Hierarchical Topic Modeling via Contrastive Learning and Hyperbolic Embedding

Zhicheng Lin[1], **Hegang Chen**[1], **Yuyin Lu**[1], **Yanghui Rao**[1,*], **Hao Xu**[2,*], **Hanjiang Lai**[1]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Laboratory for Big Data and Decision, National University of Defense Technology, Changsha, China
{linzhch26, chenhg25, luyy37}@mail2.sysu.edu.cn
{raoyangh, laihanj3}@mail.sysu.edu.cn, xuhao@nudt.edu.cn

Abstract

Hierarchical topic modeling, which can mine implicit semantics in the corpus and automatically construct topic hierarchical relationships, has received considerable attention recently. However, the current hierarchical topic models are mainly based on Euclidean space, which cannot well retain the implicit hierarchical semantic information in the corpus, leading to irrational structure of the generated topics. On the other hand, the existing Generative Adversarial Network (GAN) based neural topic models perform satisfactorily, but they remain constrained by pattern collapse due to the discontinuity of latent space. To solve the above problems, with the hypothesis of hyperbolic space, we propose a novel GAN-based hierarchical topic model to mine high-quality topics by introducing contrastive learning to capture information from documents. Furthermore, the distinct tree-like property of hyperbolic space preserves the implicit hierarchical semantics of documents in topic embeddings, which are projected into the hyperbolic space. Finally, we use a multi-head self-attention mechanism to learn implicit hierarchical semantics of topics and mine topic structure information. Experiments on real-world corpora demonstrate the remarkable performance of our model on topic coherence and topic diversity, as well as the rationality of the topic hierarchy. Our code is available at `https://github.com/Adrian-LZC/hHTM`.

Keywords: Hierarchical topic modeling, contrastive learning, hyperbolic embedding

1. Introduction

Recently, Hierarchical Topic Models (HTMs), which can mine the implicit topic hierarchy in documents, have received increasing attention. HTMs aim to interpret topics in a coherent word co-occurrence pattern and capture the hierarchical semantic between topics to build a rational topic structure (Zhang et al., 2022). HTMs has been successfully applied to tasks such as hierarchical classification of Web pages (Ming et al., 2010) and the discovery of hierarchical relationships in academic repositories (Paisley et al., 2015), and is emerging as one of the most powerful tools for automatic text analysis (Rubin et al., 2012; Wang et al., 2018; Jelodar et al., 2020).

Neural Hierarchical Topics Models (NHTMs) based on Neural Variational Inference (NVI) are gaining huge attention owing to their high efficiency and scalability (Chen et al., 2021; Zhang et al., 2022; Duan et al., 2021a). For example, a Tree-Structured Neural Topic Model (TSNTM) (Isonuma et al., 2020) was proposed to learn hierarchical semantic by parameterizing the hierarchical topic distribution. Chen et al. (2021) proposed a non-parametric model named nTSNTM by introducing the dependency matrix to mine topic structure based on TSNTM. However, both TSNTM and nTSNTM generate a topic tree only, which limits the extensibility of the topic structure. To enrich the information of topic structure, Zhang et al. (2022) proposed a forest-like topic distribution (nFNTM) and introduced a self-attention mechanism (Vaswani et al., 2017) to mine the relationships between topics. To emphasize symmetrical dependencies between topics at the same level, Chen et al. (2023) proposed NSEM-GMHTM with a Gaussian mixture prior distribution to improve the model's ability to adapt to sparse data, which explicitly models hierarchical and symmetric relations between topics through the introduced dependency matrices and nonlinear structural equations. In addition, SawETM (Duan et al., 2021a), which exploits a sawtooth connection module to mitigate the problem of posterior collapse, and TopicNet (Duan et al., 2021b), which introduces external hierarchical prior knowledge, both target at optimizing the rationality of topic relations without addressing the drawback of topic redundancy. Nevertheless, insufficient information regarding the prior distribution has significant impacts on the training quality of NVI-based neural topic models.

The Generative Adversarial Network (GAN) based architecture introduces a separate neural network module to fit the difference between real and fake data distributions, which avoids the complex derivation in the variational inference approach and generates topics of higher quality than the framework based on NVI. ATM (Wang et al.,

*The corresponding authors.

LREC-COLING 2024, pages 8133–8143

2019) assumes that the topic distribution obeys a dirichlet distribution, for which, the generator projects the topic distribution randomly sampled from a dirichlet prior distribution onto the document-word distribution, and then ATM uses a discriminator to distinguish the true document-word distribution from the document-word distribution generated by the generator. However, ATM fails to infer the corresponding topic distribution from a given document (Wang et al., 2020). BAT (Wang et al., 2020) generates flat topics by introducing an encoder module with bi-directional training that combines the document-topic distribution and the document as inputs to the discriminator, enabling it to capture differences in high and low dimensional distributions. Although the above GAN-based topic models have achieved satisfactory performance, the training of GAN needs to find non-convex solutions under continuous high-dimensional parameters, and the existing gradient descent methods usually can only converge to the locally optimal solution, which is prone to problems such as pattern collapse (Lei et al., 2019). It leads to the difficulty of the generator to fully fit the probability distribution of the training data.

As a kind of representation learning methods, contrastive learning has been widely studied in both computer vision (Chen et al., 2020; He et al., 2020) and natural language processing (Gao et al., 2021; Wu et al., 2022b). Constrative learning can be effective in mitigating the pattern collapse problem, and improves the generator's ability to capture key information about real data to generate high-quality pseudo-data. For instance, Yang et al. (2021) proposed InsGen, which aids the discriminator in learning the implicit features of the data by introducing contrastive learning, thereby improving the discriminative ability. Li et al. (2022) proposed FakeCLR, which utilizes contrastive learning to slove latent discontinuty in GANs, resulting in improved generative performance of the generator. Additionally, Nguyen and Luu (2021) proposed CLNTM to capture the mutual information between document prototypes and positive samples by modeling the relationship between the contrasting augmentation samples. Wu et al. (2022a) proposed TSCTM, which utilizes a topic semantics-based sampling strategy to generate samples as a way to alleviate the data sparsity problem so that document relationships can be properly modeled. Both of the aforementioned models demonstrate the effectiveness of contrastive learning in solving the topic quality problem. However, the existing topic models based on contrastive learning only focus on mining topic information and ignore modeling of topic hierarchical relationships.

