Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts

Yu Zhang*, Yunyi Zhang*, Martin Michalski*, Yucheng Jiang*, Yu Meng*, Jiawei Han
University of Illinois at Urbana-Champaign
{yuz9, yzhan238}@illinois.edu
{martinm6, yj17}@illinois.edu
{yumeng5, hanj}@illinois.edu

ABSTRACT
Instead of mining coherent topics from a given text corpus in a completely unsupervised manner, seed-guided topic discovery methods leverage user-provided seed words to extract distinctive and coherent topics so that the mined topics can better cater to the user's interest. To model the semantic correlation between words and seeds for discovering topic-indicative terms, existing seed-guided approaches utilize different types of context signals, such as document-level word co-occurrences, sliding window-based local contexts, and generic linguistic knowledge brought by pre-trained language models. In this work, we analyze and show empirically that each type of context information has its value and limitation in modeling word semantics under seed guidance, but combining three types of contexts (i.e., word embeddings learned from local contexts, pre-trained language model representations obtained from general-domain training, and topic-indicative sentences retrieved based on seed information) allows them to complement each other for discovering quality topics. We propose an iterative framework, SeedTopicMine, which jointly learns from the three types of contexts and gradually fuses their context signals via an ensemble ranking process. Under various sets of seeds and on multiple datasets, SeedTopicMine consistently yields more coherent and accurate topics than existing seed-guided topic discovery approaches.

CCS CONCEPTS
• Information systems → Clustering. Document topic models;
• Computing methodologies → Natural language processing.

KEYWORDS
topic discovery; text embedding

ACM Reference Format:

1 INTRODUCTION
To efficiently grasp the information in a large collection of documents, it is of great interest to automatically discover a set of coherent topics from the corpus. Besides capturing meaningful structures in massive text data [11], topic discovery also widely benefits downstream text mining tasks such as taxonomy construction [20] and document classification [5].

Unsupervised topic models, from LDA [4] to embedding-based [8, 44] and pre-trained language model-enhanced [29, 39] approaches, have been extensively studied for decades as the mainstream approach to topic discovery. Despite their efficacy in uncovering prominent themes of a corpus, such models tend to retrieve semantically general topics that may not align well with users’ specific interests, as explained in [13, 27]. Motivated by this, rather than finding arbitrary topics in a fully unsupervised manner, seed-guided topic discovery [10, 13, 15, 27, 47] aims to extract topics along a certain dimension based on user-provided seeds, and the top-ranked words under each topic should be discriminatively relevant to the corresponding seed. For example, given a collection of restaurant reviews, if a user would like to explore topics of food types (e.g., by providing the seeds “noodles”, “steak”, and “pizza”), then a seed-guided topic discovery model should discover topic-indicative terms for each input seed (e.g., “ramen” and “pasta” under “noodles”), instead of finding terms that are relevant to multiple seeds (e.g., “beef”) or retrieving topics along other dimensions (e.g., “good” or “bad” as topics of sentiments).

Recent studies on seed-guided topic discovery [13, 20, 27, 47] have been focusing on utilizing different types of context information so that they can go beyond the “bag-of-words” generative assumption in LDA and learn more accurate word semantics for topic discovery. To be specific, there are three major types of context signals used in related studies.

Skip-Gram Word Embeddings. Different from LDA which infers topics based on the global document-word frequency matrix, skip-gram embedding learning [31] assumes that words occurring in similar local contexts (e.g., ±5 words) tend to have similar semantic properties. Following this assumption, each word in the corpus can be represented by an embedding vector in a latent space. To incorporate skip-gram signals in topic discovery, the word embeddings can be injected into the LDA backbone [13] or jointly learned by viewing documents and seeds also as contexts [27]. However, skip-gram embeddings are less helpful in disambiguating word meanings because only one vector is learned for each word given
the whole corpus. Indeed, Sia et al. [39] show that clustering skip-gram embeddings underperforms clustering output representations of contextualized language models such as BERT [7] in unsupervised topic modeling.

**Pre-trained Language Model Representations.** Pre-trained language models (PLMs) [7, 23, 33] have revolutionized the text mining field by learning contextualized word embeddings. The Transformer architecture [43] used in many PLMs can capture long-range and high-order context signals, and the knowledge learned by PLMs from web-scale corpora can complement contexts in the input corpus in topic discovery [47]. Meanwhile, related studies have observed several cases where PLMs generate noticeably bad topics. For example, Meng et al. [29] show that PLM representations suffer from the curse of dimensionality and do not form clearly separated clusters; Thompson and Mimno [42] find that GPT-2 representations [33] work well only if the outputs of certain layers are taken, and RoBERTa-induced topics [23] are consistently of poor quality.

**Topic-Indicative Documents.** Although skip-gram embeddings and PLMs are powerful in representing each word based on its contexts, neither of them considers whether the contexts they use are topic-indicative (i.e., semantically close to a certain seed). In fact, skip-gram embedding learning always takes the whole words as contexts, regardless of whether they are relevant to any seed; a PLM will always output the same representation for a word if the input corpus is fixed, no matter what the seeds are. To tackle this problem, supervised topic models [17, 25] propose to leverage document-level training data (i.e., each document belongs to which seed or semantic category). However, such information relies on massive human annotation, which may be difficult to obtain in practical applications (e.g., weakly supervised text classification [26, 28, 46]). Moreover, a document may be too broad to be viewed as a context unit because each document can be relevant to multiple topics simultaneously.

To summarize, each type of context signals has its specific advantages and disadvantages. Therefore, a topic discovery method purely relying on one type of context information may not be robust across different datasets or seed dimensions. Meanwhile, it is worth noting that the three types of contexts strongly complement each other. For example, PLMs have contextualization power which skip-gram embeddings are short of; skip-gram embeddings usually have fewer dimensions than PLM representations and are less prone to the curse of dimensionality; topic-indicative documents are not naturally available, but they can be retrieved by applying skip-gram embeddings and PLMs.

