$.

The generative process of KGE-LDA(b) is similar to KGE-LDA(a). The only difference is the entity embeddings are generated by topics of words in the same document:

- For each of the N_{e_d} entities in document d:
  1. Draw a topic $z'_{dm} \sim \text{Unif}(z_{d1}, \dots, z_{dN_{w_d}})$.
  2. Draw entity embedding $e_{dm} \sim \text{vMF}(\mu_{z'_{dm}}, \kappa_{z'_{dm}})$.

Inference and Parameter Learning

We use Gibbs sampling to infer latent topic assignments z_{dn} and z'_{dm}. The Gibbs sampling equation for z_{dn} in both two proposed models is defined as:

$$p(z_{dn} = k | w_{dn}, \mathbf{w}_{-dn}, \mathbf{z}_{-dn}, \mathbf{z}', \alpha, \beta) \propto \frac{n_{dk} + \alpha}{N_{w_d} + N_{e_d} + K\alpha} \times \frac{n_{kw_{dn}} + \beta}{n_k + V\beta} \qquad (2)$$

where k is a topic, $\mathbf{w}_{-dn}$ are all words except w_{dn}, $\mathbf{z}_{-dn}$ are topic assignments for all words except w_{dn}, $\mathbf{z}'$ are topic assignments for all entities, n_{dk} is the number of times topic k is assigned to a word or an entity in document d, $n_{kw_{dn}}$ is the number of times w_{dn} is assigned to topic k and n_k is the number of times any word is assigned to topic k.

The Gibbs sampling equations for z'_{dm} are similar to Gibbs sampling equations in (Gopal and Yang 2014). We can develop efficient sampling techniques by using the fact that vMF distributions are conjugate. This enables us to completely integrate out μ_k and update the model only by maintaining the topic assignment variable z'_{dm} and the concentration parameters κ_k. The inference equations for KGE-LDA(a) are given by:

$$p(z'_{dm} = k | e_{dm}, \mathbf{z'}_{-dm}, \mathbf{e}_{-dm}, \mathbf{z}, \kappa_k, \alpha, \mu_0, C_0, m, \sigma) \propto \frac{n_{dk} + \alpha}{N_{w_d} + N_{e_d} + K\alpha} \times C_L(\kappa_k) \times \frac{C_L(\|\kappa_k \sum_{j \neq dm, z'_j = k} e_j + C_0\mu_0\|)}{C_L(\|\kappa_k \sum_{j:z'_j = k} e_j + C_0\mu_0\|)} \tag{3}$$

$$p(\kappa_k | \kappa_{-k}, ..) \propto \frac{C_L(C_0) C_L(\kappa_k)^{n'_k}}{C_L(\|\kappa_k \sum_{j:z'_j = k} e_j + C_0\mu_0\|)} \times \text{logNormal}(\kappa_k | m, \sigma^2) \tag{4}$$

where $\mathbf{z'}_{-dm}$ are topic assignments for all entities except e_{dm}, $\mathbf{e}_{-dm}$ are embeddings of all entities except e_{dm}, $\mathbf{z}$ are topic assignments for all words, n'_k is the number of times any entity is assigned to topic k. $C_L(\kappa_k)$ has the same meaning as $C_l(\kappa)$ in equation (1) if we replace L and κ_k with l and κ. Since κ_k is drawn from the logNormal distribution, we first sample some κ_k samples (100 samples in our experiment, we also try other numbers, but don't find much difference) from logNormal$(\kappa_k | m, \sigma^2)$, then sample the final κ_k from these samples using equation (4).

The inference equations for KGE-LDA(b) are very similar to KGE-LDA(a), the only distinction is:

$$p(z'_{dm} = k | e_{dm}, \mathbf{z'}_{-dm}, \mathbf{e}_{-dm}, \mathbf{z}, \kappa_k, \alpha, \mu_0, C_0, m, \sigma) \propto \frac{n_{dk}}{N_{w_d}} \times C_L(\kappa_k) \times \frac{C_L(\|\kappa_k \sum_{j \neq dm, z'_j = k} e_j + C_0\mu_0\|)}{C_L(\|\kappa_k \sum_{j:z'_j = k} e_j + C_0\mu_0\|)} \tag{5}$$

With Gibbs sampling, we can estimate θ_d and ϕ_k using the two factors in equation (2).

$$\theta_{dk} = \frac{n_{dk} + \alpha}{N_{w_d} + N_{e_d} + K\alpha} \tag{6}$$

$$\phi_{kw_{dn}} = \frac{n_{kw_{dn}} + \beta}{n_k + V\beta} \tag{7}$$

Experiment

In this section we evaluate our Knowledge Graph Embedding LDA on two experimental tasks. Specifically we wish to determine:

- Can our model find coherent and meaningful topics?
- Can our model learn better topic distribution for document classification?

