Discovering Coherent Topics with Entity Topic Models

Mehdi Allahyari
Computer Science Department
University of Georgia, Athens, GA, USA
Email: mehdi@cs.uga.edu

Krys Kochut
Computer Science Department
University of Georgia, Athens, GA, USA
Email: kochut@cs.uga.edu

***Abstract*—Probabilistic topic models are powerful techniques which are widely used for discovering topics or semantic content from a large collection of documents. However, because topic models are entirely unsupervised, they may lead to topics that are not understandable in applications. Recently, several knowledge-based topic models have been proposed which primarily use word-level domain knowledge in the model to enhance the topic coherence and ignore the rich information carried by entities (e.g persons, location, organizations, etc.) associated with the documents. Additionally, there exists a vast amount of prior knowledge (background knowledge) represented as ontologies and Linked Open Data (LOD), which can be incorporated into the topic models to produce coherent topics. In this paper, we introduce a novel entity-based topic model, called *EntLDA*, to effectively integrate an ontology with an entity topic model to improve the topic modeling process. Furthermore, to increase the coherence of the identified topics, we introduce a novel ontology-based regularization framework, which is then integrated with the EntLDA model. Our experimental results demonstrate the effectiveness of the proposed model in improving the coherence of the topics.**

***Keywords*—Statistical learning, Topic modeling, Topic coherence, Semantic Web, Ontologies**

I. Introduction

Probabilistic topic models such as Latent Dirichlet Allocation [1] have been shown to be powerful techniques to analyze the content of documents and extract the underlying topics represented in the collection. Topic models usually assume that individual documents are mixtures of one or more topics, while topics are probability distributions over the words. These models have been extensively used in a variety of text processing tasks, such as word sense disambiguation [2], [3], relation extraction [4], text classification [5], [6], and information retrieval [7]. Thus, topic models provide an effective framework for extracting the latent semantics from the unstructured text collection.

However, due to the fact that topic models are entirely unsupervised, purely statistical and data driven, they may produce topics that are not meaningful and understandable to humans. Recently, several *knowledge-based* topic models haven been proposed to cope with this issue [8]–[11]. These models incorporate domain knowledge to guide the topic identification process. Chemudugunta et al. [8], for example, combined human-defined concepts with unsupervised topic modeling. [9] proposed a model to incorporate the domain knowledge in the form of must-links and cannot-links into LDA. A must-link indicated that two words should be in the same topic, whereas a cannot-link stated that two words should not be in the same topic. [11]–[13] introduced models that utilize prior knowledge in the form of seed words to direct the topic coherency. [14] proposed a model which uses semantic-sets provided by the user to enhance topic coherency in a new domain. A semantic-set is a set of words sharing the same semantic meaning in a domain which is similar to must-links. In [15], the authors described a model that employs specific types of lexical knowledge called lexical semantic relations. Some of the lexical semantic relations are synonymy, antonym, hyponymy, adjective-attribute, etc. They used synonym, antonym and adjective-attribute relations and show the advantages of utilizing these relations for discovering coherent topics.

Although the aforementioned topic model approaches use some kinds of prior knowledge, they essentially treat documents as bag of words and integrate solely word-level prior knowledge into the topic models. However, documents are associated with richer aspects. For instance, news articles convey information about people, locations or events, research articles are linked with authors and venues, and social posts are associated with geo-locations and timestamps.

There already exist topic models that deal with the various aspects of documents. For example, in [16], authors integrate the authorship information into the topic model and discover a topic mixture over the documents and authors. [17] proposed a model to learn the relationship between the topics and entities mentioned in the articles. In [18], authors introduced a topic model in order to link entity mentions in documents to their corresponding entities in the knowledge base.

Moreover, existing topic models do not utilize the vast amount of existing external knowledge bases which are available in the form of ontologies, such as DBpedia [19] and Linked Open Data (LOD)[1].

Another line of work combines topic modeling with graph structure of the data. [20] proposed a method to integrate a topic model with a harmonic regularizer [21] based on the network structure of the data. In [22], the authors introduced

[1] http://linkeddata.org/

a topic model that incorporates heterogeneous information network. Our work differs from previous works in a way that we utilize the *semantic graph* of the entities in the ontology in order to regularize the topic model and discover coherent topics. The underlying intuition is that the entities classified into the same or similar domains in the ontology are semantically closely related to each other and should have similar topics. Accordingly, entities (i.e. ontology concepts and instances) occurring in a document along with the relationships between them create a semantic graph where can be combined with the entity topic model and a regularization framework for improving topic coherence.

In this paper, we propose a topic model which utilizes the DBpedia ontology to enhance the topic modeling process. Our aim is to leverage the semantic graph of concepts in DBpedia and combine their various properties with unsupervised topic models, such as LDA, in a well-founded manner. Although there are existing knowledge-based topic models [8] that use human-defined concepts hierarchies along with topic models, they basically focus on simple aspects of ontologies, i.e. associated vocabulary of concepts and hierarchical relations between concepts, and do not consider the rich aspects of ontology concepts such as *non-hierarchical* relations.

