ETM: Entity Topic Models for Mining Documents Associated with Entities

Hyungsul Kim, Yizhou Sun, Julia Hockenmaier and Jiawei Han
University of Illinois at Urbana-Champaign
{*hkim21, sun22, juliahmr, hanj*}*@illinois.edu*

***Abstract*—Topic models, which factor each document into different topics and represent each topic as a distribution of terms, have been widely and successfully used to better understand collections of text documents. However, documents are also associated with further information, such as the set of real-world entities mentioned in them. For example, news articles are usually related to several people, organizations, countries or locations. Since those associated entities carry rich information, it is highly desirable to build more expressive, entity-based topic models, which can capture the term distributions for each topic, each entity, as well as each topic-entity pair. In this paper, we therefore introduce a novel *Entity Topic Model* (ETM) for documents that are associated with a set of entities. ETM not only models the generative process of a term given its topic and entity information, but also models the correlation of entity term distributions and topic term distributions. A Gibbs sampling-based algorithm is proposed to learn the model. Experiments on real datasets demonstrate the effectiveness of our approach over several state-of-the-art baselines.**

***Keywords*-topic models; data mining; entity;**

I. Introduction

Starting with the great success of Probabilistic Latent Semantic Analysis (PLSA) [8] and Latent Dirichlet Allocation (LDA) [4], there have been numerous proposals for topic models that identify patterns of word occurrences in large collections of documents which reflect the underlying topics represented in the collection, and can then be used to organize, search, index and browse large collection of documents [21].

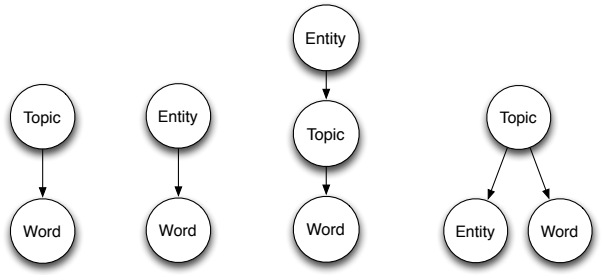
While traditional topic models treat each document as a bag of words, documents are in fact associated with richer attributes: for example, news articles are associated with people, organizations or locations, many tweets are associated with geo-locations and timestamps, research articles are associated with authors, and webpages are associated with link information. This has opened up interesting opportunities and challenges for document analysis. To deal with the different types of attributes associated with documents, different topic models have been proposed: (1) Topic Over Time [18] and Dynamic Topic Models [3] are designed for documents with timestamps, (2) GeoFolk [13] and Latent Geographical Topic Analysis [19] are proposed for documents with GPS information, (3) Author Models [9] and Autor Topic Models [12] deal with documents with author lists, and (4) Link-LDA [6] and Block-LDA [1] are designed for dealing with documents with hyperlinks, citations, and other forms of link information.

In this paper, we are particularly interested in topic analysis for collections of documents associated with sets of entities. The ability to capture the association of documents with real-world entities or concepts holds great promise over traditional keyword-based approaches (cf. Google's "knowledge graph", which enhances search results by linking documents to entities[1]). In a similar vein, we argue that it is also highly desirable to build topic models that can capture the complex patterns involving the entities associated with documents. Almost any document is associated with some set of real-worlds entities. For instance, news articles may mention people, organizations or locations, research papers are associated with authors, medical records are associated with patients, doctors, diseases and so on. Many documents are explicitly associated with entities such as authors or publications via metadata. But since we are now quite successful at wide-coverage named-entity extraction from raw text [5], [7], we can also capture the implicit associations of documents to the entities mentioned in them.

In addition to the term distributions for each topic, we may therefore also wish to know the term distributions for each entity, or topic-entity pair. Namely, letting z, e, and w denote a topic, an entity, and a word, respectively, we want to design a topic model that can answer the following queries: $P(w|z)$, $P(w|e)$, and $P(w|e,z)$.

For example, in a collection of computer science research articles, we may want to find a topic called data mining, and understand it by browsing its word distribution $P(w|\text{Data Mining})$. If we want to identify the topics that a specific researcher, e.g. Judea Pearl, the 2011 winner of the A.M. Turing Award, has worked on, we may want to browse the word distribution $P(w|\text{Judea Pearl})$. We can also have better understanding of his contribution to specific areas such as data mining or artificial intelligence through $P(w|\text{Judea Pearl, Data Mining})$ or $P(w|\text{Judea Pearl, A.I.})$, or the difference of focus of his data mining related works from data mining in general by comparing $P(w|\text{Judea Pearl, Data Mining})$ with $P(w|\text{Data Mining})$. We may also wish to compare his artificial intelligence related works with another leading researcher in that field, e.g. Michael Jordan, by comparing $P(w|\text{Judea Pearl, A.I.})$ with $P(w|\text{Michael Jordan, A.I.})$.

[1] http://www.google.com/insidesearch/features/search/knowledge.html



(a) $w|z$ (b) $w|e$ (c) $w|z$ and $z|e$ (d) $e|z$ and $w|z$

Figure 1. Different dependencies among topic(z), entity(e), and word(w)

As another example, in a collection of news articles about Japan's Tsunami in 2011, we can find frequently mentioned words related to relief efforts by $P(w|\text{Relief Efforts})$, related to the United States by $P(w|\text{United States})$, and the term distribution related to the relief efforts of the United States by $P(w|\text{United States, Relief Efforts})$. Also, we can learn about Naoto Kan who was Japan's Prime Minister at the time by $P(w|\text{Naoto Kan})$, his actions on the tsunami disaster by $P(w|\text{Naoto Kan, Tsunami})$, and his actions on the economic damages by $P(w|\text{Naoto Kan, Economic Damages})$.

To the best of our knowledge, there are no previous studies that have modeled $P(w|e,z)$ directly: they assume either $P(w|e,z) = P(w|z)$ or $P(w|e,z) = P(w|e)$ by introducing different types of conditional dependency relations among topics, entities, and words. In Figure I, we summarize the dependency structures among these variables in several well-known topic models. In LDA (Figure 1(a)), words are drawn for a given topic, and entities are not modeled. In the Author Model, words are drawn for a given author, and topics are not modeled (Figure 1(b)). In the Author Topic Model, topics are drawn for a given author, and words are drawn for a given topic (Figure 1(c)). In Link-LDA, entities are drawn for a given topic, and words are drawn for a given topic (Figure 1(d)). In Figure 1(a), Figure 1(c), and Figure 1(d), $P(w|e,z) = P(w|z)$ is assumed whereas in Figure 1(b), $P(w|e,z) = P(w|e)$ is assumed.