Baselines. We compare our KGE-LDA with six state-of-the-art topic models:

- LDA (Blei, Ng, and Jordan 2003), the most widely used topic model.
- Corr-LDA (Blei and Jordan 2003), a classical entity topic model.
- CI-LDA (Newman, Chemudugunta, and Smyth 2006), a classical entity topic model.
- Concept-topic model (CTM) (Chemudugunta et al. 2008), a knowledge-based topic model that can exploit word semantic category knowledge.
- GK-LDA (Chen et al. 2013), a knowledge-based topic model that can process must-link knowledge. It can automatically deal with wrong knowledge[1].
- LF-LDA (Nguyen et al. 2015), a knowledge-based topic model using word embeddings trained from large external data[2].

Datasets. We run our experiments[3] on three widely used datasets 20-Newsgroups (20NG), NIPS and the Ohsumed corpus. The 20NG dataset[4] ("bydate" version) contains 18,846 documents evenly categorized into 20 different categories. 11,314 documents are in the training set and 7,532 documents are in the test set. The NIPS dataset[5] contains 1,740 papers from the NIPS conference. The Ohsumed corpus is from the MEDLINE database, which is a bibliographic database of important medical literature maintained by the National Library of Medicine. In this study, we consider the 13,929 unique Cardiovascular diseases abstracts in the first 20,000 abstracts of the year 1991[6]. Each document in the set has one or more associated categories from the 23 disease categories. As we focus on single-label text classification, the documents belonging to multiple categories are eliminated so that 7,400 documents belonging to only one category remain. 3,357 documents are in the training set and 4,043 documents are in the test set.

The datasets are tokenized with Stanford CoreNLP. After standard pre-processing of removing stop words, low frequency words (appearing less than 10 times) and words do not appear in pre-trained word embeddings (see next paragraph), there are 20,881 distinct words in the 20NG dataset, 14,482 distinct words in the NIPS dataset and 8,446 distinct words in the Ohsumed dataset.

External Knowledge. 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 are conceptual-semantic and lexical relations. In this work, we use a subset of WordNet (WN18) introduced in (Bordes et al. 2013) [7]. WN18 contains 151,442 triplets with 40,943 entities and 18 relations. We link tokenized words to entities in WN18 via NLTK[8]. For CI-LDA and Corr-LDA, the linked entities are the external knowledge. For KGE-LDA, we pretrain 50-dimensional entity vectors using TransE, we find experimental results are not sensitive to dimensions of entity vectors. For CTM, we treat the linked entities in WN18 as concepts of words (in vocabulary). For GK-LDA, we view

[1] https://github.com/czyuan/GKLDA

[2] https://github.com/datquocnguyen/LFTM

[3] We released the implementation of this paper at the first author's GitHub: https://github.com/yao8839836/KGE-LDA.