Furthermore, most NHTMs mine hierarchical semantics of topics in the Euclidean space. Despite the commendable achievements, it leads to a fundamental limitation in that their ability to model complex patterns (similar to knowledge graphs, and topic hierarchical structure) is limited by the property of the embedding space. Hyperbolic space, a non-Euclidean space with constant negative curvature, has received much attention in recent years due to its ability to express hierarchical structure (Xu et al., 2022). Separately, Ganea et al. (2018) introduced hyperbolic space into the training of neural networks by defining arithmetical operations. HyperMiner (Xu et al., 2022) introduced hyperbolic space to topic embeddings and word embeddings. Unfortunately, HyperMiner does not explicitly exploit the correlation between topics and the generated topic structure is not flexible sufficiently.

In light of these considerations, with the hypothesis of **h**yperbolic space, we propose a novel **H**ierarchical **T**opic **M**odel (hHTM) based on the framework of GAN. To address the problem of pattern collapse in GAN-based topic models, the hHTM introduces contrastive learning which improves the ability of the generator to capture information from the corpus and enables the model to generate higher quality topics. Moreover, to better mine the topic hierarchy in documents, we model topic relations in the hyperbolic space using a multi-head attention mechanism and introduce the directed acyclic graph (DAG) constraints to make our topic hierarchy more reasonable and flexible. To the best of our knowledge, it's the first time that contrastive learning and hyperbolic space are utilized to mine high-quality topics and model more sensible hierarchical topic relationships. Experiments show that hHTM outperforms state-of-the-art baselines on widely-used quantitative metrics, which validates that our model captures a more rational topic hierarchy. In addition, the validity of our approach is further demonstrated through extensive qualitative analysis.

2. Related Work

2.1. Hierarchical Topic Model

In recent years, several emerging methods have attempted to mine high-quality topic hierarchies. Isonuma et al. (2020) proposed TSNTM, which utilized doubly recurrent neural networks (DRNN) (Alvarez-Melis and Jaakkola, 2017) to parameterize the topic distribution. Based on the above study, Chen et al. (2021) introduced neural variational inference (NVI) and non-parameterized the topic distribution, which made the model more flexible for topic mining. Zhang et al. (2022) proposed a forest-like neural topic model (nFNTM) and used self-attention mechanism (Vaswani et al., 2017) to mine latent relationships between topics, so that the hierarchy of topics is not limited to a tree. Also, non-negative matrix factorization (NMF) (Lee and Seung, 2000) is used in the hierarchical topic modeling task, where CluHTM (Viegas et al., 2020) employed NMF to generate hierarchical topics in a DAG structure. In addition, SawETM (Duan et al., 2021a) exploited a sawtooth connection module to mitigate the problem of posterior collapse. NSEM-GMHTM (Chen et al., 2023) ameliorated the problem of data sparsity by introducing a Gaussian mixture prior distribution and focused on the relationships between topics in the same layer.

2.2. Contrastive Learning

Contrastive learning is often used to learn high-quality data representations by contrasting the data with positive and negative samples. Wang and Isola (2020) demonstrated that contrastive learning possesses both alignment and uniformity properties, including: (a) Similar data representations are closer together in distribution space, while divergent data representations are farther apart. (b) Data representations can be more uniformly distributed in the distribution space. Recently, contrastive learning has also been applied to neural topic models. Nguyen and Luu (2021) captured the relationship between samples from the data perspective and proposed a new contrasting goal to help the model uncover more meaningful topics. Wu et al. (2022a) proposed a semantic contrastive-based neural topic model named TSCTM, which introduced an efficient sampling strategy of positive and negative samples to mitigate data sparsity for short documents. However, the aforementioned methods only focus on generating flat topics without exploring the relationship between topics.

2.3. Hyperbolic Embedding

Hyperbolic geometry is a non-Euclidean geometry with a constant negative curvature. The ability to characterize Euclidean space will be limited when the distribution of documents exhibits non-Euclidean geometry. Hyperbolic space shows exponential expansion with increasing radius, and it can be seen as a continuous tree-structured space (Ganea et al., 2018), which allows the hyperbolic space to preserve the hierarchical structure implied by documents well. The classical models in hyperbolic space include Poincaré Ball Model (Nickel and Kiela, 2017) and Lorentz Model (Nickel and Kiela, 2018). Ganea et al. (2018) proposed a set of operators for hyperbolic spaces, which allowed the training of neural networks in hyperbolic spaces to become a reality. In topic modeling, HyperMiner (Xu et al., 2022) projected topic embeddings and word embeddings into hyperbolic spaces to mine the hierarchical semantics in the original corpora. Different from it, we model the topic structure by projecting the topics into the hyperbolic space under the premise of exploiting contrastive learning to sufficiently mine high-quality topics. Moreover, a multi-head self-attention mechanism is combined with hyperbolic embeddings to exploit the implicit hierarchical semantics better.

3. Methodology

In this section, we describe in detail all the modules of hHTM and the corresponding way of working. As shown in Figure 1, our model is divided into three parts: encoder, decoder, and discriminator.

3.1. Encoder

We introduce contrastive learning into the encoder, for which, two symmetric feedforward neural networks are employed to learn the alignment and uniformity of data representations and consequently learn diverse document-topic distributions for real data. Let $D_r = \{d \mid d \in \mathbb{R}^{n_V}\}$ denotes the set of document vectors in the form of TF-IDF, where n_V represents the vocabulary size. Each document vector has a relative positive sample $d^+ = T(d)$ and a batch of negative samples $\{d_i^-\}_{i=1}^{N_{neg}}$, where $T(\cdot)$ represents the data augmentation function and N_{neg} is the number of negative samples. We employ random masking of some words to achieve the effect of data augmentation. For the construction of negative samples, we follow the approach of Moco (He et al., 2020). Through contrastive learning between samples, we aim to learn the high quality latent distribution π of each document vector d as well as the latent distributions π^+ and π^- of d^+ and d^-, respectively. Therefore, our contrastive loss function is given below:

$$\begin{aligned}&\mathcal{L}_{\mathcal{CON}}(\pi, \pi^+, \{\pi_i^-\}_{i=1}^{N_{neg}}, \vartheta_q, \vartheta_k, \tau)\\ &= -\log \frac{e^{sim(\pi,\pi^+)/\tau}}{e^{sim(\pi,\pi^+)/\tau} + \sum_{i=1}^{N_{neg}} e^{sim(\pi,\pi_i^-)/\tau}},\end{aligned} \quad (1)$$

where ϑ_q and ϑ_k are respectively the parameters of feedforward neural network $f_q(\cdot)$ and $f_k(\cdot)$, $\tau > 0$ is temperature coefficient, and $sim(\cdot, \cdot)$ is the function of cosine similarity. Moreover, since the encoder aims to learn the topic distribution of real documents, its optimisation direction should be aligned with that of the discriminator. Therefore, the discriminative loss for real documents $\mathbb{E}_{d\sim D_r}[D(d, E(d))]$ is added to the encoder, where $D(\cdot, \cdot)$ and $E(\cdot)$ represent the discriminator and the encoder, respectively. The objective cost function of the encoder is given below:

$$\mathcal{L}_E = \alpha_E \cdot \mathcal{L}_{\mathcal{CON}} + \beta_E \cdot \mathbb{E}_{d\sim D_r}[\underbrace{D(d, E(d))}_{D_{in}}], \quad (2)$$

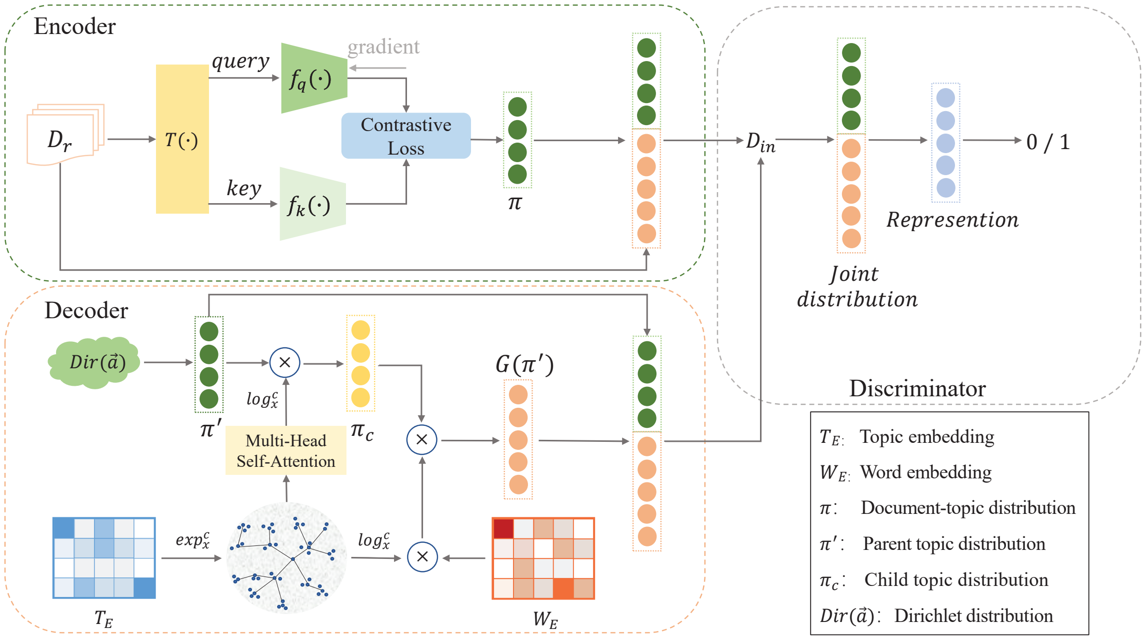

Figure 1: The framework of hHTM.

where α_E and β_E represent the weights of the loss terms, respectively. In this paper, we set $\alpha_E = 100$ and $\beta_E = 1$.

3.2. Decoder

In decoder, we sample the topic distribution $\pi' \in \mathbb{R}^{n_k}$ of each fake documents from the Dirichlet prior distribution, where n_k represents the number of topics. We project both topics and words into the embedding space and estimate the topic-word distribution through the correlation between embeddings. In addition, we introduce the pre-trained GloVe model (Pennington et al., 2014) to obtain the initialization for word embeddings $\boldsymbol{W_E} \in \mathbb{R}^{n_V \times n_t}$ and randomly initialize topic embeddings $\boldsymbol{T_E} \in \mathbb{R}^{n_k \times n_t}$, where n_t is the embedding size.

Poincaré Ball Model We introduce a classical model in the hyperbolic space: Poincaré Ball Model. Assuming that the Poincaré Ball is n-dimensional and has curvature c (i.e., radius $\frac{1}{\sqrt{c}}$), it can be denoted as $\mathbb{P}_c^n = \{z \in \mathbb{R}^n \mid ||z||^2 < 1\}$ with its metric given by $g_z^c = \lambda_z^2 g^E$, where $\lambda_z = \frac{z}{1-c||z||^2}$ and g^E is the regular Euclidean metric tensor. Intuitively, when a point z is near the boundary, its hyperbolic distance from a neighboring point z' will grow at the rate of $\frac{1}{1-c||z||^2} \to \infty$. This property plays a significant role in learning the topic hierarchy implied by documents. Note that when $c \to 0$, the model will recover back to Euclidean space $\mathbb{R}^n$.