This general unified framework has many advantages, such as linking text documents to knowledge bases and LOD and discovering more coherent topics. We demonstrate the usefulness of our approach by performing a series of experiments.

II. Related Work

Probabilistic topic models, such as the Latent Dirichlet Allocation (LDA) [1] have been proved to be effective and widely applied in various text processing tasks. The nature of these topic models is that they are unsupervised and entirely statistical, therefore, they do not exploit any prior knowledge in the models.

Recently, several approaches have been introduced that incorporate prior knowledge to direct the topic modeling process. For example, [2], [9], [11], [12] integrate word-level knowledge into topic models. [9] proposed DF-LDA that uses word-level domain knowledge in the form of must-links and cannot-links in LDA. [12] leverages word features as side information to boost topic cohesion. [11] described a topic model which uses word-level prior knowledge as the form of sets of seed words in order to find coherent topics. Seed words are user provided words that represent the topics underlying the corpus. [2] proposed a model that allows the user to incorporate knowledge interactively during the topic modeling process.

Some other related works include [14] which introduces the MDK-LDA model to use multiple domains knowledge to guide the topic modeling process. The knowledge is called s-set (semantic-set) and refers to a set of words sharing the same semantic meaning in the domain. In [15], the authors proposed GK-LDA, general knowledge-based model, which exploits general knowledge of lexical semantic relations in the topic model. Our work presented in this paper differs from previous works, because none of the aforementioned works have used ontologies as their background knowledge in the topic models.

Other related works combine ontological concepts with statistical topic models. [3] introduces LDAWN topic model which leverages WordNet knowledge for the word-sense disambiguation task. [8] describe CTM, Concept-Topic model, which combines human-defined concepts with LDA. The key idea in their framework is topics from the statistical topic models and concepts of the ontology are both represented by a set of "focused" words and they use this similarity in their model. In [23], the authors extended the work and proposed HTCM, Hierarchical Concept-Topic model, in order to leverage the known hierarchical structure among concepts.

Other related work [20], [22] combined statistical topic modeling with network analysis by regularizing the topic model with a regularization framework based on the network structure.

Our proposed approach is somewhat similar to a few previous works, particularly [3], [8], [23] in terms of exploiting ontologies in the topic models and with [20], [22] in terms of regularizing the topic model, yet it differs from all of them. In [3], the task is word-sense disambiguation whereas we are going to find coherent topics. CTM [8], uses tree-structured Open Directory Project[2] as ontology and also rely on simple aspect of this concept hierarchy which is the set of words associated to concepts, and HCTM [23] additionally utilizes the hierarchical structures of ontology concepts to direct the topic model. In [20], [22], models do not consider the entities mentioned in the documents in the topic models.

In this paper we propose a novel entity-based topic model, EntLDA, which incorporates DBpedia ontology into the topic model in a systematic manner. In our model we exploit various properties of concepts and not only hierarchical relations but also *lateral (other than hierarchical) relations* between ontology concepts. EntLDA also accounts for entity mentions in documents and their corresponding DBpedia entities as labeled information in the generative process to constrain Dirichlet prior of document-entity and entity-topic in order to effectively improve topic coherency.

III. Problem Statement

In this section, we formally describe the proposed entity topic model and its learning process. We then define the entity network regularization and investigate how to join the entity topic model with the regularization framework.

Many topic models like LDA are based on the idea that documents are made up of topics, while each topic is a probability distribution over the vocabulary. Therefore, they ignore the entities associated with the documents in the modeling process. Unlike LDA, in EntLDA, each document is a distribution over the entities (of the ontology) where each entity is a multinomial distribution over the topics and each topic is a probability distribution over the words. For example,

[2] http://www.dmoz.org

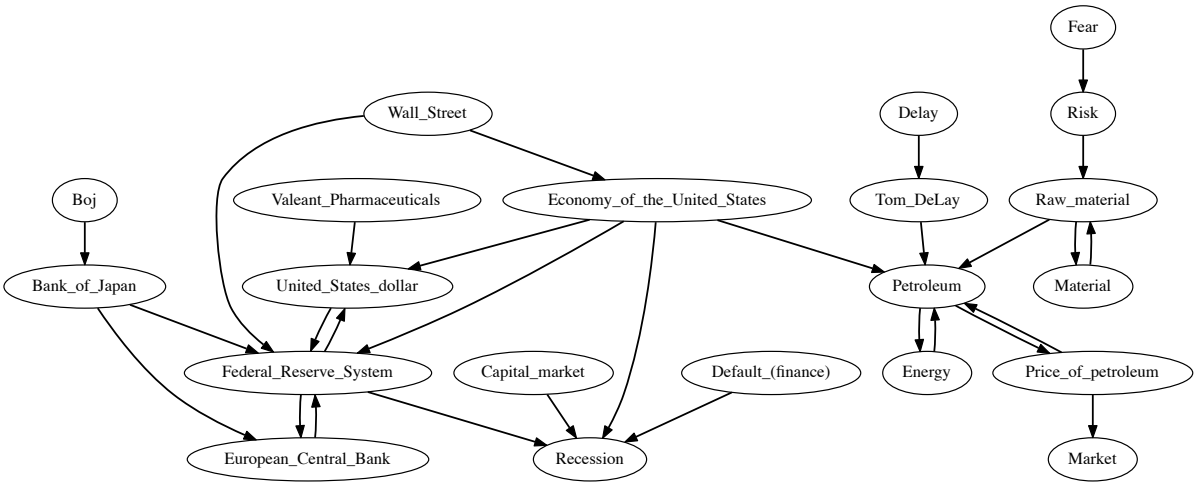

Fig. 1. A fragment of the semantic graph from the example text

in DBpedia ontology, each entity has a number of topics (categories) assigned to it. Hence, each entity is a mixture of different topics with various probabilities.