[4] http://qwone.com/~jason/20Newsgroups/

[5] http://www.cs.nyu.edu/~roweis/data.html

[6] http://disi.unitn.it/moschitti/corpora.htm

[7] available at https://everest.hds.utc.fr/doku.php?id=en:transe

[8] http://www.nltk.org/howto/wsd.html

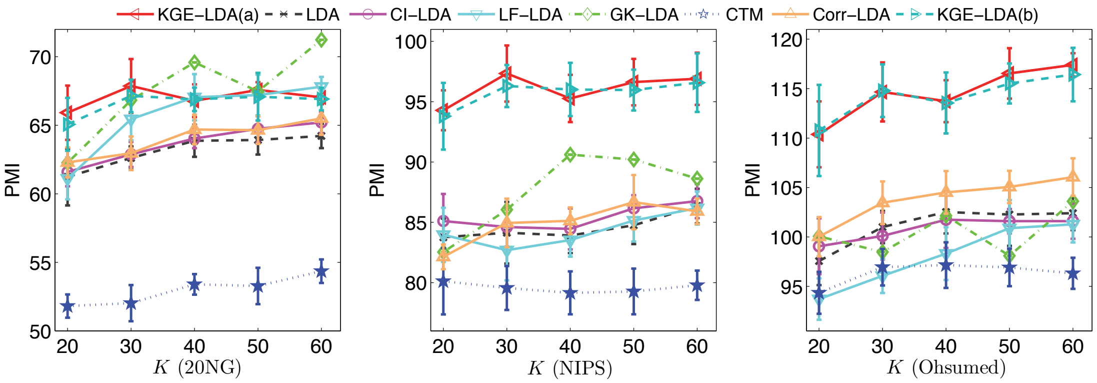


Figure 2: Average PMI Topic Coherence of all models on three datasets with different number of topics K. We run all models 10 times and report the mean ± standard deviation. A higher PMI score implies a more coherent topic. Improvements of KGE-LDA(a) and KGE-LDA(b) over LDA, CI-LDA and Corr-LDA are significant ($p < 0.01$) based on student t-test.

words (in vocabulary) linking to the same entity as must-links. For LF-LDA, we pre-train 50-dimensional Skip-Gram word2vec (Mikolov et al. 2013) word vectors using a full snapshot of the English-language edition of Wikipedia in Feb 2015 [9]. The corpus has 4,776,093 articles.

Settings. For all the methods in comparison, we set the hyperparameters as $\alpha = 50/K$, $\beta = 0.01$, a commonly used setting which has often been employed in prior work (Steyvers and Griffiths 2007). For KGE-LDA, we initialize each dimension of μ_0 with a Gaussian distribution $N(0, 1)$ and then normalize μ_0 into a unit norm vector, we set $C_0 = 0.01$, $m = 0.01$, $\sigma = 0.25$, we also experiment with other settings of these priors but do not find much difference. We set other parameters as the recommended settings in baseline papers, i.e., entity (concept) topic hyperparameter $\beta' = 0.01$ for CI-LDA, Corr-LDA and CTM, $\lambda = 0.6$ for LF-LDA and $\lambda = 2000, \varepsilon = 0.07$ for GK-LDA. All models are trained using 1000 Gibbs sampling iterations. The only exception is that 1200 iterations (1000 initial iterations with LDA model + 200 iterations with LF-LDA) are run for LF-LDA. This setting can lead to a fair comparison with LDA because the differences are in the 200 LF-LDA iterations.

Topic Coherence

Quantitative Analysis. We evaluate topics produced by each model based on point-wise mutual information (PMI) Topic Coherence (Newman et al. 2010). Typically topic models are evaluated based on perplexity. Unfortunately, perplexity on the held-out test set does not reflect the interpretability of topics and may be contrary to human judgments (Chang et al. 2009). Alternatively, the Topic Coherence metric has been shown to correlate well with human judging (Lau, Newman, and Baldwin 2014). Since our goal is to discover coherent or meaningful topics, Topic Coherence is more suitable for our evaluation. The PMI Topic Coherence of a topic k is defined as:

$$PMI(k) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{8}$$

where N is the number of top words of k, $p(w_i)$ is the probability that word w_i appears in a document, $p(w_i, w_j)$ is the probability that word w_i and w_j co-occur in the same document. A higher PMI score implies a more coherent topic. We compute the average Topic Coherence of top $N = 10$ words of each topic. We use top 10 words because they show the most important semantic information of a topic, and many papers using PMI (including the original work by (Newman et al. 2010)) consider top 10 words. In order to compute PMI, we need a large external corpus. In our experiments, we use the 4,776,093 Wikipedia articles.

Figure 2 plots average PMI Topic Coherence of all models on three datasets. We can observe that CI-LDA and Corr-LDA perform slightly better than LDA, which means using entities in text can improve topic interpretability, but the improvements are not significant. When using entity embeddings instead of entities, the topic coherence scores are significantly improved because entity embeddings encode more information about knowledge graph structure. CTM does not perform well, this maybe because the number of unique concepts (several thousand in our experiments) is far more than the number of topics, which skews the original topic space. GK-LDA can produce some very good results on 20NG and NIPS, which shows its ability to utilize must-link knowledge. However, the results on Ohsumed are not satisfactory, for the generated must-links using WN18 may not help topic modeling in medical domain. LF-LDA performs well on 20NG but not on NIPS and Ohsumed, which means word vectors learned from Wikipedia can help topic modeling in general domain, but may not be useful for specific domains such as machine learning and medicine.

