Self-supervised Topic Taxonomy Discovery in the Box Embedding Space

Yuyin Lu[◊], **Hegang Chen**[◊], **Pengbo Mao**[◊], **Yanghui Rao**[*◊],
Haoran Xie[♡], **Fu Lee Wang**[♣], and **Qing Li**[♠]
[◊]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[♡]School of Data Science, Lingnan University, Hong Kong SAR
[♣]School of Science and Technology, Hong Kong Metropolitan University, Hong Kong SAR
[♠]Department of Computing, the Hong Kong Polytechnic University, Hong Kong SAR
{luyy37,chenhg25,maopb}@mail2.sysu.edu.cn,
raoyangh@mail.sysu.edu.cn, hrxie@ieee.org,
pwang@hkmu.edu.hk, csqli@comp.polyu.edu.hk

Abstract

Topic taxonomy discovery aims at uncovering topics of different abstraction levels and constructing hierarchical relations between them. Unfortunately, most prior work can hardly model semantic scopes of words and topics by holding the Euclidean embedding space assumption. What's worse, they infer asymmetric hierarchical relations by symmetric distances between topic embeddings. As a result, existing methods suffer from problems of low-quality topics at high abstraction levels and inaccurate hierarchical relations. To alleviate these problems, this paper develops a Box embedding-based Topic Model (BoxTM) that maps words and topics into the box embedding space, where the asymmetric metric is defined to properly infer hierarchical relations among topics. Additionally, our BoxTM explicitly infers upper-level topics based on correlation between specific topics through recursive clustering on topic boxes. Finally, extensive experiments validate high-quality of the topic taxonomy learned by BoxTM.

1 Introduction

Taxonomy knowledge discovery, the process of extracting latent semantic hierarchies from text corpora, is a crucial yet challenging research field. For text mining applications, it can serve as the foundation of complex question answering (Luo et al., 2018) and recommendation systems (Xie et al., 2022). An important line of research focuses on learning word-level or entity-level taxonomies (Miller, 1995; Jiang et al., 2022), but such products may encounter problems of low coverage, high redundancy, and limited information (Zhang et al., 2018). Since a topic can cover the semantics of a set of coherent words, some works propose to use topics as the basic taxonomic units. Taking the topic taxonomy of the arXiv website as an example, ‘‘*computer science*’’ is an academic discipline highlighted by general keywords of ‘‘*information*’’, ‘‘*computation*’’, and ‘‘*automation*’’. It involves various sub-fields such as ‘‘*computation and language*’’ and ‘‘*computer vision*’’, which have specific keywords of ‘‘*language*’’ and ‘‘*image*’’, respectively. With this topic taxonomy, users can readily retrieve papers of interest and explore related research fields.

Early methods for topic taxonomy discovery (Blei et al., 2003a; Kim et al., 2012; Mimno et al., 2007) take a probabilistic perspective originated from LDA (Blei et al., 2003b). In these approaches, each topic is a distribution across words. A document is generated by sampling topics in different levels, and then sampling words from the selected topics iteratively. As a more flexible and efficient solution compared with probabilistic models, the Hierarchical Neural Topic Models (HNTMs) that adopt deep generative models and Neural Variational Inference (NVI) have been developed in recent years (Isonuma et al., 2020). With remarkable developments of text representation learning (Pennington et al., 2014; Devlin et al., 2019; Vilnis, 2021), mining topic taxonomy in the high-quality embedding space has become a promising idea. Particularly, the latest HNTMs (Chen et al., 2021b; Duan et al., 2021a) extend the Embedded Topic Modeling (ETM) (Dieng et al., 2020) method to topic taxonomy discovery. With the assumption that topics and their keywords are close in the embedding space, these models utilize dot products between topic and word embeddings to infer topic-word distributions.

[*] Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1401–1416, 2024. https://doi.org/10.1162/tacl_a_00712
Action Editor: Ivan Titov. Submission batch: 3/2024; Revision batch: 7/2024; Published 11/2024.

In parallel, some other methods conduct recursive clustering on word embeddings to construct topic taxonomy directly (Zhang et al., 2018; Grootendorst, 2022). Such clustering-based methods often train the word embedding space on local contexts, which helps them capture accurate word semantics. Unfortunately, they have difficulty in exploiting global statistics of word occurrences, such as Bag-of-Words and TF-IDF representations. As a result, topics mined by these methods are highly coherent but may not be representative of the entire corpus. Due to this flaw of clustering-based methods, HNTMs persist as the prevailing paradigm for topic taxonomy discovery.

Despite the impressive performance of existing HNTMs, they suffer from the following problems. **(1)** ***Suboptimal representation*****:** Most of these methods are limited in modeling semantic scopes of words and topics at different abstraction levels using classic point embeddings (Pennington et al., 2014). Instead, geometric embeddings such as hyperbolic and box embeddings are more effective representations for structured data, including knowledge graphs and taxonomies (Bai et al., 2021; Abboud et al., 2020). Although HyperMiner (Xu et al., 2022) attempts to uncover topic taxonomy within a geometric embedding space, it simply replaces point embeddings in traditional HNTMs with hyperbolic embeddings and lacks in-depth analysis. This makes HyperMiner suffer from the following problems. **(2)** ***Topic collapse*****:** Prior models struggle to learn high-quality topics, especially at higher abstraction levels. In particular, their top-level topics often degenerate into clusters of meaningless common words (Wang et al., 2023; Wu et al., 2023). **(3)** ***Inaccurate hierarchy relations*****:** Many existing HNTMs rely on the symmetric distance metric (i.e., dot product) to infer the asymmetric hierarchy relations among topics. Such approximation results in an inaccurate hierarchical topic structure.

Considering the above challenges, we propose to learn topic taxonomy in the box embedding space (Vilnis et al., 2018) and develop a Box embedding-based Topic Model (BoxTM)[1] following the framework of NVI. Figure 1 shows the differences of the topic taxonomy discovery processes in the point embedding space and the box embedding space, which are adopted by most existing HNTMs and our BoxTM, respectively. And the topic taxonomy discovery process in the hyperbolic embedding space is similar to that in the point embedding space. Specifically, BoxTM represents a topic or word as a hyperrectangle instead of a point, whose volume is proportional to the size of its semantic scope. In other words, the box embedding of a general topic covers a relatively larger region than that of a specific topic. Additionally, we conduct recursive clustering on the box embeddings of the lower-level topics to extract the upper-level topics. This approach leverages the connection between descendant topics to precisely capture the semantics of the upper-level topics, which can address the topic collapse problem caused by unguided upper-level topic mining. Intuitively, we employ symmetry and asymmetry distance metrics defined in the box embedding space respectively to capture similarity and hierarchy relations among topics. In summary, the main contributions of this paper are as follows:

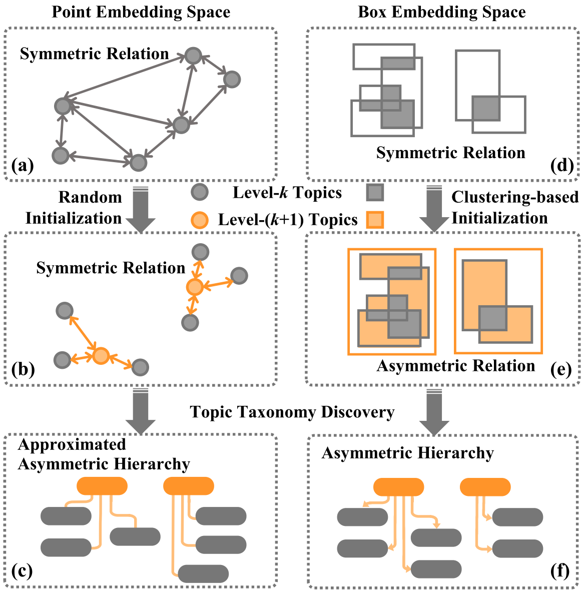


Figure 1: The topic taxonomy discovery processes in the **point embedding space (a–c)** and the **box embedding space (d–f)** of most existing HNTMs and the proposed BoxTM, respectively.

- We propose representing topics and words as box embeddings to capture their semantic scopes and accurately infer the hierarchical relations among these topics.
- We propose to conduct recursive clustering on leaf topics to mine upper-level topics,

[1] The source code of our model is available in public at: https://github.com/luyy9apples/BoxTM.

which is an interpretable and effective way to capture the semantics of upper-level topics.

- We conduct intrinsic evaluation, extrinsic evaluation, human evaluation, and qualitative analysis to validate the effectiveness of our model compared to state-of-the-art baselines.

2 Related Work

2.1 Document Generation-based Methods

The classic topic model, i.e., LDA (Blei et al., 2003b), uses a document generative process under the framework of probabilistic graphical models to extract flat topics. As an extension of LDA to topic taxonomy discovery, a series of hierarchical topic models has been proposed, such as nCRP (Blei et al., 2003a) and rCRP (Kim et al., 2012). Despite their popularity, they suffer from high complexity of posterior inference. Recently, HNTMs (Isonuma et al., 2020; Chen et al., 2021a), based on NVI and deep generative model, have been developed to tackle this problem.