**Contributions.** Motivated by the complementarity of context signals, in this paper, we propose **SeedTopicMine**, an effective seed-guided topic discovery framework by integrating multiple types of contexts. **SeedTopicMine** iteratively retrieves and updates the set of topic-discriminative terms for each seed. In each iteration, we first jointly leverage seed-guided skip-gram embeddings and PLM-based representations to discover a set of topic-indicative terms. Then, using these terms, we retrieve a set of topic-indicative sentences. Here, we consider sentences rather than documents because each sentence, as a more fine-grained unit, is more likely to concentrate on one topic. Finally, the derived topic-indicative sentences and the other two types of contexts are cooperatively utilized through an ensemble ranking process, after which the topic-discriminative terms will be updated and used for the next iteration.

**Extensive experiments on real-world datasets show that SeedTopicMine effectively discovers discriminative terms under each seed to form coherent topics.** Our human evaluation quantitatively validates the superiority of **SeedTopicMine** over baselines that rely on a single type of contexts. In the ablation study, we observe that even in the same dataset, if we consider different dimensions of seeds, the contributions of different context signals vary significantly, which confirms our key motivation that any single type of context signal is insufficient for discovering seed-discriminative topics stably.

### 2 FRAMEWORK

In this section, we first review various types of contexts utilized by previous studies for topic discovery. Then, we present our framework, **SeedTopicMine**, that iteratively ensembles these types of signals.

#### 3.1 Types of Context Information

Previous studies on topic discovery, either unsupervised, seed-guided, or supervised, propose to leverage different types of information such as skip-gram embeddings [8, 13, 27, 30, 44], pre-trained language model representations [3, 29, 39, 42, 47], and topic-indicative documents [17, 20, 25]. We now introduce three major types of information sources, which are illustrated in Figure 1, and how we propose to use them in **SeedTopicMine**.

**3.1.1 Seed-Guided Text Embeddings.** Previous embedding-based topic models [8, 13, 44] propose to incorporate word embeddings to make up for the representation deficiency of the “bag-of-words” generation assumption in LDA. The intuition of text embedding learning is based on the hypothesis that semantically similar terms share similar contexts. In unsupervised topic discovery, the contexts of a term may refer to its skip-grams [31] and the documents it appears in [19, 40]. In our task of seed-guided topic discovery, we can further leverage seeds in the embedding learning process by viewing the category that a term belongs to as its context. To facilitate this goal, in **SeedTopicMine**, we follow [27, 47] and unify the three types of contexts into one objective. To be specific, we aim to maximize the likelihood of observing a term’s skip-gram, document, and category contexts given that term. Formally, the...
Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts

Figure 1: Overview of the SeedTopicMine framework.

The seed-guided text embedding learning objective is:

$$ J_{emb} = \log \prod_{d \in D} \prod_{w \in \text{seed}} \prod_{c \in C(w)} p(w_j | w_i) $$

$$ + \log \prod_{d \in D} \prod_{w \in \text{seed}} p(d | w) + \log \prod_{i \leq j \leq |S|} \prod_{w \in \mathcal{T}_i} p(c_i | w). $$

(1)

Here, $C(w_i)$ is the set of terms in $w_i$’s skip-gram window. For example, given a text sequence $w_1, w_2, ..., w_X$, we have $C(w_i) = \{w_j | l - x \leq j \leq l + x, j \neq i\}$, where $x$ is the skip-gram window size. $T_i$, as mentioned in Definition 2.1, is the set of terms related to the seed $s_i$ (i.e., belong to the semantic category $c_i$). We adopt an iterative framework to gradually expand $T_i$. At the very beginning, $T_i = \{s_i\}$ (i.e., the seed initially belongs to its corresponding topic). After each iteration, terms close to the category $c_i$ in the embedding space will be added to $T_i$.

There are various ways to define each likelihood in Eq. (1). Following previous studies on topic modeling [2, 16, 21, 27], we adopt the von Mises-Fisher (vMF) distribution.

$$ p(w_j | w_i) = \frac{\exp(\kappa_{w_0} \cos(u_{w_0}, v_{w_0}))}{\sum_{w'} \exp(\kappa_{w_0} \cos(u_{w_0}, v_{w'}))} = \text{vMF}(v_{w_j} | u_{w_j}, \kappa_{w_0}). $$

(2)

where $\kappa_{w_0} \geq 0$ is the concentration parameter, indicating the semantic specificity of $w_i$; $u_{w_j}$ and $v_{w_j}$ are the embeddings of $w_i$ and $w_j$, respectively. The vMF distribution can be viewed as an analogue of the Gaussian distribution on a sphere. In Eq. (2), the distribution concentrates around the mean direction $u_{w_j}$, and is more concentrated if $\kappa_{w_0}$ is larger (i.e., $w_i$ is a more specific term). Similar to Eq. (2), the other two likelihood terms in Eq. (1) can be defined as

$$ p(d | w) = \frac{\exp(\kappa_{w_d} \cos(u_{w_d}, v_{d}))}{\sum_{d'} \exp(\kappa_{w_d} \cos(u_{w_d}, v_{d'}))} = \text{vMF}(v_{d} | u_{w_d}, \kappa_{w_d}), $$

$$ p(c_i | w) = \frac{\exp(\kappa_{w_0} \cos(u_{w_0}, v_{c_i}))}{\sum_{c'} \exp(\kappa_{w_0} \cos(u_{w_0}, v_{c'}))} = \text{vMF}(v_{c_i} | u_{w_0}, \kappa_{w_0}), $$

(3)

where $v_{d}$ and $v_{c_i}$ are the embedding vectors of document $d$ and category $c_i$, respectively.

