Grounding Topic Models with Knowledge Bases

Zhiting Hu[‡*], **Gang Luo**[§], **Mrinmaya Sachan**[‡], **Eric Xing**[‡], **Zaiqing Nie**[†]
[†]Microsoft Research, Beijing, China [§]Microsoft, California, USA
[‡]School of Computer Science, Carnegie Mellon University
{zhitingh,mrinmays,epxing}@cs.cmu.com, {gluo,znie}@microsoft.com

Abstract

Topic models represent latent topics as probability distributions over words which can be hard to interpret due to the lack of grounded semantics. In this paper, we propose a structured topic representation based on an entity taxonomy from a knowledge base. A probabilistic model is developed to infer both hidden topics and entities from text corpora. Each topic is equipped with a random walk over the entity hierarchy to extract semantically grounded and coherent themes. Accurate entity modeling is achieved by leveraging rich textual features from the knowledge base. Experiments show significant superiority of our approach in topic perplexity and key entity identification, indicating potentials of the grounded modeling for semantic extraction and language understanding applications.

1 Introduction

Probabilistic topic models [Blei *et al.*, 2003] have been one of the most popular statistical frameworks to identify latent semantics from large text corpora. The extracted topics are widely used for human exploration [Chaney and Blei, 2012], information retrieval [Wei and Croft, 2006], machine translation [Mimno *et al.*, 2009], and so forth.

Despite their popularity, topic models are weak models of natural language semantics. The extracted topics are difficult to interpret due to incoherence [Chang *et al.*, 2009] and lack of background context [Wang *et al.*, 2007]. Furthermore, it is hard to grasp semantics merely as topics formulated as word distributions without any grounded semantics [Song *et al.*, 2011; Gabrilovich and Markovitch, 2009]. Though recent research has attempted to exploit various knowledge sources to improve topic modeling, they either bear the key weakness of representing topics merely as distribution over words or phrases [Mei *et al.*, 2014; Boyd-Graber *et al.*, 2007; Newman *et al.*, 2006] or sacrifice the flexibility of topic models by imposing a one-to-one binding of topics to pre-defined knowledge base (KB) entities [Gabrilovich and Markovitch, 2009; Chemudugunta *et al.*, 2008].

This paper aims to bridge the gap, by proposing a new structured representation of latent topics based on *entity taxonomies* from KBs. Figure 1 illustrates an example topic extracted from a news corpus. Entities organized in the hierarchical structure carry salient context for human and machine interpretation. For example, the relatively high weight of entity *Amy Winehouse* can be attributed to the fact that Winehouse and Houston were both prominent singers who have passed from drug-related causes. In addition, the varying weights associated with taxonomy nodes ensure flexibility to express the gist of diverse corpora. The new modeling scheme poses challenges for inference as both topics and entities are hidden from observed text and the topics are regularized by hierarchical knowledge. We develop Latent Grounded Semantic Analysis (LGSA), a probabilistic generative model, to infer both topics and entities from text corpora. Each topic is equipped with a random walk over the taxonomy which naturally integrates the structure to ground the semantics as well as leverages the highly-organized knowledge to capture entity correlations. For accurate entity modeling, we augment bag-of-word documents with entity mentions and incorporate rich textual features of entities from KBs. To keep inference over large corpora and KBs practical, we use ontology pruning and dynamic programming.

Extensive experiments validate the effectiveness of our approach. LGSA improves topic quality in terms of perplexity significantly. We apply the model to identity key entities of documents (e.g., the dominant figures of a news article). LGSA achieves 10% improvement (precision@1 from 80% to 90%) over the best performing competitors, showing strong potential in semantic search and knowledge acquisition. To our knowledge, this is the first work to combine statistical topic representation with structural entity taxonomy. Our probabilistic model that incorporates rich world knowledge provides a potentially useful scheme to accurately induce grounded semantics from natural language data.

2 Related Work

Topic modeling Probabilistic topic models such as LDA [Blei *et al.*, 2003] identify latent topics purely based on observed data. However, it is well known that topic models are only a weak model of semantics. Hence, a large amount of recent work has attempted to incorporate domain knowledge [Foulds *et al.*, 2015; Yang *et al.*, 2015; Mei *et al.*, 2014;

[*]This work was done when the first two authors were at Microsoft Research, Beijing.

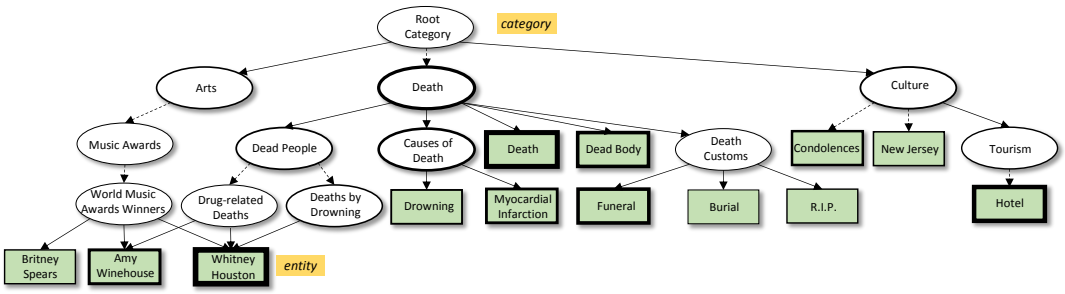

Figure 1: Structured representation of topic "Death of Whitney Houston" from our model. Entities (leaf rectangular nodes) and categories (internal elliptical nodes) with highest probabilities are shown. The thickness of each node's outline is proportional to the weight of that node. The dashed links denote multiple edges that have been consolidated.

Chen and Liu, 2014; Andrzejewski *et al.*, 2011] or relational information [Chang and Blei, 2009; Hu *et al.*, 2015b] as regularization in topic modeling. Yet, a clear shortcoming in these models is that topics are simply modeled as word distributions without grounded meanings. LGSA mitigates this by grounding topics to KB entities. Another line of research attempts to *explicitly* map document semantics to human-defined concepts [Gabrilovich and Markovitch, 2009; Chemudugunta *et al.*, 2008]. These methods assume one-to-one correspondence between topics and a small set of ontological concepts. Though enjoying clear meaning, this work sacrifices expressiveness and compactness of latent semantic models. LGSA ensures both interpretability and flexibility by extracting latent yet grounded topics. Topic models have also been used to model KB entities for the entity linking task [Han and Sun, 2012; Kataria *et al.*, 2011] which has a different focus from ours.

Using hierarchical knowledge Semantic hierarchies are key knowledge sources [Resnik, 1995; Hu *et al.*, 2015a]. Few generative models have been developed for specific tasks which integrate hierarchical structures through random walks [Kataria *et al.*, 2011; Hu *et al.*, 2014; Boyd-Graber *et al.*, 2007]. E.g., [Boyd-Graber *et al.*, 2007] exploits WordNet-Walk [Abney and Light, 1999] for word sense disambiguation. Our work is distinct in that we use entity taxonomy to construct a representation of topics; moreover, we infer hidden entities from text, leading to unique inference complexity. We propose an efficient approach to tackle this issue. Note that our work also differs from hierarchical topic models [Griffiths and Tenenbaum, 2004; Movshovitz-Attias and Cohen, 2015] which aim to infer latent hierarchies from data rather than ground latent semantics to existing KBs.

3 Latent Grounded Semantic Analysis

Model Overview: LGSA is an unsupervised probabilistic model that goes beyond the conventional word-based topic modeling, and represents latent topics based on the highly-organized KB entity taxonomies. We first augment the conventional bag-of-word documents with entity *mentions* in order to capture salient semantics (§3.1). An entity is modeled as distributions over both mentions and words. Here we leverage entities' rich textual features from KBs for accurate entity modeling (§3.3). Finally, each topic is associated with a root-to-leaf random walk over the entity taxonomy. This endows the topic with a semantic structure, as well as captures the valuable entity correlation knowledge from the taxonomy (§3.2). A well-defined generative process combines the above components in a joint framework (§3.4). Next, we present the details of each component (Table 1 lists key notations; Figure 2 illustrates a running example of LGSA; and the graphical model representation of LGSA is shown in Figure 3).

Symbol	Description
$\mathcal{E}, \mathcal{M}, \mathcal{V}$	the set of entities, mentions, and words
D, K, E, M, V	#docs, #topics, vocabulary sizes of entities/mentions/words
M_d, V_d	# of mention and word occurrences in doc d
m_{dj}	the jth mention in doc d
w_{dl}	the lth word in doc d
z_{dj}	the topic associated with m_{dj}
e_{dj}	the entity associated with m_{dj}
$\boldsymbol{r}_{dj}$	the path in taxonomy associated with m_{dj}
y_{dl}	the entity index associated with w_{dl}
$\boldsymbol{\theta}_d, \boldsymbol{\theta}'_d$	multinomial distribution over topics and entities of doc d
$\boldsymbol{\Lambda}_k$	random walk transition distributions of topic k
$\boldsymbol{\phi}_k$	multinomial distribution over entities of topic k
$\boldsymbol{\tau}_k$	probabilities over category nodes of topic k
$\boldsymbol{\eta}_e, \boldsymbol{\zeta}_e$	multinomial distribution over mentions and words of entity e
$\boldsymbol{p}_e^\eta, \boldsymbol{p}_e^\zeta$	base measures of priors over $\boldsymbol{\eta}_e$ and $\boldsymbol{\zeta}_e$ extracted from KBs
$\lambda_e^\eta, \lambda_e^\zeta$	concentration parameters (prior strengths)

Table 1: Notations used in this paper.

3.1 Document Modeling

Topic models usually represent a document as a bag of words. However, language has rich structure and different word classes perform different functions in cognitive understanding of text. The noun class (e.g., entity mentions in a news article) is an important building block and carries much of the salient semantics in a document. Semanticists have often debated the *cognitive economy* of the noun class [Kemp and Regier, 2012]. If every mention had a unique name, this would eliminate ambiguities. However, our memory is constrained and it is impossible to remember all mention names. Hence, ambiguity is essential. Our work grounds mentions to KBs, leading to consistent interpretation of entities.

As the first step, our model augments the bag-of-word representation with mentions. Each document $d \in \{1, \dots, D\}$ is

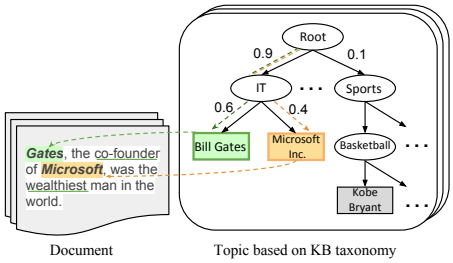

Figure 2: Model overview. Mentions *Gates* and *Microsoft* (highlighted) in the document refer to entities in the KB taxonomy. Words *co-founder* and *wealthiest* (underlined) are describing entity *Bill Gates*. A topic random walk is parameterized by the parent-to-child transition probabilities.

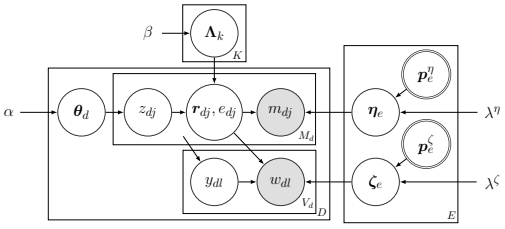

Figure 3: Graphical representation of LGSA. The entity e_{dj} is the leaf of the path $\boldsymbol{r}_{dj}$. Document's entity distribution $\boldsymbol{\theta}'_d$ is derived from other variables and thus does not participate in the generative process directly.

now represented by a set of words $\boldsymbol{w}_d = \{w_{dl}\}_{l=1}^{N_d}$ as well as a set of mentions $\boldsymbol{m}_d = \{m_{dj}\}_{j=1}^{M_d}$ occurring in d. Mentions can be automatically identified using existing mention detection tools. E.g., the document in Figure 2 contains mentions {Gates, Microsoft} and words {co-founder, ...}.

Each document d is associated with a topic distribution $\boldsymbol{\theta}_d = \{\theta_{dk}\}_{k=1}^{K}$ and an entity distribution $\boldsymbol{\theta}'_d = \{\theta'_{de}\}_{e=1}^{E}$, based on which the entity groundings can be identified. LGSA simulates a generative process in which the entity mentions are determined first and the content words come later to describe the entities' attributes and actions (e.g., in Figure 2, *wealthiest* characterizes *Gates*). This leads to the differential treatment of mentions and words in the generative procedure: each mention m_{dj} is associated with a topic z_{dj} and an entity e_{dj} (drawn from z_{dj} as described next), while each word w_{dl} is associated with an index y_{dl} indicating w_{dl} is describing the y_{dl}-th mentioned entity (i.e., $e_{dy_{dl}}$).

3.2 Topic Random Walk on Entity Taxonomy

We now present the taxonomy-based modeling of latent topics, from which the underlying entities $\{e_{dj}\}$ of the mentions $\{m_{dj}\}$ are drawn. A KB entity taxonomy is a hierarchical structure that encodes rich knowledge of entity correlations, e.g., nearby entities tend to be relevant to the same topics. To capture this useful information through a generative procedure, we model each topic as a root-to-leaf random walk over the entity taxonomy. Let $\mathcal{E}$ be the set of entities from KB and $\mathcal{H}$ be the hierarchical taxonomy where entities are leaf nodes assigned to one or more categories; categories are further organized into a hierarchical structure in a generic-to-specific manner. For each category node c, we denote the set of its immediate children (subcategories or leaf entities) as $C(c)$.

The topic random walk over $\mathcal{H}$ (denoted as $\boldsymbol{\Lambda}_k$) for topic k is parameterized by a set of parent-to-child transitions, i.e., $\boldsymbol{\Lambda}_k = \{\boldsymbol{\Lambda}_{k,c}\}_{c\in\mathcal{H}}$ where $\boldsymbol{\Lambda}_{k,c} = \{\Lambda_{k,cc'}\}_{c'\in C(c)}$ is the transition distribution from c to its children. Starting from the root category c_0, a child is selected according to $\boldsymbol{\Lambda}_{kc_0}$. The process continues until a leaf entity node is reached. Hence the random walk assigns each generated entity e_{dj} a root-to-leaf path $\boldsymbol{r}_{dj}$. A desirable property of the random walk is that entities with common ancestors in the hierarchy share sub-paths starting at the root and thus tend to have similar generating probabilities. This effectively encourages clustering highly-correlated entities and produces semantically coherent topics. For example, entities *Bill Gates* and *Microsoft Inc.* in Figure 2 share the sub-path from root to category *IT*, which carries a transition probability of 0.9. Thus the two entities are likely to both have high generating probabilities in the specific topic, while the less relevant *Kobe Bryant* will have a low probability. Based on $\boldsymbol{\Lambda}_k$, we can compute the probability of the random walk reaching each of the entities, and hence obtain a distribution over entities, $\boldsymbol{\phi}_k$. Similarly, for each category node c we can compute a probability τ_{kc} indicating the possibility of c being included in a random walk path. The set of parameters $\{\boldsymbol{\Lambda}_k, \boldsymbol{\phi}_k, \boldsymbol{\tau}_k\}$ together forms a structured representation of the latent topic k, which has grounded meaning.

3.3 Entity Modeling on Mentions and Words

As described before, we learn entity representations in both mention and word spaces. Moreover, since the rich textual features of entities in KBs encode relevance between entities and mentions/words, we leverage them to construct informative priors for accurate entity modeling.

Specifically, each entity $e \in \mathcal{E}$ has a distribution over mention vocabulary $\mathcal{M}$, denoted as $\boldsymbol{\eta}_e$, along with a distribution over word vocabulary $\mathcal{V}$, denoted as $\boldsymbol{\zeta}_e$. Intuitively, $\boldsymbol{\eta}_e$ captures the relatedness between e and other entities, e.g., mention *Gates* tends to have high probability in entity *Microsoft Inc.*'s mention distribution; while $\boldsymbol{\zeta}_e$ characterizes the attributes of entity e, e.g., word *wealthiest* for entity *Bill Gates*. The informative priors over $\boldsymbol{\eta}_e$ and $\boldsymbol{\zeta}_e$ are derived from the frequency of mentions and words in entity e's Wikipedia page. Let $\boldsymbol{p}_e^{\eta}$ be the prior mention distribution over $\boldsymbol{\eta}_e$, with each dimension p_{em}^{η} proportional to the frequency of mention m in e's page. The prior word distribution $\boldsymbol{p}_e^{\zeta}$ over $\boldsymbol{\zeta}_e$ is built in a similar manner. To reflect the confidence of the prior knowledge, we introduce scaling factors λ^{η} and λ^{ζ} with a larger value indicating a greater emphasis on the prior.

Note that in LGSA the mention distribution of an entity (e.g. *Microsoft Inc.*) can put mass on not only its referring mentions (e.g. *Microsoft*), but also other related mentions (e.g. *Gates*). This captures the intuition that, for instance, the observation of *Gates* can promote the probability of the document being about *Microsoft Inc.*. This differs from previous entity linking methods and improves the detection of document's key entities, as shown in our empirical studies.

3.4 Generative Process

We summarize the generative process of LGSA in Algorithm 1 that combines all the above components. Given the mentions $\boldsymbol{m}_d$ and words $\boldsymbol{w}_d$ of a document d, each mention is first assigned a topic according to the topic distribution $\boldsymbol{\theta}_d$. The topics in turn generate entities for each mention through the random walks. For each word, one of the above entities is uniformly selected.

Algorithm 1 Generative Process for LGSA

- For each topic $k = 1, 2, \ldots, K$,
  1. For each category $c \in \mathcal{H}$, sample the transition probabilities, $\boldsymbol{\Lambda}_{kc}|\beta \sim \text{Dir}(\beta)$.
- For each entity $e = 1, 2, \ldots, E$,
  1. Sample the mention distribution: $\boldsymbol{\eta}_e|\lambda^\eta, \boldsymbol{p}_e^\eta \sim \text{Dir}(\lambda^\eta \boldsymbol{p}_e^\eta)$.
  2. Sample the word distribution: $\boldsymbol{\zeta}_e|\lambda^\zeta, \boldsymbol{p}_e^\zeta \sim \text{Dir}(\lambda^\zeta \boldsymbol{p}_e^\zeta)$.
- For each document $d = 1, 2, \ldots, D$,
  1. Sample the topic distribution: $\boldsymbol{\theta}_d|\alpha \sim \text{Dir}(\alpha)$.
  2. For each mention $m_{dj} \in \boldsymbol{m}_d$,
     (a) Sample a topic indicator: $z_{dj}|\boldsymbol{\theta}_d \sim \text{Multi}(\boldsymbol{\theta}_d)$.
     (b) Initialize path $\boldsymbol{r}_{dj} = \{c_0\}$, and $h = 0$.
     (c) While leaf not reached
        i. Sample the next node: $c_{h+1} \sim \text{Multi}(\boldsymbol{\Lambda}_{z_{dj}, c_h})$.
        ii. If c_{h+1} is a leaf node, then the corresponding entity $e_{dj} = c_{h+1}$; otherwise, $h = h + 1$.
     (d) Sample a mention $m_{dj}|\boldsymbol{\eta}_{e_{dj}} \sim \text{Multi}(\boldsymbol{\eta}_{e_{dj}})$.
  3. For each word $w_{dl} \in \boldsymbol{w}_d$,
     (a) Sample an index $y_{dl} \sim \text{Unif}(1, \ldots, M_d)$.
     (b) $e'_{dl} := e_{dy_{dl}}$
     (c) Sample a word $w_{dl}|\boldsymbol{\zeta}_{e'_{dl}} \sim \text{Multi}(\boldsymbol{\zeta}_{e'_{dl}})$.

4 Model Inference

Exact inference for LGSA is intractable due to the coupling between hidden variables. We exploit *collapsed Gibbs sampling* [Griffiths and Steyvers, 2004] for approximate inference. As a widely used *Markov chain Monte Carlo* algorithm, Gibbs sampling iteratively samples latent variables ($\{\boldsymbol{z}, \boldsymbol{r}, \boldsymbol{e}, \boldsymbol{y}\}$ in LGSA) from a Markov chain whose stationary distribution is the posterior. The samples are then used to estimate the distributions of interest: $\{\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\Lambda}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{\eta}, \boldsymbol{\zeta}\}$. We directly give the sampling formulas and provide the detailed derivations in the supplementary materials.