Inspired by the Embedded Topic Model (ETM) (Dieng et al., 2020), nTSNTM (Chen et al., 2021b), and SawETM (Duan et al., 2021a) project topics and words into the same Euclidean embedding space and construct topic taxonomy via the symmetric distances between topic and word points. Due to the advantage of hyperbolic space in modeling tree-structured data (Nickel and Kiela, 2017), HyperMiner (Xu et al., 2022) adopts a hyperbolic embedding space to discover topic taxonomy. However, HyperMiner still uses the symmetric distance metric (i.e., dot product) to infer the complex relations among topics and randomly initializes topic embeddings, following prior HNTMs. Such approximation of asymmetric relations and ''cold start'' of embedding learning result in a risk of top-level topics collapsing into meaningless common words. To alleviate the latter problem, C-HNTM (Wang et al., 2023) attempts to learn topics of different levels using different semantic patterns. Specifically, C-HNTM learns level-2 topics by clustering on word embeddings, and it adopts ETM to mine leaf topics. Unfortunately, C-HNTM lacks the flexibility to learn topic taxonomies of different depths.

2.2 Clustering-based Methods

Since pre-trained embedding models (Devlin et al., 2019; Pennington et al., 2014) have boosted the performance of many text mining tasks in recent years, a branch of research attempts to mine flat (Sia et al., 2020; Meng et al., 2022) or hierarchical topics (Zhang et al., 2018; Grootendorst, 2022) from high-quality embedding spaces directly. As a representative clustering-based method, TaxoGen (Zhang et al., 2018) conducts hierarchical clustering to group similar words into clusters (topics) and split coarse clusters (topics) into specific ones. Additionally, it ranks the importance of each word to its topic by some manually designed metrics, such as the symmetric distance between a word and its cluster centroid. Importantly, most clustering-based methods train word embedding spaces on local contexts, which enables them to capture accurate semantics of words but hinders them from getting high-quality topics, because the boundaries between clusters are blurred in such delicate embedding spaces. Regardless, since topics are semantic summaries of corpora, global semantic information is more critical for topic mining compared to local contexts. However, clustering-based methods have trouble in utilizing the global statistics of word occurrences effectively. For example, both BERTopic (Grootendorst, 2022) and TaxoGen (Zhang et al., 2018) simply apply TF-IDF information as weights for topic keyword ranking.

2.3 Supervised Methods

Apart from self-supervised topic taxonomy discovery, another line of research tries to adopt a word-level knowledge graph (Lee et al., 2022; Meng et al., 2020) or manually built topic hierarchy (Duan et al., 2021b) as the ''framework'' of the topic taxonomy. As a representative method of supervised HNTMs, TopicNet (Duan et al., 2021b) adopts prior knowledge from WordNet (Miller, 1995). Specifically, TopicNet discovers each topic and each topic hierarchical relation guided by a seed word and the hypernym-hyponym relation between seed words, respectively. Similarly, a clustering-based method called TaxoCom (Lee et al., 2022) uses manually defined seed words as centers of topic clusters. Unfortunately, there may be a semantic gap between the general knowledge graph and the target corpus, and it's difficult and costly to determine a complete topic hierarchy manually. Therefore, self-supervised topic taxonomy discovery is more flexible and versatile, since it does not rely on prior knowledge.

3 Background Knowledge

As a representative geometric embedding technology, the box embedding method represents a word or topic as a box (i.e., axis-aligned hyperrectangle) instead of a point in the traditional Euclidean embedding method. With extra degrees of freedom, box embeddings can capture semantic scopes and asymmetric relations of objects (Vilnis et al., 2018; Li et al., 2019; Dasgupta et al., 2020).

Definition 1 (*box embedding*)**.** A D-dimensional box is determined by its minimum and maximum coordinates in each axis, parameterized by a pair of vectors $(\boldsymbol{x}_m, \boldsymbol{x}_M)$, where $\boldsymbol{x}_m, \boldsymbol{x}_M \in [0,1]^D$ and $\boldsymbol{x}_{m,i} \leq \boldsymbol{x}_{M,i}$, for $\forall i \in \{1 \dots D\}$.

Definition 2 (*box operations*)**.** Let $\mathrm{Box}(A) := (\boldsymbol{x}_m^A, \boldsymbol{x}_M^A)$, $\mathrm{Box}(B) := (\boldsymbol{x}_m^B, \boldsymbol{x}_M^B)$ denote box embeddings of objects A and B, respectively. The basic box operations are defined as follows:

Definition 2.1 (*volume*)**.** The volume of $\mathrm{Box}(A)$ is defined as $\mathrm{Vol}(\mathrm{Box}(A)) := \prod_{i=1}^{D}(\boldsymbol{x}_{M,i}^A - \boldsymbol{x}_{m,i}^A)$.

Definition 2.2 (*intersection*)**.** If there is an overlap between $\mathrm{Box}(A)$ and $\mathrm{Box}(B)$, their intersection box is defined to be $\mathrm{Box}(A) \wedge \mathrm{Box}(B) := (\max(\boldsymbol{x}_m^A, \boldsymbol{x}_m^B), \min(\boldsymbol{x}_M^A, \boldsymbol{x}_M^B))$; otherwise, it is defined to be $\mathrm{Box}(A) \wedge \mathrm{Box}(B) := \perp$.

Definition 2.3 (*union*)**.** The union box of $\mathrm{Box}(A)$ and $\mathrm{Box}(B)$ is defined as $\mathrm{Box}(A) \vee \mathrm{Box}(B) := (\min(\boldsymbol{x}_m^A, \boldsymbol{x}_m^B), \max(\boldsymbol{x}_M^A, \boldsymbol{x}_M^B))$.

Note that box embeddings are closed under the intersection and union operations. For simplicity, the base box operations are described above, while in practice we adopt the Gumbel version that is more stable for training (Dasgupta et al., 2020).

In this work, we consider the volume of a topic or word box as its size of semantic scope, i.e., a more general concept covers a larger region in the latent semantic space. The union box of topics and words is a generalization of their semantics. For the symmetric affinity, denoted as R_1, there is $\forall A, B : AR_1B \Rightarrow BR_1A$. We estimate R_1 with the volume of the intersection between topic and word boxes (R_s), which is defined as follows:

$$\mathrm{R}_s(A,B) = \mathrm{Vol}\left(\mathrm{Box}(A) \wedge \mathrm{Box}(B)\right). \quad (1)$$

Accordingly, we have $\mathrm{R}_s(A,B) = \mathrm{R}_s(B,A)$. To mitigate the bias towards large boxes, we can regularize the $\mathrm{R}_s(A,B)$ metric through division by $\mathrm{Vol}\left(\mathrm{Box}(A)\right) \cdot \mathrm{Vol}\left(\mathrm{Box}(B)\right)$ in practice.

For the asymmetric hierarchical relation between topics of adjacent levels, denoted as R_2, there is $\forall t^i, t^j \in \mathcal{T} : t^i R_2 t^j \Rightarrow \neg t^j R_2 t^i$, which means ''if t^i is a sub-topic of t^j, then t^j is NOT a sub-topic of t^i''. We reflect R_2 by the ratio of the volume of their intersection box to the upper-level topic box (R_a), that is,

$$\mathrm{R}_a\left(t_k^i \middle| t_{k+1}^j\right) = \frac{\mathrm{Vol}\left(\mathrm{Box}(t_k^i) \wedge \mathrm{Box}(t_{k+1}^j)\right)}{\mathrm{Vol}\left(\mathrm{Box}(t_{k+1}^j)\right)}, \quad (2)$$

where $t_k^i \in \mathcal{T}_k$ and $t_{k+1}^j \in \mathcal{T}_{k+1}$ denote topics of the k-th and (k+1)-th level, respectively. Unlike $\mathrm{R}_s(\cdot,\cdot)$, $\mathrm{R}_a(\cdot|\cdot)$ has the property that $\mathrm{R}_a(A|B) = \mathrm{R}_a(B|A) \neq 0$ iff. $\mathrm{Vol}\left(\mathrm{Box}(A)\right) = \mathrm{Vol}\left(\mathrm{Box}(B)\right)$. Thus $\mathrm{R}_a(\cdot|\cdot)$ can better model the hierarchical relation that is asymmetric.

Discussion of Box Embeddings for Taxonomy Learning

Most of the previous works (Vilnis et al., 2018; Lees et al., 2020; Dasgupta et al., 2020) learn box embeddings of pre-defined entities or words for taxonomy completion in a supervised manner. For instance, Vilnis et al. (2018) first proposed to train box embeddings for words on the incomplete ontology, in order to infer missing hypernym relations. Unlike these supervised methods, this paper aims at self-supervised topic taxonomy construction from unstructured text via box embeddings. This research problem poses new challenges for box embedding learning. Accordingly, we propose a recursive clustering algorithm for self-supervised box embedding learning, which is integrated with a VAE framework to provide an efficient solution for topic taxonomy construction based on box embeddings.