Hyperbolic Operations To learn the representation of data in the hyperbolic space, we need to implement hyperbolic operations, including vector addition, exponential map, logarithmic map and parallel transport. Following the framework of *gyrovector spaces* (Ungar, 2009), we can obtain the addition of two points $z, z' \in \mathbb{P}_c^n$ by Möbius addition as follows:

$$z \oplus_c z' = \frac{1 + 2c\langle z, z'\rangle + c||z'|| + (1 - c||z||^2)z'}{1 + 2c\langle z, z'\rangle + c^2||z||^2||z'||^2}, \tag{3}$$

where $\oplus_c$ denotes the Möbius addition symbol.

For tangent space computations, according to (Ganea et al., 2018), given any point $x \in \mathbb{P}_c^n$, the exponential map and the logarithmic map are defined for $v \neq 0$ and $y \neq x$ by:

$$\begin{aligned}\exp_x^c(v) &= x \oplus_c (\tanh(\sqrt{c}\frac{\lambda_x^c\,||v||}{2})\frac{v}{\sqrt{c}\,||v||}),\\ \log_x^c(y) &= \frac{2}{\sqrt{c}\lambda_x^c}\tanh^{-1}(\sqrt{c}\,||\varphi_{x,y}||)\frac{\varphi_{x,y}}{||\varphi_{x,y}||},\end{aligned} \tag{4}$$

where $\varphi_{x,y} = -x \oplus_c y$. Besides, the parallel transport can map a vector $v \in T_0\mathbb{P}_c^n$ to another tangent space $T_x\mathbb{P}_c^n$ is given by the following isometry:

$$\mathrm{P}_{0\to x}^c(v) = \log_x^c(x \oplus_c \exp_0^c(v)) = \frac{\lambda_0^c}{\lambda_x^c}v. \tag{5}$$

Topic Relation To mine the structural semantics implied between topics, we project $\boldsymbol{T_E}$ into the hyperbolic space and mine the structural semantics using multi-head self-attention mechanism (Vaswani et al., 2017), as follows:

$$\boldsymbol{Q_i} = (\frac{\lambda_0^c}{\lambda_x^c}\boldsymbol{T_E})\boldsymbol{W_i^Q}, \quad \boldsymbol{K_i} = (\frac{\lambda_0^c}{\lambda_x^c}\boldsymbol{T_E})\boldsymbol{W_i^K}, \tag{6}$$

$$
\begin{aligned}
\boldsymbol{Q} &= \text{Concat}(\boldsymbol{Q_1}, \ldots, \boldsymbol{Q_{n_h}}), \\
\boldsymbol{K} &= \text{Concat}(\boldsymbol{K_1}, \ldots, \boldsymbol{K_{n_h}}),
\end{aligned}
\tag{7}
$$

$$
\boldsymbol{C} = \text{Softmax}\left(\frac{\frac{\boldsymbol{Q}\boldsymbol{K}^T}{n_Q}}{\tau_c}\right), \tag{8}
$$

where $\boldsymbol{Q} \in \mathbb{R}^{n_k \times n_Q}$ and $\boldsymbol{K} \in \mathbb{R}^{n_k \times n_Q}$ are learnable parameters, $n_Q = n_t/n_h$, n_h is the number of attention heads, τ_c denotes the temperature value, $\boldsymbol{C} \in \mathbb{R}^{n_k \times n_k}$ is the relationship matrix, which implies a hierarchical relationship between topics. Each element of the relationship matrix $\boldsymbol{C}$ represents the degree of relevance of the parent-child relationship between topics. The Softmax operation guarantees the discretization of the relationship matrix $\boldsymbol{C}$ and avoids redundancy in the topic hierarchy. However, a reasonable hierarchy should be DAG-structured. According to (Zheng et al., 2018), we ensure that the relational weight matrix $\boldsymbol{C}$ is a structure of DAG if and only if $h(\boldsymbol{C}) = \text{tr}(e^{(\boldsymbol{C} \circ \boldsymbol{C})}) - n_k = 0$, where $\circ$ is the Hadamard product.

Data Reconstruction Intuitively, the semantic information of both parent and child topics should be available for generating complete documents. First, we compute the topic word distribution $\Phi = \text{Softmax}(\boldsymbol{T_E} \cdot \boldsymbol{W_E}^T)$. Then, we compute the parent topic distribution $\pi_p = \pi'$ and child topic distribution $\pi_c = \pi' \times \boldsymbol{C}$ simultaneously. Under the constraint of $h(\boldsymbol{C}) = 0$, we also need to ensure that documents constructed by parent topic distributions and child topic distributions are as similar as possible, with the following objective cost function:

$$
\begin{aligned}
\min_{\boldsymbol{C}} \mathcal{L}_{\mathcal{C}} = &\frac{1}{2} ||(\pi_p - \pi_c) \times \Phi||_F^2 \\
&+ \frac{\rho}{2} |h(\boldsymbol{C})^2| + \epsilon h(\boldsymbol{C}),
\end{aligned}
\tag{9}
$$

where ρ is a penalty parameters and ϵ is the Lagrange multiplier. We follow (Zheng et al., 2018) to update ρ and ϵ, i.e.,

$$
\begin{aligned}
\rho_i &= 2\rho_{i-1}, \\
\epsilon_i &= \epsilon_{i-1} + \rho h_{i-1},
\end{aligned}
\tag{10}
$$

where $\rho_0 = 1$, $\epsilon_0 = 0$, and h is the value of $h(\boldsymbol{C})$.

To summarize, in decoder, our objective cost function is given below:

$$
\mathcal{L}_{De} = -\mathbb{E}_{\pi' \sim Dir(\overrightarrow{\alpha})}[D(\underbrace{G(\pi'), \pi'}_{D_{in}})] + \mathcal{L}_C, \tag{11}
$$

where $-\mathbb{E}_{\pi' \sim Dir(\overrightarrow{\alpha})}[D(G(\pi'), \pi')]$ is the fake loss (Arjovsky et al., 2017) and $G(\cdot)$ is the generator.

