On the Affinity, Rationality, and Diversity of Hierarchical Topic Modeling

Xiaobao Wu[1], **Fengjun Pan**[1], **Thong Nguyen**[2],
Yichao Feng[1], **Chaoqun Liu**[1, 3], **Cong-Duy Nguyen**[1], **Anh Tuan Luu**[1]
[1]Nanyang Technological University, Singapore
[2]National University of Singapore, Singapore
[3]DAMO Academy, Alibaba Group, Singapore
{xiaobao002,panf0004,yichao002,chaoqun001,nguyentr003,anhtuan.luu}@ntu.edu.sg, e0998147@u.nus.edu

Abstract

Hierarchical topic modeling aims to discover latent topics from a corpus and organize them into a hierarchy to understand documents with desirable semantic granularity. However, existing work struggles with producing topic hierarchies of low affinity, rationality, and diversity, which hampers document understanding. To overcome these challenges, we in this paper propose Transport Plan and Context-aware Hierarchical Topic Model (TraCo). Instead of early simple topic dependencies, we propose a transport plan dependency method. It constrains dependencies to ensure their sparsity and balance, and also regularizes topic hierarchy building with them. This improves affinity and diversity of hierarchies. We further propose a context-aware disentangled decoder. Rather than previously entangled decoding, it distributes different semantic granularity to topics at different levels by disentangled decoding. This facilitates the rationality of hierarchies. Experiments on benchmark datasets demonstrate that our method surpasses state-of-the-art baselines, effectively improving the affinity, rationality, and diversity of hierarchical topic modeling with better performance on downstream tasks.

Introduction

Instead of traditional flat topic models, hierarchical topic models strive to discover a topic hierarchy from documents (Griffiths et al. 2003; Teh et al. 2004). Each topic is interpreted as relevant words to represent a semantic concept. The hierarchy captures the relationships among topics and organizes them by semantic granularity: child topics at lower levels are relatively specific to parent topics at higher levels. Therefore hierarchical topic models can provide a more comprehensive understanding of complex documents with desirable granularity. Due to this advantage, they have been applied in various downstream applications like document retrieval (Weninger, Bisk, and Han 2012), sentiment analysis (Kim et al. 2013), and text summarization (Celikyilmaz and Hakkani-Tur 2010) or generation (Guo et al. 2020).

Existing hierarchical topic models have two categories. The first category is conventional models like hLDA (Griffiths et al. 2003) and its variants (Kim et al. 2012; Paisley et al. 2013). They infer parameters through Gibbs sampling or Variational Inference. But they cannot well handle large-scale datasets due to their high computational cost (Chen et al. 2021b, 2023). The second category is neural models including HNTM (Chen et al. 2021a), HyperMiner (Xu et al. 2022), and others (Isonuma et al. 2020; Chen et al. 2021b, 2023; Duan et al. 2021). They generally follow VAE frameworks and enjoy back-propagation for faster parameter inferences (Wu, Nguyen, and Luu 2023).

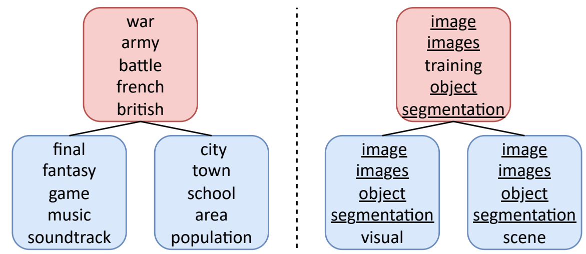


Figure 1: Illustration of low affinity (left), and low rationality and diversity issues (right). Each rectangle is the top related words of a topic from HyperMiner (Xu et al. 2022). Repetitive words are underlined.