4 Proposed Method

In this section, we introduce the proposed BoxTM in detail. Firstly, we propose the box embedding-based document generative process in Section 4.1, which is the main framework of BoxTM. In general, BoxTM infers topic distributions via the symmetric affinities and semantic scopes of topics and words in the box embedding

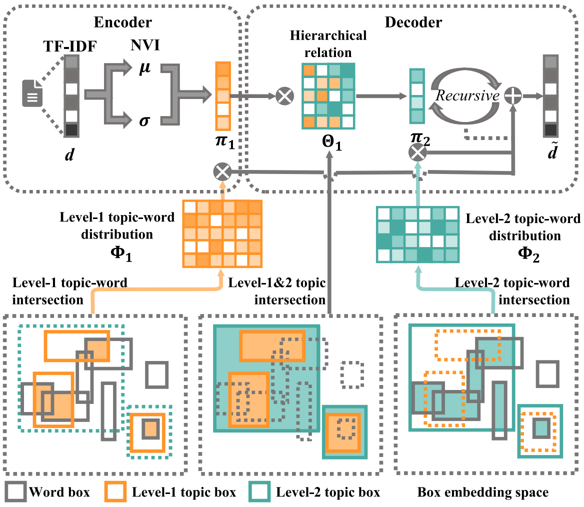


Figure 2: The main framework of BoxTM.

space. Additionally, the hierarchical relations are modeled by the values of the asymmetric metric between topic boxes. Subsequently, we introduce more detailed designs of BoxTM, including a novel workflow of recursive topic clustering for upper-level topic mining (Section 4.2) and two self-training tasks for modeling the semantic scopes of words and topics better (Section 4.3). Finally, we introduce the learning strategy of BoxTM in Section 4.4.

4.1 Document Generative Process

BoxTM holds the assumption that a document is generated by any topics in the topic taxonomy and adopts a bottom-up hierarchical topic discovery method following Chen et al. (2021b). For NVI, BoxTM adopts a classic Variational AutoEncoder (VAE) with a logistic normal distribution $\mathcal{LN}(\mathbf{0}, \boldsymbol{I})$ (Atchison and Shen, 1980) as the prior of topic proportion. A VAE consists of an encoder that learns hierarchical topic proportions given document representations and a decoder that reconstructs documents based on hierarchical topic proportions and topic distributions. Figure 2 shows the main framework of BoxTM.

Given a corpus $\mathcal{D}$ and a vocabulary $\mathcal{V}$, BoxTM firstly encodes the TF-IDF representation $\boldsymbol{d} \in \mathbb{R}^{|\mathcal{V}|}$ of each document into a latent distribution, from which the latent feature $\boldsymbol{z}$ is sampled. After transforming $\boldsymbol{z}$ to acquire the leaf topic proportion $\boldsymbol{\pi}_1$, we infer upper-level topic proportions $\{\boldsymbol{\pi}_{>1}\}$ based on the asymmetric relations $\{\Theta_k\}$ of topics in the box embedding space. Specifically, $\Theta_k \in \mathbb{R}^{|\mathcal{T}_k| \times |\mathcal{T}_{k+1}|}$ between level-k topics $\mathcal{T}_k$ and the upper-level topics $\mathcal{T}_{k+1}$ are estimated by the asymmetric metric $\mathrm{R}_a(\cdot|\cdot)$, i.e.,

$$\Theta_k^{ij} = \log \mathrm{R}_a\left(t_k^i \middle| t_{k+1}^j\right), \tag{3}$$

where $t_k^i \in \mathcal{T}_k$ and $t_{k+1}^j \in \mathcal{T}_{k+1}$. The encoding process of BoxTM is defined as follows:

$$\boldsymbol{h} = f_{\boldsymbol{h}}(\boldsymbol{d}), \tag{4}$$

$$\boldsymbol{z} \sim \mathcal{N}\left(f_{\boldsymbol{\mu}}(\boldsymbol{h}), f_{\boldsymbol{\sigma}}(\boldsymbol{h})\right), \tag{5}$$

$$\boldsymbol{\pi}_1 = \mathrm{Softmax}\left(f_{\boldsymbol{\pi}}(\boldsymbol{z})\right), \tag{6}$$

$$\boldsymbol{\pi}_{k+1} = \mathrm{Softmax}\left(\boldsymbol{\pi}_k \Theta_k\right), \tag{7}$$

where $f_{\boldsymbol{h}}(\cdot)$, $f_{\boldsymbol{\mu}}(\cdot)$, $f_{\boldsymbol{\sigma}}(\cdot)$, and $f_{\boldsymbol{\pi}}(\cdot)$ are feedforward neural networks. As the sampling process for the latent feature $\boldsymbol{z}$ is not differentiable, we adopt the reparameterization trick (Rezende et al., 2014) to make the gradient descent possible. Specifically, the sampled feature $\boldsymbol{z}$ can be expressed by a standard normal distribution, i.e., $\boldsymbol{z} = f_{\boldsymbol{\mu}}(\boldsymbol{h}) + \epsilon \cdot f_{\boldsymbol{\sigma}}(\boldsymbol{h}), \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

For the decoding process of BoxTM, we apply normalization before document reconstruction to enhance the generation power of weak topic levels (Hung et al., 2019), which is defined as follows:

$$\tilde{\boldsymbol{d}} = \sum_{k=1}^{K} (\boldsymbol{\pi}_k \cdot \Phi_k) \circ \mathrm{CV}_{\Phi_k} / Z_k, \tag{8}$$

where K is the depth of the topic taxonomy and $\circ$ denotes the element-wise multiplication. $\Phi_k \in [0,1]^{|\mathcal{T}_k| \times |\mathcal{V}|}$ is topic-word distributions of the k-th level and $Z_k = ||(\boldsymbol{\pi}_k \cdot \Phi_k) \circ \mathrm{CV}_{\Phi_k}||_2$ is a 2-norm term. To weaken the impact of common words on document generation, we adopt the Coefficient of Variation (CV) (Brown, 1998) to sharpen all topic-word distributions $\{\Phi_k\}$. Specifically, the j-th element of $\mathrm{CV}_{\Phi_k} \in \mathbb{R}^{|\mathcal{V}|}$ is the ratio of the standard deviation to the mean of the j-th column in Φ_k, which is defined by $\mathrm{CV}_{\Phi_k}^j = \sigma(\Phi_k^{:,j}) / \mu(\Phi_k^{:,j})$.

Notably, BoxTM infers topic-word distributions over the vocabulary $\mathcal{V}$ via the normalized symmetric affinity between topic and word boxes.

For the i-th topic t_k^i at level-k and the j-th word w_j in $\mathcal{V}$,

$$\Phi_k^{ij} = \text{Softmax}\left(\log \frac{\mathrm{R}_s(t_k^i, w_j)}{\mathrm{Vol}(t_k^i) \cdot \mathrm{Vol}(w_j)}\right), \quad (9)$$

which enables abstract topics to bias toward general words, and vice versa.

In summary, we describe the document generative process of BoxTM as follows:

▷ *For global topics, $k \in \{1, \ldots, K\text{-}1\}$:*

1. Infer the hierarchical relations between level-k and level-(k+1) topics Θ_k by Eq. (3).
2. Infer the topic-word distribution Φ_k by Eq. (9).

▷ *For each document:*

1. Draw the leaf topic proportion $\boldsymbol{\pi}_1 \sim \mathcal{LN}(\mathbf{0}, \boldsymbol{I})$.
2. Infer the upper-level topic proportion $\boldsymbol{\pi}_{k+1}$ by Eq. (7), for level $k \in \{1, \ldots, K\text{-}1\}$.
3. For each word w_j in the document:

   a. Draw topic level $k \sim$ **Uniform**(K).

   b. Draw topic assignment $t_k^i \sim$ **Cat**$(\boldsymbol{\pi}_k)$.

   c. Draw word $\hat{w}_j \sim$ **Cat**$(\Phi_k^{i,:})$.