The underlying intuition behind our model is that documents are associated with entities carrying rich information about the topics of the document. We assume that entities occurring in the document together with the relationships among them can determine the document's topics. Thus, utilizing this plentiful information is of great interest and can potentially improve the topic modeling and topic coherence.

For instance, following is a fragment of a recent news article:

*Stocks in major **markets** fell on Tuesday, weighed by basic **materials** **shares** after **weak** U.S. **retail** sales **data** and as the **Bank of Japan** painted a bleaker picture of the **world**'s third-largest **economy** without immediately adding to its **stimulus**. The yen **rose** sharply against the **U.S. dollar**, **crude oil** and **copper** **prices** dropped, and **emerging market** **shares** fell the most in more than a month.*

*U.S. **retail** sales fell less than expected in February, but a **sharp** downward **revision** to January's sales could reignite concerns about the U.S. **economy**'s **growth** **prospects**. **Wall Street** opened down but **stocks** were off the session's lows. **Energy** **shares** fell alongside the **price of oil** and **healthcare** weighed the most on the S&P 500 hurt by a more than 40 **percent** **drop** in **shares** of Valeant. The Canadian drug-maker **slashed** its 2016 **revenue** **forecast** and said a delay in filing its **annual report** could mean a **debt default**. Following the **BOJ** **statement** and last week's **European Central Bank** action, **investors** seemed to be prepared for a hawkish tone when the **Federal Reserve** ends its two-day **meeting** on Wednesday.*

*Recent **data** suggested the **U.S. economy** is strengthening, however, with fears of **recession** much diminished compared with earlier this year. ''The **Fed** **meeting** is important because ... there is a **risk** of a hawkish **statement**,'' RIA **Capital Markets** bond **strategist** Nick Stamenkovic said. "Investors will wait for the **statement** and the (interest **rate** and **economic growth** expectations) before taking new **positions**."*[...]

We could identify the entity mentions (underlined) in the document and using the information from the DBpedia ontology, induce relationships among them. This would then lead to creation of a semantic graph of connected entities that were recognized in the document, which is illustrated in Figure 1. We combine the graph structure of the semantic network of entities with the probabilistic topic models in order to enhance the coherence of the discovered topics.

A. The EntLDA Topic Model

The novel idea of EntLDA is to include entity information contained in the each document and to integrate ontology concepts and *lateral* relations between them into the topic model and exploit this prior knowledge in order to produce coherent topics automatically.

The graphical representation of EntLDA is shown in Figure 2 and the generative process is defined in Algorithm 1.

It should be noted that in the generative process for each document d, instead of selecting an entity uniformly from $\boldsymbol{E}_d$ as in the author-topic model [16], we draw an entity from a document-specific multinomial distribution ζ_d over $\boldsymbol{E}_d$. The reason is based on the assumption that each entity in $\boldsymbol{E}_d$ contributes differently in generating the document d. $\boldsymbol{E}_d$ is a vector containing all the entities of the document d.

B. Inference Using Gibbs Sampling

In our EntLDA model, two sets of unknown parameters need to be estimated: (1) the E entity-topic distributions θ,

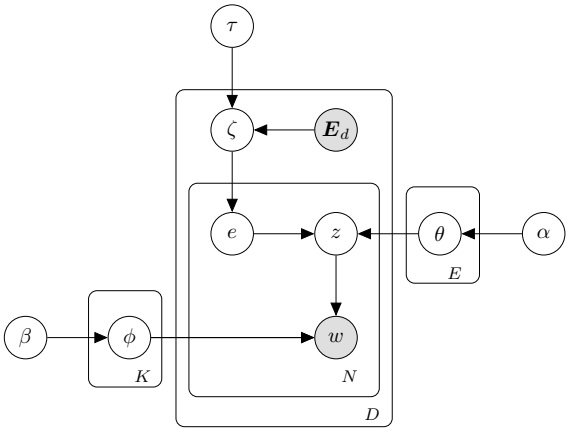

Fig. 2. Graphical representation of EntLDA; symbols explained in Alg. 1

Algorithm 1: EntLDA Topic Model

foreach topic $k \in \{1, 2, \ldots, K\}$ **do**
  Draw a word distribution $\phi_k \sim \text{Dir}(\beta)$
end
foreach entity $e \in \{1, 2, \ldots, E\}$ **do**
  Draw a topic distribution $\theta_e \sim \text{Dir}(\alpha_e)$
end
foreach document $d \in \{1, 2, \ldots, D\}$ **do**
  Draw $\zeta_d \sim \text{Dir}(\tau; \boldsymbol{E}_d)$
  foreach word w of document d **do**
    Draw an entity $e \sim \text{Mult}(\zeta_d)$
    Draw a topic $z \sim \text{Mult}(\theta_e)$
    Draw a word w from topic $z, w \sim \text{Mult}(\phi_z)$
  end
end

and K topic-word distributions; and (2) the assigned topic z_i and assigned entity e_i for each word w_i. There are a variety of techniques to estimate the parameters of the topic models such as variational EM [1] and Gibbs sampling [24]. In this paper we will utilize the collapsed Gibbs sampling algorithm for EntLDA. Collapsed Gibbs sampling is a Markov Chain Monte Carlo (MCMC) [25] algorithm to sample from posterior distribution over the latent variables. Instead of estimating the model parameters directly, we evaluate the posterior distribution on just e and z and then use the results to infer θ and ϕ.

The posterior inference is defined as follows:

$$P(\boldsymbol{e}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{E}_d, \alpha, \beta, \tau) = \frac{P(\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{E}_d, \alpha, \beta, \tau)}{\sum_e \sum_z P(\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{E}_d, \alpha, \beta, \tau)} \propto P(\boldsymbol{e})P(\boldsymbol{z}|\boldsymbol{e})P(\boldsymbol{w}|\boldsymbol{z}) \quad (1)$$

$$P(e_i = j, z_i = k|w_i = w, \boldsymbol{z}_{-i}, \boldsymbol{e}_{-i}, \boldsymbol{w}_{-i}, \boldsymbol{E}_d, \alpha, \beta, \tau) \propto \frac{C^{ED}_{jd,-i} + \tau_j}{\sum_{j'} (C^{ED}_{j'd,-i} + \tau_{j'})} \times \frac{C^{TE}_{kj,-i} + \alpha_k}{\sum_{k'} (C^{TE}_{k'j,-i} + \alpha_{k'})} \times \frac{C^{WT}_{wk,-i} + \beta_w}{\sum_{w'} (C^{WT}_{w'k,-i} + \beta_{w'})} \quad (2)$$

where C^{WT}_{wk} is the number of times word w is assigned to topic k. C^{TE}_{kj} is the number of times topic k is assigned to entity e and C^{ED}_{jd} is the number of word tokens assigned to entity e. Subscript $-i$ denotes the contribution of the current word w_i being sampled is removed from the counts.

After Gibbs sampling, we can easily estimate the topic-word distributions ϕ, entity-topic distributions θ and document-entity distributions ζ by:

$$\zeta_{dj} = \frac{C^{ED}_{jd} + \tau_j}{\sum_{j'} (C^{ED}_{j'd} + \tau_{j'})} \quad (3)$$

$$\theta_{jk} = \frac{C^{TE}_{kj} + \alpha_k}{\sum_{k'} (C^{TE}_{k'j} + \alpha_{k'})} \quad (4)$$

$$\phi_{kw} = \frac{C^{WT}_{wk} + \beta_w}{\sum_{w'} (C^{WT}_{w'k} + \beta_{w'})} \quad (5)$$

where ζ_{dj} is the probability of an entity given a document, θ_{jk} is the probability of a topic given an entity and ϕ_{kw} is the probability of a word given a topic.

C. Regularization Framework for Topic Models

With the advent of Web, text documents are not only getting richer, but also extensively interconnected with users and other types of objects and create network of data where in addition to textual information of documents, we have access to the associated network structure in the data. Bibliographic data and social networks are such examples where we have both textual documents and a network multi-typed objects.

Although topic models have demonstrated to be useful for document analysis, they usually only consider the textual information and ignore the network structure of the data. Interactions between objects of the network plays an important role in revealing the rich semantics of the network. Topic modeling with network structure (regularized topic modeling) has shown to be effective to extract topics and discover topical communities [20], [22], [26]. The basic idea is to combine topic modeling and social network analysis, and leverage the power of both topic models and discrete regularization, which optimizes the likelihood of the generation of topics and topic smoothness on the graph together.

In the following section, we propose an ontology-based regularization framework that combines network structure of the entities in the documents with the topic models.

D. Ontology-based Regularization

In this section we describe the regularization framework that combines the EntLDA topic model with the semantic graph structure of the entities occurring in the documents. The key idea is the entities in the ontology that are semantically closely related to each other, are categorized under the same or similar topics. Thus, we leverage the information in the individual documents including entities mentioned in the document text and join it with graph structure of the ontology by regularizing the topic model based on the entity network. Particularly, entities appearing in a document that are semantically related to each other in the ontology should have similar topics.

Entity Network: An entity network associated with a collection of documents D, is a graph $G = \langle V, E \rangle$, where V is a set of entities occurring in the corpus and E is a set of ontology relations (properties). Each entity e_u is considered as a node in the graph. There is an edge $\langle e_u, e_v \rangle$ between e_u and e_v if both co-occur in the same document and there is a *relation* p in the ontology $\mathcal{O}$ that connects them. Even though edges in the ontology are directed, in this paper we only consider the undirected case. Thus, we define the regularized data likelihood of the EntLDA as follows:

$$O_\xi(D, G) = -(1-\xi)L(D) + \xi R(D, G) \tag{6}$$

where $L(D)$ is the log likelihood of the collection D to be generated by EntLDA topic model, $R(D, G)$ is a harmonic regularizer defined on the entity network G and ξ is the controlling factor of the two terms. The harmonic regularizer can further defined as:

$$R(D, G) = \frac{\lambda}{2} \sum_{\langle e_u, e_v \rangle \in E} w(e_u, e_v) \sum_{j=1}^{k} \big(p(z_k|e_u) - p(z_k|e_v)\big)^2 \tag{7}$$

where $p(z_k|e_i)$ denotes the probability that entity e_i belongs to topic z_k. $w(e_u, e_v)$ is the weight of the edge $\langle e_u, e_v \rangle$. We define $w(e_u, e_v)$ as the *semantic relatedness* between e_u and e_v. We adopt the Wikipedia Link-based Measure (WLM) introduced in [27]. Given two DBpedia entities e_u and e_v, the semantic relatedness between them is defined as:

$$w(e_u, e_v) = 1 - \frac{log(max(|E_u|, |E_v|)) - log(|E_u \bigcap E_v|)}{log(|Y|) - log(min(|E_u|, |E_v|))} \tag{8}$$

where E_u and E_v are sets of DBpedia entities that link to e_u and e_v respectively and Y is the set of all entities in DBpedia.

In order to minimize Eq. 6, we have to find a probabilistic topic model that fits the text collection D and also smoothes the topic distributions between the entities in the entity network. In the special case that $\xi = 0$, the objective function boils down to log-likelihood function of EntLDA with no regularization term. But for the general case $\xi > 0$, there is no closed-form solution for the complete likelihood function [20]. Thus, we use a two-step algorithm to learn all the parameters in Eq. 6. In the first step, we train the model parameters (ζ, θ, ϕ) using the objective function $O_1(D, G) = -L(D)$ by the Gibbs sampling algorithm. We set the Dirichlet prior for each entity e_i as:

$$\alpha_{e_i k} = (1-\xi)\alpha + \xi \frac{K}{|G_{e_i}|} \sum_{\langle e_i, e_j \rangle \in G} \theta_{e_j k}$$

where $|G_e|$ is the number of neighbors of entity e in the entity network G. In the second step, we fix ϕ and ζ, and re-estimate parameters θ to minimize O_ξ by running an iterative process to obtain the new θ for each entity e as:

$$p_{t+1}^{(n+1)}(z_k|e_u) = (1-\gamma)p_{t+1}^{(n)}(z_k|e_u) + \gamma \frac{\sum_{\langle e_u, e_v \rangle \in G} w(e_u, e_v) p_{t+1}^{(n)}(z_k|e_v)}{\sum_{\langle e_u, e_v \rangle \in G} w(e_u, e_v)} \tag{9}$$

where $\theta_{uk}^{(n+1)} = p_{t+1}^{(n+1)}(z_k|e_u)$ and γ is a coefficient to smooth the topic distribution. The learning algorithm has also been used previously in [20], [22], [28]. We summarized the fitting approach in algorithm 2.

Algorithm 2: Parameter Estimation

Input : A collection of documents D and entity network, $\Xi = \{\mathbf{E_d}, \alpha, \beta, \tau, \pi\}$
Output: $\phi = \{p(w_i|z_k)\}$, $\theta = \{p(z_k|e_u)\}$ and $\zeta = \{p(e_u|d_j)\}$

Initialize the parameters ϕ, θ and ζ randomly;
$t \leftarrow 0$;
while $t < MaxIteration$ **do**
  /* Train the model parameters ϕ, θ and ζ with a Gibbs sampling to calculate $O_1(D, G) = -L(D)$ */
  Compute the probability $P(e, z|w, \Xi)$ as in Eq. 2;
  /* Fix ϕ, ζ, and re-estimate parameters θ to minimize $O_\xi(D, G)$ */
  Re-estimate $p(z_k|e_u)_{t+1}$ as in Eq. 4;
  $p(z_k|e_u)_{t+1}^{(1)} \leftarrow p(z_k|e_u)_{t+1}$;
  Compute $p(z_k|e_u)_{t+1}^{(2)}$ (i.e. $\theta_{t+1}^{(2)}$) using Eq. 9 in the paper;
  while $O_\xi(\theta_{t+1}^{(2)}) \geq O_\xi(\theta_{t+1}^{(1)})$ **do**
    $p(z_k|e_u)_{t+1}^{(1)} \leftarrow p(z_k|e_u)_{t+1}^{(2)}$;
    Compute $\theta_{t+1}^{(2)}$;
  end
  if $O_\xi(\theta_{t+1}^{(1)}) \geq O_\xi(\theta_t)$ **then**
    $p(z_k|e_u)_{t+1} \leftarrow p(z_k|e_u)_{t+1}^{(1)}$;
  else
    Keep current θ parameters
  end
  $t \leftarrow t + 1$;
end

IV. Experiments

In this section we evaluate our EntlDA model with regularization framework which we call it ETMR (entity topic model with regularization) and compare it with three baseline models: LDA [1], EntLDA without regularization framework and GK-LDA [15]. LDA is the basic topic model to learn

TABLE I
TOPIC COHERENCE ON TOP T WORDS. A HIGHER COHERENCE SCORE MEANS MORE COHERENT TOPICS.

		TopWords				
	T	**5**	**10**	**15**	**20**	ξ
$K = 20$	**LDA**	−236.966	−1187.90	−3039.50	−6018.90	–
	GK − LDA	−266.103	−1304.60	−3181.30	−6102.20	–
	EntLDA	−264.204	−1239.10	−3072.70	−5999.10	0
	ETMR	**−220.659**	**−1162.70**	**−2919.70**	**−5663.10**	0.9
$K = 25$	**LDA**	−385.643	−1930.50	−4795.30	−9361.30	–
	GK − LDA	−311.517	−1697.50	−3916.10	−7785.70	–
	EntLDA	−392.535	−1947.90	−4894.0	−9123.20	0
	ETMR	**−318.669**	**−1593.50**	**−3859.0**	**−7444.90**	0.9

the topics from the corpus. EntLDA without regularization is just the proposed model excluding the regularization term (i.e. $\xi = 0$ in Eq. 6). GK-LDA is a model that uses word-level lexical knowledge (i.e. synonyms and antonyms) from dictionaries to improve the topic coherence. Therefore, it aims to constrain the words to appear under the topics according to the lexical relations between the words. GK-LDA is the most recent work and the closest to our method in terms of leveraging prior knowledge in the model and discovering topic coherence, which is why we selected it for our experiments.

A. Data Sets

We evaluated the proposed approach on a text corpus from Reuters[3] news articles. The text collection contains $D = 1,243$ news articles categorized into six primary topics: *Business*, *Health*, *Politics*, *Science*, *Sports* and *Technology*. This data collection consists of $239,009$ words and the size of the initial vocabulary is $24,695$. We used DBpedia ontology as our background knowledge and extracted $5,887$ entities (named entities) mentioned in the corpus and used these entities in the experiments.

B. Experimental Setup

We pre-processed the dataset by removing punctuation, stopwords, numbers, and words occurring fewer than 10 times in the corpus. Then, we created a $W = 5,226$ vocabulary. For GK-LDA, we followed the data preparation method explained in [15], ran Stanford POS Tagger[4] and recognized nouns and adjectives. We then utilized WordNet [29] to produce LR-sets (lexical relation sets). We used GK-LDA implementation from the author website.

We implemented the LDA model with the Mallet toolkit[5]. The number of topics was empirically set $K = \{15, 20, 25, 30, 35\}$ and the hyperparameters α, β and τ were set with $\alpha = 50/K, \beta = 0.01$ and $\tau = 0.01$ respectively. Additionally, we empirically set both coefficients ξ and γ for topic smoothing and the effectiveness of background knowledge as 0.9 because they produced better results. Parameter analysis is further described in the section IV-C2. For all the models, we ran the Gibbs sampling algorithm for 500 iterations and computed the posterior inference after the last sampling iteration.

[3] http://www.reuters.com/
[4] http://nlp.stanford.edu
[5] http://mallet.cs.umass.edu/

TABLE II
AVERAGE TOPIC COHERENCE ON TOP T WORDS FOR VARIOUS NUMBER OF TOPICS.

	K = 15	**K = 20**	**K = 25**	**K = 30**	**K = 35**
LDA	−2019.096	−2620.816	−3353.964	−4118.186	−4884.294
GK − LDA	−2091.456	−2713.551	−3427.704	−4301.466	−5009.519
EntLDA	−2002.758	−2643.776	−3331.678	−4121.085	−4839.843
ETMR	**−1870.453**	**−2491.540**	**−3304.017**	**−4040.859**	**−4782.272**

C. Experimental Results

In this section, we comprehensively evaluate our ETMR model (EntLDA with regularization) with the baseline models. We choose the topic coherence metric to evaluate the quality of the topics. Topic models are often evaluated using the perplexity measure [1] on held-out test data. According to Newman et al. [30], perplexity has limitations and may not reflect the semantic coherence of topics learned by the model. Also, in Chang et al. [31], the authors indicate that human judgements can sometimes be contrary to perplexity measure. Additionally, perplexity only provides a measure of how well a model fits the data, which is different from our goal of finding coherent topics.

1) Quantitative Analysis: For quantitative evaluation, we apply a metric, namely ***topic coherence score***, proposed by [32] for measuring the topics quality. This has become the most commonly used topic coherence evaluation method. Given a topic z and its top T words $V^{(z)} = (v_1^{(z)}, \cdots, v_T^{(z)})$ ordered by $P(w|z)$, the coherence score is defined as:

$$C(z; V^{(z)}) = \sum_{t=2}^{T} \sum_{l=1}^{t-1} \log \frac{D(v_t^{(z)}, v_l^{(z)}) + 1}{D(v_l^{(z)})} \qquad (10)$$

where $D(v)$ is the document frequency of word v and $D(v, v')$ is the number of documents in which words v and v' co-occurred. It has been demonstrated that the coherence score is highly consistent with human-judged topic coherence [32]. Higher coherence scores indicate higher quality of topics.

We performed several experiments with various $0 < \xi \leq 1$ for different number of topics $K = \{15, 20, 25, 30, 35\}$, and for majority of experiments $\xi \geq 0.8$ consistently produced better results. Thus, we empirically set $\xi = 0.9$ for all the experiments. Table I shows the results for two topics $K = \{20, 25\}$ where the number of top words ranges from 5 to 20. ETMR receives the highest coherence score which suggests that it outperforms other models significantly.

Table II illustrates the average topic coherence for the top words (ranges from 5 to 20) among all the models with different number of topics. ETMR improves significantly (p-value < 0.01 by t-test) over the LDA, GK-LDA and EntLDA without regularization models.

Figure 3 shows that our ETMR model consistently achieves higher topic coherence score over the baselines. Among the baseline models, LDA works better than GK-LDA and EntLDA without regularization (i.e. $\xi = 0$) which strengthens

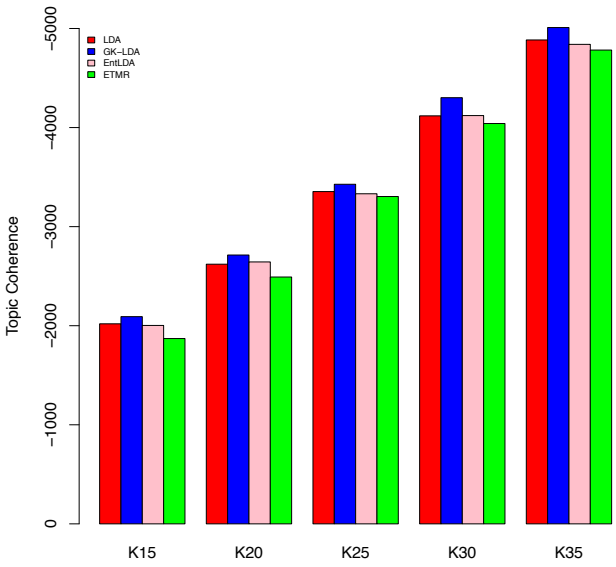

Fig. 3. Average Topic Coherence for all models

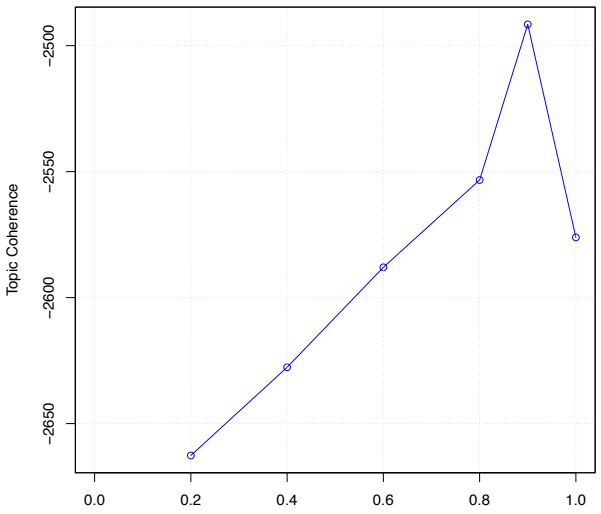

Fig. 4. The effect of varying parameter ξ in the regularization framework for $K = 20$.

the impact of background knowledge, particularly at the concept level in the topic model. The reason that non-regularized EntLDA did not outperform LDA is that we added all the entities corresponding to entity mentions in document d to $\boldsymbol{E}_d$ without doing any explicit entity disambiguation on it. Therefore, $\boldsymbol{E}_d$ might have multiple entities for a single entity mention (i.e., ambiguous entities), which add noise to the topic modeling process. More interestingly, GK-LDA did not work well in our experiments which might be because of the nature of our corpus which is much different than the corpus in [15].

2) Parameter Analysis: In our method, we use the underlying regularization parameter ξ which impacts the ETMR model effectively.

Figure 4 shows the performance of the ETMR varies with the regularization parameter ξ. As we mentioned in section III-D, parameter ξ controls the balance between the data likelihood and the smoothness of the topic distribution over the entity network. When $\xi = 0$, no background knowledge is integrated with the topic model. When $\xi > 0$, the regularization framework considers the topic consistency and semantic relatedness between the entities in the documents which accordingly enhances coherence of the topics. As we increase the value of ξ, we rely more on the integration of the background knowledge into the topic model and receive better topic coherence scores. We also set $\xi = 1$ to see whether we achieve better performance if we solely rely on the entity network. But as Figure 4 illustrates, the topic coherence decreases significantly. Thus, we empirically set $\xi = 0.9$ in all the experiments.