However, in many documents collections, these independence assumptions are not valid. For example, Judea Pearl published many papers in several different domains, including artificial intelligence and data mining. On the one hand, with the assumption of $P(w|e,z) = P(w|e)$, $P(w|\text{Judea Pearl, A.I.}) = P(w|\text{Judea Pearl, Data Mining})$[2], but obviously papers from different topics may not use the same terms. On the other hand, with the assumption of $P(w|e,z) = P(w|z)$, $P(w|\text{Judea Pearl, A.I.}) = P(w|\text{Michael Jordan, A.I.})$, but different authors usually use different terms even in the same research area. Therefore there is a necessity for us to model the correlation of words between a pair of an entity and a topic by directly modeling $P(w|e,z)$. In order to solve this problem, we propose a novel topic model named Entity Topic Model (ETM) for analyzing a given collection of documents with given entities. ETM not only models the generative process of a term given its topic and entity information, but also models the correlation of entity-term and topic-term distributions. We show that LDA and the Author Model are special cases of our model with different parameter settings. A Gibbs sampling-based algorithm is proposed to learn the model. Experiments on real datesets demonstrate the effectiveness of our approach over several state-of-the-art baselines.

[2]This problem cannot be solved by simple counting, as we are not sure who has contributed to a particular term when a paper is written by multiple authors.

The major contributions of this paper are summarized in the following.

1) We identify a general type of task for topic modeling, i.e. designing topic models for documents with entity information.
2) We propose a novel Entity Topic Model (ETM) which solve this task by explicitly modeling the term correlation between entities and topics. We also define a Gibbs sampling-based algorithm to learn the model.
3) We demonstrate the power of our new model over several state-of-the-art baselines by using two real-world datasets.

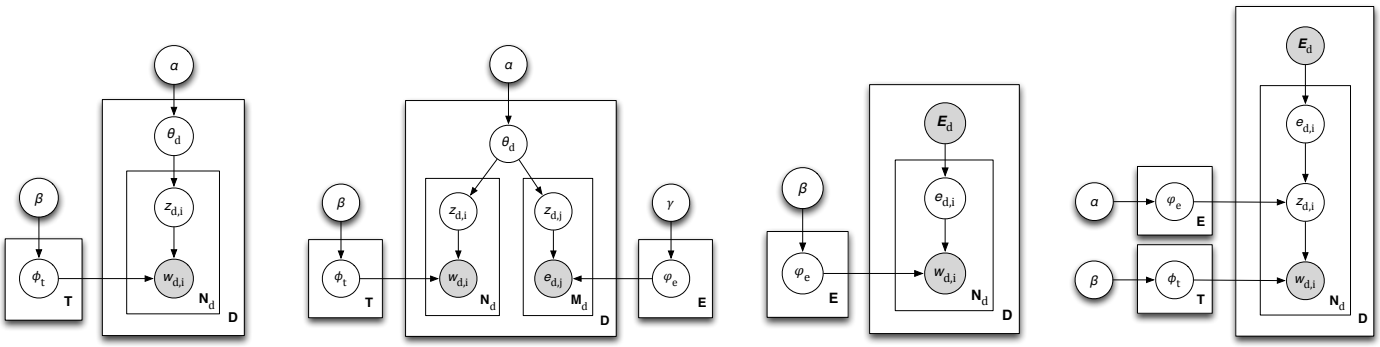
II. Related Work

In this section, we summarize several of topic models that are most closely related to the Entity Topic Model (ETM).

Latent Dirichlet Allocation (LDA) [4] is one of the most well-known topic models (Figure 2(a)). It assumes that a document is generated via a mixture of topics. In its generative process, for each document d, a multinomial distribution θ_d over topics is drawn from a Dirichlet prior with α. Then for each word, a topic $z_{d,i}$ is drawn from θ_d, and a word $w_{d,i}$ is generated by randomly sampling from a topic-specific multinomial distribution $\phi_{z_{d,i}}$. Since LDA is originally proposed for documents without any associated entities, no entities are involved in its generative process.

Link-LDA [6] (Figure 2(b)) is proposed for scientific publication with citations. In this model, documents consist of a bag of words and a bag of citations. For each document d, a multinomial distribution θ_d over topics is drawn from a Dirichlet prior with α. Then, for each word, a topic $z_{d,i}$ is drawn from θ_d, and a word $w_{d,i}$ is generated from randomly choosing from a topic specific multinomial distribution $\phi_{z_{d,i}}$ over words. For each citation, a topic $z_{d,j}$ is drawn from θ_d, and a citation $e_{d,i}$ is generated from randomly sampling from a topic specific multinomial distribution $\varphi_{z_{d,j}}$ over citations. Having similar dependencies among annotations, words, and topics, correspondence LDA [2] was also proposed.

The Author Model (AM) [9] (Figure 2(c)) is originally proposed for multi-labeled documents, where each label could represent a class or an entity. In other words, for each



(a) LDA (b) Link-LDA (c) Author Model (d) Author Topic Model

Figure 2. Four related models with different dependencies among topics(z), entities(e), and words(w)

Table I
SUMMARY OF DEPENDENCY RELATIONS EACH MODEL CAN CAPTURE

	LDA	Link-LDA	AM	ATM	ETM
$P(w\|z)$	Yes	Yes	No	Yes	Yes
$P(w\|e)$	No	No	Yes	Indirect	Yes
$P(w\|e,z)$	No	No	No	No	Yes

document d, the set of associated labels, $\boldsymbol{E}_d$, is given. For each word, a label $e_{d,i}$ is uniformly chosen from $\boldsymbol{E}_d$, and $w_{d,i}$ is generated by randomly sampling from a label-specific multinomial distribution $\varphi_{e_{d,i}}$. However, the AM only captures term distributions for each entity without investigating further the hidden patterns (topics) in documents.

The Author-Topic Model (ATM) [12] (Figure 2(d)) is an extension of Author Model by introducing topics in Author Model. It is used for modeling documents associated with multiple authors like research articles. In its generative process of document d, for each word, an author $e_{d,i}$ is uniformly sampled from the author set $\boldsymbol{E}_d$, a topic $z_{d,i}$ is drawn from author-specific multinomial distribution $\varphi_{e_{d,i}}$, and a word $w_{d,i}$ is generated by randomly sampling from a topic-specific multinomial distribution $\phi_{z_{d,i}}$. However, ATM assumes conditional independence between terms (w) and authors (e), given the knowledge of the topic (z), i.e., $P(w|e,z) = P(w|z)$, which is usually not the case.

We summarize the dependency structures that these four models as well as our ETM model can capture in Table I. ETM is the only model that can capture the correlation of term distributions between entities and topics by explicitly modeling $P(w|e,z)$.

III. ENTITY TOPIC MODEL

In this section, we formally define our problem, introduce our topic model, and finally provide a Gibbs sampling-based learning algorithm.

A. Overview of the Problem

The input to the ETM model is a collection of documents in which each document has a set of associated entities. A document d is associated with a term vector, $\boldsymbol{w}_d$, where each $w_{d,i}$ is chosen from the vocabulary of W, and an entity vector $\boldsymbol{E}_d$, chosen from a set of entities of size E. A collection of D documents is defined by $\mathcal{D} = \{\langle \boldsymbol{w}_1, \boldsymbol{E}_1 \rangle, \ldots, \langle \boldsymbol{w}_D, \boldsymbol{E}_D \rangle\}$. (The notation used in this paper is summarized in Table II). The goal is to discover word patterns for each pair of an entity and a topic. In other words, we want to discover the hidden topics in the documents, as well as the word distributions for a given entity e and a topic z, $P(w|e,z)$, which follows a multinomial distribution with parameter $\psi_{e,z}$.

The biggest challenge is that there are too many parameters to be estimated when modeling $P(w|e,z)$ directly. With E entities, T topics, and W words in a given collection, we need to estimate $O(ETW)$ parameters, which will most

Table II
NOTATION USED IN THIS PAPER

Symbol	Description
D	number of documents
T	number of topics
W	number of words
E	number of entities
N_d	number of word tokens in document d
θ_d	multinomial distribution of topics specific to document d
ϑ_d	multinomial distribution of entities specific to document d
$\boldsymbol{w}_d$	bag of words associated with document d
$\boldsymbol{E}_d$	list of entities associated with document d
$z_{d,i}$	topic associated with the ith token in document d
$e_{d,i}$	entity associated with the ith token in document d
$w_{d,i}$	ith token in document d
ϕ_z	asymmetric Dirichlet prior for topic t
φ_e	asymmetric Dirichlet prior for entity e
$\psi_{e,z}$	multinomial distribution of words specific to entity e and topic t
$\mathcal{D}$	set of all documents
$\mathcal{Z}$	set of all topic assignments $\{e_{d,i}\}$
$\mathcal{E}$	set of all entity assignments $\{e_{d,i}\}$
Φ	set of all parameters in the model

likely cause overfitting. In order to solve this problem, we propose a novel parameter smoothing method by designing hierarchical Dirichlet priors for the multinomial distribution of $P(w|e,z)$, where intuitively $P(w|e,z)$ is determined by the term distribution for the entity $P(w|e)$ and the term distribution for the topic $P(w|z)$. In particular, we use a weighted linear combination of ϕ_z and φ_e as the Dirichlet prior for $\psi_{e,z}$, where ϕ_z is an asymmetric Dirichlet parameter vector for each topic z, and φ_e is an asymmetric Dirichlet parameter vector for each entity e.

In the ETM model, we design a process for generating all the terms in a document that is associated with a given set of entities. Note that the entities that a document is associated with are not generated, but are assumed to be given. This assumption is also used in the author model and author topic model. However, in contrast to these models, we no longer assume entities are generated uniformly, but follow a multinomial distribution ϑ_d.

B. The ETM Model

The graphical representation for ETM is shown in Figure 3, and the detailed explanations are given in the following.

1) Generative Process: The hypothesis at the heart of our model is that different entities are described with different word patterns or word distributions, and that the words used to describe an entity can change with the topic. In other words, $P(w|e_i,z) \neq P(w|e_j,z)$ if $e_i \neq e_j$ and $P(w|e,z_i) \neq P(w|e,z_j)$ if $z_i \neq z_j$.

As shown in Algorithm 1, for each document d, a multinomial distribution θ_d over topics is drawn from a Dirichlet prior with α_0, and another multinomial distribution ϑ_d over the associated entity set $\boldsymbol{E}_d$ is drawn from a Dirichlet prior with α_1. Note that instead of selecting an entity uniformly from $\boldsymbol{E}_d$ as in the author model and author topic model [9], [12], we draw it from a document-specific multinomial distribution ϑ_d over $\boldsymbol{E}_d$. This is due to the assumption that each entity in $\boldsymbol{E}_d$ has a different weight in generating a document d. For example, when writing a research article, different authors make different contributions. Then, to generate each word, a topic $z_{d,i}$ is drawn from θ_d, an entity $e_{d,i}$ is drawn from ϑ_d, and word $w_{d,i}$ is generated by randomly sampling from an entity and topic specific multinomial distribution $\psi_{e_{d,i},z_{d,i}}$. That is, each term is associated with a entity-topic pair, and the generation of term is dependent on both factors.

2) Shared Asymmetric Dirichlet Priors: In this section, we will explain how to model the word distributions $P(w|e,z)$ for each entity-topic pair (e,z). As addressed in Section III-B1, our model uses two contexts, entity e and topic z, to generate word w. One of the important issues for statistical language models and topic models is data sparsity, which is the phenomenon of not observing enough data in a corpus to learn accurate model parameters. Effective smoothing techniques [20] are required to alleviate this problem. A well-known smoothing technique is to use symmetric Dirichlet priors with fixed, uniform concentration parameters. This allows any topic to generate any word with non-zero probability. However, recent studies [16] have shown that the quality of topic models can be significantly enhanced by considering asymmetric Dirichlet priors, an idea we adopt in ETM. We use the intuition that the word distribution for (e,z) pair should be dependent on word distributions for both entity e and topic z, and share some similarity with both of them. For example, the word distribution for *Judea Pearl* in *Data Mining* should be similar to the word distribution for *Judea Pearl* and the word distribution for *Data Mining* separately. Therefore, the

