Enhancing Topic Models by Incorporating Explicit and Implicit External Knowledge

Yang Hong YANG-HONG@ALUMNI.SJTU.EDU.CN
Xinhuai Tang ✉ TANG-XH@CS.SJTU.EDU.CN
Tiancheng Tang TANGTIAN_CHENG@SJTU.EDU.CN
Yunlong Hu HUYUNLONGSJTU@SJTU.EDU.CN
School of Software, Shanghai Jiao Tong University
Jintai Tian JINTAI@MIDUCHINA.COM
Institute of Sina WRD Big Data

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

Topic models are widely used for extracting latent features from documents. Conventional count-based models like LDA focus on co-occurrence of words, neglecting features like semantics and lexical relations in the corpora. To overcome this drawback, many knowledge-enhanced models are proposed, attempting to achieve better topic coherence with external knowledge. In this paper, we present novel probabilistic topic models utilizing both explicit and implicit knowledge forms. Knowledge of real-world entities in a knowledge base/graph are referred to as explicit knowledge. We incorporate this knowledge form into our models by entity linking, a technique for bridging the gap between corpora and knowledge bases. This helps solving the problem of token/phrase-level synonymy and polysemy. Apart from explicit knowledge, we utilize latent feature word representations (implicit knowledge) to further capture lexical relations in pretraining corpora. Qualitative and Quantitative evaluations are conducted on 2 datasets with 5 baselines (3 probabilistic models and 2 neural models). Our models exhibit high potential in generating coherent topics. Remarkably, when adopting both explicit and implicit knowledge, our proposed model even outperforms 2 state-of-the-art neural topic models, suggesting that knowledge-enhancement can highly improve the performance of conventional topic models.

Keywords: knowledge-enhanced topic models, entity linking, latent feature word representations

1. Introduction

Probabilistic topic modeling algorithms like probabilistic latent semantic allocation (pLSA) (Hofmann (1999)) and latent Dirichlet allocation (LDA) (Blei et al. (2003)) extract latent topic features from a corpus, which are widely used in topic modeling and other text analysis and mining tasks.

Count-based topic models like LDA treat documents as collections of tokens and generate tokens independently according to a set of multinomial distributions. In this generating process, only word co-occurrence is taken into consideration, leaving problems of synonymy and polysemy unsolved.

Blending various forms of external knowledge into existing models provides a potential solution for these problems. Some researchers attempt to use explicit domain knowledge in the knowledge bases. Two representative knowledge bases are Wikipedia[1] and WordNet(Miller (1995)). Models using Wikipedia(Yao et al. (2016)) focus on detecting and recognizing named entities in a corpus while those using WordNet(Chen et al. (2013a); Yao et al. (2017)) emphasize on word correlations. Apart from explicit knowledge, some research(Nguyen et al. (2015); Xie et al. (2015)) uses implicit knowledge which is embedded in structures like latent feature word representations.

A majority of previous research adopts only one knowledge form. However, since each knowledge form has its emphasis, none of them can be powerful enough to cover all knowledge requirements alone. Rather than relying merely on explicit domain knowledge or latent features, our models take advantage of both of them. In specific, we use entity linking to retrieve knowledge of real-world entities and pretrained word vectors to capture semantic and lexical relations of words.

(1) Explicit knowledge acquirement by entity linking. Different from previous research, our method acquires domain knowledge by entity linking. It is a technique that links word mentions to certain knowledge base entities. Entity linking brings advantage to topic modeling for 2 reasons. First, it recognizes real-world entities and retrieves external knowledge with high accuracy. Second, as an independent NLP task, this technique decouples topic modeling and domain knowledge acquirement. This feature provides us a more flexible scheme relying on which we can customize the selection of topic models and domain knowledge retrieving systems. As will be shown in the following parts of this paper, an existing model may not need too many alterations to incorporate this technique. In contract, models taking other approaches may suffer from highly complex structures and difficulties in inference.

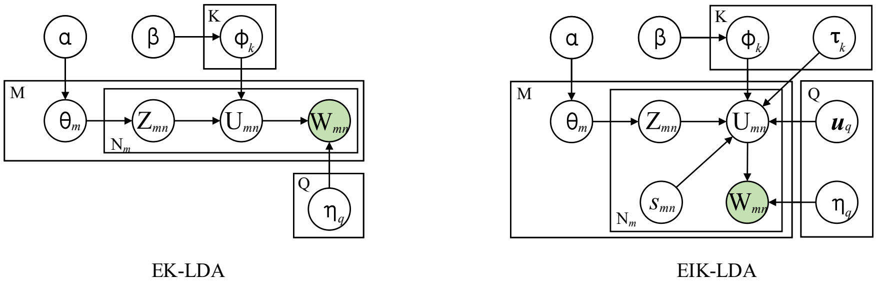
(2) Implicit knowledge acquirement by latent feature word vectors. Though powerful, entity linking is not perfect, for it cannot acquire external knowledge of tokens like verbs, adjectives or adverbs, which may encode valuable lexical or semantic information under certain conditions. To overcome this drawback, we use latent feature word representations carrying implicit knowledge to further improve the performance of our model.

This paper contributes on the following 3 aspects.

- We present 2 knowledge-enhanced LDA variants, EK-LDA (**E**xplicit External **K**nowledge Enhanced **LDA**) and EIK-LDA (**E**xplicit and **I**mplicit External **K**nowledge Enhanced **LDA**). EK-LDA uses entity linking to acquire external knowledge. Based on EK-LDA, EIK-LDA further utilizes implicit knowledge provided by pretrained word vectors.
- We provide Gibbs-sampling-based inference methods for the two models.
- We conduct evaluations on 20-Newsgroups and NIPS dataset with 5 baselines. Experimental results reveal that our models achieve good results on topic coherence and text classification.