3) Qualitative Analysis: In this section, we describe some qualitative results to give an intuitive feeling about the performance of different models. We selected a sample of topics with the top-10 words for each topic from an experiment with number of topics $K = 20$. Table III presents the top words of each topic for LDA and our proposed models. Although both LDA and ETMR represent the top words for each topic, the ***topic coherence*** under ETMR is qualitatively better than LDA. For each topic, we italicized and marked in red the wrong topical words. We can see that ETMR model produces much better topics than LDA does.

V. Conclusion and Future Work

In this paper, we proposed an entity topic model, EntLDA with a regularization framework, to integrate the probabilistic topic models with the knowledge graph of the ontology in order to investigate the task of discovering coherent topics. The proposed model effectively utilizes the semantic graph of the ontology including entities and the relations among them and combines this knowledge with the topic model to produce more coherent topics. We demonstrated the effectiveness of this model by conducting thorough experiments.

There are many interesting future directions of this work. It would be interesting to investigate the usage of this model for "automatic topic labeling" and text categorization tasks. Additionally, this model can be potentially used for entity classification and entity topic summarization tasks.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[2] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine Learning*, vol. 95, no. 3, pp. 423–469, 2014.

[3] J. L. Boyd-Graber, D. M. Blei, and X. Zhu, "A topic model for word sense disambiguation." in *EMNLP-CoNLL*. Citeseer, 2007, pp. 1024–1033.

[4] L. Yao, A. Haghighi, S. Riedel, and A. McCallum, "Structured relation discovery using generative models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1456–1466.

TABLE III

EXAMPLE TOPICS FROM TWO DOMAINS, ALONG WITH TOP-10 WORDS UNDER LDA AND ETMR MODELS. THE FIRST ROW PRESENTS THE MANUALLY GENERATED LABELS. ITALICIZED RED WORDS INDICATE THE WORDS THAT ARE *not likely* TO BE RELEVANT TO THE TOPICS.

HEALTHCARE STUDY		INTERNET COMPANIES		SPORTS		NATIONAL SECURITY		U.S POLITICS	
LDA	**ETRM**	**LDA**	**ETRM**	**LDA**	**ETRM**	**LDA**	**ETRM**	**LDA**	**ETRM**
drug	drug	company	company	season	season	*china*	government	obama	obama
patients	cancer	million	technology	game	team	data	security	house	house
drugs	patients	billion	mobile	win	win	security	officials	state	president
treatment	drugs	mobile	google	play	game	government	agency	president	state
fda	treatment	google	apple	league	play	information	data	washington	washington
reuters	fda	market	market	home	league	agency	information	law	republican
percent	study	business	internet	*back*	*editing*	states	national	*court*	government
trial	virus	apple	based	team	cup	national	companies	republican	administration
reporting	cases	*reuters*	online	club	final	*defense*	intelligence	healthcare	law
editing	people	corp	business	match	won	*chinese*	nsa	administration	healthcare

[5] S. Hingmire and S. Chakraborti, "Topic labeled text classification: a weakly supervised approach," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 385–394.

[6] J. Li, C. Cardie, and S. Li, "Topicspam: a topic-model based approach for spam detection." in *ACL (2)*, 2013, pp. 217–221.

[7] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 178–185.

[8] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers, "Modeling documents by combining semantic concepts with unsupervised statistical learning," in *The Semantic Web-ISWC 2008*. Springer, 2008, pp. 229–244.

[9] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via dirichlet forest priors," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 25–32.

[10] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht, "A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 1171.

[11] J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 204–213.

[12] J. Petterson, W. Buntine, S. M. Narayanamurthy, T. S. Caetano, and A. J. Smola, "Word features for latent dirichlet allocation," in *Advances in Neural Information Processing Systems*, 2010, pp. 1921–1929.

[13] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 81–88.

[14] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Leveraging multi-domain prior knowledge in topic models," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 2071–2077.

[15] ——, "Discovering coherent topics using general knowledge," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 209–218.

[16] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 487–494.

[17] D. Newman, C. Chemudugunta, and P. Smyth, "Statistical entity-topic models," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 680–686.

[18] X. Han and L. Sun, "An entity-topic model for entity linking," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 105–115.

[19] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia-a crystallization point for the web of data," *Web Semantics: science, services and agents on the world wide web*, vol. 7, no. 3, pp. 154–165, 2009.

[20] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 101–110.

[21] X. Zhu, Z. Ghahramani, J. Lafferty *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, 2003, pp. 912–919.

[22] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin, "Probabilistic topic models with biased propagation on heterogeneous information networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1271–1279.

[23] C. Chemudugunta, P. Smyth, and M. Steyvers, "Combining concept hierarchies and statistical topic models," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 1469–1470.

[24] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[25] C. P. Robert and G. Casella, *Monte Carlo statistical methods*. Citeseer, 2004, vol. 319.

[26] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 287–296.

[27] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, 2008, pp. 25–30.

[28] J. Tang, H.-f. Leung, Q. Luo, D. Chen, and J. Gong, "Towards ontology learning from folksonomies." in *IJCAI*, vol. 9, 2009, pp. 2089–2094.

[29] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[30] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," in *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 2010, pp. 215–224.

[31] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

[32] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.