3.3. Discriminator

In discriminator, we consider documents and topic distributions as inputs to the discriminator, and while training the discriminator, we also prompt the generator to generate documents that better match real topic distributions. Following (Arjovsky et al., 2017), the objective cost function of the discriminator is given below:

$$
\begin{aligned}
\mathcal{L}_D = &\mathbb{E}_{\pi' \sim Dir(\overrightarrow{\alpha})}[D(G(\pi'), \pi')] \\
&- \mathbb{E}_{d \sim D_r}[D(d, E(d))].
\end{aligned}
\tag{12}
$$

Our algorithm is shown in Algorithm 1.

Algorithm 1: Algorithm of hHTM

Input : The embedding of words $\boldsymbol{W_E}$ and documents $\{d_1, \ldots, d_{n_D}\}$;
Output : Topic-word distribution $\boldsymbol{\Phi}$, topic relationship matrix $\boldsymbol{C}$

```
Randomly initialize query matrices Q, key
  matrices K, and topic embeddings T_E.
repeat
  for documents d ∈ {d_1, ..., d_{n_D}} do
    Obtain π by the encoder E;
    Sample document-topic distribution
      π' ~ Dir(α⃗);
    Project T_E to the hyperbolic space by
      Eq. (5);
    Compute C by Eqs. (6-8);
    Compute L_D by Eq. (12);
    Update the discriminator D by
      RMSprop;
    for l ∈ D do
      Update the l − th layer weights of
        D by spectral normalization;
      W_D^l = W_D^l / σ(W_D^l);
    Compute L_E by Eq. (2);
    Compute L_C by Eq. (9);
    Compute L_De by Eq. (11);
    Update the encoder E;
    Update the decoder De;
until convergence;
Topic structure are built from C and Φ.
```

---

4. Experiments

4.1. Experimental Setting

Datasets We validate the effectiveness of our model on three widely used benchmark corpora, including NIPS (Tan et al., 2017), 20News (Miao et al., 2017) and Wikitext-103 (Merity et al., 2017). These datasets have been processed to remove stop words and filter low frequency words by following Chen et al. (2023). Table 1 summarizes the statistics of the three corpora.

Baselines In order to make a comprehensive evaluation for our model, the benchmark models

Dataset	#Docs (Train)	#Docs (Test)	Vocabulary size
NIPS	1,350	149	3,531
20News	11,314	7,531	3,997
Wikitext-103	28,472	120	20,000

Table 1: The statistics of corpora.

mainly include hierarchical topic models with tree, forest, and DAG structures.

SawETM[1] (Duan et al., 2021a): The hierarchical topic model which introduces a sawtooth connection module to mitigate the problem of posterior collapse.

HyperMiner[2] (Xu et al., 2022): The hierarchical topic model which exploits hyperbolic embeddings for topic and word representations.

nTSNTM[3] (Chen et al., 2021): The tree-like topic model that introduces non-parameterization in the number of topics.

nFNTM[4] (Zhang et al., 2022): The forest topic model which employs the self-attention mechanism to capture parent-child topic relations.

CluHTM[5] (Viegas et al., 2020): The DAG-structured topic model based on non-negative matrix factorization.

NSEM-GMHTM[6](Chen et al., 2023): A deep topic model with a Gaussian mixture prior distribution and nonlinear structural equations to capture topic relations.

Hyperparameter Settings In our experiments, for the nonparametric models (i.e., nTSNTM and nFNTM), we set their maximum number of topics to 200. For all parametric models (i.e., SawETM, HyperMiner, CluHTM, NSEM-GMHTM, and hHTM), the number of topics is uniformly set to 200. All other hyperparameters of those baselines are set according to the original paper. For hHTM, we set the weight parameter d_t to 300 for the self-attention module and the temperature τ to 0.07. The optimisation of hHTM is achieved by $rmsprop$ with a learning rate of 5e-4 and batch size of 256.

4.2. Quantitative Analysis of Topic Hierarchy

To quantitatively compare the performance of our model and other baselines, we employ the Normalized Pointwise Mutual Information (NPMI) (Zhang et al., 2022; Chen et al., 2021), the Cross-Level Normalized Point-wise Mutual Information (CLNPMI) (Chen et al., 2021), the Topic Uniqueness (TU) (Nan et al., 2019), the Topic Quality (TQ) (Dieng et al., 2020) and the Topic Specialization (TS) (Kim et al., 2012) as the evaluation metrics on the quality of model-mined hierarchical topics from different perspectives.

Interpretability of Topics The topic hierarchy generated by an exceptional hierarchical topic model should have the following properties. First, the semantics of individual topics should ensure high coherence. Second, there is some similarity between the child topic and the corresponding parent topic. Therefore, we employ NPMI scores to evaluate the coherence between individual topics and CLNPMI scores to evaluate the similarity between parent and child topics. NPMI (Zhang et al., 2022), a widely adopted metric in the field of topic modeling, allows assessing the interpretability of the generated topics. CLNPMI is proposed by Chen et al. (2021) to measure the subordination of topics by calculating the average NPMI scores of parent-child topics, as follows: CLNPMI$(W_p, W_c) = \sum_{w_i \in W_p'} \sum_{w_j \in W_c'} \frac{\text{NPMI}(w_i,w_j)}{|W_p'||W_c'|}$, where $W_p^{'} = W_p - W_c$ and $W_c^{'} = W_c - W_p$, in which, W_p and W_c represent the top N words of the parent topic and its child topics, respectively.