The rest of our paper is organized as follows. The proposed models, their inference algorithms and approaches for incorporating external knowledge into our models are elaborated

[1] https://www.wikipedia.org/



EK-LDA

EIK-LDA

Figure 1: Graphical models of EK-LDA and EIK-LDA

Table 1: Mathematical notations in the models

Model	Symbol	Description
EK-LDA and EIK-LDA	M	Document number.
	N$_m$	Token number in document m.
	K	Topic number.
	Q	Semantic unit number.
	$\mathbf{Z}$	Matrix of topic labels for tokens in documents.
	$\mathbf{U}$	Matrix of semantic units.
	$\mathbf{W}$	Matrix of tokens (words).
	η	Word distribution over semantic units.
	θ	Topic distribution over documents.
	ϕ	Semantic unit distribution over topics.
	α	Dirichlet prior on distribution θ.
	β	Dirichlet prior on distribution ϕ.
EIK-LDA	τ	Latent feature weight for topics.
	$\boldsymbol{u}$	Latent feature representations for semantic units.
	s	0-1 indicator for choosing multinomial distribution or latent feature word vectors to generate semantic units.

in §2. Related work of knowledge-enhanced topic models and entity linking are introduced in §3. Experimental results and their analysis are covered in §4. A conclusion is made in §5.

2. Methodology

In this section, we elaborate the design and implementation of two novel knowledge enhanced topic models. We call them EK-LDA (**E**xplicit External **K**nowledge Enhanced **LDA**) and EIK-LDA (**E**xplicit and **I**mplicit External **K**nowledge Enhanced **LDA**). Their graphical models are illustrated in Figure 1. Mathematical notations in this figure are defined in Table 1.

Both the two models are external-knowledge-enhanced. Instead of generating words ($\mathbf{W}_{mn}$ in Figure 1) directly from topic labels ($\mathbf{Z}_{mn}$ in Figure 1) like LDA, our models firstly generate semantic units (denoted by $\mathbf{U}_{mn}$) from topic labels, and then generate tokens from semantic units. Semantic units represent tokens or phrases with similar or same meanings. They serve as an interface through which topic models interact with knowledge retrieving systems. With semantic units, topic models can support various knowledge sources and forms, which decouples specific topic modeling algorithms and knowledge retrieving strategies.

In the rest of this section, we will introduce the generating process and inference algorithm of the two models, and we will also describe in detail how we incorporate both explicit and implicit knowledge into the models with entity linking and latent feature word representations.

2.1. EK-LDA and Its Generative Process

As is illustrated in Figure 1, we design EK-LDA to be a more general form of LDA. The proof is as follows.

Suppose each semantic unit maps exactly to a unique word in the vocabulary, then semantic unit size Q will be equal to vocabulary size V. This one-one mapping between semantic units and tokens leads to the fact that *any* semantic unit will generate a specific word with probability 1. In this condition, the model is reduced to LDA. With this case, we prove LDA is an instance of EK-LDA.

EK-LDA generates a document according to the following steps. For each document m , draw a topic distribution θ_m. Then generate a topic label Z_{mn} for each token in this document from distribution θ_m. For each topic k, draw a multinomial distribution ϕ_k of semantic units. Generate semantic units $\mathbf{U}_{mn}$ with topic labels $\mathbf{Z}_{mn}$ and distribution ϕ_k. Draw a word from distribution $\eta_{\mathbf{U}_{mn}} \in \mathrm{R}^{\mathrm{V}}$. $\eta_{\mathbf{U}_{mn}}$, the distribution of semantic units over tokens are defined by entity linking, which will be explained in §2.2. The generating process can be represented as:

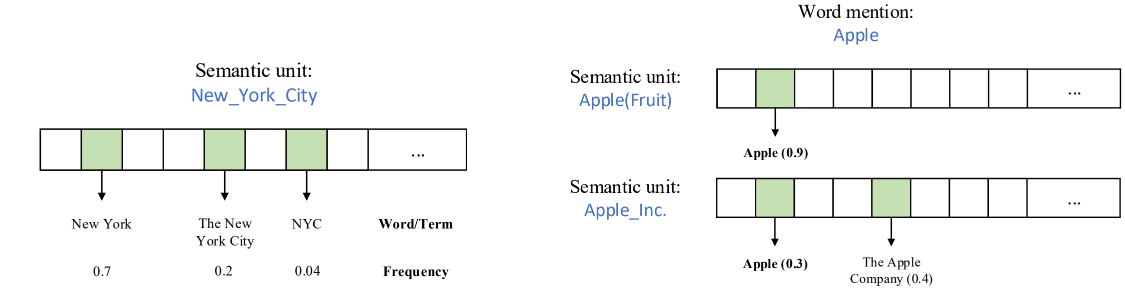
$$
\begin{array}{ll}
\theta_m \sim \mathrm{Dir}(\alpha) & \mathbf{Z}_{mn} \sim \mathrm{Mult}\,(\theta_m) \\
\phi_{\mathbf{Z}_{mm}} \sim \mathrm{Dir}(\beta) & \mathbf{U}_{mn} \sim \mathrm{Mult}\,(\phi_{\mathbf{Z}_{mn}}) \\
\mathbf{W}_{mn} \sim \mathrm{Mult}\,(\eta_{\mathbf{U}_{mn}}) &
\end{array}
$$

2.2. Explicit Knowledge Incorporation for EK-LDA

In this section, we describe in detail how we incorporate explicit knowledge into EK-LDA and how EK-LDA uses this knowledge. Explicit knowledge incorporation is supported by semantic units and entity linking.

Semantic units $\mathbf{U}_{mn}$ make it possible for our model to handle synonymy and polysemy of words and phrases. As is illustrated in Figure 2(a), words or phrases with the same meaning (*New York*, *the New York City*, etc.) are allocated to one specific semantic unit. For words or phrases with more than one meaning, EK-LDA defines semantic units for each of their unique meanings. As is illustrated in Figure 2(b), EK-LDA generates 2 semantic units for word mention *Apple* corresponding to its possible meanings (a kind of fruit and a name for a company).

Semantic units provide a structure with which our model has the potential of utilizing explicit knowledge. Yet it is entity linking that decides the value, or more precisely, the probability distribution of each semantic unit. In this paper, we adopt the entity linking system presented by (Gupta et al. (2017)) which links word mentions to Wikipedia entries according to their meanings, for we are especially interested in knowledge of real-world entities. Since in entity linking, potential word mentions are linked to knowledge base



(a) several words to one semantic unit (b) one word to several semantic units

Figure 2: 2 cases for adding words to specific semantic units

entities, semantic units are naturally derived in this process. For words with no entity to link, they are semantic units themselves. The model also records the frequencies that each word appears in each semantic unit. Denoted by η_{qv} (q for semantic unit ID and v for word ID), this frequency suggests the probability of word v having meaning q, which is vital for our model to generate words from semantic units.

The above description reveals a two-step approach that EK-LDA takes to conduct topic modeling. First, the entity linking system recognizes potential entities. Through this process the explicit knowledge is encoded in the semantic units and their corresponding word frequencies. And then, the probabilistic model is trained to optimize its parameters with the encoded knowledge and the original corpus.

What is noteworthy is that we choose this simple and intuitive two-step approach for several reasons. Compared to models that couple knowledge acquirement with topic modeling (we refer to them as “tight-coupled models”), EK-LDA has less parameters and can be inferenced with less effort. This is especially a good news for latent probabilistic models. Another advantage this strategy brings about is that EK-LDA is open to the diversity of entity linking systems. Although in this paper, we adopt the system proposed by (Gupta et al. (2017)) which is an instance of neural entity linking models, EK-LDA has the potential of integrating various entity linking systems and accessing various knowledge sources. Last but not least, this strategy is intuitive yet effective. Empirical results in the following sections reveal our model is as much effective as “tight-coupled models” and even outperforms several neural topic models.

2.3. Inference for EK-LDA

We develop a collapsed Gibbs sampling algorithm for the inference of EK-LDA. Let $i = (m, n)$ be the notation for the nth token in document m, the complete conditional of $\mathbf{Z}_i$ and $\mathbf{U}_i$, i.e. $\mathrm{P}\left(\mathbf{Z}_i | \mathbf{Z}_{-i}, \mathbf{U}, \mathbf{W}\right)$ and $\mathrm{P}\left(\mathbf{U}_i | \mathbf{U}_{-i}, \mathbf{Z}, \mathbf{W}\right)$ can be calculated by integrating out θ and ϕ. The two complete conditionals are defined as:

$$\mathrm{P}\left(\mathbf{Z}_i = k | \mathbf{Z}_{-i}, \mathbf{U}, \mathbf{W}, \alpha, \beta\right) \propto \left(\mathrm{n}_{m,-i}^{(k)} + \alpha_k\right) \times f(k, q) \times \left(\mathrm{K} \frac{\beta_q}{\sum_{q=1}^{\mathrm{Q}} \beta_q}\right) \tag{1}$$

$$\mathrm{P}\left(\mathbf{U}_i = q|\mathbf{U}_{-i}, \mathbf{Z}, \mathbf{W}, \alpha, \beta\right) \propto \eta_{qv} \times f(k, q) \tag{2}$$

$$f(k, q) = \frac{\mathrm{n}_{k,-i}^{(q)} + \beta_q}{\sum_{q=1}^{\mathrm{Q}} \mathbf{n}_{k,-i}^{(q)} + \beta_q} \tag{3}$$

where k is topic ID and q is semantic unit ID. α_k and β_q are the kth and qth component of hyperparameter α and β respectively. $\mathrm{n}_{m,-i}^{(k)}$ is a component of a rank-3 tensor, referring to the number of words in document m with topic k (except for word i). $\mathrm{n}_{k,-i}^{(q)}$ is also a rank-3 tensor, referring to the number of words classified to topic k with semantic unit ID q (except for word i). η_{qv} is the vth component of vector η_q. It represents the probability for a word v to have its meaning allocated to semantic unit q. Algorithm 1 illustrates the inference process.

Algorithm 1: Inference for EK-LDA

```
initialize:
each $\mathbf{Z}_{mn}$ in $\mathbf{Z}$ by randomly assigning it a topic ID $k \in \{1, 2, \ldots, \mathrm{K}\}$;
each $\mathbf{U}_{mn}$ in $\mathbf{U}$ by randomly assigning it a semantic unit ID $q \in \{1, 2, \ldots, \mathrm{Q}\}$;
let $i = (m, n), i \in \{1, 2, \ldots, \mathrm{M}\} \times \{1, 2, \ldots, \mathrm{N}\}$;
repeat
    foreach $i$ do
        sample $\mathbf{Z}_i$ from Equation (1);
        sample $\mathbf{U}_i$ from Equation (2);
    end
until convergence;
```

2.4. EIK-LDA and Its Generative Process

In EK-LDA, semantic uints are defined by entity linking and are generated from a multinomial distribution. A drawback of using entity linking is that not every word in the vocabulary can be linked to an entity. For instance, entity linking systems using Wikipedia have difficulty in linking verbs, adjectives or adverbs to specific entities. For these words, the only feature can be used by EK-LDA is their frequencies in each topic. To address this potential problem, we introduce pretrained word representations into EK-LDA so that the model can access every word's latent features. We name this upgraded model EIK-LDA.

In EIK-LDA, we denote semantic unit representation by $\boldsymbol{u}_q$, which is a weighted sum of its corresponding words:

$$\boldsymbol{u}_q = \sum_{v \in \mathrm{T}_q} \eta_{qv} \cdot \boldsymbol{w}_v \tag{4}$$

$\boldsymbol{w}_v$ is the word representation for the vth word in the vocabulary, and $\mathrm{T}_q = \{i|\eta_{qi} \neq 0, 0 \leq i \leq \mathrm{V}\}$ represents a set of word IDs that appear in semantic unit q. η_{qv} is the frequency that word v is allocated to semantic unit q.

As is illustrated by the graphical model in Figure 1, EIK-LDA generates semantic units by a combination of two approaches. We firstly define an indicator $\mathbf{s}_{mn} \sim \mathrm{Ber}(1/2)$ for each word n in document m. Then we generate $\mathbf{U}_{mn}$ according to the value of $\mathbf{s}_{mn}$:

$$\mathrm{P}\left(\mathbf{U}_{mn} = q|\mathbf{Z}_{mn} = k, \alpha, \beta\right) = \begin{cases} \phi_{kq} & \text{for } \mathbf{s}_{mn} = 0 \\ \frac{\exp\left(\tau_k^T \cdot u_q\right)}{\sum_q \exp\left(\tau_k^T \cdot u_q\right)} & \text{for } \mathbf{s}_{mn} = 1 \end{cases} \tag{5}$$

As can be seen from equation (5), we choose semantic unit for a word by a multinomial distribution and a softmax-based scoring mechanism. When $\mathbf{s}_{mn} = 0$, value of $\mathbf{U}_{mn}$ is decided by distribution ϕ_k, which is identical to that of EK-LDA. When $\mathbf{s}_{mn} = 1$, value of $\mathbf{U}_{mn}$ is decided by semantic representation $\boldsymbol{u}_q$ and topic weights τ. The softmax value $\frac{\exp\left(\tau_k^T \cdot u_q\right)}{\sum_q \exp\left(\tau_k^T \cdot u_q\right)}$ serves as a compatibility score which indicates the closeness between semantic unit q and topic k.

The generative process of EIK-LDA is as follows. For each document m , draw distribution θ_m of topics. Generate a topic label Z_{mn} for each token n in document m. For each topic k, draw a distribution ϕ_k of semantic units. Generate $\mathbf{s}_{mn}$ from Bernoulli distribution with parameter 0.5. For position n in document m, generate a semantic unit $\mathbf{U}_{mn}$ according to Dirichlet multinomial or latent features, which is decided by the value of $\mathbf{s}_{mn}$. Based on semantic unit $\mathbf{U}_{mn}$, sample a word from multinomial distribution $\eta_{\mathbf{U}_{mn}}$. This can be written as:

$$\begin{array}{ll} \theta_m \sim \mathrm{Dir}(\alpha) & \mathbf{Z}_{mn} \sim \mathrm{Mult}\left(\theta_m\right) \\ \phi_{\mathbf{Z}_{mn}} \sim \mathrm{Dir}(\beta) & \mathbf{s}_{mn} \sim \mathrm{Ber}(1/2) \\ \mathbf{U}_{mn} \sim \mathrm{P}\left(\mathbf{U}_{mn} = q|\mathbf{Z}_{mn} = k, \alpha, \beta\right) & \\ \mathbf{W}_{mn} \sim \mathrm{Mult}\left(\eta_{\mathrm{U}_{mn}}\right) & \end{array}$$

2.5. Inference for EIK-LDA

In this section, we introduce a collapsed Gibbs sampling algorithm for EIK-LDA. Let $i = (m, n)$ be the notation for the nth token in document m. Complete conditional $\mathrm{P}\left(\mathbf{Z}_i|\mathbf{Z}_{-i}, \mathbf{U}, \mathbf{s}, \tau, \boldsymbol{u}, \mathbf{W}\right)$, $\mathrm{P}\left(\mathbf{U}_i|\mathbf{U}_{-i}, \mathbf{Z}, \mathbf{s}, \tau, \boldsymbol{u}, \mathbf{W}\right)$ and probability $\mathrm{P}\left(\mathbf{s}_i|\mathbf{Z}_i, \mathbf{U}_i, \mathbf{s}_{-i}, \tau, \boldsymbol{u}, \mathbf{W}\right)$ can be calculated by integrating out θ and ϕ , which are shown in the equations below:

$$\mathrm{P}\left(\mathbf{Z}_i = k|\mathbf{Z}_{-i}, \mathbf{U}, \mathbf{s}, \tau, \boldsymbol{u}, \mathbf{W}, \alpha, \beta\right) \propto \left(\mathbf{n}_{m,-i}^{(k)} + \alpha_k\right) \times g(k, q) \times \left(\mathrm{K}\frac{\beta_q}{\sum_{q-1}^{\mathrm{Q}} \beta_q}\right) \tag{6}$$

$$\mathrm{P}\left(\mathbf{U}_i = q|\mathbf{U}_{-i}, \mathbf{Z}, \mathbf{s}, \tau, \boldsymbol{u}, \mathbf{W}, \alpha, \beta\right) \propto \eta_{qv} \times g(k, q) \tag{7}$$

$$\mathrm{P}(\mathbf{s}_i, |\mathbf{Z}_i = k, \mathbf{U}_i = q, \mathbf{s}_{-i}, \tau, \boldsymbol{u}, \mathbf{W}, \alpha, \beta) \propto (1 - \mathbf{s}_i) \cdot f(k, q) + \mathbf{s}_i \cdot \frac{\exp\left(\tau_k^T \cdot u_q\right)}{\sum_q \exp\left(\tau_k^T \cdot u_q\right)} \tag{8}$$

where

$$g(k,q) = \frac{1}{2}\left[f(k,q) + \frac{\exp\left(\tau_k^T \cdot u_q\right)}{\sum_q \exp\left(\tau_k^T \cdot u_q\right)}\right] \tag{9}$$

and $f(k,q) is$ defined in equation (3).