Sampling topic z_{dj} for mention m_{dj} according to:

$$p(z_{dj} = z|e_{dj} = e, \boldsymbol{r}_{-dj}, .) \propto (n_d^{(z)} + \alpha) \cdot \sum\nolimits_{\boldsymbol{r}(e \in \boldsymbol{r})} p(\boldsymbol{r}|\boldsymbol{r}_{-dj}, z_{dj} = z, .), \quad (1)$$

where $n_d^{(z)}$ denotes the number of mentions in document d that are associated with topic z. Marginal counts are represented with dots; e.g., $n_d^{(\cdot)}$ is obtained by marginalizing $n_d^{(z)}$ over z. The second term of Eq.(1) is the sum over the probabilities of all paths that could have generated entity e, conditioned on topic z. Here the probability of a path $\boldsymbol{r}$ is the product of the topic-specific transition probabilities along the path from root c_0 to leaf $c_{|\boldsymbol{r}|-1}$ (i.e. entity e):

$$p(\boldsymbol{r}|\boldsymbol{r}_{-dj}, z_{dj} = z, .) = \prod_{h=0}^{|\boldsymbol{r}|-2} \frac{n_{c_h, c_{h+1}}^{(z)} + \beta}{n_{c_h, \cdot}^{(z)} + |C(c_h)|\beta}, \quad (2)$$

where $n_{c_h, c_{h+1}}^{(z)}$ is the number of paths in topic z that go from c_h to c_{h+1}. All the above counters are calculated with the mention m_{dj} excluded.

Sampling path $\boldsymbol{r}_{dj}$ and entity e_{dj} for mention m_{dj} as:

$$p(\boldsymbol{r}_{dj} = \boldsymbol{r}, e_{dj} = e|z_{dj} = z, m_{dj} = m, .) \propto p(\boldsymbol{r}|\boldsymbol{r}_{-dj}, z_{dj} = z, .) \cdot \frac{n_e^{(m)} + \lambda^\eta p_{em}^\eta}{n_e^{(\cdot)} + \lambda^\eta} \cdot (\frac{n_d^{(e)} + 1}{n_d^{(e)}})^{q_d^{(e)}}, \quad (3)$$

where $n_e^{(m)}$ is the number of times that mention m is generated by entity e; $n_d^{(e)}$ is the number of mentions in d that are associated with e; and $q_d^{(e)} = \sum_l \mathbb{1}(e_{y_{dl}} = e)$ is the number of words in d that are associated with e. All the counters are calculated with the mention m_{dj} excluded.

Sampling index y_{dl} for word w_{dl} according to:

$$p(y_{dl} = y|e_{dy} = e, w_{dl} = w, .) \propto \frac{n_e^{(w)} + \lambda^\zeta p_{ew}^\zeta}{n_e^{(\cdot)} + \lambda^\zeta}, \quad (4)$$

where $n_e^{(w)}$ is the number of times that word w is generated by entity e, and is calculated with w_{dl} excluded.

The Dirichlet hyperparameters are set as fixed values: $\alpha = 50/K$, $\beta = 0.01$, a common setting in topic modeling. We investigate the effects of λ^η and λ^ζ in our empirical studies.

Efficient inference in practice: The inference on large text corpora and KBs can be complicated. To ensure efficiency in practice, we use ontology pruning, dynamic programming, and careful initialization: (a) The total number of all entities' paths can be very large, rendering the computation of Eq.(3) for all paths prohibitive. We make the observation that in general only a few entities in $\mathcal{E}$ are relevant to a document, and these are typically ones with their name mentions occurring in the document [Kataria *et al.*, 2011]. Hence, we select candidate entities for each document using a name-to-entity dictionary [Han and Sun, 2011], and only the paths of these entities are considered when sampling. Our experiments show the approximation has negligible impact on modeling performance, while dramatically reducing the sampling complexity, making the inference practical. (b) We further reduce the hierarchy depth by pruning low-level concrete category nodes (whose shortest root-to-node path lengths exceed a threshold). We found that such a "coarse" entity ontology is sufficient to provide strong performance. (c) To compute the probabilities of paths (Eq.(2)) we use dynamic programming to avoid redundant computation. (d) We initialize the entity and path assignments to ensure a good starting point. The entity assignment of a mention is sampled from the prior entity-mention distributions $\boldsymbol{p}^\eta$; based on the assignments, a path leading to the respective entity is then sampled according to an initializing transition distribution where the probability of transitioning from a category c to its child c' is proportional to the total frequency of descendant entities of c'.

5 Experiments

We evaluate LGSA's modeling performance on two news corpora. We observe that LGSA reduces topic perplexity significantly. In the task of key entity identification, LGSA improves over competitors by 10% in precision@1. We also explore the effects of entity textual priors.

Dataset	Text corpus			Wikipedia KB (pruned)		
	#doc	#word	#mention	#entity	#category	#layer
TMZ	3.2K	150K(4.6K)	71K(15K)	72K	102K	11
NYT	0.3M	130M(169K)	13M(71K)	100K	7.1K	4

Table 2: Statistics of two datasets. The numbers in parentheses are the vocabulary sizes. The average #path to each entity in TMZ and NYT KBs are 300 and 25, respectively.

Datasets: We evaluate on two news corpora (Table 2): **(a) TMZ news** is collected from TMZ.com, a popular celebrity gossip website. Each news article is tagged with one or more celebrities which serve as ground truth in the task of key entity identification; **(b) NYT news** is a widely-used large corpus from LDC[1]. For both datasets, we extract the mentions of each article using a mention annotation tool *The Wiki Machine*[2]. We use the Wikipedia snapshot of 04/02/2014 as our KB. In Wikipedia, entities correspond to Wikipedia pages which are organized as leaf nodes of a category hierarchy. We pruned irrelevant entities and categories for each dataset.

Baselines: We compare the proposed LGSA with the following competitors (Table 3 lists their differences):
(a) ConceptTM (CnptTM) [Chemudugunta *et al.*, 2008] employs ontological knowledge by assuming one-to-one correspondence between human-defined entities and latent topics. Thus each topic has identifiable transparent semantics. **(b) Entity-Topic Model (ETM)** [Newman *et al.*, 2006] models both words and mentions of documents by word topic and mention topic, respectively. No external knowledge is incorporated in ETM. **(c) Latent Dirichlet Allocation (LDA)** [Blei *et al.*, 2003] is a bag-of-words model and represents each latent topic as a word distribution. Following [Gabrilovich and Markovitch, 2009], LDA can be used for identifying key entities by measuring the similarity between the document's and the entity Wikipedia page's topic distributions. **(d) Explicit Semantic Analysis (ESA)** [Gabrilovich and Markovitch, 2009] is a popular Wikipedia-based method aimed at finding relevant entities as semantics of text. Features including content words and Wikipedia link structures are used to measure the relatedness between documents and entities. **(e) Mention Annotation & Counting (MA-C)**. We map each mention to its referent entity, and rank the entities by the frequency they are mentioned. The priority of occurrence is further incorporated to break the tie. We use The Wiki Machine in the mention-annotation step. **(f) LGSA without Hierarchy (LGSA-NH)**. To directly measure advantage of structured topic representation, we design the intrinsic competitor that models latent topic as a distribution over entities without incorporating the entity hierarchical structure.

[1]https://www.ldc.upenn.edu

[2]http://thewikimachine.fbk.eu

Topic Perplexity: We evaluate the quality of extracted topics by topic *perplexity* [Blei *et al.*, 2003]. As a widely used metric in text modeling, perplexity measures the predictive power of a model in terms of predicting words in unseen held-out documents [Chemudugunta *et al.*, 2008]. A lower perplexity means better generalization performance.

Method	Features			Tasks	
	word	mention	structured knowledge	topic extraction	key entity identification
CnptTM	√		√	√	√
ETM	√	√		√	
LDA	√			√	√
ESA	√		√		√
MA-C	√	√	√		√
LGSA-NH	√	√		√	√
LGSA	√	√	√	√	√

Table 3: Feature and task comparison of different methods

We use 5-fold cross validation testing. Figure 4a and 4b show the perplexity values on the TMZ and NYT corpora respectively using different number of topics. We see that LGSA consistently yields the lowest perplexity, indicating the highest predictive quality of extracted topics. We further observe that: (a) ETM and LDA perform inferior to CnptTM and LGSA, showing that without the guidance of human knowledge, purely data-driven method is incapable of accurately modeling text latent semantics. (b) Compared to LGSA, CnptTM has an inferior performance in that it bounds each topic with one pre-defined concept, which is not flexible enough to represent diverse corpus semantics. LGSA avoids the pitfall by associating an entity distribution with each topic, which is both expressive and interpretable. (c) Comparing LGSA and LGSA-NH further reveals the advantage of the structured topic representation—LGSA reduces perplexity by 6.5% on average. (d) Even without taxonomy structure, LGSA-NH still outperforms the baselines. This is because our model goes beyond the bag-of-words assumption and accounts for the mentions and underlying entities, which captures salient text semantics. (e) On the NYT dataset, LDA and ETM perform best at $K = 400$, where our method yields 17.9% and 5.03% lower perplexity, respectively. This again validates the benefits of incorporating world knowledge.

Key Entity Identification: Identifying key entities in documents (e.g., the persons that a news article is mainly about) serves to reveal fine-grained semantics as well as map documents to structured ontologies, which in turn facilitates downstream applications such as semantic search and document categorization. Our next evaluation tries to measure the precision of LGSA in key entity identification. We test on the TMZ dataset since the ground truth (usually a celebrity) is available. Given a document d, LGSA infers its entity distribution $\boldsymbol{\theta}'_d$ and ranks entities accordingly.

Figure 4c shows the Precision@R (proportion of test instances where a correct key entity is included in the top-R predictions) based on 5-fold cross validation. Here, both LGSA and LDA achieves their best performance by setting #topic $K = 30$. From the figure, we can see that LGSA consistently outperforms all other methods, and achieves 90% pre-

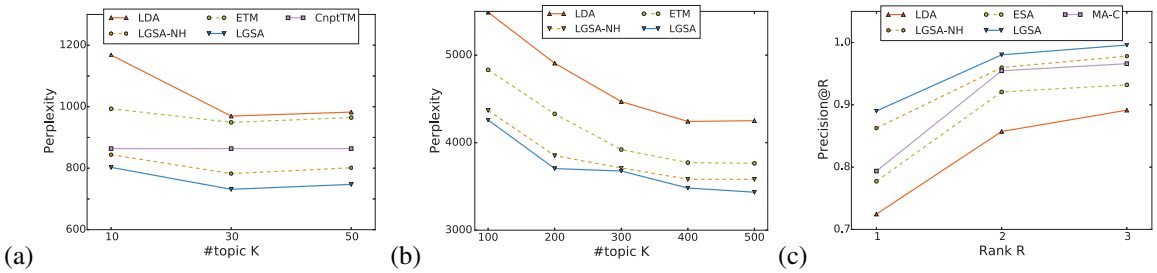

Figure 4: (a) Perplexity on TMZ dataset, (b) Perplexity on NYT, and (c) Precision@R of key entity identification on TMZ.

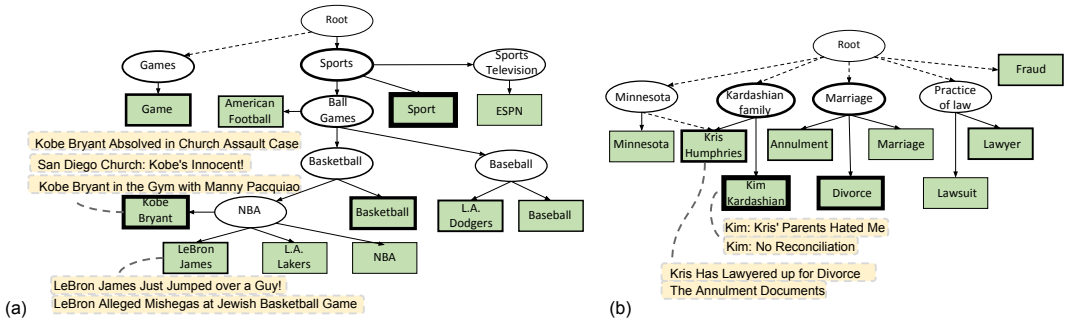

Figure 5: Topics (a) "Sports" and (b) "Kardashian and Humphries' Divorce", showing top entities (by entity distributions ϕ) and categories (by the probabilities of reaching category nodes through the random walks $\mathbf{\Lambda}$). Titles of several news are attached to their top-1 key entities.

cision at rank-1. The results reveal that: (a) MA-C has an inferior performance than LGSA, which can be attributed to the improper decoupling of candidate selection (i.e., mention annotation) and ranking (i.e., counting). In particular, for instance, though the observation of mention *Gates* may help to correctly annotate mention *MS* as referring to entity *Microsoft Inc.* it cannot directly promote the weight of *Microsoft Inc.* in the document. In contrast, LGSA captures this useful signal by allowing each entity to associate weights with all relevant mentions (Sec.3.3). (b) Our proposed model also outperforms ESA and LDA. Indeed, LGSA essentially combines these two lines of work (i.e., the *explicit* and *latent* semantic representations), by stacking the latent topic layer over the explicit entity knowledge. This ensures the best of both worlds: the flexibility of latent modeling and the interpretability of explicit modeling. (c) LGSA-NH is superior over previous methods while falling behind the full model. This confirms the effect of incorporating grounded hierarchical knowledge.

Qualitative Analysis: We now qualitatively investigate the extracted topics, illustrating the benefits of semantically-grounded modeling as well as revealing potential directions for future improvements. Figure 5 shows two example topics from the TMZ corpus. We can see the top-ranked entities and categories are semantically coherent and the highly-organized structure provides rich context and relations between the top entities (e.g., *Kobe Bryant* and *LeBron James* are both from *NBA*), helping topic interpretation. More importantly, the extracted topics show the benefits of entity grounding in latent semantic modeling. Figure 5 also demonstrates example news titles and their key entities inferred by LGSA. This naturally links documents to KBs, showing strong potential in semantic search and automatic knowledge acquisition. It is also noticeable that there exists no single entity or category in Wikipedia that directly corresponds to the topic of Kardashian and Humphries' divorce. In contrast, the full meaning is constituted through the combination of a priori unrelated ones. This validates the superior expressiveness of LGSA compared to CnptTM and ESA which rely on pre-defined concepts. The analysis also reveals some potential improvement space of our work. E.g., the actions of Kardashian and Humphries are captured in entity *Divorce*, while incorporating action representations (e.g., verbs with grounded meaning) would help to characterize the full semantics more directly. We consider this as a future work.

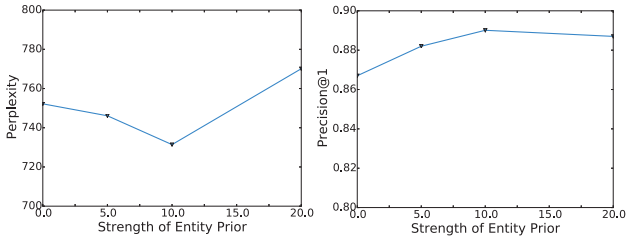

Figure 6: Effect of entity prior strength λ on TMZ dataset.