Qualitative Analysis. Table 2 shows some example topics with their PMI scores learned from the three corpora by LDA and our KGE-LDA(a) model. We try to find the best possible matches from the topics. We can note that

[9] http://deepdive.stanford.edu/opendata/

20NG			NIPS			Ohsumed		
LDA								
information	car	card	model	control	learning	cancer	gene	treatment
list	cars	video	distribution	system	examples	tumor	dna	therapy
group	buy	apple	data	motor	generalization	tumors	analysis	dose
send	engine	monitor	probability	model	training	carcinoma	protein	effects
mail	article	memory	gaussian	position	error	breast	normal	drug
address	writes	mac	models	trajectory	set	cases	region	days
posting	oil	ram	parameters	controller	space	primary	genetic	study
questions	subject	speed	mixture	robot	algorithm	malignant	found	day
book	organization	organization	bayesian	figure	vector	lesions	mrna	effective
internet	dealer	drivers	likelihood	learning	support	local	mutation	placebo
67.181	57.443	58.594	130.151	72.418	76.053	136.831	88.789	97.110
KGE-LDA(a)								
internet	car	drive	distribution	control	kernel	cancer	gene	treatment
mail	cars	windows	bayesian	trajectory	support	tumor	dna	therapy
email	engine	dos	gaussian	robot	xi	survival	protein	dose
list	oil	card	prior	controller	vector	tumors	region	drug
send	miles	disk	posterior	arm	margin	carcinoma	genetic	effects
e-mail	dealer	mac	probability	model	examples	breast	analysis	placebo
information	speed	scsi	variables	forward	set	stage	mutation	trial
address	buy	memory	markov	motor	kernels	malignant	sequence	oral
fax	ford	system	distributions	trajectories	svm	chemotherapy	molecular	mg
network	drive	apple	approximation	inverse	machines	primary	mrna	effective
89.216	63.679	84.378	154.842	86.199	88.283	149.913	107.788	106.555

Table 2: Example topics learned from three datasets by LDA and our KGE-LDA(a) model with $K = 30$. The last row for each model is the topic coherence (PMI) computed using the 4,776,093 Wikipedia documents as reference.

KGE-LDA(a) finds more closely related words in a topic. For 20NG, KGE-LDA(a) finds "e-mail" and "network" in the first email topic which are not discovered by LDA, and "questions" and "book" in LDA topic are not related words. In the second car topic, the closely related words "speed", "miles", "ford" and "drive" are in KGE-LDA(a) topic, and noisy words "article", "writes", "subject" and "organization" are in LDA topic. In the third computer topic, LDA still shows a noisy word "organization" while KGE-LDA(a) words are all related to the topic. The results are similar for NIPS and Ohsumed. For NIPS, KGE-LDA(a) presents more related words in the Bayesian topic (e.g., "prior" and "posterior"), the robotics topic (e.g., "arm" and "trajectories") and the Support Vector Machines topic (e.g., "kernel" and "svm"). For Ohsumed, more words in KGE-LDA(a) are related to the cancer topic (e.g., "chemotherapy" and "survival"), the gene topic (e.g., "sequence" and "molecular") and the drug topic (e.g., "oral" and "mg"). The observations are consistent with quantitative results.

Document Classification Evaluation

We perform document classification with the learned θ_d as the feature vector of document d, and employ a linear kernel Support Vector Machines (SVM) classifier LIBLINEAR (Fan et al. 2008).

Table 3 gives the classification accuracy on the two labeled datasets. We can see that Corr-LDA and CI-LDA perform similarly to LDA, which means using entities only could not help to distinguish documents. This is probably because entities in WordNet usually have the same name with the linked words, which may not provide additional information for the task. The best variation of KGE-LDA significantly outperforms LDA, which shows using entity embeddings can also lead to a more distinguishable document representation. CTM could not produce good results as in the topic coherence evaluation, because the skewed topic vectors are difficult to be classified. GK-LDA and LF-LDA also perform similarly to LDA, which shows must-links and pre-trained word vectors may not be helpful for document topic representation. The only exception is GK-LDA on Ohsumed with K = 20 and 25, the highest accuracies maybe because the number of LR-sets (must-links) for Ohsumed is small, so when K is small, the statistics for topic–LR-set–word and the word correlation computing are more sufficient in GK-LDA.