In Equation (6) − (8), k is topic ID, q is semantic unit ID. α_k and β_q are the kth and qth component of hyperparameter α and β respectively. $\mathrm{n}_{m,-i}^{(k)}$ is a component of a rank-3 tensor, referring to the number of words in document m with topic k (except for word i). $\mathrm{n}_{k,-i}^{(q)}$ is also component of a rank-3 tensor, referring to the number of words classified to topic k with semantic unit ID q (except for word i). Algorithm 2 illustrates the inference process. Considering the complexity of Algorithm 2, we choose topic weight τ_k using MAP by minimizing the negative log likelihood of $\mathrm{P}\left(\tau_k|\mathbf{Z},\mathbf{U}\right)$ instead of conducting Gibbs sampling.

Algorithm 2: Inference for EIK-LDA

```
initialize:
each $\mathbf{Z}_{mn}$ in $\mathbf{Z}$ by randomly assigning it a topic ID $k \in \{1,2,\ldots,\mathrm{K}\}$;
each $\mathbf{U}_{mn}$ in $\mathbf{U}$ by randomly assigning it a semantic unit ID $q \in \{1,2,\ldots,\mathrm{Q}\}$;
let $i = (m,n), i \in \{1,2,\ldots,\mathrm{M}\} \times \{1,2,\ldots,\mathrm{N}\}$;
repeat
    for $k$ in $\{1,2,\ldots,\mathrm{K}\}$ do
        choose $\tau_k = \arg\min_{\tau_k} \; -\log \mathrm{P}\left(\tau_k|\mathbf{Z},\mathbf{U}\right)$;
    end
    foreach $i$ do
        sample $\mathbf{Z}_i$ from Equation (6);
        sample $\mathbf{U}_i$ from Equation (7);
        sample $\mathbf{s}_i$ from Equation (8);
    end
until convergence;
```

3. Related Work

3.1. Knowledge enhanced topic modeling

A lot of recent work has contributed to knowledge-enhanced topic modeling. Though adopting similar ideas, the presented models take advantage of different knowledge sources and focus on various knowledge forms.

(1) Knowledge base/graph enhanced topic models. Knowledge bases and graphs are common knowledge sources adopted by researchers. (Boyd-Graber et al. (2007)) embeds WordNet hierarchy into LDA. By including hidden meanings in word generation, the model can carry the notion of sense, improving the performance of word sense disambiguation. MDK-LDA(Chen et al. (2013b)) represents topics as distributions of domain knowledge in WordNet, which includes synonyms and antonyms of verbs, nouns and adjectives. By concept clustering, (Yao et al. (2015)) treats facts from Probase as asymmetric Dirichlet priors of LDA, which improves topic coherence.

Knowledge graph embedding is also introduced to topic modeling by researchers. (Yao et al. (2017)) presents KGE-LDA, which incorporates embeddings of WordNet entities and relations into a Dirichlet multinomial topic model. By taking this approach, topic models are capable of analyzing lexical relations of words, with better interpretability at the same time. (Wang et al. (2019)) develops KGETM, a novel topic model using self constructed domain knowledge graph as its knowledge source and TransE(Bordes et al. (2013)) as knowledge embedding method.

(2) Topic modeling with latent feature word representations. Apart from incorporating explicit knowledge into topic modeling, researchers also take advantage of latent feature word representations to acquire implicit knowledge.

Some work attempts to integrate word embeddings into probabilistic topic models. MRF-LDA (Xie et al. (2015)) encodes word correlations with a Markov random field (MRF) on LDA's topic layer. It uses Web Eigenwords[1] as knowledge source, where words are represented by real-valued vectors with semantic information. (Nguyen et al. (2015)) leverage pretrained word vectors for the generation of words, taking advantage of both Dirichlet multinomial models and latent features.

Another portion of work turns to neural topic models. In (Cao et al. (2015)), a neural topic model (NTM) is presented. NTM is an neural network interpretation of a probablistic topic model. It integrates word and document representations by adding an n-gram embedding layer to the network. (He et al. (2017)) develops an attention-based approach for aspect extraction, which is often regarded as an application of topic modeling. With pretrained word representations and self-attention, the presented model is capable of extracting more coherent aspects(topics) than conventional topic models.

3.2. Knowledge base entity linking

Entity linking bridges the gap between documents and knowledge bases. With this technique, real-world entities can be recognized from word mentions in documents, facilitating various NLP tasks where entity information is needed (e.g. topic modeling and event detection on news). The emergence and thriving of deep learning inspires research on neural entity linking(He et al. (2013); Tsai and Roth (2016)). (Sun et al. (2015)) introduces word mention context into entity linking. It constructs mention-context representations by concatenating vectors of word mention and its context, and develops a cosine-similarity-based scoring mechanism for ranking potential entities. (Francis-Landau et al. (2016)) uses convolutional neural networks (CNN) to capture features of word mentions and contexts on different granularities, and integrates this technique into a scoring-based entity linking scheme. Apart from context of word mentions and entity descriptions, (Gupta et al. (2017)) utilizes entity types defined in knowledge bases, which trains and encodes word mention representations, word mention context, entity descriptions and entity types jointly.

[1] http://www.cis.upenn.edu/ ungar/eigenwords/

4. Experiments

To evaluate EK-LDA and EIK-LDA, we conducted experiments on public datasets with several baseline models. We focused on two tasks: topic coherence evaluation and document classification. Topic coherence indicates the closeness among topic words by several metrics, e.g. PMI (Pointwise Mutual Information) (Newman et al. (2010)) and NPMI (Normalized Pointwise Mutual Information) (Lau et al. (2014)). A high coherence score indicates the words allocated to a topic are relevant to each other. Document classification evaluates the latent document representations learned by a topic model. High classification accuracy indicates the topic assignment for each document is reasonable.