However, these work struggles with producing low-quality topic hierarchies due to three issues: (**i**) *Low Affinity*: child topics are *not* affinitive to their parents (Kim et al. 2012). As exemplified in the left of Figure 1, the parent topic relates to “army”, whereas its child topics contain irrelevant words “game music” and “school”. Such low-affinity hierarchies capture inaccurate relationships among topics. (**ii**) *Low Rationality*: child topics are excessively similar to their parent topics instead of being specific to them as expected (Viegas et al. 2020). The right part of Figure 1 shows the parent and its child topics all focus on “image segmentation” with the same granularity. So low-rationality hierarchies provide topics with less comprehensive granularity. (**iii**) *Low Diversity*: sibling topics are repetitive instead of being diverse as expected (Zhang, Zhang, and Rao 2022). In the right part of Figure 1, the two sibling topics repeat each other and become redundant, implying other undisclosed latent topics. Thus low-diversity hierarchies produce less informative and incomplete topics. Due to these issues, existing hierarchical topic models generate low-quality hierarchies, which impedes document understanding and thus damages their interpretability and performance on downstream applications.

To address these challenges, we in this paper propose a novel neural hierarchical topic model, called **Tra**nsport Plan and **Co**ntext-aware Hierarchical Topic Model (**TraCo**). First, to address the low affinity and diversity issues, we propose a new **Transport Plan Dependency** (**TPD**) approach. Instead of unconstrained dependencies as previous work (Chen et al. 2021b; Duan et al. 2021; Xu et al. 2022), TPD models dependencies of hierarchical topics as optimal transport plans between them, which constrains the dependencies to ensure their sparsity and balance. Guided by the constrained dependencies, TPD additionally regularizes the building of topic hierarchies: it pushes a child topic only close to its parent and away from others, and avoids gathering excessive sibling topics together. As a result, this improves the affinity between child and parent topics and the diversity of sibling topics in learned hierarchies.

Second, to solve the low rationality issue, we further propose a novel **Context-aware Disentangled Decoder** (**CDD**). Rather than entangled decoding in early work (Chen et al. 2021a, 2023; Li et al. 2022), CDD decodes input documents using topics at each level individually, leading to disentangled decoding. In addition, the decoding of each level incorporates a bias containing topical semantics from its contextual levels. This incorporation forces topics at each level to cover semantics different from their contextual levels. In consequence, CDD can distribute different semantic granularity to topics at different levels, which therefore enhances the rationality of hierarchies. We conclude the contributions of this paper as follows [1]:

- We propose a novel neural hierarchical topic model with a new transport plan dependency method that regularizes topic hierarchy building with sparse and balanced dependencies, mitigating the low affinity and diversity issues.
- We further propose a new context-aware disentangled decoder, which explicitly distributes different semantic granularity to topics at different levels and thus alleviates the low rationality issue.
- We conduct extensive experiments on benchmark datasets and demonstrate that our model surpasses state-of-the-art baselines and significantly improves the affinity, rationality, and diversity of topic hierarchies.

Related Work

Conventional Hierarchical Topic Models Instead of flat topics like LDA (Blei, Ng, and Jordan 2003; Wu and Li 2019), Griffiths et al. (2003) propose hLDA to generate topic hierarchies with a nested Chinese Restaurant Process (nCRP). To relieve the single-path formulation of nCRP, Paisley et al. (2013) propose a nested Hierarchical Dirichlet Process. More variants are also explored (Mimno, Li, and McCallum 2007; Blei, Griffiths, and Jordan 2010; Perotte et al. 2011; Kim et al. 2012). Alternatively, Viegas et al. (2020) use NMF (Liu et al. 2018) with cluster word embeddings; Shahid et al. (2023) extend it with hyperbolic word embeddings. But they do *not* involve inferring topic distributions of documents.

[1] Our code is available at https://github.com/bobxwu/TraCo.

Neural Hierarchical Topic Models Recently, neural hierarchical topic models have emerged in the framework of VAE (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014; Nguyen and Luu 2021; Wu et al. 2020b; Wu, Luu, and Dong 2022; Wu et al. 2023a; