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 work was done when the first two authors were at Microsoft Research, Beijing.

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 identify 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;
Utilizing 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]. For instance, [Boyd-Graber et al., 2007] exploits WordNet-Walk [Abney and Light, 1999] for the task of word sense disambiguation. Our work is distinct in that we use entity taxonomy to construct a representation of topics directly; moreover, we also infer hidden entities from text, which leads to unique complexity for inference. We propose an efficient approach to tackle this issue. Note that our work also differs from hierarchical topic models, e.g., [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 lever-
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 \(\theta_d = \{\theta_{dc}\}_{c=1}^C\) and an entity distribution \(\vartheta_d = \{\vartheta_{dc}\}_{c=1}^C\), 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, \textit{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 \(w_{dj}\) is associated with an index \(y_{dl}\) indicating \(w_{dj}\) 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 \(E\) 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 \(\Lambda_{k}\)) for topic \(k\) is parameterized by a set of parent-to-child transitions, i.e., \(\Lambda_k = \{\Lambda_{ke,c}\}_{c \in C(c)}\) where \(\Lambda_{ke,c}\) is the transition distribution from \(e\) to its children. Starting from the root category \(c_0\), a child is selected according to \(\Lambda_{ke,c_0}\). The process continues until a leaf node (i.e., an entity), is reached. Hence the random walk assigns each generated entity \(e_{dj}\) a root-to-leaf path \(r_{dj}\). A desirable property of the random walk is that entities with common ancestors in the hierarchy share sub-paths starting at the root and therefore tend to have similar generating probabilities in the same topics. This effectively encourages clustering highly-correlated entities and produces semantically coherent topics. For example, entities \textit{Bill Gates} and \textit{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 \(\Lambda_k\), we can compute the probability of the random walk reaching each of the entities, and hence obtain a distribution over entities, \(\phi_k\). Similarly, for each category node \(c\) we can compute a probability \(\tau_{kc}\) indicating the possibility of \(c\) being included in a random walk path. The set of parameters \(\{\Lambda_{kc}, \phi_k, \tau_{kc}\}\) 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 \(e \in E\) has a distribution over mention vocabulary \(M\), denoted as \(\eta_e\), along with a distribution over word vocabulary \(V\), denoted as \(\zeta_e\). Intuitively, \(\eta_e\) captures the relatedness between \(e\) and other entities, e.g., mention \textit{Gates} tends to have high probability in entity \textit{Microsoft Inc.}’s mention distribution; while \(\zeta_e\) characterizes the attributes of entity \(e\), e.g., word \textit{wealthiest} for entity \textit{Bill Gates}. The informative priors over \(\eta_e\) and \(\zeta_e\) are derived from the frequency of mentions and words in entity \(e\)’s Wikipedia page. Let \(p^\eta_e\) be the prior mention distribution over \(\eta_e\), with each dimension \(p^\eta_{em}\) proportional to the frequency of mention \(m_e\) in \(e\)’s page. The prior word distribution \(p^\zeta_e\) over \(\zeta_e\) is built in a similar manner. To reflect the confidence of the prior knowledge, we introduce scaling factors \(\lambda^\eta\) and \(\lambda^\zeta\) 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 \( m_d \) and words \( w_d \) of a document \( d \), each mention is first assigned a topic according to the topic distribution \( \theta \).

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 \( c \in \mathcal{C} \), sample the transition probabilities, \( \Lambda_{kc} \sim \text{Dir}(\beta) \).
- For each entity \( e = 1, 2, \ldots, E \),
  1. Sample the mention distribution: \( \eta_e \sim \text{Dir}(\lambda^y, p^y_k) \).
  2. Sample the word distribution: \( \zeta_e \sim \text{Dir}(\lambda^z, p^z_k) \).
- For each document \( d = 1, 2, \ldots, D \),
  1. Sample the topic distribution: \( \theta_d \sim \text{Dir}(\alpha) \).
  2. For each mention \( m_d \in m_d \),
     a. Sample a topic indicator: \( z_{dj} \sim \text{Multi}(\theta_d) \).
     b. Initialize path \( r_{dj} = \{c_{0j}\} \), and \( h = 0 \).
     c. While leaf not reached
        i. Sample the next node: \( c_{h+1} \sim \text{Multi}(\Lambda_{z_h,c_h}) \).
        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 \( m_d \mid \eta_{m_d} \sim \text{Multi}(\eta_{m_d}) \).
  3. For each word \( w_{dl} \in w_d \),
     a. Sample an index \( y_{dl} \sim \text{Unif}(1, \ldots, M_d) \).
     b. \( e_{dl} := e_{y_{dl}} \).
     c. Sample a word \( w_{dl} \mid \zeta_{e_{dl}} \sim \text{Multi}(\zeta_{e_{dl}}) \).

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 \( \{z, r, e, y\} \) in LGSA from a Markov chain whose stationary distribution is the posterior. The samples are then used to estimate the distributions of interest: \( \{\theta, \theta', \Lambda, \phi, \tau, \eta, \zeta\} \).

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:

\[
p(z_{dj} = z | e_{dj} = e, r_{d-1}, \ldots) \propto (n_d^{(z)} + \alpha) \cdot \sum_{r(e) \epsilon r} p(r | r_{d-1}, z_{dj} = z, \ldots),
\]

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., \( n_d^{(z)} \) 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 \( r \) is the product of the topic-specific transition probabilities along the path from root \( c_0 \) to child \( c_{|r|-1} \) (i.e. entity \( e \)):

\[
p(r | r_{d-1}, z_{dj} = z, \ldots) = \prod_{h=0}^{|r|-2} \frac{n_{c_h,c_{h+1}} + \beta}{n_{c_h} + |C(c_h)| \beta},
\]

where \( n_{c_h,c_{h+1}} \) is the number of paths of 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** \( r_{d} \) **and entity** \( e_{dj} \) **for mention** \( m_{dj} \) as:

\[
p(r_{dj} = r, e_{dj} = e | z_{dj} = z, m_{dj} = m, \ldots) \propto p(r | r_{d-1}, z_{dj} = z, \ldots) \cdot \sum_{r(e)} \frac{n_d^{(e)} + \lambda^y p_{e|m} \cdot \lambda^z p_{e|w}}{n_d^{(e)} + \lambda^y + \lambda^z},
\]

where \( n_d^{(e)} \) is the number of times that mention \( m \) is generated by entity \( e \); \( n_d^{(w)} \) is the number of mentions in \( d \) that are associated with \( e \); and \( q_d^{(e)} = \sum_r I(e_{y_{dl}} = e) \) 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:

\[
p(y_{dl} = y | e_{dl} = e, w_{dl} = w, \ldots) \propto n_d^{(w)} + \lambda^y p_{e|w} \cdot \lambda^z, \frac{n_d^{(w)} + \lambda^y + \lambda^z}{n_d^{(w)} + \lambda^y + \lambda^z},
\]

where \( n_d^{(w)} \) 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 \( \lambda^y \) and \( \lambda^z \) 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 \( E \) 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 \( p^2 \); 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 conduct quantitative and qualitative experiments to 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@R. We also explore the effects of entity textual priors.

### Statistics of two datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Text Corpus</th>
<th>(Pruned) Wikipedia KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMZ</td>
<td>3.2K</td>
<td>150K (4.6K)</td>
</tr>
<tr>
<td>NYT</td>
<td>330K</td>
<td>130M (169K)</td>
</tr>
</tbody>
</table>

Table 2: Statistics of two datasets. The numbers in parentheses are the respective vocabulary sizes. For efficiency we pruned the category hierarchy for NYT to 4 layers. The average number of paths to the respective vocabulary sizes. For efficiency we pruned the category hierarchy for NYT to 4 layers. The average number of paths to

during we pruned the category hierarchy for NYT to 4 layers. The average number of paths 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

---

### Table 3: Feature and task comparison of different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>CnptTM</td>
<td>word, mention</td>
<td>structured knowledge</td>
</tr>
<tr>
<td>ETM</td>
<td>word, mention</td>
<td>topic extraction</td>
</tr>
<tr>
<td>LDA</td>
<td></td>
<td>key entity identification</td>
</tr>
<tr>
<td>ESA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA-C</td>
<td>word, mention</td>
<td></td>
</tr>
<tr>
<td>LGSA-NH</td>
<td>word, mention</td>
<td></td>
</tr>
<tr>
<td>LGSA</td>
<td>word, mention</td>
<td></td>
</tr>
</tbody>
</table>

#### 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.

---

**Figure 4a and 4b** show the perplexity values on the TMZ and NYT corpora respectively using different number of topics. We can see that LGSA consistently yields the lowest perplexity, indicating the highest predictive quality of extracted topics. We can further make the following observations: (a) The perplexity values of ETM and LDA are higher than those of 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 our 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-specific semantics. LGSA avoids the pitfall by associating an entity distribution with each topic, which ensures expressive power while remaining interpretable. (c) The comparison between LGSA and LGSA-NH further reveals the advantage of the structured topic representation: LGSA reduces perplexity by 6.5% on average. By integrating the rich world knowledge encoded in the entity taxonomy, LGSA is able to cluster highly-correlated entities in the same topic and uncover coherent semantics. (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 achieve perform 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 \(d\), for LGSA, we infer its entity distribution \(\theta^e_d\) 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
Figure 4: (a) Topic perplexity on TMZ dataset, (b) Topic perplexity on NYT, and (c) Precision at rank 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 Λ). Titles of several news are attached to their top-1 key entities.

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% precision 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. in the document. In contrast, LGSA captures this valuable 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) The performance of LGSA-NH is superior over previous methods while falling behind the full model. This confirms the effect of incorporating grounded entity knowledge and hierarchical structure.

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 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. Moreover, 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 $\lambda$ on topic perplexity and key entity identification on TMZ dataset.

Impact of Entity Prior Strengths: LGSA leverages mention/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 $\lambda^v = \lambda^c = \lambda$ for simplicity. We can see that LGSAs performs best with an modest $\lambda$ value (i.e. 10.0) in both tasks. The improvement of performance as $\lambda$ 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


A  Inference via Collapsed Gibbs Sampling

Here we describe the inference algorithm for LGSA based on collapsed Gibbs Sampling.

Given a document corpus \( D \), the informative priors over entities \( \{ p^0, p^\zeta, \lambda^0, \Lambda^\zeta \} \), and the hyperparameters \( \{ \alpha, \beta \} \), LGSA specifies the following full posterior distribution:

\[
\begin{align*}
&\quad p(\Lambda, \eta, \zeta, r, e, y|D, p^0, p^\zeta, \lambda^0, \Lambda^\zeta, \alpha, \beta) \propto \\
&\quad \left( p(\Lambda|\beta)p(\eta|p^0, \lambda^0)p(\zeta|p^\zeta, \Lambda^\zeta)p(\theta|\alpha)p(z|\theta) \right) \\
&\quad p(r, e|z, \Lambda)p(m_D|e, \eta)p(y|M_D)p(w_D|y, e, \zeta).
\end{align*}
\] (A.1)

where \( m_D \) and \( w_D \) are the mentions and words in the document collections, respectively; \( M_D = \{ M_1, \ldots, M_D \} \) is the counts of mentions of the documents. The constant of proportionality is the marginal likelihood of the observed data.

The task of posterior inference for LGSA is to determine the probability distribution of the hidden variables given the observed mentions and words. However, exact inference is intractable due to the difficulty of calculating the normalizing constant in the above posterior distribution.

We use collapsed Gibbs Sampling, a well-established Markov chain Monte Carlo (MCMC) technique for approximate inference. In collapsed Gibbs Sampling, the distributions \( \Phi = \{ \Lambda, \eta, \zeta, \theta \} \) are first marginalized (collapsed), a Markov chain over the latent indicators \( \{ z, r, e, y \} \) is then constructed, whose stationary distribution is the posterior. We obtain samples of latent variables from the Markov chain. Point estimates for the collapsed distributions \( \Phi \) can then be computed given the samples, and predictive distributions are computed by averaging over multiple samples.

Sampling Procedure

Gibbs Sampler repeatedly samples each latent variable conditioned on the current states of other hidden variables and observations; a configuration of latent states of the system is then obtained. Next we provide the derivation of the sampling formulas (i.e. Eqs.(1-4) in the paper).

By marginalizing out \( \Phi \) in Eq.(A.1), we obtain:

\[
\begin{align*}
p(z, r, e, y, \cdot) &\propto p(z|\alpha)P(r, e|z, \beta)p(m_D|e, p^0, \lambda^0)p(y|M_D) \\
&\quad \cdot p(w_D|y, e, p^\zeta, \Lambda^\zeta) \\
&= \int p(\theta|\alpha)P(z|\theta)d\theta \int p(\Lambda|\beta)p(r, e|z, \Lambda)d\Lambda \quad \text{(A.2)} \\
&\quad \cdot \int p(\eta|p^0, \lambda^0)p(m_D|e, \eta)d\eta \cdot p(y|M_D) \\
&\quad \cdot \int p(\zeta|p^\zeta, \Lambda^\zeta)p(w_D|y, e, \zeta)d\zeta.
\end{align*}
\]

The conditional of \( z_{dj} \) can be computed as:

\[
p(z_{dj} = z|e_{dj} = e, r_{-dj}, z_{-dj}, \cdot) \propto \\
p(z_{dj} = z|z_{-dj}, \cdot)p(e_{dj} = e|z_{dj} = z, r_{-dj}, \cdot)
\] (A.3)

The first term of Eq.(A.3) is:

\[
p(z_{dj} = z|z_{-dj}, \cdot) = \frac{p(z_{dj} = z, z_{-dj} | \cdot)}{p(z_{-dj} | \cdot)}.
\] (A.4)

As we assume each \( z \) is generated from a multinomial distribution \( \theta \), and the hyperparameter for conjugate Dirichlet prior is \( \alpha \), we have:

\[
p(z|\alpha) = \int P(\theta|\alpha)P(z|\theta)d\theta
\]

\[
= \int \prod_d \Gamma(K_\alpha) \prod_z \Gamma(\alpha) \prod_z \theta_{z_{-dj}}^{\alpha - 1} \cdot \prod_z \prod_z \theta_{z_{-dj}z}^{n^{(z)}} d\theta
\]

\[
= \prod_d \Gamma(K_\alpha) \prod_z \Gamma(\alpha) \cdot \prod_z \Gamma(n_{z_{-dj}z}^{(z)} + \alpha)/\Gamma(n_{z_{-dj}z}^{(z)} + K_\alpha),
\]

where \( n_{d,z}^{(z)} \) is the number of times that topic \( z \) has been associated with a mention of document \( d \). Marginal counts are represented with dots (i.e. \( n_{d,z}^{(z)} \) is obtained by marginalizing \( n_{d,z}^{(z)} \) over \( z \)). Combining the above equation with Eq.(A.4) leads to:

\[
\begin{align*}
\frac{p(z_{dj} = z, z_{-dj} | \cdot)}{p(z_{-dj} | \cdot)} &= \frac{\Gamma(n_{d,z_{-dj}z}^{(z)} + \alpha)}{\Gamma(n_{d,z_{-dj}z}^{(z)} + K_\alpha)} \cdot \frac{\Gamma(n_{d,z_{-dj}z}^{(z)} + \alpha)}{\Gamma(n_{d,z_{-dj}z}^{(z)} + K_\alpha)}
\end{align*}
\] (A.5)
where the count with subscript \(-ij\) denotes a quantity with the current instance (i.e. mention \(m_{dij}\)) excluded. Here we use the identity \(\Gamma(x + 1) = x\Gamma(x)\).

The second term of Eq.(A.3) is the probability of generating entity \(e\) conditioned on topic \(z\), which requires summing over the probabilities of all paths in \(z\) that could have generated \(u\):

\[
p(e_{dij} = e|z_{dij} = z, r_{-dij}, \cdot) = \sum_{r(e \in r)} p(r|z_{dij} = z, \cdot).
\]

(A.6)

The probability of a path \(r\) is the product of the topic-specific transition probabilities along the path from root \(c_0\) to leaf \(c_{|r|−1}\) (i.e. entity \(e\)):

\[
p(r|z_{dij} = z, \cdot) = \prod_{h=0}^{|r|−2} p(c_{h+1}|c_h, z_{dij} = z, r_{-dij}).
\]

Here \(p(c_{h+1}|c_h, z_{dij} = z, r_{-dij})\) can be derived analogously to Eqs.(A.4-A.5), where the Dirichlet-Multinomial conjugates ensure the tractability of the integrals. We then obtain:

\[
p(c_{h+1}|c_h, z_{dij} = z, r_{-dij}) = \frac{n_{c_h,c_{h+1}} + \beta}{\sum_{c_h,c_{h+1}} n_{c_h,c_{h+1}} + |C(c_h)|\beta},
\]

(A.7)

where \(n_{c_h,c_{h+1}}\) is the number of paths in topic \(z\) that go from \(c_h\) to \(c_{h+1}\), with the path of mention \(m_{dij}\) excluded.

Finally, by combining Eqs.(A.3-A.7) we obtain the sampling formula for \(z_{dij}\) as Eq.(1-2) in the paper. Note that we omit the subscripts/superscripts \(-ij\) in Eq.(1-2) to avoid cluttering of notation. Eq.(3) and Eq.(4) are derived in a similar manner.

**Distribution Estimation**

After a sufficient number of sampling iterations as described above, we obtain a set of samples. The unknown distributions can then be computed by integrating across the samples. Specifically, for any single sample we can estimate \(\theta, \theta', \Lambda, \eta\) and \(\zeta\) as:

\[
\hat{\phi}_{ze} = \sum_{r(e \in r)} \prod_{h=0}^{|r|−2} \hat{\Lambda}_{z_{c_h},c_{h+1}},
\]

\[
\hat{\tau}_{ze} = \sum_{r(e \in r)} \prod_{h=0}^{|r|−2} \hat{\Lambda}_{z_{c_h},c_{h+1}}.
\]

Finally, based on the estimated \(\hat{\Lambda}\), we can compute the top-
It is straightforward to see that \(\sum_e \hat{\phi}_{ze} = 1\). That is, \(\hat{\phi}_z\) is a distribution over entities.

**Inference on New Documents**

Given a new-arriving document \(d\), we can infer its topic distribution \(\theta'\) and entity distribution \(\theta'_e\) to reveal its major themes and entities. The inference can be carried out using the Gibbs Sampling described above, but this time with the topic and entity statistics (i.e. \(\Lambda, \eta\) and \(\zeta\)) fixed.