4.1. Experiment Setup

4.1.1. External Knowledge

EK-LDA obtains external knowledge by knowledge base entity linking, while EIK-LDA utilizes both entity linking and pretrained word representations. In this paper, for knowledge base entity linking, we adopted neural-el[1], a neural entity linking system presented in (Gupta et al. (2017)). We used the author-provided CDT model as our linking system, which uses Wikipedia with dump date 2016/09/20 as source of entity, and Glove (Pennington et al. (2014)) for word embedding. For pretrained word vectors, we used Google's Word2vec with 300 dimensions, which were trained using Wikipedia[2] (English Edition).

4.1.2. Datasets

We chose 20-Newsgroups[3] and NIPS[4] as our evaluation datasets. The 20-Newsgroups dataset contains nearly 20,000 documents distributed over 20 categories. We used the original 20-Newsgroups dataset for evaluation. NIPS dataset is a collection of papers published at NIPS from 2000 to 2012, which contains about 1500 documents. Non-alphabetic characters, words with less than 4 characters, along with low frequency words (#appears $<$ 10) were removed from these datasets. Statistics is shown in Table 2. Both topic coherence and document classification tasks were conducted on 20-Newsgroups dataset. Since NIPS does not contain labels for classification, we evaluated topic coherence on it.

Table 2: Statistics of two datasets

Dataset	20-Newsgroups	NIPS
#Documents	19294	1740
#Vocabulary	13464	10118

4.1.3. Baselines

We compared EK-LDA and EIK-LDA with 5 baseline models: LDA, LF-LDA (Nguyen et al. (2015)), KGE-LDA (Yao et al. (2017)), ProdLDA (Srivastava and Sutton (2017))

[1] Available at https://nitishgupta.github.io/neural-el/
[2] Available at http://i.stanford.edu/hazy/opendata/
[3] Available at http://qwone.com/ jason/20Newsgroups/
[4] Available at https://cs.nyu.edu/ roweis/data/

and ABAE (He et al. (2017)). LDA is one of the most famous topic models based on Dirichlet multinomial. It is a count-based model. LF-LDA is a topic model that combines Dirichlet multinomial and pretrained latent feature word representations. It incorporates implicit knowledge into LDA. KGE-LDA is a topic model that incorporates knowledge graph embeddings into LDA. It improves semantic coherence and model interpretability by utilizing knowledge graph. ProdLDA is a VAE (variational auto-encoders) (Kingma and Welling (2014)) based neural topic model which is designed to incorporate expert knowledge. ABAE is a self-attention based neural model which extracts topics with attention and latent feature word representations.

4.1.4. Other Settings

Following previous research, we set $\alpha = \frac{50}{\mathrm{K}}$ (Lu et al. (2011)) and $\beta = 0.01$ (Griffiths and Steyvers (2004)) for corresponding Dirichlet multinomial hyperparameters in EK-LDA, EIK-LDA, LDA, LF-LDA and KGE-LDA. For LF-LDA, we set mixture weight $\lambda = 0.5$. For KGE-LDA, we initialized $\mu_0 \sim N(0, 1)$ with normalization, and set $C_0 = 0.01$, $m = 0.01$ and $\sigma = 0.25$ respectively. 1000 iterations of Gibbs sampling was conducted for the approximate inference of all 5 models. To train ProdLDA, we used Adam optimizer with “High Learning Rate” option ($\beta_1 > 0.8$, $0.1 > learning\_rate > 0.001^6$). For ABAE, we initialized topic matrix $\mathbf{T}$ with K-means centroids of pretrained word vectors. Word vectors used in this model was trained by a Word2vec implementation provided by Gensim[1]. We used Adam to train this model with $learning\_rate = 0.001$ for 150 epochs.

4.2. Topic Coherence

Topic coherence evaluates the coherence of high-frequency words within a topic. Models with a higher topic coherence score tend to have a more interpretable topic-word distribution. In this section, we conducted topic coherence evaluations for our new models on 20-Newsgroups and NIPS, with 5 baselines mentioned in §4.1.3. Both quantitative and qualitative analysis are included in the evaluation.

4.2.1. Quantitative Analysis

We used pointwise mutual information (PMI) for quantitative analysis. It is defined in Equation (10).

$$\text{PMI-score}(k) = \sum_{1 \leq i < j \leq \mathrm{N}} \log \frac{\mathrm{P}(w_i, w_j)}{\mathrm{P}(w_i) \cdot \mathrm{P}(w_j)} \tag{10}$$

where k is the topic whose coherence is to be tested. N is the top N word used for evaluation. w_i and w_j are words. $\mathrm{P}(w_i, w_j)$ is the probability of the two words' co-existence in one document. $\mathrm{P}(w_i)$ and $\mathrm{P}(w_j)$ are the probabilities of w_i and w_j appearing in a document.

We chose top 10 words to calculate PMIs and repeated the task for 10 times for each model. Figure 3 illustrates the average PMI score and the corresponding standard derivations of 7 models (2 for evaluation and 5 baselines) on 2 datasets (20-Newsgroups and NIPS), with parameter K ranging from 20 to 40.