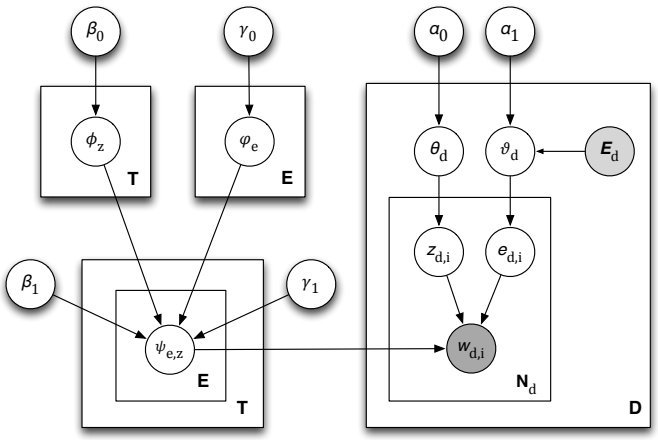


Figure 3. A graphical representation of ETM

Algorithm 1 Entity Topic Models

for each topic z **do**
  Draw $\phi_z \sim Dir(\beta_0)$
end for
for each entity e **do**
  Draw $\varphi_e \sim Dir(\gamma_0)$
end for
for each (e,z) **do**
  Draw $\psi_{e,z} \sim Dir(\beta_1\phi_z + \gamma_1\varphi_e)$
end for
for each document d **do**
  Draw $\theta_d \sim Dir(\alpha_0)$
  Draw $\vartheta_d \sim Dir(\alpha_1; \boldsymbol{E}_d)$
  for each $i \in 1,\ldots,N_d$ **do**
    Draw $z_{d,i} \sim Multi(\theta_d)$
    Draw $e_{d,i} \sim Multi(\vartheta_d)$
    Draw $w_{d,i} \sim Multi(\psi_{e_{d,i},z_{d,i}})$
  end for
end for

prior for $\psi_{e,z}$ could be designed as some function of word distributions for e and z. More specifically, suppose that we have some common word patterns φ_e for an entity e across topics, and ϕ_z for a topic z across entities. We use a linear combination of φ_e and ϕ_z as Dirichlet prior of $\psi_{e,z}$:

$$\psi_{e,z} \sim Dir(\beta_1\phi_z + \gamma_1\varphi_e)$$

Since such common word patterns are not necessary symmetric, their linear combination is asymmetric. By sharing common word patterns as priors, we can get better word smoothing, and with a much smaller parameter space, i.e., EW for φ_e and TW for ϕ_z.

C. Model Learning

We use Gibbs sampling to learn the model. Specifically, we repeatedly sample the entity-topic pair for each word in the document collection, given the entity-pair of assignment to all the rest words $(\mathcal{Z}, \mathcal{E})$ as well as the priors (Φ). This conditional posterior of assignment $(e_{d,i}, z_{d,i})$ to the ith word $w_{d,i}$ in document d is:

$$\begin{aligned} & P(z_{d,i}, e_{d,i} | w_{d,i}, \mathcal{Z}_{\setminus d,i}, \mathcal{E}_{\setminus d,i}, \Phi) \\ \propto \quad & P(w_{d,i} | z_{d,i}, e_{d,i}, \mathcal{Z}_{\setminus d,i}, \mathcal{E}_{\setminus d,i}, \Phi) \\ & P(z_{d,i} | \mathcal{Z}_{\setminus d,i}, \Phi) \\ & P(e_{d,i} | \mathcal{E}_{\setminus d,i}, \Phi) \end{aligned} \tag{1}$$

where sub- or super-script "$\setminus d, i$" denotes a quantity excluding data from position i in document d.

The second and third terms on the right-hand side are straightforward:

$$P(z_{d,i} | \mathcal{Z}_{\setminus d,i}, \Phi) \propto \frac{N^{\setminus d,i}_{z_{d,i}|d} + \frac{\alpha_0}{T}}{N_d - 1 + \alpha_0} \tag{2}$$

$$P(e_{d,i} | \mathcal{E}, \Phi) \propto \frac{N^{\setminus d,i}_{e_{d,i}|d} + \frac{\alpha_1}{|\boldsymbol{E}_d|}}{N_d - 1 + \alpha_1} \tag{3}$$

where $N^{\setminus d,i}_{z_{d,i}|d}$ is the number of word tokens assigned with topic $z_{d,i}$ except ith token in document d, and $N^{\setminus d,i}_{e_{d,i}|d}$ is the number of word tokens assigned with entity $e_{d,i}$ except ith token in document d.

In order to better understand the first term on the right-hand side, we describe its generative process[3]. Figure 4 depicts the process of drawing nine words from the Dirichlet-multinomial $\psi_{e,z}$ that has $\beta_1\phi_z + \gamma_1\varphi_e$ as its prior. This process introduces a set of internal draws $\{\sigma_1, \sigma_2, \dots\}$. Those internal draws are chosen when a word is generated from $\psi_{e,z}$. When drawing the first word, there are no previous internal draws, and σ_1 is drawn from either ϕ_z with probability $\frac{\beta_1}{\beta_1+\gamma_1}$ or φ_e with probability $\frac{\gamma_1}{\beta_1+\gamma_1}$. In the example of Figure 4, σ_1 is drawn from ϕ_z. The second word is drawn by selecting σ_1 with probability proportional to the number of previous words that are from σ_1, a new draw from ϕ_z with probability proportional to β_1, or a new draw from φ_e with probability proportional to γ_1. In the case of Figure 4, the second word is drawn by the new draw σ_2 from ϕ_z. The next words are drawn with the same procedure.

[3]Our word generative process is an extension of the generative process described in [16], where they have only one base measure while ours has two base measures.

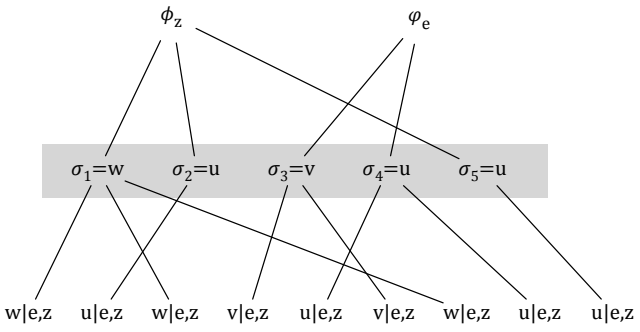


Figure 4. The generative process of nine words from $\psi_{e,z}$ that has $\beta_1\phi_z + \gamma_1\varphi_e$ as its prior

Let $N_{w|e,z}$ denote the number of word-w tokens assigned with the pair (e, z), $\hat{N}_{w|z}$ denote the number of internal draws in $\{\sigma_1, \sigma_2, \dots\}$ whose values are w drawn from ϕ_z, and $\hat{N}_{w|e}$ denote the number of internal draws in $\{\sigma_1, \sigma_2, \dots\}$ whose values are w drawn from φ_e. Also, let $N_{\cdot|e,z} = \sum_{w \in W} N_{w|e,z}$, $\hat{N}_{\cdot|z} = \sum_{w \in W} \hat{N}_{w|z}$, and $\hat{N}_{\cdot|e} = \sum_{w \in W} \hat{N}_{w|e}$. Then, the predictive probability of word w in given z, e, $\mathcal{Z}$, $\mathcal{E}$, and Φ is:

$$\begin{aligned} & P(w | z, e, \mathcal{Z}, \mathcal{E}, \Phi) \\ & = \frac{N_{w|e,z} + \beta_1 \frac{\hat{N}_{w|z} + \frac{\beta_0}{W}}{\hat{N}_{\cdot|z} + \beta_0} + \gamma_1 \frac{\hat{N}_{w|e} + \frac{\gamma_0}{W}}{\hat{N}_{\cdot|e} + \gamma_0}}{N_{\cdot|e,z} + \beta_1 + \gamma_1} \end{aligned} \tag{4}$$

By combining Equation 2, 3, and 4, we can compute Equation 1.