As shown in Table 2, our model achieves the best NPMI score on Wikitext-103 as well as sub-optimal results on the other two datasets. On the other hand, our model received the best CLNPMI score relative to the other benchmark models. Compared to our model, although NSEM-GMHTM captures more consistent topics, it performs much worse than our model in terms of CLNPMI and TU scores, which suggests that the topics mined by NSEM-GMHTM are redundant to a certain extent, and also do not capture reasonable topic hierarchies well. Comparing with HyperMiner, our model achieves better performance on all metrics, which suggests that contrastive learning (Wang and Isola, 2020) leads to a better distribution of topics generated by the model on the hypersphere. It is worth mentioning that our model and HyperMiner, based on the hyperbolic space assumption, achieve optimal and sub-optimal performance on CLNPMI, respectively, which verifies that learning topic embeddings in the hyperbolic space allows topics to retain more information about the semantic structure implicit in the corpus. In summary, these results show that our model guarantees high quality topics while better capturing the semantic relationships between parent and child topics, which fully demonstrates that our model can mine more reasonable topic hierarchies.

[1]https://github.com/BoChenGroup/SawETM
[2]https://github.com/NoviceStone/HyperMiner
[3]https://github.com/hostnlp/nTSNTM
[4]https://github.com/Angr4Mainyu/nFNTM
[5]https://github.com/feliperviegas/cluhtm
[6]https://github.com/nbnbhwyy/NSEM-GMHTM

Dataset	Metric	SawETM	HyperMiner	nTSNTM	nFNTM	CluHTM	NSEM-GMHTM	hHTM
NIPS	NPMI↑	0.135	0.134	0.100	0.113	0.137	**0.147**	0.137
	CLNPMI↑	0.071	0.060	0.022	0.025	0.027	0.028	**0.097**
	TU↑	0.659	0.640	0.373	0.765	0.554	0.719	**0.766**
	TQ↑	0.089	0.086	0.037	0.086	0.076	**0.106**	0.105
20News	NPMI↑	0.256	0.266	0.284	0.246	0.219	**0.307**	0.288
	CLNPMI↑	0.137	0.164	0.156	0.150	0.164	0.146	**0.215**
	TU↑	0.380	0.471	0.757	0.844	0.577	0.811	**0.864**
	TQ↑	0.097	0.125	0.215	0.208	0.126	0.249	**0.249**
Wikitext-103	NPMI↑	0.243	0.239	0.225	0.228	-	0.255	**0.274**
	CLNPMI↑	0.131	0.137	0.121	0.147	-	0.090	**0.175**
	TU↑	0.533	0.640	0.662	0.739	-	0.797	**0.912**
	TQ↑	0.130	0.153	0.149	0.168	-	0.203	**0.250**

Table 2: The performance of all hierarchical topic models, where - indicates that the model has not converged after 48 hours of training.

Topic Diversity In the real world, in addition to the semantic consistency of the topics, it is equally important that the topics found are diverse. If the topics are redundant, the topic structure is unreasonable and the resulting topics are less meaningful. Therefore, we adopt topic uniqueness (TU) to evaluate the diversity of hierarchical topics generated, which is calculated as follows:

$$\mathrm{TU} = \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{1}{\mathrm{cnt}(n,k)}, \tag{13}$$

where K represents the number of topics and $\mathrm{cnt}(n,k)$ is the total number of times the n_{th} top word in the k_{th} topic appears in the top N words of all topics. As shown in Table 2, on all datasets, our model outperforms benchmark models in terms of topic diversity, which is mainly due to the uniformity property of contrastive learning (Wang and Isola, 2020). It facilitates the model learn high-quality latent spaces and alleviates the problem of discontinuities in latent space, thus improves the performance of the generator and generates more diverse topics.

Topic Quality Intuitively, higher NPMI scores imply that the correlations within topics are better, which may then result in increased redundancy between topics, and thus the TU scores will become lower. Conversely, higher TU scores tend to be accompanied by lower NPMI scores, because for most of topics with higher TU scores, they tend to be marginal topics (Wu et al., 2020b), which are often represented by less coherent words. Therefore, in order to provide a more comprehensive evaluation of the overall topic quality, we use topic quality (TQ) to evaluate the quality of topics, which is calculated as follows:

$$\mathrm{TQ} = \mathrm{NPMI} \times \mathrm{TU}. \tag{14}$$

As shown in Table 2, on 20News and Wikitext-103, our model achieves the best performance, and sub-optimal results on NIPS. This illustrates the relatively high quality of the topics generated by our model.

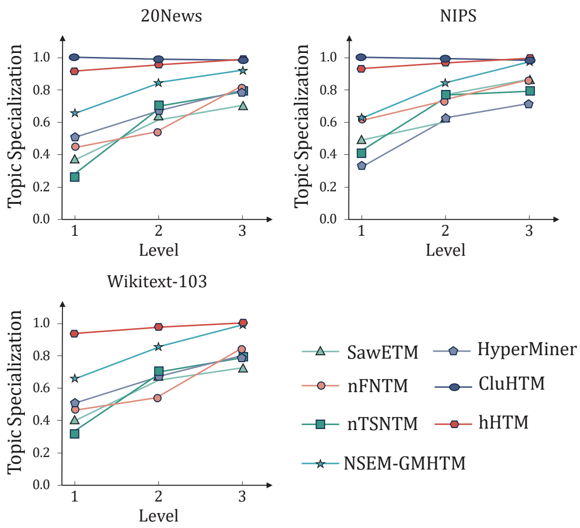

Figure 2: Topic specialization of different topic structure generated on all datasets.

Topic Structure Rationality For the hierarchical topic model, topics closer to the root node should be more general, while topics closer to the leaf node should be more specific. Topic specialization (Kim et al., 2012) score measures the generalization of topics by comparing the word distribution of each topic with that of the entire corpus. A topic with a higher score indicates a more specific semantic. The formula for topic specialization is given below:

$$\mathrm{TS}(\Phi) = 1 - cos(\Phi, \Phi_{\text{Norm}}) = 1 - \frac{\Phi \cdot \Phi_{\text{Norm}}}{|\Phi||\Phi_{\text{Norm}}|}, \tag{15}$$

where Φ and Φ_{Norm} denote a topic-word distribution and the word distribution of the entire corpus, respectively.