To summarize, the seed-guided text embedding learning process is cast as the following optimization problem:

$$ \max J_{emb} \text{ s.t. } ||u_{w_0}|| = ||v_{w_0}|| = ||v_{c_i}|| = ||v_{d}|| = 1, \kappa_{w_0} \geq 0. $$

(4)

We follow the optimization process of [27] to optimize Eq. (4).

After embedding learning, for each term $w$, we obtain two vectors $u_w$ and $v_w$, which, as they do in previous studies [27, 31, 40], carry the semantics of $w$ when it is viewed as a center term and a context term, respectively. Given a term $w$ and a seed $s_i$, we calculate the cosine similarity between their learned embeddings as the first criterion of their semantic proximity, which will later be used in topic discovery.

$$ \text{sim}_{emb}(w, s_i) = \cos(u_w, u_{s_i}). $$

(5)

3.1.2 Pre-trained Language Model Representations. Recently, PLMs such as BERT [7] have achieved great success in a wide spectrum of text mining tasks. The Transformer architecture [43] used in many PLMs is capable of capturing long-range and high-order context signals. Moreover, the generic knowledge learned by PLMs from web-scale corpora (e.g., Wikipedia) can complement the information one can get from the input corpus. To utilize such signals in topic discovery, for each term appearing in the input corpus, we employ a PLM to derive its representation.

Suppose a term $w$ appears $M$ times in the corpus $D$. For each of its mentions $w^i (1 \leq i \leq M)$, we feed the sentence containing this mention into a PLM. Note that $w^i$ may be segmented into multiple word pieces $w^i_1, w^i_2, ..., w^i_{T_i}$ according to the PLM tokenizer [36, 37], and each word piece $w^i_j$ will have an output representation vector $\text{PLM}(w^i_j)$ after PLM encoding. Following previous studies on topic discovery [39, 42], we take the average of these word piece representations as the representation of the mention.

$$ \text{PLM}(w^i) = \frac{1}{L} \sum_{j=1}^{L} \text{PLM}(w^i_j). $$

(6)

The mention representation is contextualized given the sentence it appears in. To get the corpus-level semantics of a term, we average
The representations of all its mentions.

\[ h_w = \frac{1}{M} \sum_{i=1}^{M} \text{PLM}(w^i). \]  

(7)

In this way, for each \( w \), we obtain a vector \( h_w \), whose dimension is given by the adopted PLM. For example, if we use BERTBase \( [7] \), then \( h_w \in \mathbb{R}^{768} \). Given a term \( w \) and a seed \( s_j \), we calculate the cosine similarity between their PLM-based representations as our second criterion of their semantic proximity.

\[ \text{sim}_{\text{PLM}}(w, s_j) = \cos(h_w, h_{s_j}). \]  

(8)

3.3 Topic-Indicative Context. Although seed-guided embedding learning and PLM encoding are both powerful tools to represent each term based on its contexts, neither of them considers whether the utilized context information is topic-indicative or not. To be specific, the PLM-based representation \( h_w \) is unaware of the seed space \( S \) (in other words, no matter what the seeds \( S = \{s_1, ..., s_N\} \) are, if the corpus \( D \) is fixed, then the same PLM always generates the same representation vector \( h_w \) for \( w \)); the embeddings \( u_w \) and \( v_w \) always take the skip-gram \( C(w) \) (i.e., \( \pm 1, ..., \pm x \) terms) and the document \( d \) containing \( w \) as contexts during learning, regardless of whether such information is relevant to a certain seed/topic. To alleviate this gap, we propose to use topic-indicative context to derive the correlation between a term \( w \) and a seed \( s_j \).

For each seed \( s_j \in S \), we assume it has a set of topic-indicative sentences \( \Theta_j = \{\theta_{j1}, ..., \theta_{jk}\} \). (Initially, \( \Theta_j \) is not given as input. We will discuss how to obtain and iteratively update \( \Theta_j \) in Section 3.3.2.) The reason that we consider sentences instead of documents here is because a document is more likely to cover multiple topics. Motivated by \([20, 41]\), we calculate the semantic closeness between \( w \) and \( \Theta_j \) according to the following two criteria: (1) **Popularity:** a term close to \( \Theta_j \) should appear frequently in the sentences in \( \Theta_j \). Formally, \( \text{pop}(w, \Theta_j) = \log(1 + \text{tf}(w, \Theta_j)) \), where \( \text{tf}(\cdot, \cdot) \) denotes term frequency and \( \text{tf}(w, \Theta_j) = \sum_{j=1}^{\Theta_j} \text{tf}(w, \theta_{ij}) \). (2) **Distinctiveness:** a term close to \( \Theta_j \) should be much more relevant to the sentences in \( \Theta_j \) than to the sentences indicating other topics. This can be characterized by the formula: \( \text{dist}(w, \Theta_j) = \exp(\text{BM25}(\text{w}, \Theta_j)) + \exp(\text{BM25}(w, \Theta_j)) \) \( + \sum_{c \in C} \exp(\text{BM25}(w, \Theta_j)) \), where \( \text{BM25}(\cdot, \cdot) \) denotes the BM25 relevance function \([34]\).

To jointly consider popularity and distinctiveness, the similarity between a term \( w \) and a category \( c_j \) based on topic-indicative sentences is defined as follows.

\[ \text{sim}_{\text{init}}(w, c_j) = \text{pop}(w, \Theta_j)^{\alpha} \cdot \text{dist}(w, \Theta_j)^{1-\alpha}, \]  

(9)

where \( 0 < \alpha < 1 \) is a hyperparameter.

3.2 The Iterative SEEDTOpicMINE Framework

We lay out our framework in Figure 1. It has three major modules: initial term ranking, topic-indicative sentence retrieval, and rank ensemble. We now introduce these modules in detail.