Once we obtain entity-topic pair assignments for each word, we can estimate the parameters in the model accordingly.

D. Discussions on Special Cases

Another advantage of our model is that it has connections to previous topic models, and it turns out that LDA and the Author Model are both special (limiting) cases of our model.

If the concentration parameter β_1 is large and γ_1 is small relative to $N_{e,z}$, then counts $N_{e,z}$ are effectively ignored, and lead to have $P(w|z, e, \mathcal{Z}, \mathcal{E}, \Phi) \approx P(w|z, \mathcal{Z}, \mathcal{E}, \Phi)$. As $\beta_1 \to \infty$ and $\gamma_1 \to 0$, the role of entities in the model becomes ignored, and our model approaches to LDA.

By contrast, if concentration parameter γ_1 is large and β_1 is small relative to $N_{e,z}$, our model will have $P(w|z, e, \mathcal{Z}, \mathcal{E}, \Phi) \approx P(w|e, \mathcal{Z}, \mathcal{E}, \Phi)$. As $\beta_1 \to \infty$ and

$\gamma_1 \to 0$, the role of topics in the model becomes ignored, and our model approaches to Author Model.

IV. Experiments

We have two different datasets to evaluate our model: a news article dataset and a DBLP dataset. In the news article dataset, we collected articles about Japan's 2011 Tsunami from NewsBank[4]. A massive 8.9-magnitude earthquake shook Japan on March 11, 2011, causing a devastating tsunami to the coast of Japan. Due to the tsunami, the nuclear power plants in Fukushima were damaged, and one of the reactors in the Fukushima No. 1 nuclear plant partially melted down in the following day. As a result, the nuclear accident caused the exposure of nuclear radiation near the plant.

We searched articles with "Japan Tsunami" keywords, and collected 42,727 articles published from Mar. 11, 2011 to Apr. 11, 2011. Since news articles do not contain associated entity set explicitly, we extracted entities mentioned in the articles. We used Zemanta[5], a high-performance online entity extraction and disambiguation service that links extracted entities to Wikipedia entries. Despite of many other available entity annotation tools, Zemanta was chosen because it has very high throughput and high precision [10]. After extracting entities, we discarded infrequent entities that appear in less than 5 documents. We also removed stop words and infrequent words that appear less than 5 documents.

The Digital Bibliography and Library Project (DBLP)[6] is a collection of bibliographic information on major computer science journals and proceedings. Each paper is represented by a bag of words that appear in the abstract and title of the paper. Also, its associated entity set is defined as the set of authors. In this experiment, we use a subset of the DBLP records that belongs to four areas: databases, data mining, information retrieval and artificial intelligence. We discard authors with less than 5 publications in our corpus. We again removed stop words and infrequent words that appear less than 5 documents. The two datasets are summarized in Table III.

Our main claim is that word distributions should depend on associated entities as well as topics. For each dataset, as a case study, we show how word distributions change over topics with a fixed entity, and over entities with a fixed topic. In addition, we show rankings of entities for each topic as by-products of our model.

Finally, we compare our model with several baselines in terms of perplexity, and investigate the parameters of our model.

For Japan's Tsunami dataset, we used $T = 20$, $\beta_1 = 100$, $\gamma_1 = 10$, and set other hyperparameters to 0.1. For the DBLP dataset, we used $T = 50$, $\beta_1 = 1000$, $\gamma_1 = 1$, and set other hyperparameters to 0.1. We used a relatively small number of topics when visually investigating word distributions from ETM. The evaluation of our model for different number of topics will be addressed in Section IV-C. The hyperparameters will be addressed in Section IV-D.

[4]http://www.newsbank.com/

[5]http://zemanta.com

[6]http://www.informatik.uni-trier.de/~ley/db/

Table III
Two Datasets with Statistics

Dataset Name	D	E	W	$avg(\lvert \boldsymbol{E}_d \rvert)$	$avg(N_d)$
Japan Tsunami	2,000	596	10,104	11.49	243.30
DBLP	20,860	3,251	11,609	1.79	96.51

A. Case Study 1: Japan's Tsunami (2011)

Since T is set to 20, we get 20 topics, including Tsunami, Nuclear Accident, Nuclear Radiation, Economic Effects, Industrial Effects, Relief Efforts, Tsunami Rescue, and so on[7].

Naoto Kan, who was the prime minister of Japan during the incident, was frequently mentioned in the corpus. He was involved in many topics like Relief Efforts, Nuclear Accident, and Economic Effects.

First, the top 20 words in the entity prior φ_e of Naoto Kan are shown in the first column in Table IV[8]. The entity prior can be interpreted as entity-related and topic-independent word distribution for Naoto Kan. Combining with topic priors, the entity prior helps to shape the word distributions ($\psi_{e,z}$) of Naoto Kan in different contexts.

To support our main claim, we compare the word distributions ($\psi_{e,z}$) for Naoto Kan across different topics. Here, we show Naoto Kan in three different topics – Relief Efforts, Nuclear Accident, and Economic Effects. The top words are listed in the rest of columns in Table IV based on their $\psi_{e,z}$ values. Note that there are "troops", "soldiers", "bodies", and "search" in Relief Efforts since the Japanese government had sent 50,000 troops for the rescue and recovery efforts, and "yukio" in Nuclear Accident refers to Yukio Edano who was the chief secretary of Japan's cabinet, leading the government to combat the aftermath of Nuclear Accident. As shown in Table IV, the word distributions ($\psi_{e,z}$) related to Naoto Kan vary significantly across the topics.

In the first column in Table V, the top 20 words of the topic prior ϕ_z of Relief Efforts are listed. The topic prior can be interpreted as topic-related and entity-independent word distribution for Relief Efforts. The topic prior help to learn the word distributions ($\psi_{e,z}$) related to the entities involved in Relief Efforts.