4.2 Recursive Topic Clustering

Unlike most HNTMs that randomly initialize embeddings of topics in different abstraction levels, BoxTM conducts recursive clustering on topic boxes to learn upper-level topics. Notably, such a method can alleviate the problem of topic collapse, since the upper-level topic mining is guided by the correlation between lower-level topics. For the selection of clustering algorithms, we adopt the Affinity Propagation (AP) (Frey and Dueck, 2007) algorithm for its flexibility and interpretability.[2]

BoxTM constructs a topic affinity graph for topics at each level, where topic nodes are connected if their boxes overlap. However, the direct correlation between topics may be sparse in the box embedding space due to the diversity of topics, i.e., $\mathrm{Vol}(\mathrm{Box}(t_k^i) \wedge \mathrm{Box}(t_k^j)) \to 0, \forall t_k^i, t_k^j \in \mathcal{T}_k$. To address this, we expand the semantic scope of each topic by merging the information of its keyword boxes. The box embedding of the processed i-th topic $\tilde{t}_k^i$ at level-k is defined as follows:

$$\mathrm{Box}(\tilde{t}_k^i) := \left[\vee_{w \in \mathcal{W}_k^i} \mathrm{Box}(w)\right] \vee \mathrm{Box}(t_k^i), \quad (10)$$

where $\mathcal{W}_k^i = \{w_j \,|\, \arg\max_j \Phi_k^{ij}\}$ with $|\mathcal{W}_k^i| = n$ denotes the set of top-n (n = 5 in our experiments) representative words of topic t_k^i. Next, the affinity between topics is measured by the value of the asymmetric metric $\mathrm{R}_a(\cdot|\cdot)$ instead of the symmetric similarity metric $\mathrm{R}_s(\cdot,\cdot)$, because $\mathrm{R}_a(\cdot|\cdot)$ can weaken the influence of hub topics in clustering and prevent over-smoothing. Formally, the affinity matrix $\mathcal{A}_k \in \mathbb{R}^{|\mathcal{T}_k| \times |\mathcal{T}_k|}$ is defined by

$$\mathcal{A}_k^{ij} = \begin{cases} \log \mathrm{R}_a(\tilde{t}_k^j | \tilde{t}_k^i) & , i \neq j; \\ 0 & , i = j. \end{cases} \quad (11)$$

Later, the union of topic boxes in each cluster is adopted as a reasonable initialization of an upper-level topic. To reduce the impact of outliers in clustering, we propose a soft union operation $\vee_\dagger$, which is defined as follows:

$$\begin{aligned} &\mathrm{Box}\left(t_{k+1}^i\right) := (x_m^i, x_M^i) = \vee_{\dagger t \in \mathcal{C}_k^i} \mathrm{Box}(\tilde{t}), \\ &x_m^i = \mu(\{x_m^t\}_{t \in \mathcal{C}_k^i}), x_M^i = \mu(\{x_M^t\}_{t \in \mathcal{C}_k^i}), \end{aligned} \quad (12)$$

where $\mathcal{C}_k^i$ is the i-th topic cluster of the k-th level and $\mu(\cdot)$ is the mean operation. Additionally, $\mathrm{Box}\left(t_{k+1}^i\right)$ is the reinitialized box embedding for the upper-level topic t_{k+1}^i. Then BoxTM infers the hierarchical relations Θ_k between level-k and level-(k+1) topics based on their box embeddings. For each topic $t_k^i \in \mathcal{T}_k$ at the k-th level, its most relevant topic at the upper level is adopted as its parent topic $t_p^i \in \mathcal{T}_{k+1}$. Formally, we have

$$t_p^i := t_{k+1}^j = \arg\max_j \Theta_k^{ij}. \quad (13)$$

After conducting (K-1) times of topic clustering recursively, BoxTM can mine topics of K levels in a bottom-up manner.

[2] Compared to the AP algorithm, centroid-based methods such as k-means++ (Arthur and Vassilvitskii, 2007) cannot accommodate non-flat geometries like the box embedding space, while density-based DBSCAN (Ester et al., 1996) is vulnerable to the setting of hyperparameters.

4.3 Semantic Scope Modeling

The effectiveness of our box embedding-based document generative process with recursive topic clustering is based on an important premise that box embeddings can accurately model the semantic scopes of words and topics. Here we propose two self-supervised tasks by means of word-level and topic-level constraints for semantic scope modeling.

4.3.1 Word-level Constraint

Importantly, the semantic scope of each word consists of its ***abstraction level*** and ***semantics***, which correspond to the ***volume*** and ***position*** of its box, respectively. Inspired by GloVe (Pennington et al., 2014), we propose to encode the (co-)occurrence patterns of words into word boxes.

Our key insight is that the marginal probability $P(w_j)$ of word w_j reveals its abstraction level. Besides, as the distributional hypothesis states that similar words w_i and w_i' tend to co-occur with the same word w_j, the joint probability $P(w_i, w_j)$ may reflect the correlation between the semantics of w_i and w_j. In practice, the joint and marginal probabilities can be estimated by $P(w_i, w_j) \sim X_{ij}$ and $P(w_j) \sim X_j$, where X_{ij} is the co-occurrence time of w_i and w_j in the corpus, and $X_j = \sum_{w_n \in \mathcal{V}} X_{jn}$. Integrating these patterns, we propose that the values of the asymmetric metric $\mathrm{R}_a(w_i|w_j)$ in the box embedding space should be consistent with the conditional probability $P_{i|j} = P(w_i|w_j) = X_{ij}/X_j$.

For the word-level constraint of semantic scope modeling, the Mean-Square Error (MSE) loss is a straightforward selection, i.e., $\mathcal{L}_{CO} = \left\| \mathrm{R}_a(w_i|w_j) - P_{i|j} \right\|_2^2$. However, the MSE loss strongly restricts the absolute volumes of word boxes, which is difficult for training. Therefore, we adopt the cross-entropy loss $H(\cdot,\cdot)$ to constrain the relative volumes of word boxes among a randomly sampled batch $\mathcal{B} = \{(w_i, w_j)|P_{i|j} > 0\}$. Formally, we denote the box volume distribution as $q_{\text{Box}}(w_i, w_j) \sim \mathrm{R}_a(w_i|w_j)$ and the co-occurrence pattern distribution as $p_{CO}(w_i, w_j) \sim P_{i|j}$. Then the loss function is defined by

$$\begin{aligned}\mathcal{L}_{CO} &= H(p_{CO}, q_{\text{Box}}) \\ &= -\sum_{(w_i,w_j)\in\mathcal{B}} p_{CO}(w_i, w_j) \log q_{\text{Box}}(w_i, w_j).\end{aligned} \tag{14}$$

4.3.2 Topic-level Constraint

In a reasonable topic taxonomy $\mathcal{S}$, the semantic scope of a parent topic t_p should cover that of its child topic t_c (Viegas et al., 2020). In other words, the box embedding of t_p should entail that of t_c. Intuitively, we can define the following loss to maximize the score of asymmetric correlation metric between t_p and t_c:

$$\begin{aligned}\mathcal{L}_{HT} &= -\sum_{(t_p,t_c)\in\mathcal{S}} \log \mathrm{R}_a(t_c|t_p) \\ &= -\sum_{(t_p,t_c)\in\mathcal{S}} \log \mathrm{R}_s(t_c, t_p) - \log \mathrm{Vol}(t_p),\end{aligned} \tag{15}$$

where the first term $\mathrm{R}_s(t_c, t_p)$ regularizes the semantic coherence between t_p and t_c. However, the second term of the above definition may lead to a trivial solution that all topic boxes collapse to points, i.e., $\mathrm{Vol}(t) \to 0$ and then $\mathrm{R}_s(t_c, t_p) \to 0$, $\forall t, t_c, t_p$. To avoid this problem, we replace the second term with a max-margin objective, which makes the box of t_p larger than that of t_c by at least the margin m. So $\mathcal{L}_{HT}$ is redefined as follows:

$$\begin{aligned}\mathcal{L}_{HT} = &-\sum_{(t_p,t_c)\in\mathcal{S}} \log \mathrm{R}_s(t_c, t_p) \\ &- \max\left[0, m - \log \mathrm{Vol}(t_p) + \log \mathrm{Vol}(t_c)\right].\end{aligned} \tag{16}$$

4.4 Learning Strategy

Similar to the training objective of VAEs, the main loss of BoxTM is to maximize the Evidence Lower BOund (ELBO). Specifically, the ELBO loss of BoxTM is defined by

$$\begin{aligned}\mathcal{L}_{ELBO} =& \mathbb{E}_{\boldsymbol{\pi}_1 \sim q_{\boldsymbol{d}}} \log p(\boldsymbol{d}|\{\boldsymbol{\pi}_k\}, \{\Phi_k\}) \\ &- D_{KL}\left[q_{\boldsymbol{d}}(\boldsymbol{\pi}_1)||p(\boldsymbol{\pi}_1)\right],\end{aligned} \tag{17}$$

which balances between maximising the expected log-likelihood (the first term) and minimising the KL divergence (the second term) of the variational distribution $q_{\boldsymbol{d}}(\boldsymbol{\pi}_1) := \mathcal{N}(f_\mu(\boldsymbol{d}), f_\sigma(\boldsymbol{d}))$ and the prior distribution $p(\boldsymbol{\pi}_1) := \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