3.2.1 Initial Term Ranking. Initially, we only have the seed \( s_j \) for each semantic category \( c_j \), and the topic-indicative sentences \( \Theta_j \) have not been derived yet. Therefore, we first use seed-guided text embeddings (derived in Section 3.1.1) and PLM-based representations (derived in Section 3.1.2) to find terms that are relevant to each category. To be specific, for each category \( c_j \), we calculate the following score for each term \( w \).

\[ \text{score}_{\text{init}}(w, c_j) = \text{sim}_{\text{emb}}(w, s_j) \cdot \text{sim}_{\text{PLM}}(w, s_j), \]  

(10)

where \( \text{sim}_{\text{emb}}(\cdot, \cdot) \) and \( \text{sim}_{\text{PLM}}(\cdot, \cdot) \) are given in Eqs. (5) and (8), respectively. As mentioned in Section 3.1.1, the set of topic-related terms \( T_j \) is expanded and updated iteratively. In later iterations, when \( T_j \) is more than just \( \{s_j\} \), Eq. (10) can be generalized to

\[ \text{score}_{\text{init}}(w, c_j) = \sum_{t_{ij} \in T_j} \text{sim}_{\text{emb}}(w, t_{ij}) \cdot \sum_{t_{ij} \in T_j} \text{sim}_{\text{PLM}}(w, t_{ij}). \]  

(11)

For each category \( c_j \), we find top-\( \tau \) terms according to \( \text{score}_{\text{init}}(w, c_j) \) to update the topic-indicative term set \( T_j \).

3.2.2 Topic-Indicative Sentence Retrieval. Based on the set of updated topic-indicative terms \( T_j \), we now retrieve the set of topic-indicative sentences \( \Theta_j \) from the input corpus so that we can calculate Eq. (9). The retrieval process is inspired by two assumptions: (1) The sentences containing many topic-indicative terms from one category and do not contain any topic-indicative term from other categories should be topic-indicative sentences. We call such sentences “anchor” sentences. (2) The “neighbor” sentences of topic-indicative “anchor” sentences should also be viewed as topic-indicative if they do not contain topic-indicative terms from other categories.
According to Assumption (1), we first retrieve "anchor" sentences by counting the number of topic-indicative terms appearing in each sentence. Formally, given a category \( c_i \), for each sentence \( \theta \) in \( D \), we calculate

\[
\text{count}(\theta, c_i) = \sum_{w \in t_j} \text{tf}(w, \theta).
\]

Because category-indicative "anchor" sentences should have a high count with \( c_i \) and a count of 0 with any other category, we rank the sentences using the following criterion:

\[
\max_{\theta \in D} \text{count}(\theta, c_i), \quad \text{where count}(\theta, c_j) = 0 \quad (\forall j \neq i).
\]

We use \( \Theta_i^A = \{ \theta_{ij}, ..., \theta_{i|\Theta_i^A|} \} \) to denote the set of selected "anchor" sentences for \( c_i \).

Then, according to Assumption (2), we find "neighbor" sentences for each "anchor" sentence \( \theta_{ij} \). To be specific, given an "anchor" sentence, we check its \( \pm 1, \pm 2, ..., \pm y \) sentences in the document (if they exist). If the \( +k \) (resp., \( -k \)) sentence contains topic-indicative terms from other categories, we view it as not topic-indicative, and we do not further check the \( + (k + 1), ..., +y \) (resp., \( - (k + 1), ..., -y \)) sentences because they may have diverged to other topics. Otherwise, we add the \( +k \) (resp., \( -k \)) sentence into the set of topic-indicative "neighbor" sentences \( \Theta_i^N \). A more formal description of this process can be found in Lines 13–24 in Algorithm 1.

Finally, the set of retrieved topic-indicative sentences is the union of "anchor" sentences and "neighbor" sentences (i.e., \( \Theta_i = \Theta_i^A \cup \Theta_i^N \)).

3.2.3 Ensemble of Multiple Types of Contexts. After obtaining topic-indicative context \( \Theta_i \) for each category \( c_i \), we can now calculate \( \text{sim}_{\text{TF}}(w, c_i) \) in Eq. (9). Then, we have a score measuring the semantic proximity between a term \( w \) and a category \( c_i \) by jointly considering all three types of contexts.

\[
\text{score}_{\text{TF}}(w, c_i) = \sum_{t_j \in T_j} \text{sim}_{\text{TF}}(w, t_j) \cdot \text{sim}_{\text{PLM}}(w, t_j) \cdot \text{sim}_{\text{SETH}}(w, c_i).
\]

By ranking all terms in a descending order of \( \text{score}_{\text{TF}}(w, c_i) \), we get a ranking list where each term \( w \) has a rank position \( r_{\text{TF}}(w, c_i) \). Besides, instead of incorporating topic-indicative context into ranking, we can consider seed-guided text embeddings alone or PLM-based representations alone. By ranking terms in a descending order of \( \sum_{t_j \in T_j} \text{sim}_{\text{TF}}(w, t_j) \) and \( \sum_{t_j \in T_j} \text{sim}_{\text{PLM}}(w, t_j) \), each term \( w \) will have two more rank positions \( r_{\text{TF}}(w, c_i) \) and \( r_{\text{PLM}}(w, c_i) \), respectively. Based on the three rank positions, we perform rank ensemble by calculating the mean reciprocal rank (MRR).

\[
\text{MRR}(w, c_i) = \frac{1}{3} \left( \frac{1}{r_{\text{TF}}(w, c_i)} + \frac{1}{r_{\text{PLM}}(w, c_i)} + \frac{1}{r_{\text{SETH}}(w, c_i)} \right).
\]

In practice, instead of ranking all terms in the vocabulary, we only check the top-\( \rho \) terms in each ranking list. If a term \( w \) is not among the top-\( \rho \) (e.g., \( r_{\text{TF}}(w, c_i) > \rho \)), we simply set its reciprocal rank to be 0 (e.g., \( r_{\text{TF}}(w, c_i) = 0 \)). Finally, we update \( T_i \) with the terms whose MRR score exceeds a certain threshold \( \eta \).

\[
T_i = \{ w | \text{MRR}(w, c_i) \geq \eta \}, \quad (1 \leq i \leq |S|).
\]

The updated term sets \( T_1, ..., T_S \) are then fed into the next iteration of \text{SEEDTOPICMINE}.