We compare the word distributions ($\psi_{e,z}$) of three entities in the context of Relief Efforts – American Red Cross, Korea, and Tokyo. Even though American Red Cross and Korea are entities that had supported the Japanese people, their word distributions ($\psi_{e,z}$) are different: Korea has "sympathy" and

[7]The topics are manually named based on their word distributions.

[8]For simplicity, we omitted parameter values, and listed the top words

Table IV
NAOTO KAN'S ENTITY PRIOR (φ_e) AND WORD DISTRIBUTIONS ($\psi_{e,z}$) OF HIS RELATED TOPICS

Naoto Kan	Relief Efforts	Nuclear Accident	Economic Effects
kan	bodies	kan	prime
minister	search	minister	rule
prime	kan	prime	bill
naoto	people	naoto	kan
government	troops	nuclear	powerful
tokyo	car	radiation	business
crisis	crisis	plant	minister
troops	prime	evacuated	naoto
friday	confirmed	yukio	mind
party	business	reactors	term
assistance	told	urged	starting
democratic	minister	time	past
asked	lost	televised	march
kans	concrete	complex	loans
house	coastal	situation	financing
situation	centers	cabinet	economic
mr	center	fears	disaster
efforts	soldiers	crippled	april
conference	naoto	indoors	kans
spokeswoman	leaks	statement	yen

Table V
RELIEF EFFORT'S TOPIC PRIOR (ϕ_z) AND WORD DISTRIBUTIONS ($\psi_{e,z}$) OF ITS RELATED ENTITIES

Relief Efforts	American Red Cross	Korea	Tokyo
japan	cross	japan	people
japanese	red	japanese	japan
people	japan	korea	friends
tsunami	american	korean	japanese
earthquake	relief	donations	tokyo
disaster	support	koreans	tsunami
world	donations	sympathy	back
relief	donation	march	earthquake
money	disaster	earthquake	home
time	raise	helping	email
country	march	hard	devastating
damage	efforts	victims	family
friends	affected	support	earthquakes
information	tsunami	collected	student
aid	victims	quake	miles
week	earthquake	personal	concerned
affected	money	people	watch
nation	thursday	news	live
march	people	money	concern
devastation	located	important	close

"personal" in the top 20 words, and American Red Cross has "efforts" and "raise". Tokyo has the words "family", "friend", "home", and "email" because many articles mentioned that many people contacted with their family or friends in Tokyo via phone and e-mail. As shown in Table V, the word distributions ($\psi_{e,z}$) related to Relief Efforts also change over the related entities and fit more to the entities.

As by-products, we can rank entities for each topic and rank topics for each entity. In contrast to ATM [12], our model does not model the relationship between entities and topics directly. Our model, however, can get their relationship indirectly for a given assignments $\mathcal{E}$ and $\mathcal{Z}$. Let $N_{\cdot|e,z}$ denote the number of words that are assigned with (e, z). Also, let $N_{\cdot|\cdot,z} = \sum_e N_{\cdot|e,z}$ and $N_{\cdot|e,\cdot} = \sum_z N_{\cdot|e,z}$. Then, $P(e|z, \mathcal{E}, \mathcal{Z}, \Phi) = \frac{N_{\cdot|e,z}}{N_{\cdot|\cdot,z}}$, and $P(z|e, \mathcal{E}, \mathcal{Z}, \Phi) = \frac{N_{\cdot|e,z}}{N_{\cdot|e,\cdot}}$. Based on $P(e|z, \mathcal{E}, \mathcal{Z}, \Phi)$ and $P(z|e, \mathcal{E}, \mathcal{Z}, \Phi)$, we can rank entities for each topic, and rank topics for each entity.

Table VI shows two topics and their entity rankings. Nuclear Accident and Nuclear Radiation have three entities in common in the top entities: Tokyo Electric Power Company, Fukushima Nuclear Power Plant, and Potassium iodide. Tokyo Electric Power Company is the operating company of Fukushima Nuclear Power Plant, and one of the nuclear reactors in Fukushima Nuclear Power Plant had been damaged and started to melt down. Potassium iodide is an inorganic compound that is used as drugs to prevent Thyroid cancer caused by radioactive chemicals. However, the rest of entities are very different. Note that there are Nuclear Regulatory Commission, U.S. Environmental Protection Agency, and Seawater in the top entities of Nuclear Accident: Nuclear Regulatory Commission oversees nuclear reactor safety, U.S. Environmental Protection Agency protects human health and the environment by enforcing related regulations, and Seawater was used to cool down the nuclear reactor. On the other hand, there are Iodine-131, Caesium, Thyroid, and Tap Water in the top entities of Nuclear Radiation: Iodine-131 and Caesium are the emitters of strong gamma radiation that causes cancers and even death. Those radioactive chemicals can dissolve in water, and people may get exposed to the radioactive chemicals by drinking Tap Water.

Table VI
ENTITY RANKINGS FOR DIFFERENT TOPICS

Nuclear Accident	Nuclear Radiation
Nuclear Regulatory Commission	Tokyo Electric Power Company
Nuclear power plant	Fukushima Nuclear Power Plant
Chernobyl disaster	Electrical grid
Japan	Tap water
Tokyo Electric Power Company	Caesium
Libya	Iodine-131
Potassium iodide	Thyroid
U.S. Environmental Protection Agency	Radiation
Seawater	Yukio Edano
Fukushima Nuclear Power Plant	Raw Milk
Barack Obama	Potassium iodide
Automotive industry	Thyroid cancer

Similarly, it is possible to analyze topic rankings for each entity.