Impact of Entity Prior Strengths: LGSA leverages men-

tion/word frequency of entities in KBs to construct informative priors over mention/word distributions. Here we study the effect of these entities priors by showing performance variation with different prior strengths. Figure 6 shows the results where we have set $\lambda^{\eta} = \lambda^{\zeta} = \lambda$ for simplicity. We can see that LGSA performs best with an modest λ value (i.e. 10.0) in both tasks. The improvement of performance as λ increases in a proper range validates that the textual features from KBs can improve modeling; while improperly strong priors can prevent the model from flexibly fitting to the data.

6 Conclusion and Future Work

We proposed a structured representation of latent topics based on an entity taxonomy from KB. A probabilistic model, LGSA, was developed to infer both hidden topics and entities from text corpora. The model integrates structural and textual knowledge from KB, grounding entity mentions to KB. This leads to improvements in topic modeling and entity identification. The grounded topics can be useful in various language understanding tasks, which we plan to explore in the future.

Acknowledgments: Zhiting Hu and Mrinmaya Sachan are supported by NSF IIS1218282, NSF IIS1447676, AFOSR FA95501010247, and Air Force FA8721-05-C-0003.

References

[Abney and Light, 1999] Steven Abney and Marc Light. Hiding a semantic hierarchy in a markov model. In *the Workshop on Unsupervised Learning in NLP, ACL*, pages 1–8, 1999.

[Andrzejewski *et al.*, 2011] David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *IJCAI*, pages 1171–1177, 2011.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

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

[Chaney and Blei, 2012] Allison June-Barlow Chaney and David M Blei. Visualizing topic models. In *ICWSM*, 2012.

[Chang and Blei, 2009] Jonathan Chang and David M Blei. Relational topic models for document networks. In *AISTATS*, pages 81–88, 2009.

[Chang *et al.*, 2009] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296, 2009.

[Chemudugunta *et al.*, 2008] Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, and Mark Steyvers. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *The Semantic Web-ISWC 2008*, pages 229–244. Springer, 2008.

[Chen and Liu, 2014] Zhiyuan Chen and Bing Liu. Topic modeling using topics from many domains, lifelong learning and big data. In *ICML*, 2014.

[Foulds *et al.*, 2015] James Foulds, Shachi Kumar, and Lise Getoor. Latent topic networks: A versatile probabilistic programming framework for topic models. In *ICML*, pages 777–786, 2015.

[Gabrilovich and Markovitch, 2009] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *JAIR*, 34(2):443, 2009.

[Griffiths and Steyvers, 2004] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101(Suppl 1):5228–5235, 2004.

[Griffiths and Tenenbaum, 2004] DMBTL Griffiths and MIJJB Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *NIPS*, 16:17, 2004.

[Han and Sun, 2011] Xianpei Han and Le Sun. A generative entity-mention model for linking entities with knowledge base. In *ACL*, pages 945–954. ACL, 2011.

[Han and Sun, 2012] Xianpei Han and Le Sun. An entity-topic model for entity linking. In *EMNLP*, pages 105–115. ACL, 2012.

[Hu *et al.*, 2014] Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan Boyd-Graber. Polylingual tree-based topic models for translation domain adaptation. In *ACL*, 2014.

[Hu *et al.*, 2015a] Zhiting Hu, Poyao Huang, Yuntian Deng, Yingkai Gao, and Eric P Xing. Entity hierarchy embedding. In *ACL*, pages 1292–1300, 2015.

[Hu *et al.*, 2015b] Zhiting Hu, Junjie Yao, Bin Cui, and Eric Xing. Community level diffusion extraction. In *SIGMOD*, pages 1555–1569. ACM, 2015.

[Kataria *et al.*, 2011] Saurabh S Kataria, Krishnan S Kumar, Rajeev R Rastogi, Prithviraj Sen, and Srinivasan H Sengamedu. Entity disambiguation with hierarchical topic models. In *KDD*, pages 1037–1045. ACM, 2011.

[Kemp and Regier, 2012] Charles Kemp and Terry Regier. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054, 2012.

[Mei *et al.*, 2014] Shike Mei, Jun Zhu, and Jerry Zhu. Robust regbayes: Selectively incorporating first-order logic domain knowledge into Bayesian models. In *ICML*, pages 253–261, 2014.

[Mimno *et al.*, 2009] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *EMNLP*, pages 880–889. ACL, 2009.

[Movshovitz-Attias and Cohen, 2015] Dana Movshovitz-Attias and William W Cohen. KB-LDA: Jointly learning a knowledge base of hierarchy, relations, and facts. In *ACL*, pages 1449–1459, 2015.

[Newman *et al.*, 2006] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *KDD*, pages 680–686. ACM, 2006.

[Resnik, 1995] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.

[Song *et al.*, 2011] Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI*, pages 2330–2336. AAAI Press, 2011.

[Wang *et al.*, 2007] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*, pages 697–702, 2007.

[Wei and Croft, 2006] Xing Wei and W Bruce Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185. ACM, 2006.

[Yang *et al.*, 2015] Yi Yang, Doug Downey, Jordan Boyd-Graber, and Jordan Boyd Graber. Efficient methods for incorporating knowledge into topic models. In *EMNLP*, 2015.