We iterate the process of initial term ranking, topic-indicative sentence retrieval, and rank ensemble for \( N \) iterations. The entire SEEDTOPICMINE framework is summarized in Algorithm 1.

### Table 1: Dataset Statistics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>NYT</th>
<th>Yelp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>Topic</td>
<td>Location</td>
</tr>
<tr>
<td>#Docs</td>
<td>31,997</td>
<td>31,997</td>
</tr>
<tr>
<td>#Seeds</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Seeds</td>
<td>arts, technology, health, education, sports, science, business, politics, real estate</td>
<td>united states, iraq, britain, japan, canada, china, france, italy, russia, germany</td>
</tr>
</tbody>
</table>

4 EXPERIMENTS

4.1 Setup

4.1.1 Datasets. Following [27], we conduct experiments on two datasets from different domains.

- **NYT** is a collection of news articles written and published by the New York Times. It has two sets of seeds along the topic and location dimensions, hence we do include them into comparison.

- **Yelp** is a corpus of restaurant reviews released by the Yelp Dataset Challenge. It has two sets of seeds along the food and sentiment dimensions, respectively.

For both datasets, we use AutoPhrase [38] to perform phrase chunking. Following [39], we adopt a 60-40 train-test split for both datasets. The training set is used as the input corpus \( D \), and the testing set is used to calculate the topic coherence metric (see evaluation metrics for details). Dataset statistics are summarized in Table 1.

4.1.2 Compared Methods. We compare our SEEDTOPICMINE with the following baselines including seed-guided topic modeling methods and seed-guided embedding learning methods.

- **SeededLDA** [15] is a seed-guided topic modeling method. It modifies the generative process of LDA by biasing each topic to generate more seeds and by biasing each document to select topics relevant to the seeds appearing in the document.

- **Anchored CorEx** [10] is a seed-guided topic modeling method. It does not rely on generative assumptions. Instead, it leverages seeds by balancing between compressing the input corpus and preserving seed-related information.

- **KeyETM** [13] is an embedding-based topic model (ETM) assisted by keyword seeds. It modifies the objective of ETM [8] to utilize seeds in the form of topic-level priors over the vocabulary.

- **CatE** [27] is a seed-guided embedding learning method for discriminative topic mining. It jointly learns term embedding and specificity from the input corpus. Terms are then selected based on both embedding similarity with the seeds and specificity.

For unsupervised topic discovery approaches (e.g., BERTopic [12] and TopClus [29]), it is difficult to match their generated topics to the given seeds, so we cannot calculate term accuracy metrics (see evaluation metrics for details) for their output, and hence we do include them into comparison.

4.1.3 Evaluation Metrics. Given the top-\( |T_i| \) discovered terms under each seed (\( |T_i| = 20 \) in our experiments), we evaluate the results based on two different criteria: topic coherence and term accuracy.

1. https://catalog.ldc.upenn.edu/LDC2008T19
Table 2: NPMI, P@20, and NDCG@20 scores of compared algorithms. NPMI measures topic coherence; P@20 and NDCG@20 measure term accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>NYT-Topic NPMI</th>
<th>NYT-Topic P@20</th>
<th>NYT-Topic NDCG@20</th>
<th>NYT-Location NPMI</th>
<th>NYT-Location P@20</th>
<th>NYT-Location NDCG@20</th>
<th>Yelp-Food NPMI</th>
<th>Yelp-Food P@20</th>
<th>Yelp-Food NDCG@20</th>
<th>Yelp-Sentiment NPMI</th>
<th>Yelp-Sentiment P@20</th>
<th>Yelp-Sentiment NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeededLDA [15]</td>
<td>0.0841</td>
<td>0.2389</td>
<td>0.2979</td>
<td>0.0801</td>
<td>0.1050</td>
<td>0.1873</td>
<td>0.0504</td>
<td>0.1200</td>
<td>0.2132</td>
<td>0.0499</td>
<td>0.1700</td>
<td>0.2410</td>
</tr>
<tr>
<td>Anchored CorEx [10]</td>
<td>0.1325</td>
<td>0.2922</td>
<td>0.3627</td>
<td>0.1283</td>
<td>0.2040</td>
<td>0.3003</td>
<td>0.1204</td>
<td>0.3725</td>
<td>0.4531</td>
<td>0.0627</td>
<td>0.1200</td>
<td>0.1997</td>
</tr>
<tr>
<td>KeyETM [13]</td>
<td>0.1254</td>
<td>0.1589</td>
<td>0.2342</td>
<td>0.1146</td>
<td>0.0700</td>
<td>0.1676</td>
<td>0.0578</td>
<td>0.1788</td>
<td>0.2940</td>
<td>0.0337</td>
<td>0.4250</td>
<td>0.4994</td>
</tr>
<tr>
<td>CatE [27]</td>
<td>0.1941</td>
<td>0.8667</td>
<td>0.8306</td>
<td>0.2165</td>
<td>0.7480</td>
<td>0.7840</td>
<td>0.2058</td>
<td>0.6812</td>
<td>0.7312</td>
<td>0.1509</td>
<td>0.7150</td>
<td>0.7713</td>
</tr>
<tr>
<td>SeedTopicMine</td>
<td>0.1947</td>
<td>0.9456</td>
<td>0.9573</td>
<td>0.2176</td>
<td>0.8360</td>
<td>0.8709</td>
<td>0.2018</td>
<td>0.7912</td>
<td>0.8379</td>
<td>0.0922</td>
<td>0.9750</td>
<td>0.9811</td>
</tr>
</tbody>
</table>

- NPMI [18] is a widely adopted metric in topic modeling to measure topic coherence inside each topic. It is defined as the average normalized pointwise mutual information of each pair of terms in $T_i$.

$$NPMI = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{|T_i|} \sum_{t_i \in T_i} \sum_{c_i \in S_i} \frac{\log \frac{P(t_i, c_i)}{P(t_i) P(c_i)}}{\log \frac{1}{P(t_i)}, t_i \in c_i), (17)$$

where $P(t_i, c_i)$ is the probability that $t_i$ and $c_i$ co-occur in a document; $P(t_i)$ is the probability that $t_i$ occurs in a document.