For modeling the semantic scopes of words and topics, we propose two constraints in

```
Algorithm 1 The $i$-th epoch of training
Input: The corpus $\mathcal{D}$ and its vocabulary $\mathcal{V}$; The word and topic box embeddings $W$ and $\{T_k\}$; The topic taxonomy $\mathcal{S}$ after prior epoch; The threshold $\gamma$ for early stop.
Output: Updated word and topic box embeddings $\tilde{W}$ and $\{\tilde{T}_k\}$; Updated topic taxonomy$\tilde{\mathcal{S}}$.
if $i < \gamma$ then
    $\tilde{\mathcal{S}}, \{\tilde{T}_k\} \leftarrow$ RecurClus$(W, T_1, K)$
else $\tilde{\mathcal{S}}, \{\tilde{T}_k\} \leftarrow \mathcal{S}, \{T_k\}$
for each batch $\mathcal{B} \subset \mathcal{D}$ do
    Infer hierarchical relations $\Theta$ by Eq. (3).
    Infer topic-word distributions $\Phi$ by Eq. (9).
    for each document $\boldsymbol{d} \in \mathcal{D}$ do
        Draw topic proportions $\{\boldsymbol{\pi}_k\} \leftarrow$ Encode$(\boldsymbol{d}, \Theta)$.
        Reconstruct document $\tilde{\boldsymbol{d}} \leftarrow$ Decode$(\{\boldsymbol{\pi}_k\}, \Phi)$.
    Compute loss $\mathcal{L} = \mathcal{L}_{ELBO} + \mathcal{L}_{Box}$.
    Update $\tilde{W}$ and $\{\tilde{T}_k\}$ by minimizing $\mathcal{L}$.
    Update $\tilde{\mathcal{S}}$ based on $\{\tilde{T}_k\}$ by Eq. (13).
return $\tilde{W}$, $\{\tilde{T}_k\}, \tilde{\mathcal{S}}$
```

Section 4.3. Accordingly, we define the regularization loss by

$$\mathcal{L}_{Box} = \alpha \cdot \mathcal{L}_{CO} + \beta \cdot \mathcal{L}_{HT}, \tag{18}$$

where α and β are weights for these losses. And the overall loss function of BoxTM is defined by

$$\mathcal{L} = \mathcal{L}_{ELBO} + \mathcal{L}_{Box}. \tag{19}$$

Then we adopt the Adam optimizer to update the network parameters of the encoder and box embeddings of topics and words. Based on the updated topic boxes, we perform a correction for the topic taxonomy using Eq. (13). The training workflow of BoxTM is shown in Algorithm 1. Intuitively, topic boxes overlap less along with the training to capture diverse semantics, which limits the effectiveness of our recursive clustering module at the late phase of training. To tackle this problem, we use the early stopping trick that stops recursive clustering after the γ-th iteration. In the following experiments, γ is set to 100.

	#document				
dataset	**#train**	**#valid**	**#test**	**#word**	**#class**
20news	9,007	2,251	7,487	1,838	20
NYT	6,279	1,569	5,233	8,171	25
arXiv	110,451	27,612	92,042	11,799	53

Table 1: Statistics of datasets.

5 Experiments

5.1 Experimental Settings

5.1.1 Datasets

We conduct comprehensive evaluations on three benchmark datasets with latent topic hierarchies: (1) **20news**[3]: A corpus consists of 20 newsgroups (Song and Roth, 2014). (2) **NYT**[4]: A set of news articles from the *New York Times*, which are categorized into 25 classes. (3) **arXiv**[5]: A set of paper abstracts covering 53 classes from arXiv website. The latter two datasets are collected by Meng et al. (2019). Table 1 shows the statistics of all datasets. After preprocessing of removing stopwords and low-frequency words, we split documents into a training set and a testing set with the ratio of 6:4. In addition, we adopt 20% of documents in the training set as a validation set.

5.1.2 Baselines

We compare our model with state-of-the-art topic taxonomy discovery models based on different frameworks, including document generation-based methods of **nTSNTM**[6] (Chen et al., 2021b), **SawETM**[7] (Duan et al., 2021a), **HyperMiner**[8] (Xu et al., 2022), and **C-HNTM**[9] (Wang et al., 2023), as well as a clustering-based method of **TaxoGen**[10] (Zhang et al., 2018). Notably, HyperMiner adopts the hyperbolic embedding space, and the others hold the Euclidean embedding space assumption.

5.1.3 Hyperparameter Settings

The maximum depth of the topic taxonomy is set to 3 for the 20news and NYT datasets following

[3] http://qwone.com/~jason/20Newsgroups/.

[4] http://developer.nytimes.com/.

[5] https://arxiv.org/.

[6] https://github.com/hostnlp/nTSNTM.

[7] https://github.com/BoChenGroup/SawETM.

[8] https://github.com/NoviceStone/HyperMiner.

[9] https://github.com/Jladygoogoo/C-HNTM.

[10] https://github.com/franticnerd/taxogen.

	20news				NYT				arXiv			
model	C	D	C*D	HC	C	D	C*D	HC	C	D	C*D	HC
nTSNTM	0.212	0.728	0.154	0.134	0.221	0.420	0.093	0.079	–	–	–	–
SawETM	0.221	0.404	0.089	0.098	0.228	0.476	0.109	0.084	0.134	0.256	0.034	0.047
HyperMiner	0.224	0.459	0.103	0.102	0.231	0.500	0.115	0.101	0.142	0.382	0.054	0.050
C-HNTM	0.196	0.633	0.124	0.090	0.152	0.458	0.070	0.036	–	–	–	–
TaxoGen	0.202	**0.789**	0.159	0.123	0.239	**0.881**	0.210	0.111	0.214	**0.681**	0.146	0.084
BoxTM	**0.301**	0.661	**0.199**	**0.159**	**0.409**	0.648	**0.265**	**0.177**	**0.257**	0.672	**0.173**	**0.113**

Table 2: Intrinsic metric scores on three datasets.

Chen et al. (2021b). To evaluate the flexibility of BoxTM and baseline models, the maximum depth for the large dataset arXiv is set to 5. Additionally, the maximum number of leaf topics $|\mathcal{T}_1|^{\max}$ of nTSNTM is 200 following the setting in its paper, which can get a reasonable number of topics adaptively based on the stick-breaking process. According to the number of active topics obtained by nTSNTM, $|\mathcal{T}_1|^{\max}$ of BoxTM and the other HNTMs is set to 50/50/100 for three datasets, respectively. For TaxoGen, the maximum number of clusters is set to 5/5/3. The embedding dimension of BoxTM is set to 50 following Vilnis et al. (2018). Since box embeddings have 2 parameters per dimension, the embedding size of baselines are set to 100 for a fair comparison.

Other hyperparameters of baselines take the optimal values reported in their papers. For BoxTM, the learning rate is 5e-3, the dimension of hidden layers is 256, and the max margin m is set to 10. The weight of $\mathcal{L}_{HT}$ gradually increases to the maximum value ($\beta^{max} = 0.005$) during training, when the constant weight of $\mathcal{L}_{CO}$ is set to 3.

5.2 Intrinsic Evaluation of Topic Taxonomy

For a reasonable topic taxonomy, each topic is a set of closely coherent words and diverse from one another. Also, keywords of a parent topic t_p and its child topic t_c are coherent but have different semantic abstraction levels. Thus we validate the quality of the topic taxonomy from the following perspectives: (1) **Topic Coherence (C)**: We adopt a classic metric NPMI (Lau et al., 2014) to quantify the coherence of mined topics. (2) **Topic Diversity (D)**: The widely used TU (Nan et al., 2019) metric is for assessing the diversity among all topics, which is calculated by the number of unique keywords among all topics. (3) **Hierarchical Coherence (HC)**: We adopt the CLNPMI (Chen et al., 2021b) metric to evaluate the hierarchical coherence between topics t_p and t_c.

Because highly overlapping topics may cause inflated coherence scores, the product of NPMI and TU are used as an integrated metric (**C*D**) for a comprehensive validation (Dieng et al., 2020). For the aforementioned metrics, we calculate the average of the scores of top-5, top-10, and top-15 topic words. Because the source code of nTSNTM and the algorithm of C-HNTM cannot adapt to topic taxonomy with more than 3 levels, their results on the arXiv dataset are not reported.

As shown in Table 2, BoxTM achieves new state-of-the-art results on most metrics across three datasets, when HyperMiner using hyperbolic embeddings outperforms SawETM. These results validate the advantage of geometric (i.e., hyperbolic and box) embeddings on topic taxonomy discovery over traditional point embeddings. Compared to C-HNTM that performs poorly on the **HC** metric, the proposed recursive topic clustering module of BoxTM can effectively learn topics of different levels. While both SawETM and HyperMiner fail to learn a deep topic taxonomy on the arXiv dataset with massive documents, BoxTM remains outstanding performance on topic quality and hierarchical coherence. It validates that BoxTM not only has scalability for large-scale data but also has flexibility to learn topic taxonomies of different structures. In terms of the clustering-based method, TaxoGen obtains high scores of topic diversity (**D**), because each word only belongs to one topic at each level in its approach. However, it neglects the polysemy of some words, i.e., a word can be the keyword of different topics, which leads to its performance decline on topic coherence. For example, the word ''*driver*'' could be the keyword of topics ''*hardware*'' and ''*motorcycles*''.