For the sake of fairness and uniformity, we calculate the average topic specialization score of the three layers generated by all models to assess the rationality of the topic structure. As shown in Figure 2, our model exhibits a gradual increase in topic specialization scores with increasing levels on all three datasets, and the closer to the leaf nodes, the more topic-specific it is relative to the other benchmark models. It is worth noting that the score of topic specialization could not be computed for CluHTM since it could not be converged by training on Wikitext-103. The topic specialization scores of CluHTM on 20News and NIPS also show a decreasing trend with the increase in the number of layers, which indicates the irrationality of topic structure.

4.3. Qualitative Analysis of Topics

Visualization of the Topic Embedding Space We show the distribution of topic embeddings in the hyperbolic space to analyze the distribution of topics in the embedding space. As shown in Figure 3, parent topics are positioned closer to the center in the embedding space, while child topics are distributed at the boundaries. Due to the characteristic of hyperbolic space, the distance between child topic embeddings is exponential, which also demonstrates the soundness of the hierarchical structure mined by our model.

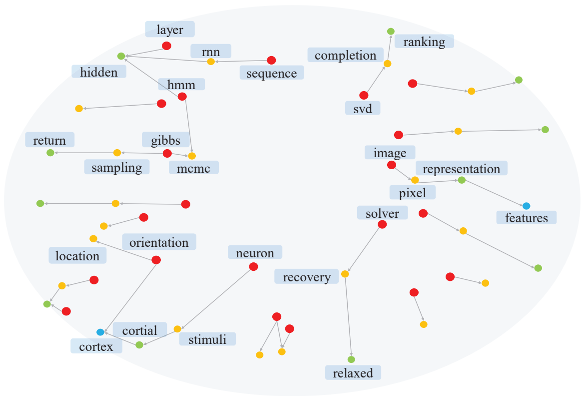

Figure 3: Visualization of the topic embedding space on NIPS.

Visualization of the Topic Structure As shown in Figure 4, we exhibit some of the topic structure of NIPS. For example, the parent topic of [objective, gradient, gradients, stochastic, descent] is about the gradient of neural networks, and its next-level child topics are about the detailed solution approach [stochastic sgd gradients gradient descent] and [newton descent update coordinate updates]. Further child topics [online batch update zt offline] and [xt zt dt ut yt] are related to specific gradient formulas. These results demonstrate that our model captures a reasonable hierarchy of topics, with parent topics being general and child topics becoming more specific with increasing depth.

Datasets	Model	NPMI↑	TU↑	TQ↑	CLNPMI↑
NIPS	Ours	0.137	**0.766**	0.105	**0.097**
	Ours w/o Con	0.141	0.726	0.102	0.077
	Ours w/o Hyper	**0.144**	0.741	**0.107**	0.032
	Ours w/o M-att	0.130	0.641	0.083	-
20News	Ours	0.288	**0.864**	**0.249**	**0.215**
	Ours w/o Con	**0.301**	0.767	0.231	0.213
	Ours w/o Hyper	0.279	0.857	0.239	0.096
	Ours w/o M-att	0.265	0.641	0.170	-
Wikitext-103	Ours	**0.274**	**0.912**	**0.250**	**0.175**
	Ours w/o Con	0.259	0.909	0.235	0.089
	Ours w/o Hyper	0.274	0.905	0.248	0.062
	Ours w/o M-att	0.271	0.867	0.235	-

Table 3: Results of ablation evaluation on all datasets.

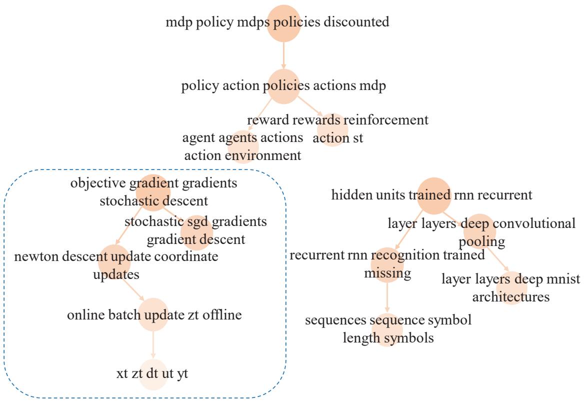

Figure 4: Topic structure visualization on NIPS.

4.4. Ablation Study

Ablation experiments can verify the role played by the different modules of our model, which is very necessary. We ablate different components in three cases: (i) Without introducing contrastive learning to the encoder (w/o Con). (ii) Without projecting topic embeddings into the hyperbolic space (w/o Hyper). (iii) Without introducing the multi-head self-attention mechanism to learn implicit hierarchical semantics of topics (w/o M-att).

As shown in Table 3, more diverse topics are effectively mined by introducing contrastive learning for complicated modeling of potential semantic relationships in documents. Contrastive learning learns a better latent space, which leads to improve performance of the generator in generating high-quality topics. Moreover, the introduction of hyperbolic space preserves the hierarchical relationship modeling, which allows our model to learn the inherent topic hierarchy of documents. Moreover, the complete model achieved optimal TQ results on 20News and Wikitext-103,and sub-optimal TQ

Metric	SawETM	nFNTM	nTSNTM	HyperMiner	NSEM-GMHTM	hHTM
Speed	5.2s	3.3s	38.6s	4.4s	3.8s	3.2s
#Params	1.9M	1.2M	0.5M	2.2M	1.5M	10.1M

Table 4: Speed and number of parameters for NHTMs on 20News.

results on NIPS. Contrastive learning focuses on improving the topic quality and has little impact on the hierarchical structure of topics. As shown in Table 3, the introduction of hyperbolic space provides a significant improvement in CLNPMI scores. This demonstrates how the module helps to generate a more rational topic hierarchy and improves the interpretability of the model. Additionally, when no multi-head self-attention mechanism is introduced, the model fails to converge in mining the structural relationships between topics, thus it is not feasible to construct a reasonable topic hierarchy of directed acyclic graphs. In conclusion, all components of our model are reasonable and effective.