Conclusion

This paper presents KGE-LDA, which combines topic model and knowledge graph embeddings, in particular LDA model and TransE. The proposed method models document-level word co-occurrence with knowledge encoded by entity vectors automatically learned from external knowledge graphs, could extract more coherent topics and better topic

Dataset	Model	$K = 20$	$K = 25$	$K = 30$	$K = 35$	$K = 40$	$K = 45$
20NG	LDA	0.556 ± 0.021	0.618 ± 0.012	0.646 ± 0.017	0.659 ± 0.019	0.681 ± 0.019	0.693 ± 0.012
	Corr-LDA	0.547 ± 0.018	0.604 ± 0.031	0.652 ± 0.024	0.671 ± 0.032	0.678 ± 0.020	0.684 ± 0.020
	CI-LDA	0.556 ± 0.022	0.598 ± 0.015	0.628 ± 0.022	0.655 ± 0.022	0.672 ± 0.017	0.689 ± 0.019
	CTM	0.193 ± 0.009	0.225 ± 0.010	0.249 ± 0.008	0.280 ± 0.008	0.295 ± 0.009	0.312 ± 0.006
	GK-LDA	0.557 ± 0.000	0.620 ± 0.000	0.634 ± 0.000	0.641 ± 0.000	0.679 ± 0.000	0.678 ± 0.000
	LF-LDA	0.541 ± 0.023	0.583 ± 0.018	0.632 ± 0.019	0.661 ± 0.019	0.662 ± 0.021	0.673 ± 0.023
	KGE-LDA(a)	**0.586** ± 0.023	**0.638** ± 0.024	0.666 ± 0.028	**0.692** ± 0.025	**0.699** ± 0.018	**0.705** ± 0.011
	KGE-LDA(b)	0.571 ± 0.025	0.634 ± 0.027	**0.670** ± 0.019	0.669 ± 0.017	0.688 ± 0.016	0.695 ± 0.015
Ohsumed	LDA	0.401 ± 0.012	0.427 ± 0.015	0.448 ± 0.008	0.452 ± 0.012	0.468 ± 0.009	0.476 ± 0.013
	Corr-LDA	0.408 ± 0.010	0.440 ± 0.018	0.453 ± 0.009	0.462 ± 0.011	0.471 ± 0.010	0.486 ± 0.008
	CI-LDA	0.411 ± 0.015	0.429 ± 0.015	0.440 ± 0.013	0.451 ± 0.013	0.458 ± 0.011	0.475 ± 0.009
	CTM	0.240 ± 0.017	0.244 ± 0.010	0.247 ± 0.007	0.252 ± 0.005	0.260 ± 0.015	0.268 ± 0.009
	GK-LDA	**0.425** ± 0.000	**0.446** ± 0.000	0.425 ± 0.000	0.463 ± 0.000	0.459 ± 0.000	0.457 ± 0.000
	LF-LDA	0.397 ± 0.009	0.424 ± 0.012	0.445 ± 0.011	0.454 ± 0.008	0.462 ± 0.016	0.479 ± 0.010
	KGE-LDA(a)	0.418 ± 0.008	0.444 ± 0.020	**0.455** ± 0.010	**0.470** ± 0.008	0.476 ± 0.008	**0.490** ± 0.014
	KGE-LDA(b)	0.418 ± 0.012	0.443 ± 0.009	0.455 ± 0.009	0.465 ± 0.011	**0.477** ± 0.011	0.490 ± 0.009

Table 3: Classification accuracy of all models on two labeled datasets with different number of topics K. We run all models 10 times and report the mean ± standard deviation. Improvements of the best variation of KGE-LDA over LDA are significant ($p < 0.05$) based on student t-test. The best results are in bold font.

representation. Experimental results on three datasets show the effectiveness of the proposed method. We plan to explore more effective ways to incorporate entity embeddings and experiment with more knowledge graphs in future work.

Acknowledgments

This work is supported by China Knowledge Centre for Engineering Sciences and Technology (No. CKCEST-2016-1-14), the National Natural Science Foundation of China (No. 61572434) and Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) (20130101110136). We thank the anonymous reviewer of IJCAI 2015.

References

Andrzejewski, D., and Zhu, X. 2009. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, 43–48.