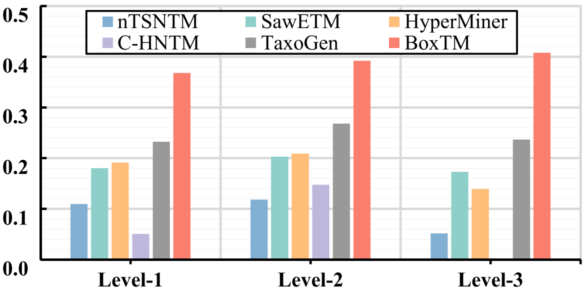


Figure 3: The **C*D** scores at each level of BoxTM and baselines on NYT.

	20news		NYT		arXiv	
model	**ARI**	**F$_\beta$**	**ARI**	**F$_\beta$**	**ARI**	**F$_\beta$**
nTSNTM	0.081	0.133	0.389	0.448	–	–
SawETM	0.074	0.123	0.452	0.494	**0.151**	**0.184**
HyperMiner	0.075	0.127	0.421	0.466	0.115	0.151
C-HNTM	0.056	0.104	0.143	0.216	–	–
TaxoGen	0.066	0.132	0.310	0.367	0.097	0.133
BoxTM	**0.117**	**0.168**	**0.541**	**0.577**	0.103	0.143

Table 3: Extrinsic metric scores on three datasets.

Furthermore, Figure 3 illustrates the **C*D** scores at each level of BoxTM and baselines on the NYT dataset. Both coherence and diversity of the level-2 topics of all models have different degrees of improvement compared to leaf topics. However, most baselines fail to learn high-quality topics at the root level, that is, they encounter the topic collapse problem. And topics mined by BoxTM remain high-quality at all levels, due to the effectiveness of the proposed recursive topic clustering module.

5.3 Extrinsic Evaluation of Topic Taxonomy

As an important application scenario for topic taxonomy discovery, the tree structure and keywords of the mined topic taxonomy can serve as auxiliary knowledge to improve the performance of hierarchical text clustering (Lee et al., 2022). Specifically, each topic is regarded as a cluster, characterized by its keywords. We utilize the topic structure and the top-15 keywords of all topics learned by our BoxTM and baseline models as the inputs of a hierarchical text clustering model named WeSHClass (Meng et al., 2019). For the evaluation metrics, we adopt two external criteria of clustering (i.e., **ARI** and **F**$_\beta$) using golden labels of documents (Steinbach et al., 2005).

Table 3 shows the results of BoxTM and baseline models on the hierarchical text clustering task. Particularly, BoxTM and other HNTMs significantly outperform C-HNTM and TaxoGen that conduct clustering on word embeddings to mine topics, which reveals the limitation of latter methods in learning document-level semantics. Among HNTMs, BoxTM achieves the best results overall (**ARI** $= 0.254$ and **F**$_\beta = 0.296$ in average), followed by SawETM (**ARI** $= 0.226$ and **F**$_\beta = 0.267$ in average). Although SawETM outperforms BoxTM on the arXiv dataset, it cannot discover coherent topics according to the intrinsic evaluation. These results show that there is a tradeoff between learning high-quality topics and document-level semantics for topic modeling methods, and our BoxTM strikes a good balance.

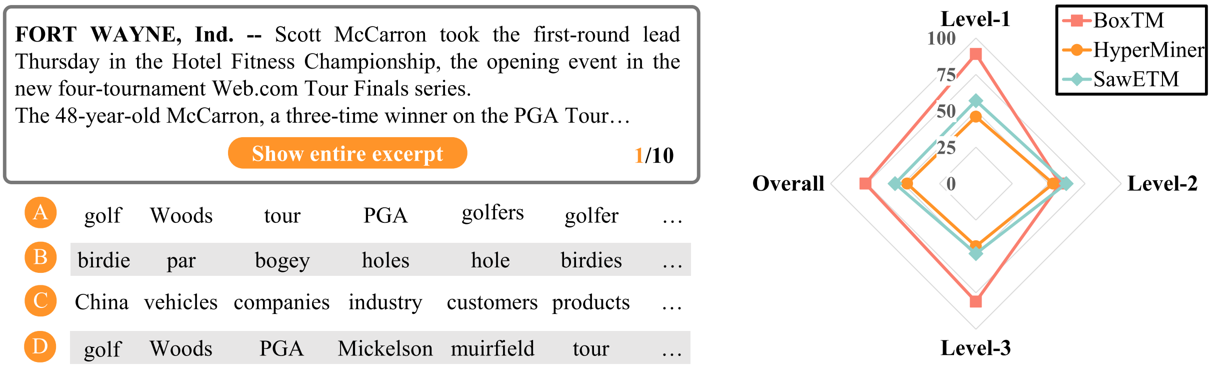
5.4 Human Evaluation

To complement the above automatic metrics, we also utilize a manual evaluation task of *topic intrusion* (Chang et al., 2009) to further validate the ability of topics at different levels to describe documents. As shown in Figure 4 (left), human raters are shown a document from the testing set of NYT, along with four topics represented by their top-10 keywords. Three of them are the top-3 topics at the same level assigned to the given document by the topic model, while the remaining *intruder topic* is sampled randomly from the other low probability topics. We recruit ten graduate students majoring in computer science as raters and instruct them to choose topics that are not relevant to the documents. For evaluation, we compare our BoxTM with two strong baselines, i.e., SawETM and HyperMiner, excluding TaxoGen that cannot infer the topic distributions of documents. According to the value of Light's kappa (Light, 2011) (κ = 0.607), the annotation results of the ten raters have a fairly high degree of agreement.

Figure 4 (right) shows the precision scores of different models on this task. The performance of all three models on the manual assessment is generally consistent with those on the extrinsic evaluation. Notably, our BoxTM achieves an overall optimal result, which indicates that it generates different levels of topics that describe documents in alignment with human judgment.

5.5 Ablation Analysis

In this section, we conduct an ablation study to analyze the roles of several key components of



FORT WAYNE, Ind. -- Scott McCarron took the first-round lead Thursday in the Hotel Fitness Championship, the opening event in the new four-tournament Web.com Tour Finals series.
The 48-year-old McCarron, a three-time winner on the PGA Tour…

Show entire excerpt 1/10

A golf Woods tour PGA golfers golfer …

B birdie par bogey holes hole birdies …

C China vehicles companies industry customers products …

D golf Woods PGA Mickelson muirfield tour …

Figure 4: Illustration of the human evaluation on the NYT dataset: An example of the topic intrusion task (left) and the average precision (%) of our BoxTM and strong baselines (right).

embedding	model	C*D	HC	ARI	F_β
box	**BoxTM**	0.265	0.177	**0.541**	**0.577**
	wo/ $\mathcal{L}_{CO}$	0.266	**0.191**	0.449	0.489
	wo/ $\mathcal{L}_{HT}$	**0.276**	0.157	0.299	0.355
	wo/ clus	0.256	0.139	0.337	0.394
point	w/ kmeans	0.201	0.174	0.397	0.441
	w/ AP	0.241	0.158	0.444	0.488
	w/ hier	0.208	0.162	0.417	0.458
	wo/ clus	0.193	0.153	0.376	0.423

Table 4: Intrinsic and extrinsic metric scores of ablation models on NYT.

BoxTM, whose results are shown in Table 4. Most importantly, the ablation models that replace box embeddings with traditional point embeddings (i.e., the **point** models), experience a drastic performance drop in both topic quality and extrinsic evaluation compared to BoxTM. Within several clustering algorithms, the **point** model using AP clustering (w/ AP) performs better than those with kmeans++ (w/ kmeans) or agglomerative clustering (w/ hier).

In terms of the proposed box embedding regularizations, BoxTM wo/ $\mathcal{L}_{HT}$ fails to capture the proper semantic scopes of topics at different levels, leading to worse performance on the **HC** metric as well as the downstream task. Though BoxTM wo/ $\mathcal{L}_{CO}$ remains competitive on intrinsic evaluation, its performance on the hierarchical text clustering task drops compared to BoxTM.

5.6 Case Study of Topic Taxonomy

In this section, we evaluate the mined topic taxonomy qualitatively via a case study. Figure 5(a) illustrates some sample topics from the 5-level topic taxonomy learned by BoxTM on the arXiv dataset. A level-4 topic about ‘‘*network*’’ branches into child topics related to ‘‘*computer communication networks*’’ (left), ‘‘*optimization algorithms*’’ (middle), and ‘‘*applications*’’ (right). Furthermore, in the field of ‘‘*applications*’’, there are sub-fields that focus on different research problems, including ‘‘*computation and language*’’ and ‘‘*computer vision and pattern recognition*’’. Moreover, Figure 5(b) shows some topics related to ‘‘*sports*’’ and ‘‘*administration*’’ mined by BoxTM on NYT.