[1] Available at https://radimrehurek.com/gensim/

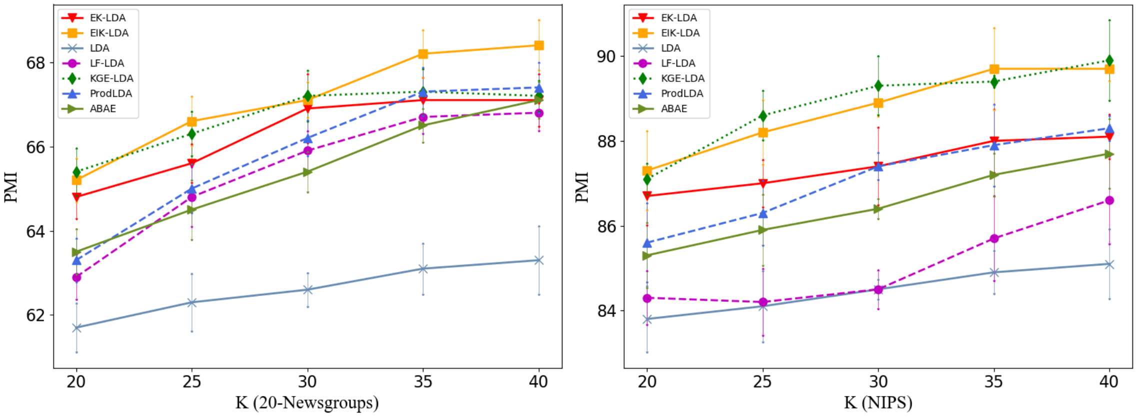


Figure 3: PMI scores of all 7 models on 20-Newsgroups and NIPS

Results of the experiments illustrate that both the EK-LDA and EIK-LDA significantly improves the PMI score in comparison to LDA, which indicates these two models' potential for improving the semantic coherence of topic modeling. They also outperform LF-LDA on each K that involved in this test.

Result analysis on 20-Newsgroups. As can be seen from Figure 3, almost all the 7 models have witnessed an increase of PMI as K gets larger. EIK-LDA has the best performance among the models with a PMI score of 68.4. Results of ProdLDA, KGE-LDA, EK-LDA, ABAE and LF-LDA are close when K = 40. We are especially interested in the performance of 3 knowledge-enhanced models: EIK-LDA, EK-LDA and KGE-LDA. The PMI score of EK-LDA is slightly lower than that of KGE-LDA. One possible explanation of this result is these two models take advantage of different knowledge sources. Emphasizing on real-world entities, EK-LDA incorporates external knowledge of Wikipedia (places, organizations, people, etc.) while KGE-LDA uses WordNet that focuses on lexical relations of words. We can see from the experiment that these two knowledge sources are both beneficial to topic coherence. From this perspective, we can also get a possible explanation of EIK-LDA's high performance. EK-LDA and KGE-LDA incorporate one knowledge source and have different emphasis while EIK-LDA take advantage of two sources. EIK-LDA covers real-world entities by entity linking to Wikipedia and incorporates word correlations by pre-trained word representations, which combines advantages of both EK-LDA and KGE-LDA. Apart from EIK-LDA, EK-LDA and KGE-LDA, results of ProdLDA and ABAE are also interesting. These models have achieved much higher PMI than LDA, but their scores are lower than that of EIK-LDA. A possible reason for ProdLDA's result is VAE-based models like ProdLDA have potential problems of KL-divergence vanishing, which may cause the model's prior to be no longer effective. When training ABAE, we found this model was easily affected by parameter initialization, especially the initialization of topic matrix, which may have influenced the model's average PMI score.

Result analysis on NIPS. On this dataset, EK-LDA performs competitively against ProdLDA and ABAE. This suggests the explicit knowledge enhancement approach we take can greatly improve a conventiaonal topic model to be able to compete with neural topic

Table 3: Top 10 words generated by 3 models on 2 datasets, with 2 topics

20-Newsgroups					
LDA		EK-LDA		EIK-LDA	
guns	information	guns	address	guns	network
used	address	police	information	crime	internet
control	**group**	**carry**	network	police	e-mail
crime	internet	killed	internet	kill	email
police	**people**	crime	e-mail	killed	address
weapons	posting	weapon	email	weapon	connection
California	**questions**	death	connect	death	mail
weapon	**book**	**people**	**questions**	control	send
firearms	network	firearms	send	deaths	list
people	connection	control	connection	firearms	connect
NIPS					
LDA		EK-LDA		EIK-LDA	
network	hidden	network	hidden	network	speech
feedforward	context	neural	Markov	neural	hidden
neural	model	feedforward	speech	learning	Markov
architecture	Markov	application	model	model	recognition
general	**dependent**	architecture	recognition	algorithm	model
introduction	recognition	**general**	probabilities	feedforward	vocabulary
learning	probabilities	system	system	input	probabilities
generalization	training	learning	**dependent**	architecture	training
fixed	**number**	training	training	**introduction**	speaker
training	system	**introduction**	processing	function	system

models. EIK-LDA and KGE-LDA are also competitive with each other. They get the highest PMI score almost alternately as K increases from 20 to 40, with an average score of 88.77 and 88.86 respectively. KGE-LDA performs slightly better than EIK-LDA. This is mostly because of the characteristics of the dataset and the models. Different from 20-Newsgroups that contains a considerable amount of real-world entities, NIPS is a dataset focusing on machine learning. As a result, articles in NIPS do not have much real-world entities inside. Apart from the characteristics of the dataset, the external knowledge the two models use may also affect the result. EIK-LDA emphasizes on linking word mentions to real-world entities in Wikipedia while KGE-LDA focuses mainly on conceptual-semantic and cross-words lexical relations, and the later one may have a more significant effect on the topic coherence of NIPS. However, this does not necessarily mean that knowledge base entity linking is not valuable for topic modeling. Instead, for corpora that contain enormous real-world entities like 20-Newsgroups, entity linking based methods EK-LDA and EIK-LDA perform quite well, which shows the value of this technique.

4.2.2. Qualitative Analysis

Table 3 illustrates the top 10 topic words generated by LDA, EK-LDA and EIK-LDA on 20-Newsgroups and NIPS. We set K = 40 and randomly chose 2 potential topics for each dataset. We asked 20 people to label the noisy words which they thought were not relevant to the topic. Words labelled by more than 10 people are treated as noisy words and are highlighted in bold.