4.5. Complexity Comparison

The training speed of the model is also an important indicator for assessing the quality of the model. A superior model needs to infer high-quality topic distributions in as little time as possible. As an illustration, we run models on a server equipped with Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz, 48 cores and 128G memory, and 2 $\times$ NVIDIA GTX 1080Ti with 2 $\times$ 12G memory. Here, we compare the time taken to train 10 epochs on the 20News dataset between our model and the benchmark model to measure the training time of the model. As shown in Table 4, our model can accommodate the largest number of parameters, and meantime spend the least amount of time for iterating 10 epochs. This is because we employ a two time-scale update rule (Heusel et al., 2017) for GAN as well as momentum update (He et al., 2020) for contrastive learning, which ensure high efficiency for each iteration. These results indicate that our model could generate high quality topics while keeping the overhead on computational resources within a reasonable range.

5. Conclusion

In this paper, we propose a GAN-based hierarchical topic model that mitigates the generation performance limited by the discontinuity of latent space through introducing contrastive learning to model the latent relations of documents, ensuring the generation of high-quality topics. The projection of topic embeddings into the hyperbolic space enables the model to learn the implicit hierarchical semantics of documents. In addition, a more rational topic hierarchy is constructed by exploiting a multi-head self-attention mechanism focusing on the multi-layer connections between topic structures and the constraints of directed acyclic graphs. The experimental results demonstrate the remarkable performance of our model on topic quality and topic structure. In the future, we will explore the potential hierarchical relationships of documents by incorporating external prior knowledge to guide the model in contrastive learning of documents, so as to generate semantically rich and hierarchically distinct topic structure better.

Acknowledgments

We are grateful to the reviewers for their constructive comments and suggestions. This work has been supported by the National Natural Science Foundation of China (62372483).

Bibliographical References

David Alvarez-Melis and Tommi S. Jaakkola. 2017. Tree-structured decoding with doubly-recurrent neural networks. In *ICLR*.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *ICML*, pages 214–223.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, pages 131:1–131:27.

Hegang Chen, Pengbo Mao, Yuyin Lu, and Yanghui Rao. 2023. Nonlinear structural equation model guided gaussian mixture hierarchical topic modeling. In *ACL*, pages 10377–10390.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607.

Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. 2021. Tree-structured topic modeling with nonparametric neural variational inference. In *ACL/IJCNLP*, pages 2343–2353.

Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguistics*, pages 439–453.

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.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. In *NeurIPS*, pages 5350–5360.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-Structured Neural Topic Model. In *ACL*, pages 800–806.

Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. 2020. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE J. Biomed. Health Informatics*, 24(10):2733–2742.

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.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *NeurIPS*, pages 556–562.

Na Lei, Yang Guo, Dongsheng An, Xin Qi, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. 2019. Mode collapse and regularity of optimal transportation maps. *CoRR*.

Ziqiang Li, Chaoyue Wang, Heliang Zheng, Jing Zhang, and Bin Li. 2022. Fakeclr: Exploring contrastive learning for solving latent discontinuity in data-efficient gans. In *ECCV*, pages 598–615.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *ICML*, pages 2410–2419.

Zhaoyan Ming, Kai Wang, and Tat-Seng Chua. 2010. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *SIGIR*, pages 2–9.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. In *ACL*, pages 6345–6381.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. In *NeurIPS*, pages 11974–11986.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, pages 6338–6347.

Maximilian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, pages 3776–3785.

John W. Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2015. Nested hierarchical dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):256–270.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Mach. Learn.*, 88(1-2):157–208.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR*.

Chenhao Tan, Dallas Card, and Noah A. Smith. 2017. Friendships, rivalries, and trysts: Characterizing relations between ideas in texts. In *ACL*, pages 773–783.

Abraham Albert Ungar. 2009. *A Gyrovector Space Approach to Hyperbolic Geometry*. Morgan & Claypool Publishers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

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.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020. Neural topic modeling with bidirectional adversarial training. In *ACL*, pages 340–350.

Rui Wang, Deyu Zhou, and Yulan He. 2019. ATM: adversarial-neural topic model. *Inf. Process. Manag.*, 56(6).

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939.

Wenlin Wang, Zhe Gan, Wenqi Wang, Dinghan Shen, Jiaji Huang, Wei Ping, Sanjeev Satheesh, and Lawrence Carin. 2018. Topic compositional neural language model. In *AISTATS*, pages 356–365.

Jiemin Wu, Yanghui Rao, Zusheng Zhang, Haoran Xie, Qing Li, Fu Lee Wang, and Ziye Chen. 2020a. Neural mixed counting models for dispersed topic discovery. In *ACL*, pages 6159–6169.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020b. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *EMNLP*, pages 1772–1782.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022a. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *EMNLP*, pages 2748–2760.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022b. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *COLING*, pages 3898–3907.

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.

Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. 2021. Data-efficient instance generation from instance discrimination. In *NeurIPS*, pages 9378–9390.

Zhihong Zhang, Xuewen Zhang, and Yanghui Rao. 2022. Nonparametric forest-structured neural topic modeling. In *COLING*, pages 2585–2597.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray L. Buntine. 2021. Neural topic model via optimal transport. In *ICLR*.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. Dags with NO TEARS: continuous optimization for structure learning. In *NeurIPS*, pages 9492–9503.

Language Resource References

Stephen Merity and Caiming Xiong and James Bradbury and Richard Socher. 2017. *Pointer Sentinel Mixture Models*. OpenReview.net. PID https://s3.amazonaws.com/fast-ai-nlp/wikitext-103.tgz.

Yishu Miao and Edward Grefenstette and Phil Blunsom. 2017. *Discovering Discrete Latent Topics with Neural Variational Inference*. PMLR. PID http://qwone.com/ jason/20Newsgroups/.

Chenhao Tan and Dallas Card and Noah A. Smith. 2017. *Friendships, Rivalries, and Trysts: Characterizing Relations between Ideas in Texts*. Association for Computational Linguistics. PID http://papers.nips.cc/.