B. Case Study 2: DBLP – Research Articles

In this section, we performed a similar analysis with the DBLP corpus. As introduced in Section I, Judea Pearl is the 2011 winner of the A.M. Turing Award for *"for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning."*[9] He is credited for inventing Bayesian networks,

[9]http://amturing.acm.org/award_winners/pearl_2658896.cfm

Table VII
JUDEA PEARL'S ENTITY PRIOR (φ_e) AND WORD DISTRIBUTIONS ($\psi_{e,z}$) OF HIS RELATED RESEARCH TOPICS

Judea Pearl	Knowledge Representation	Reasoning	Bayesian Network
causal	reasoning	logic	causal
revisited	default	dependencies	distributions
optimality	causal	probabilistic	models
markovian	formal	graphs	markovian
counterfactual	systems	directed	semi
explanations	computational	representing	identification
symbolic	specificity	dags	characterization
independence	causality	programs	recursive
path	diagnostic	reasoning	joint
specificity	inheritance	conditional	variables
scout	representation	bases	data
independencies	model	based	effects
dependence	knowledge	efficient	algorithm
proven	system	networks	based
embracing	inference	programming	clustering
dags	common	probability	network
tolerating	rule	undirected	arbitrary
economy	embracing	causal	graph
states	coherence	inference	bayesian
counterfactuals	belief	belief	networks

Table VIII
KNOWLEDEGE REPRESETNTATION'S TOPIC PRIOR (ϕ_z) AND WORD DISTRIBUTIONS ($\psi_{e,z}$) OF ITS RELATED ENTITIES

Knowledge Representation	Pedro Domingos	Benjamin Kuipers	Marzena Kryszkiewicz
knowledge	logic	qualitative	frequent
reasoning	markov	simulation	patterns
system	networks	reasoning	representation
representation	learning	knowledge	free
based	world	quantitative	disjunction
design	mlns	systems	generalized
systems	knowledge	incomplete	concise
qualitative	order	mechanism	based
logic	structure	abstraction	generators
theory	logical	envisionment	representations
learning	real	physical	negations
domain	models	behavior	oriented
planning	unifying	causal	support
problem	ilp	system	knowledge
agent	systems	description	sets
expert	purely	expert	condensed
model	representation	process	reasoning
approach	reasoning	behaviors	survey
models	viewing	logic	system
support	mln	model	borders

and several inference methods in the models. He later developed a theory of causal and counterfactual inference based on structural models.

First, the top 20 words in the entity prior φ_e of Judea Pearl are shown in the first column in Table VII. The entity prior can be interpreted as the word distribution of his general methodologies, approaches, or research interests for Judea Pearl. There are "casual" and "counterfactual" in his entity prior φ_e, indicating his research interests are casual and counterfactual inference across his research topics. Combining with topic priors, his entity prior helps to shape the word distributions $\psi_{e,z}$ of Judea Pearl in different research topics.

Based on $P(z|\text{Judea Pearl}, \mathcal{E}, \mathcal{Z})$, we selected his top 3 research topics: Knowledge Represetation, Reasoning, and Bayesian Network[10]. With his general approaches "casual" and "counterfactual", he has involved in these research topics. The top words of the word distributions ($\psi_{e,z}$) are listed in the rest of the columns in Table VII, indicating how his approach is applied in different research topics.

For Knowledge Representation, we select the top three authors based on $P(e|\text{Knowledge Representation}, \mathcal{E}, \mathcal{Z})$. Even though they have published papers on Knowledge Representation, their approaches are very different from each other. Pedro Domingos has focused on learning Markov logic networks, Benjamin Kuipers has developed qualitative models to express states of incomplete knowledge about continuous mechanisms and QSIM algorithm for qualitative simulation. Marzena Kryszkiewicz has taken very different approaches that use frequent patterns to generate rules as knowledge. Even though the three authors have very different approaches for the same research topic, and thus have very different word distributions, our model aligns them under Knowledge Representation topic to make them comparable. This comparison is not possible without modeling $P(w|e, z)$.

As we did for Japan's Tsunami dataset, we can rank entities for each topic, and topics for each entity, as shown in other studies [12], [14]. Due to the space limit, we will not show them in this study.

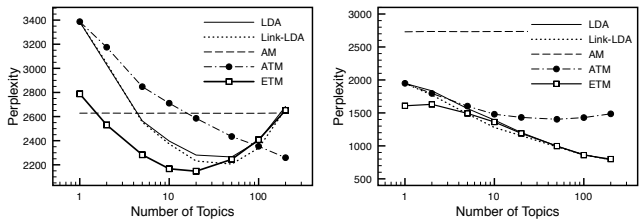
C. Perplexity Analysis

We compare our model with several baselines introduced in Section II: LDA [4], Link-LDA [6], AM [9], and ATM [12]. Their hyperparameters are set to 0.1 except ATM, where the author suggested its hyperparameter settings: $\alpha = \frac{50}{T}$ and $\beta = 0.01$ in Figure 2(d). Perplexity is a standard measure for estimating the performance of a probabilistic model. We evaluate our model and compare with the baselines by estimating the perplexity of unseen held-out documents given some training documents. A better model will have a lower perplexity of held-out documents, on average. Perplexity is defined as $exp(-\frac{\log P(\mathcal{D}^{\text{test}}|\mathcal{D}^{\text{train}})}{\sum_{d\in\mathcal{D}^{\text{test}}} N_d})$. Let Φ denote the set of all parameters in a topic model. Then,

$$P(\mathcal{D}^{\text{test}}|\mathcal{D}^{\text{train}}) = \int P(\mathcal{D}^{\text{test}}|\Phi)P(\Phi|D^{\text{train}})\mathrm{d}\Phi$$

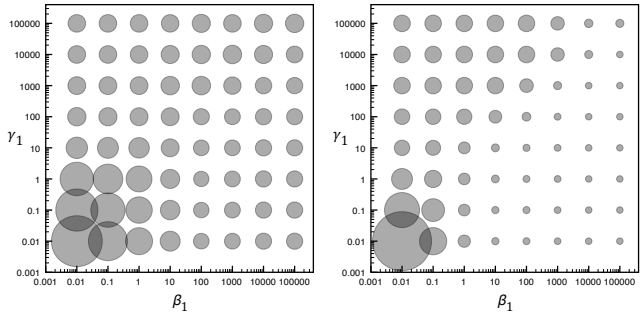
This integral can be approximated by averaging $P(\mathcal{D}^{\text{test}}|\Phi)$ under samples from $P(\Phi|D^{\text{train}})$. We used a Gibbs sampling to get 20 samples of Φ and *left-to-right* evaluation algorithm [17] to approximate $P(\mathcal{D}^{\text{test}}|\Phi)$. Note that AM, ATM, and ETM have generative processes of words for a given set of entities. Thus, $P(\mathcal{D}^{\text{test}}|\Phi)$ is defined as follows:

$$P(\mathcal{D}^{\text{test}}|\Phi) = \prod_{d\in\mathcal{D}^{\text{test}}} P(\boldsymbol{w}_d|\boldsymbol{E}_d, \Phi)$$



(a) Japan's Tsunami (b) DBLP

Figure 5. Perplexity values for different number of topics



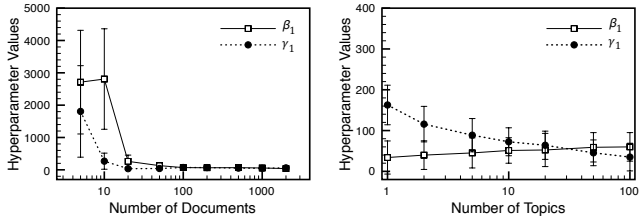
(a) Japan's Tsunami (b) DBLP

Figure 6. Perplexity values for different β_1 and γ_1. The size of circle at each data point is proportional to its perplexity value.