- **P@k (also called MACC) [27]** is a metric for term accuracy. It measures the proportion of retrieved terms $t_i$ that actually belong to the semantic category $c_i$.

$$P@k = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{|T_i|} \sum_{t_i \in T_i} \sum_{c_i \in S_i} 1(t_i \in c_i), (18)$$

where $1(t_i \in c_i)$ is the indicator function of whether $t_i$ belongs to $c_i$ (i.e., whether $t_i$ is discriminatively relevant to the seed $s_i$). This relies on human judgment, so we invite five annotators to perform independent annotation. The reported P@k score is the average P@k of the five annotators. A high inter-annotator agreement is observed, with Fleiss’ kappa [9] being 0.896, 0.928, 0.800, and 0.909 on NYT-Topic, NYT-Location, Yelp-Food, and Yelp-Sentiment, respectively. As mentioned above, we set $|T_i| = 20$ in our experiments, so we report P@20.

- **NDCG@k** is another metric for term accuracy. It gives higher weights to higher-ranked terms by applying a logarithmic discount.

$$DCG_k = \sum_{j=1}^{|T_i|} \frac{1(t_i \in c_i)}{\log(j+1)}$$

$$NDCG@k = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{DCG_k}{IDCG_k}$$

Following the case of P@k, we calculate NDCG@k based on human annotations, and we report NDCG@20.

### 4.1.4 Hyperparameters and Implementation.
Hyperparameter settings of **SeedTopicMine** are as follows. In seed-guided embedding learning, the context window size $x = 5$; the embedding dimension is 100. In PLM encoding, we use BERTBase [7] as the PLM. When computing $\text{sim}_{	ext{mssn}}(w, t_i)$, we set $a = 0.2$. In initial-term ranking, we select $\tau = 20$ terms for each seed. In topic-indicative sentence retrieval, we retrieve $|\Theta|^A = 500$ “anchor” sentences; the “neighbor” sentence window size $y = 4$. In rank ensemble, there are $\rho = 20$ terms in each ranking list; the MRR threshold $\eta = 0.1$. We run **SeedTopicMine** for $N = 4$ iterations.

The code, datasets, and annotation results are available at https://github.com/yzhan238/SeedTopicMine.

### 4.2 Performance Comparison
Table 2 shows the NPMI, P@20, and NDCG@20 scores of compared algorithms on the two datasets. We can observe that: (1) On NYT, **SeedTopicMine** consistently achieves the best performance in terms of all metrics. Among all the baselines, CatE is the most effective one, significantly outperforming “bag-of-words”-based topic models such as SeededLDA and Anchored CorEx. However, since CatE only uses one type of context information (i.e., skip-gram embeddings), **SeedTopicMine** can improve CatE by an evident margin on term accuracy through integrating multiple types of signals. (2) On Yelp, **SeedTopicMine** underperforms CatE in terms of NPMI but significantly outperforms CatE in terms of P@20 and NDCG@20. Note that NPMI is an automatically computed metric, and the other two metrics rely on human annotation. Indeed, a recent study [14] shows that automatic metrics such as NPMI may not align well with human evaluation. From this perspective, we claim that **SeedTopicMine** performs better than CatE on Yelp, and our qualitative analysis below will validate this claim.

Besides quantitative evaluation, we show the qualitative comparison in Table 3. We randomly select two seeds from NYT-Location, NYT-Topic, Yelp-Food, and Yelp-Sentiment, respectively. For each seed, we show top-5 terms retrieved by each method. A term is marked as incorrect ($\times$) if and only if at least 3 of the 5 annotators judge the term as irrelevant to the seed. Table 3 demonstrates that: (1) SeededLDA, Anchored CorEx, and KeyETM tend to find irrelevant or very general terms. For example, both Anchored CorEx and KeyETM retrieve the term “fish” under the seed “sushi”, but “fish” is also relevant to the seed “seafood”, thus it does not discriminatively belong to the sushi category. (2) Most terms discovered by CatE are accurate. However, CatE still makes mistakes in all four dimensions in Table 3. In contrast, **SeedTopicMine** achieves higher accuracy. If we further check the mistakes made by CatE, we can find general terms such as “also” and “savoury”, which may co-occur frequently with other top-ranked terms. This possibly explains why CatE achieves higher NPMI than **SeedTopicMine** on Yelp since NPMI is based on the co-occurrence of retrieved terms.

### 4.3 Ablation Study
One key design in **SeedTopicMine** is the ensemble of multiple types of contexts. Specifically, we utilize the context information from three sources: seed-guided text embeddings (Emb), pre-trained language model representations (PLM), and topic-indicative sentences (Sntn). Now, we validate their contribution to the whole framework through an ablation analysis. Specifically, we can ignore one of the three sources while keeping all the other modules unchanged. This yields three ablation versions: **SeedTopicMine-NoEmb**, **SeedTopicMine-NoPLM**, and **SeedTopicMine-NoSntn**.
Table 3: Top-5 terms retrieved by different algorithms. ×: At least 3 of the 5 annotators judge the term as irrelevant to the seed.