5.7 Analysis of Taxonomy Depth

In the aforementioned experiments, we set the maximum depth to the same value for all models by following Chen et al. (2021b). As a complement, Figure 6 illustrates the performance of our BoxTM compared to the top-2 best performing baselines (i.e., TaxoGen and HyperMiner) for different settings of taxonomy depth. In most cases, BoxTM outperforms baselines with the same taxonomy depth. Nevertheless, how to determine an appropriate taxonomy depth in the real-life applications is a valuable but challenging problem.

Considering that the automatic metrics (e.g., **C** and **HC**) may be sensitive to the taxonomy depth, we also conduct a qualitative analysis to discuss the influence of taxonomy depths on our BoxTM. As shown in Figure 7, the leaf topic about ‘‘*Galerkin methods*’’ is assigned to the parent topic related to ‘‘*numerical analysis*’’ for $K = 3$. And when $K = 4$, BoxTM further extracts a level-4 topic that is related to ‘‘*general algorithm*’’. Interestingly, when the structure of the taxonomy continues to deepen ($K = 5$), BoxTM identifies that ‘‘*Galerkin methods*’’ is commonly applied in the field of ‘‘*physics*’’ as a classic PDE solver. Overall, our BoxTM can discover

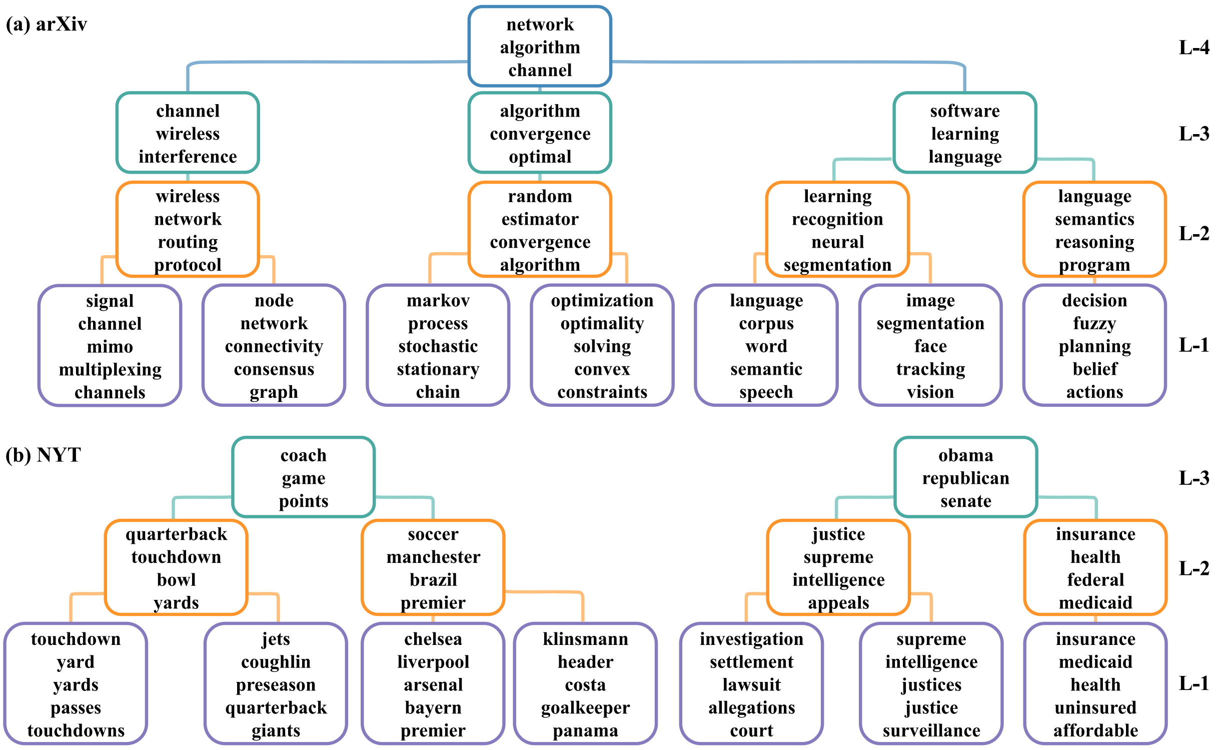

(a) arXiv
network algorithm channel
L-4
channel wireless interference
algorithm convergence optimal
software learning language
L-3
wireless network routing protocol
random estimator convergence algorithm
learning recognition neural segmentation
language semantics reasoning program
L-2
signal channel mimo multiplexing channels
node network connectivity consensus graph
markov process stochastic stationary chain
optimization optimality solving convex constraints
language corpus word semantic speech
image segmentation face tracking vision
decision fuzzy planning belief actions
L-1
(b) NYT
coach game points
obama republican senate
L-3
quarterback touchdown bowl yards
soccer manchester brazil premier
justice supreme intelligence appeals
insurance health federal medicaid
L-2
touchdown yard yards passes touchdowns
jets coughlin preseason quarterback giants
chelsea liverpool arsenal bayern premier
klinsmann header costa goalkeeper panama
investigation settlement lawsuit allegations court
supreme intelligence justices justice surveillance
insurance medicaid health uninsured affordable
L-1

Figure 5: Illustration of the partial topic taxonomy learned by BoxTM on arXiv **(a)** and NYT **(b)**.

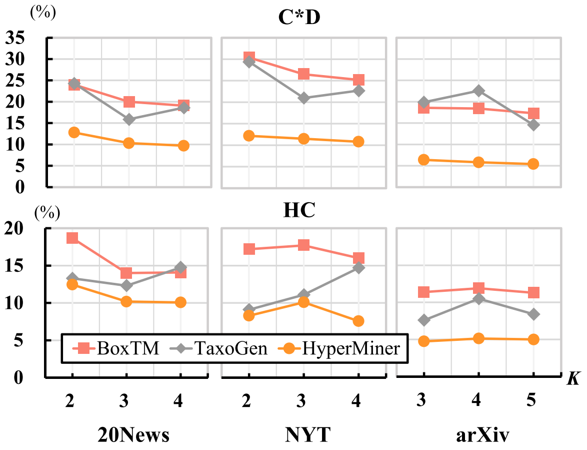


Figure 6: The **C*D** and **HC** scores of BoxTM, TaxoGen, and HyperMiner with different settings of taxonomy depth (i.e., K).

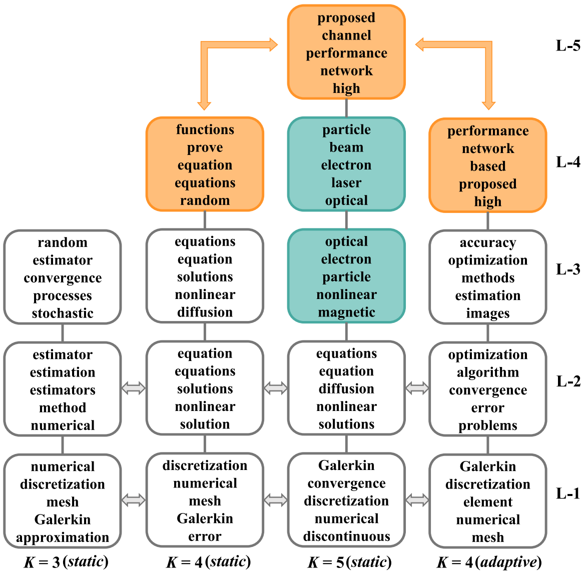

proposed channel performance network high
L-5
functions prove equation equations random
particle beam electron laser optical
performance network based proposed high
L-4
random estimator convergence processes stochastic
equations equation solutions nonlinear diffusion
optical electron particle nonlinear magnetic
accuracy optimization methods estimation images
L-3
estimator estimation estimators method numerical
equation equations solutions nonlinear solution
equations equation diffusion nonlinear solutions
optimization algorithm convergence error problems
L-2
numerical discretization mesh Galerkin approximation
discretization numerical mesh Galerkin error
Galerkin convergence discretization numerical discontinuous
Galerkin discretization element numerical mesh
L-1
K = 3 (static)
K = 4 (static)
K = 5 (static)
K = 4 (adaptive)

Figure 7: Pathways of the leaf topic about ‘‘*Galerkin methods*’’ obtained by BoxTM on the arXiv dataset, when the taxonomy depth (i.e., K) is set to different values.

topics with different granularity and the hierarchical relations under varying settings of taxonomy depth. Therefore, users can set the taxonomy depth according to their practical requirements.