Andrzejewski, D.; Zhu, X.; Craven, M.; and Recht, B. 2011. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *IJCAI*, volume 22, 1171.

Andrzejewski, D.; Zhu, X.; and Craven, M. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, 25–32. ACM.

Batmanghelich, K.; Saeedi, A.; Narasimhan, K.; and Gershman, S. 2016. Nonparametric spherical topic modeling with word embeddings. In *ACL*, 537–542.

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *SIGIR*, 127–134. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bordes, A.; Weston, J.; Collobert, R.; and Bengio, Y. 2011. Learning structured embeddings of knowledge bases. In *AAAI*.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, 2787–2795.

Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y. 2014. A semantic matching energy function for learning with multi-relational data. *Machine Learning* 94(2):233–259.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-graber, J. L.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*, 288–296.

Chemudugunta, C.; Holloway, A.; Smyth, P.; and Steyvers, M. 2008. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *International Semantic Web Conference*, 229–244. Springer.

Chen, Z., and Liu, B. 2014. Mining topics in documents: standing on the shoulders of big data. In *KDD*, 1116–1125.

Chen, Z.; Mukherjee, A.; Liu, B.; Hsu, M.; Castellanos, M.; and Ghosh, R. 2013. Discovering coherent topics using general knowledge. In *CIKM*, 209–218.

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

Doshi-Velez, F.; Wallace, B. C.; and Adams, R. 2015. Graph-sparse lda: A topic model with structured sparsity. In *AAAI*, 2575–2581.

Du, J.; Jiang, J.; Song, D.; and Liao, L. 2015. Topic modeling with document relative similarities. In *IJCAI*.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research* 9(Aug):1871–1874.

Garcia-Duran, A.; Bordes, A.; Usunier, N.; and Grandvalet, Y. 2016. Combining two and three-way embedding models for link prediction in knowledge bases. *Journal of Artificial Intelligence Research* 55:715–742.

Gopal, S., and Yang, Y. 2014. Von mises-fisher clustering models. In *ICML*, 154–162.

Guo, S.; Wang, Q.; Wang, B.; Wang, L.; and Guo, L. 2015. Semantically smooth knowledge graph embedding. In *ACL*, 84–94.

He, S.; Liu, K.; Ji, G.; and Zhao, J. 2015. Learning to represent knowledge graphs with gaussian embedding. In *CIKM*, 623–632.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57. ACM.

Hu, Z.; Luo, G.; Sachan, M.; Xing, E.; and Nie, Z. 2016. Grounding topic models with knowledge bases. In *IJCAI*.

Hu, Y.; Boyd-Graber, J.; and Satinoff, B. 2011. Interactive topic modeling. In *ACL*, 248–257. Association for Computational Linguistics.

Jagarlamudi, J.; Daumé III, H.; and Udupa, R. 2012. Incorporating lexical priors into topic models. In *EACL*, 204–213.

Ji, G.; He, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Knowledge graph embedding via dynamic mapping matrix. In *ACL*, 687–696.

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

Mardia, K. V., and Jupp, P. E. 2009. *Directional statistics*, volume 494. John Wiley & Sons.

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

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

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *EMNLP*, 262–272.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *NAACL*.

Newman, D.; Bonilla, E. V.; and Buntine, W. 2011. Improving topic coherence with regularized topic models. In *NIPS*, 496–504.

Newman, D.; Chemudugunta, C.; and Smyth, P. 2006. Statistical entity-topic models. In *KDD*, 680–686.

Nguyen, D. Q.; Billingsley, R.; Du, L.; and Johnson, M. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics* 3:299–313.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. *Handbook of latent semantic analysis* 427(7):424–440.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014a. Knowledge graph and text jointly embedding. In *EMNLP*, 1591–1601.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014b. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 1112–1119.

Wang, Z.; Wang, H.; Xiao, Y.; and Wen, J.-R. 2014c. How to make a semantic network probabilistic. *TechReport. MSR-TR-2014-59*.

Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M. 2016. Representation learning of knowledge graphs with entity descriptions. In *AAAI*.

Xie, P.; Yang, D.; and Xing, E. 2015. Incorporating word correlation knowledge into topic modeling. In *NAACL*.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

Yao, L.; Zhang, Y.; Wei, B.; Qian, H.; and Wang, Y. 2015. Incorporating probabilistic knowledge into topic models. In *PAKDD*, 586–597. Springer.