<table>
<thead>
<tr>
<th>Method</th>
<th>NYT-Topic</th>
<th>NYT-Location</th>
<th>Yelp-Food</th>
<th>Yelp-Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeedLDA</td>
<td>said (x)</td>
<td>said (x)</td>
<td>roll (x)</td>
<td>place (x)</td>
</tr>
<tr>
<td></td>
<td>dr (x)</td>
<td>percent (x)</td>
<td>food (x)</td>
<td>food (x)</td>
</tr>
<tr>
<td></td>
<td>new (x)</td>
<td>company (x)</td>
<td>good (x)</td>
<td>us (x)</td>
</tr>
<tr>
<td></td>
<td>hospital</td>
<td>billion (x)</td>
<td>place (x)</td>
<td>order (x)</td>
</tr>
<tr>
<td>Anchored</td>
<td>case (x)</td>
<td>employees (x)</td>
<td>rolls (x)</td>
<td>definitely (x)</td>
</tr>
<tr>
<td>CorEx</td>
<td>court (x)</td>
<td>advertising (x)</td>
<td>also (x)</td>
<td>one (x)</td>
</tr>
<tr>
<td></td>
<td>patients</td>
<td>media (x)</td>
<td>saihimi (x)</td>
<td>well (x)</td>
</tr>
<tr>
<td></td>
<td>cases (x)</td>
<td>businessmen (x)</td>
<td>education (x)</td>
<td>fish (x)</td>
</tr>
<tr>
<td></td>
<td>lawyer (x)</td>
<td>commerce (x)</td>
<td>tempura (x)</td>
<td>try (x)</td>
</tr>
<tr>
<td>KeyETM</td>
<td>team (x)</td>
<td>percent (x)</td>
<td>saihimi (x)</td>
<td>food (x)</td>
</tr>
<tr>
<td></td>
<td>game (x)</td>
<td>japan (x)</td>
<td>rolls (x)</td>
<td>great (x)</td>
</tr>
<tr>
<td></td>
<td>players (x)</td>
<td>year (x)</td>
<td>rolls (x)</td>
<td>delicious (x)</td>
</tr>
<tr>
<td></td>
<td>games (x)</td>
<td>japanese (x)</td>
<td>roll (x)</td>
<td>place (x)</td>
</tr>
<tr>
<td></td>
<td>play (x)</td>
<td>economy (x)</td>
<td>place (x)</td>
<td>amazing (x)</td>
</tr>
<tr>
<td>CatE</td>
<td>public health</td>
<td>diversifying (x)</td>
<td>city (x)</td>
<td>people (x)</td>
</tr>
<tr>
<td></td>
<td>health care</td>
<td>clients (x)</td>
<td>saihimi (x)</td>
<td>food (x)</td>
</tr>
<tr>
<td></td>
<td>medical corporate</td>
<td>maiden (x)</td>
<td>spanish (x)</td>
<td>british columbia</td>
</tr>
<tr>
<td></td>
<td>hospitals</td>
<td>investment banking</td>
<td>belgium (x)</td>
<td>manitoba</td>
</tr>
<tr>
<td></td>
<td>doctors</td>
<td>executives</td>
<td>program (x)</td>
<td>time (x)</td>
</tr>
<tr>
<td></td>
<td>play (x)</td>
<td>economy</td>
<td>program (x)</td>
<td>time (x)</td>
</tr>
<tr>
<td>SEEDTopicMine</td>
<td>medical</td>
<td>companies</td>
<td>french (x)</td>
<td>canadian</td>
</tr>
<tr>
<td></td>
<td>hospitals</td>
<td>businesses</td>
<td>paris (x)</td>
<td>quebec</td>
</tr>
<tr>
<td></td>
<td>public health</td>
<td>corporations</td>
<td>philippe (x)</td>
<td>montréal</td>
</tr>
<tr>
<td></td>
<td>patients</td>
<td>corporate</td>
<td>french state</td>
<td>toronto</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>frenchman</td>
<td>ottawa</td>
</tr>
</tbody>
</table>

Table 4: Ablation study on different types of context information used in SEEDTopicMine.

<table>
<thead>
<tr>
<th>Method</th>
<th>Yelp-Food</th>
<th>Yelp-Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEEDTopicMine</td>
<td>0.7912</td>
<td>0.9711</td>
</tr>
<tr>
<td>SEEDTopicMine-NoEmb</td>
<td>0.4488</td>
<td>0.9500</td>
</tr>
<tr>
<td>SEEDTopicMine-NoPLM</td>
<td>0.6962</td>
<td>0.8029</td>
</tr>
<tr>
<td>SEEDTopicMine-NoSent</td>
<td>0.7488</td>
<td>0.9500</td>
</tr>
</tbody>
</table>

Figure 2: Effect of the number of iterations (N) on NPMI.

Table 4 demonstrates the term accuracy scores of the full SEEDTopicMine model and the three ablation versions. We can observe that: (1) SEEDTopicMine consistently outperforms all three ablation versions, which implies the positive contribution of the three types of context signals. (2) Even for the same dataset (i.e., Yelp), the contribution of a certain type of context information varies significantly with the input seeds. For example, in the food dimension, SEEDTopicMine-NoEmb performs the worst, which indicates that seed-guided embeddings are the most helpful signals. Meanwhile, in the sentiment dimension, the contribution of embeddings becomes the smallest. In comparison, pre-trained language model representations have the largest offering. This observation validates the motivation of this work that each type of context information has its specific value and limitation in topic discovery. None of them can dominate the others across all dimensions. Therefore, it becomes necessary to integrate them together, and our results show that the integration does achieve consistently the best performance.