We randomly sample 80% of the data as $\mathcal{D}^{\text{training}}$ and use the remaining 20% as $\mathcal{D}^{\text{test}}$. Figure 5 shows the perplexity values of our model and the baselines for different number of topics. Note that because AM does not have topics in its model, it has the same value regardless of the number of topics. Also, because LDA does not have entities in its model, LDA cannot take advantage of given associated entity sets. Generally, Link-LDA is slightly better than LDA because it uses the given associated entity sets as extra information to learn topic distributions in documents. Since ATM models a document generative process for a given set of entities, it is expected to have lower perplexity values than LDA. However, their experiments [12] with the corpus of NIPS papers showed that ATM has higher perplexity values than LDA because ATM model has large number of parameters to be estimated, limiting its generalization performance. For DBLP dataset, ATM also has higher perplexity values than LDA as shown in Figure 5(b).

In Japan's Tsuanmi dataset, the perplexity values of LDA and Link-LDA decrease until they reach the lowest values at $T = 50$, and then begin to increase. When $T > 50$, LDA and Link-LDA have too many parameters in their models, causing an overfitting problem. The perplexity value of ETM decrease until it reaches to the lowest value at $T = 20$, and then begin to increase due to the overfitting problem like LDA and Link-LDA. Until $T = 20$, ETM outperforms the four baselines, and its lowest perplexity value is lower than the lowest perplexity values of the other models.

[10]These topics are manually named based on their topic priors.



(a) For different number of documents (b) For different number of topics

Figure 7. The changes of the sampled parameters β_1 and γ_1 over the number of documents and the number of topics in Japan's Tsunami dataset

In the DBLP dataset, ETM has similar perplexity values as LDA and Link-LDA. The main reason is that most of the words in the research articles are related to research topics, and entity-specific topic-independent words are relatively rare in the corpus. For example, some coined words by an author can be entity-specific and topic-independent words, but such words are relatively rare unless the terms become popular in their related research communities. ETM, however, is still the best among all the models when the number of topics is small, and comparative to LDA and Link-LDA when number of topics is increasing, and much better than ATM for all the settings.

D. Parameter Studies

Among the six hyperparameters in our model, β_1 and γ_1 play the most important role. Depending on their values, our model slides between LDA and AM. For a given collection of documents, these parameters can be tuned by the perplexity analysis. Figure 6 shows the perplexity values for different values of β_1 and γ_1. For each pair of β_1 and γ_1, the size of circle is proportional to its perplexity value (smaller is better). For Japan's Tsunami dataset, ETM has the lowest perplexity value at $\beta_1 = 100$ and $\gamma_1 = 10$. For DBLP dataset, ETM has the lowest perplexity value at $\beta_1 = 1000$ and $\gamma_1 = 1$. With Figure 6, we can find appropriate parameter values for β_1, and γ_1. In addition, we can understand the characteristics of the corpus: topic-related words are dominant in DBLP dataset, while topic-related words and entity-related words are relatively balanced in Japan's Tsunami dataset.

Instead of enumerating parameter values and evaluating to find appropriate values, we can estimate them directly from a given corpus by sampling. As many studies suggested [16], [15], concentration parameters like β_1 and γ_1 can be given broad Gamma priors and inferred using slice sampling [11].

For Japan's Tsunami dataset, we sampled β_1 and γ_1. First, we sample them for different number of documents. In Figure 7(a), when training documents are very few, the sampled hyperparameters β_1 and γ_1 become large, leading to reduce its parameter space by weighting more on priors.

Next, we sampled β_1 and γ_1 for different number of topics. As shown in Figure 7(b), β_1 increases as the number of topics increases. This is due to the quality of topics. When the model has better quality of topics, the word distributions $P(w|e, z)$ depend more on the topic z than the entity e.

V. Conclusions

In this paper, we identify the problem of designing topic models for documents with entity information. A novel Entity Topic Model (ETM) is proposed to solve the problem, which can explicitly model the word co-occurrences in pairs of a topic and entity, with smartly designed priors. Having shared asymmetric Dirichlet priors, our model reduces the size of its parameter space while learning a large number of parameters. A Gibbs sampling-based algorithm is proposed to learn the model.

We demonstrate the power of our model in two case studies with real-world datasets. The case studies show that entities and topics are correlated with words, and our model captures the patterns among entities, topics, and words. Even though ETM does not model the relationship between entities and topics directly, the case studies show that such patterns can be computed indirectly and used for ranking entities for each topic and topics for each entity in terms of their correlation.

Finally, we compare our model with several state-of-the-art baselines in terms of the predictability and generalizability. The evaluation shows that our model is better than the baselines for small number of topics and comparable for large number of topics while providing richer patterns among entities, topics, and words.

VI. Acknowledgements

The work was supported in part by U.S. National Science Foundation grants IIS-0905215, CNS-0931975, CCF-0905014, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and DTRA. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] R. Balasubramanyan and W. W. Cohen. Block-lda: Jointly modeling entity-annotated text and entity-entity links. In *SDM '11*, pages 450–461, 2011.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03*, pages 127–134, 2003.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] W. W. Cohen and S. Sarawagi. Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods. In *KDD '04*, pages 89–98, 2004.

[6] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5220–5227, 2004.

[7] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, Dec. 2008.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57, 1999.

[9] A. Mccallum. Multi-label text classification with a mixture model trained by em. In *AAAI '99 Workshop on Text Learning*, 1999.

[10] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *I-Semantics '11*, pages 1–8, 2011.

[11] R. M. Neal. Slice sampling. *Annals of Statistics*, 31:705–767, 2003.

[12] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI '04*, pages 487–494, 2004.

[13] S. Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM '10*, pages 281–290, 2010.

[14] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD '09*, pages 797–806, 2009.

[15] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):pp. 1566–1581, 2006.

[16] H. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In *NIPS '09*, 2009.

[17] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *ICML '09*, 2009.

[18] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD '06*, pages 424–433, 2006.

[19] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *WWW '11*, pages 247–256, 2011.

[20] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[21] D. Zhang, C. Zhai, and J. Han. Topic cube: Topic modeling for olap on multidimensional text databases. In *SDM '09*, pages 1123–1134, 2009.