Moreover, unlike most HTMs that require a fixed taxonomy depth, the recursive topic clustering module in BoxTM provides a promising solution for determining the taxonomy depth adaptively. Specifically, BoxTM can halt topic clustering when the number of topics at the top level is smaller than a threshold, which is easier to determine compared to the taxonomy depth. Figure 7 (*adaptive*) illustrates the topic pathway mined by BoxTM when the threshold is set to 10.

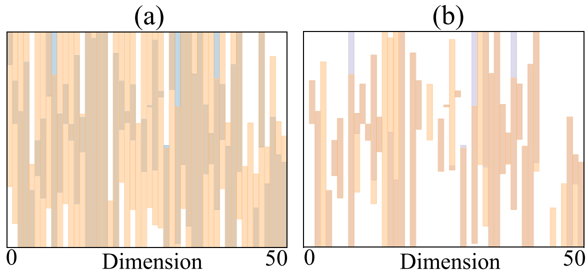


Figure 8: **(a)** Visualization of parent topic 2–5 (yellow) and child topic 1–13 (blue) boxes. **(b)** Visualization of intersection boxes of hierarchical topics (i.e., 1–13 and 2–5) (yellow) as well as irrelevant topics (i.e., 1–13 and 2–11) (purple).

5.8 Qualitative Analysis of Box Embeddings

In this section, we examine whether box embeddings can reflect the asymmetric relation between parent and child topics. For example, topic 2-5 (i.e., the 5-th topic at level-2) learned by BoxTM on NYT is related to ''*religion*'' and topic 1-13 is one of its children, while topic 1-27 is about ''*hardware*'', characterized by keywords such as ''*drive*'' and ''*controller*''. As shown in Figure 8(a), the boxes of upper-level topics entail those of their children. Besides, Figure 8(b) illustrates that the box embedding of child topic 1–13 has a larger overlap with its parent topic 2–5 compared to a randomly sampled topic 2–11, with $p = 0.007 < 0.05$ according to the paired sample t-test.

6 Conclusion

This paper proposes a novel model called BoxTM for self-supervised topic taxonomy discovery in the box embedding space. Specifically, BoxTM embeds both topics and words into the same box embedding space, where the symmetric and asymmetric metrics are defined to infer the complex relations among topics and words properly. Additionally, instead of initializing topic embeddings randomly, BoxTM uncovers upper-level topics via recursive clustering on topic boxes.

While our BoxTM has achieved state-of-the-art performance in multiple evaluation experiments, it also exhibits a limitation in efficiency. The **point** model, a variant of BoxTM that replaces the box embeddings with point embeddings, is trained for 0.22 GPU (GTX 1080 Ti) hour on the 20news dataset. Due to the extra computation of box operations compared to dot product, BoxTM costs about 1.0 hour, which reveals the research space for efficient computation of box embeddings.

Acknowledgments

We express our profound gratitude to the action editor and reviewers for their valuable comments and suggestions. This research has been supported by the National Natural Science Foundation of China (62372483), the Faculty Research Grants (DB24A4 and DB24C5) of Lingnan University, Hong Kong, the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19), and the Hong Kong Research Grants Council under the General Research Fund (project no. PolyU 15200021).

References

Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. Boxe: A box embedding model for knowledge base completion. In *NeurIPS*, pages 9649–9661.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *SODA*, pages 1027–1035.

J. Atchison and Sheng M. Shen. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272. `https://doi.org/10.1093/BIOMET/67.2.261`

Yushi Bai, Zhitao Ying, Hongyu Ren, and Jure Leskovec. 2021. Modeling heterogeneous hierarchies with relation-specific hyperbolic cones. In *NeurIPS*, pages 12316–12327.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS*, pages 17–24.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003b. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Charles E. Brown. 1998. Coefficient of variation. *Applied Multivariate Statistics in Geohydrology and Related Sciences*, pages 155–157, Springer. `https://doi.org/10.1007/978-3-642-80328-4_13`

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

Ziye Chen, Cheng Ding, Yanghui Rao, Haoran Xie, Xiaohui Tao, Gary Cheng, and Fu Lee Wang. 2021a. Hierarchical neural topic modeling with manifold regularization. *World Wide Web*, 24:2139–2160. https://doi.org/10.1007/s11280-021-00963-7

Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. 2021b. Tree-structured topic modeling with nonparametric neural variational inference. In *ACL/IJCNLP*, pages 2343–2353. https://doi.org/10.18653/v1/2021.acl-long.182

Shib Sankar Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. 2020. Improving local identifiability in probabilistic box embeddings. In *NeurIPS*, pages 182–192.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453. https://doi.org/10.1162/tacl_a_00325

Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. 2021a. Sawtooth factorial topic embeddings guided gamma belief network. In *ICML*, pages 2903–2913.

Zhibin Duan, Yishi Xu, Bo Chen, Dongsheng Wang, Chaojie Wang, and Mingyuan Zhou. 2021b. Topicnet: Semantic graph-guided topic discovery. In *NeurIPS*, pages 547–559.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976. https://doi.org/10.1126/science.1136800

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *CoRR*, abs/2203.05794.

Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. 2019. Scops: Self-supervised co-part segmentation. In *CVPR*, pages 869–878. https://doi.org/10.1109/CVPR.2019.00096

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-structured neural topic model. In *ACL*, pages 800–806. https://doi.org/10.18653/v1/2020.acl-main.73

Minhao Jiang, Xiangchen Song, Jieyu Zhang, and Jiawei Han. 2022. Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations. In *WWW*, pages 925–934. https://doi.org/10.1145/3485447.3511935

Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *CIKM*, pages 783–792. https://doi.org/10.1145/2396761.2396861

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539. https://doi.org/10.3115/v1/E14-1056

Dongha Lee, Jiaming Shen, Seongku Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *WWW*, pages 2819–2829. https://doi.org/10.1145/3485447.3512002

Alyssa Lees, Chris Welty, Shubin Zhao, Jacek Korycki, and Sara Mc Carthy. 2020. Embedding semantic taxonomies. In *COLING*, pages 1279–1291. https://doi.org/10.18653/v1/2020.coling-main.110

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. 2019. Smoothing the geometry of probabilistic box embeddings. In *ICLR*.

Richard J. Light. 2011. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76(5):365–377. https://doi.org/10.1037/h0031643

Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Q. Zhu. 2018. Knowledge base question answering via encoding of complex query graphs. In *EMNLP*, pages 2185–2194. https://doi.org/10.18653/v1/D18-1242

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *AAAI*, pages 6826–6833. https://doi.org/10.1007/978-3-031-01914-2_5

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *WWW*, pages 3143–3152. https://doi.org/10.1145/3485447.3512034

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical topic mining via joint spherical tree and text embedding. In *KDD*, pages 1908–1917. https://doi.org/10.1145/3394486.3403242

George A. Miller. 1995. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41. https://doi.org/10.1145/219717.219748

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *ICML*, pages 633–640. https://doi.org/10.1145/1273496.1273576

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. In *ACL*, pages 6345–6381. https://doi.org/10.18653/v1/P19-1640

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *NIPS*, pages 6341–6350.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! In *EMNLP*, pages 1728–1736. https://doi.org/10.18653/v1/2020.emnlp-main.135

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *AAAI*, pages 1579–1585. https://doi.org/10.1609/aaai.v28i1.8938

M. Steinbach, V. Kumar, and P. Tan. 2005. Cluster analysis: Basic concepts and algorithms. *Introduction to Data Mining*, 1st edn. Pearson Addison Wesley.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira De Souza Júnior, Leonardo Rocha, and Marcos André Gonçalves. 2020. Cluhtm - semantic hierarchical topic modeling based on cluwords. In *ACL*, pages 8138–8150. https://doi.org/10.18653/v1/2020.acl-main.724

Luke Vilnis. 2021. Geometric representation learning. Doctoral Dissertation, University of Massachusetts Amherst. https://doi.org/10.7275/20638273

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *ACL*, pages 263–272. https://doi.org/10.18653/v1/P18-1025

Ningjing Wang, Deqing Wang, Ting Jiang, Chenguang Du, Chuyu Fang, and Fuzhen Zhuang. 2023. Hierarchical neural topic model with embedding cluster and neural variational inference. In *SDM*, pages 936–944. https://doi.org/10.1137/1.9781611977653.ch105

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *ICML*, pages 37335–37357.

Ruobing Xie, Qi Liu, Liangdong Wang, Shukai Liu, Bo Zhang, and Leyu Lin. 2022. Contrastive cross-domain recommendation in matching. In *KDD*, pages 4226–4236. https://doi.org/10.1145/3534678.3539125

Yishi Xu, Dongsheng Wang, Bo Chen, Ruiying Lu, Zhibin Duan, and Mingyuan Zhou. 2022. Hyperminer: Topic taxonomy mining with hyperbolic embedding. In *NeurIPS*, pages 31557–31570.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *KDD*, pages 2701–2709. https://doi.org/10.1145/3219819.3220064