4.4 Parameter Study

Another key design in SEEDTopicMine is the iterative framework. To verify the contribution of multiple iterations, we conduct a parameter study by showing the NPMI of the discovered topics if we run SEEDTopicMine for different numbers of iterations (i.e., N). Figure 2 shows the effect of N on NPMI across all four dimensions.

From Figure 2, we see that: (1) When N is small (e.g., N ≤ 4), NPMI increases with N in most cases. When we run SEEDTopicMine for only one iteration, the performance is always significantly lower than that when we run 3–4 iterations. This finding validates our design choice that iteratively revising each topic can boost the topic coherence. (2) When N becomes larger, the NPMI curve starts to fluctuate, and the performance gain of running more iterations is subtle. Moreover, more iterations will result in longer
We have shown the top-5 terms retrieved by different algorithms in Table 3. One may ask about the quality of lower-ranked terms in each topic. Thus, we conduct an extended case study by showing the 8 lowest-ranked terms among the top 20. These terms are listed in Table 5. Due to space limit, we only show the results of our SeedTopicMine model and the strongest baseline CatE, and two topics are selected for NYT-Topic and Yelp-Food, respectively.

From Table 5, we observe that the accuracy of CatE deteriorates for lower-ranked terms. For example, under the “politics” seed from NYT-Topic, all of the 8 shown terms discovered by CatE are judged as irrelevant. By contrast, SeedTopicMine only makes one mistake under the same seed. This observation implies that the efficacy of SeedTopicMine can be generalized to the relatively lower part of the retrieved term list, which also reflects the robustness of SeedTopicMine by integrating multiple types of contexts.

### 5 RELATED WORK

#### Seed-Guided Topic Discovery

Different from supervised topic models (e.g., [17, 25]) that rely on a large number of human-annotated documents, seed-guided topic discovery only requires a set of user-interested seeds to find corresponding topics. In [1], seeds are incorporated as prior of topic modeling using must-link and cannot-link constraints. SeedLDA [15] uses seeds to bias topics to produce seed terms and documents to select terms containing them. Anchored CorEx [10] discovers informative topics with correlation maximization and leverages seeds by selecting corpus compression and seed-indicative information. Recent studies also incorporate embedding learning techniques to obtain more accurate semantic representations. For instance, CatE [27] learns category-guided text embeddings by enforcing distinctiveness among seeds in the embedding space; SeeTopic [47] further utilizes the power of pre-trained language models for better text representations and the ability to handle out-of-vocabulary seeds.

#### Representation-Enhanced Topic Discovery

With the rapid development in text representation learning, recent topic discovery methods incorporate distributed representations to enhance the modeling of text semantics. Earlier approaches incorporate context-free word embeddings [31] into classic probabilistic topic models (e.g., LDA [4]), including Gaussian LDA [6], LFTM [32], Spherical HDP [2], and CGTM [45]. TWE [22] learns embeddings based on associations between words and latent topics obtained by LDA, CLM [44] collaboratively models topics and learns word embeddings by considering both global and local contexts. ETM [8] learns topic embeddings in the word embedding space to improve LDA for a better fit of a large vocabulary. More recent studies leverage the contextualized representations generated by pre-trained language models (e.g., BERT [7]) to facilitate the discovery of coherent topics. These contextualized representations can be used either at token-level for clustering to form topics [29, 39, 42, 49] or at document-level for modeling document-topic correlations [3, 12].

### 6 CONCLUSIONS AND FUTURE WORK

In this work, we study seed-guided topic discovery by learning from multiple types of contexts, including skip-gram embeddings based on local contexts, pre-trained language model representations upon general-domain pre-training, and topic-indicative sentences retrieved according to seed-distinctive terms. Our proposed SeedTopicMine framework jointly leverages these contexts via an ensemble process for robust topic discovery under different types of seeds. On two real-world datasets and across four sets of seeds, SeedTopicMine consistently outperforms existing seed-guided topic discovery approaches in terms of topic coherence and term accuracy.

For future studies, the promising topic discovery results achieved by SeedTopicMine may further benefit keyword-based text classification [26, 48] via expanding the seed word semantics and prompt-based methods [35] via enriching their verbalizers. Also, SeedTopicMine can be extended to model input seeds organized in a hierarchical manner by injecting hierarchy regularization or discovering topics beyond the provided seeds by incorporating latent topic learning in the corpus modeling process.

### ACKNOWLEDGMENTS

We thank Yichen Liu and Ruining Zhao for their help with annotation and anonymous reviewers for their valuable and insightful feedback. Research was supported in part by the IBM-Illinois Discovery Accelerator Institute, US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.
REFERENCES


[16] Yu Zhang, Frank F Xu, Sha Li, Yu Meng, Xuan Wang, Qi Li, and Jiawei Han. 2019. Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. In NAACL'22. 323–333.


[18] Guangxu Xun, Yaliang Li, Jiawei Han, and Jiawei Zhang. 2017. Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts. In KDD'17. 535–543.


[20] Guangxu Xun, Yaliang Li, Jiawei Han, and Jiawei Zhang. 2018. Automated phrase mining from massive text corpora. IEEE TKDE 30, 10 (2018). 1825–1837.


[22] Guangxu Xun, Yuanmin Li, Jiawei Han, and Jiawei Zhang. 2019. HiGitClass: Keyword-driven hierarchical classification of github repositories. In KDD'19. 3876–3885.

[23] Guangxu Xun, Yaliang Li, Jiawei Han, and Jiawei Zhang. 2018. Weakly-supervised neural text classification. In CIKM'18. 983–992.

[24] Guangxu Xun, Yuanmin Li, Jiawei Han, and Jiawei Zhang. 2016. Hierarchical topic mining via joint spherical tree and text embedding. In KDD'16. 1908–1917.

[25] Guangxu Xun, Yuanmin Li, Jiawei Han, and Jiawei Zhang. 2020. Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. In NAACL'22. 3886–3993.