We can see that for the first topic in 20-Newsgroups, words generated by EIK-LDA have the best representativeness, which indicates this topic is about guns and crimes. In

Table 4: Classification accuracy of all 7 models with K ranging from 20 to 40

Models	**K = 20**	**K = 25**	**K = 30**	**K = 35**	**K = 40**
LDA	0.552±0.017	0.594±0.012	0.629±0.023	0.652±0.015	0.683±0.014
LF-LDA	0.570±0.009	0.583±0.019	0.638±0.015	0.660±0.020	0.687±0.016
ABAE	0.579±0.156	0.622±0.195	0.647±0.132	0.668±0.218	0.696±0.177
ProdLDA	0.575±0.033	0.594±0.063	0.651±0.034	0.677±0.029	0.692±0.046
KGE-LDA	**0.583±0.024**	0.629±0.021	0.662±0.019	**0.689±0.010**	0.701±0.013
EK-LDA	0.578±0.025	0.611±0.010	0.659±0.022	0.674±0.031	0.698±0.018
EIK-LDA	0.582±0.012	**0.630±0.007**	**0.669±0.016**	0.686±0.028	**0.705±0.021**

comparison, LDA generates 3 noisy words: *used*, *California* and *people*. These 3 words are replaced by *kill*, *killed* and *death*, which have a stronger relation with the topic. Results of the other 3 groups of words are like the first one. To sum up, words generated by EIK-LDA have a strong relationship with the potential topic. In this test, only 1 noisy word is generated by this model. EK-LDA has the second-best performance, which generates 6 noisy words in 4 potential topics. In comparison, LDA generates a total of 12 noisy words.

4.3. Document Classification

In this test, we used vectors θ_m as document representations for EIK-LDA, EK-LDA, ProdLDA, LF-LDA and LDA. In ABAE, we represent each document as a weighted sum of sentence embeddings. We performed classification with support vector machine provided by Scikit-learn[1]. Since NIPS does not offer labels for classification, we conducted this task on 20-Newsgroups. For each model, we repeated the task for 10 times. Table 4 shows the classification accuracy (mean and standard derivation) of each model with K ranging from 20 to 40. The highest accuracy on each K is marked with bold font.

Result shows that EIK-LDA and EK-LDA outperforms ProdLDA, ABAE, LF-LDA and LDA. In specific, EIK-LDA increases the accuracy by 3% to 4% in comparison to LDA. Its accuracy is also better than that of KGE-LDA when K takes the value of 25, 30 and 40. In fact, EIK-LDA has the highest average classification accuracy which is 0.6544, and the average accuracy of KGE-LDA is 0.6528. We also notice that the classification accuracy of EK-LDA is compatitve with ProdLDA and ABAE which again suggests effectiveness of knowledge-enhancement.

5. Conclusion

In this paper, we present two knowledge enhanced topic models: EK-LDA and EIK-LDA. EK-LDA uses entity linking to obtain explicit knowledge of real-world entities, providing a potential solution for the problem of synonymy and polysemy in topic modeling. To further obtain semantic information for verbs, adjectives and adverbs that are excluded in knowledge bases like Wikipedia, we present EIK-LDA which integrates pretrained word representations into EK-LDA. Experimental results show that these two models exhibit high potential in generating coherent topics. EIK-LDA even outperforms 2 state-of-the-art neural topic models, which suggests that knowledge enhancement can highly improve the performance of topic models, even for very conventional ones like LDA.

[1] https://scikit-learn.org/

Acknowledgments

This research is sponsored by National Key R&D Program of China (No. 2018YFB1003801).

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 2013*, pages 2787–2795, 2013.

Jordan L. Boyd-Graber, David M. Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *EMNLP-CoNLL 2007*, pages 1024–1033. ACL, 2007.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2210–2216. AAAI Press, 2015.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. Discovering coherent topics using general knowledge. In *CIKM'13*, pages 209–218. ACM, 2013a.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. Leveraging multi-domain prior knowledge in topic models. In *IJCAI 2013*, pages 2071–2077. IJCAI/AAAI, 2013b.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. Capturing semantic similarity for entity linking with convolutional neural networks. In *NAACL HLT 2016*, pages 1256–1261. ACL, 2016.

T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 5235–5288, 2004.

Nitish Gupta, Sameer Singh, and Dan Roth. Entity linking via joint encoding of types, descriptions, and context. In *EMNLP 2017*, pages 2681–2690. ACL, 2017.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *ACL 2017*, pages 388–397. ACL, 2017.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. Learning entity representation for entity disambiguation. In *ACL 2013*, pages 30–34. ACL, 2013.

Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57. ACM, 1999.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539. ACL, 2014.

Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Inf. Retr.*, 14(2):178–203, 2011.

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

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *HLT-NAACL*, pages 100–108. ACL, 2010.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *Trans. Assoc. Comput. Linguistics*, 3:299–313, 2015.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP 2014*, pages 1532–1543. ACL, 2014.

Akash Srivastava and Charles A. Sutton. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Modeling mention, context and entity with neural networks for entity disambiguation. In *IJCAI 2015*, pages 1333–1339. AAAI Press, 2015.

Chen-Tse Tsai and Dan Roth. Cross-lingual wikification using multilingual embeddings. In *NAACL HLT 2016*, pages 589–598. ACL, 2016.

Xinyu Wang, Ying Zhang, Xiaoling Wang, and Jin Chen. A knowledge graph enhanced topic modeling approach for herb recommendation. In *DASFAA 2019*, volume 11446 of *Lecture Notes in Computer Science*, pages 709–724. Springer, 2019.

Pengtao Xie, Diyi Yang, and Eric P. Xing. Incorporating word correlation knowledge into topic modeling. In *NAACL HLT 2015*, pages 725–734. ACL, 2015.

Liang Yao, Yin Zhang, Baogang Wei, Hongze Qian, and Yibing Wang. Incorporating probabilistic knowledge into topic models. In *PAKDD 2015*, volume 9078 of *Lecture Notes in Computer Science*, pages 586–597. Springer, 2015.

Liang Yao, Yin Zhang, Baogang Wei, Lei Li, Fei Wu, Peng Zhang, and Yali Bian. Concept over time: the combination of probabilistic topic model with wikipedia knowledge. *Expert Syst. Appl.*, 60:27–38, 2016.

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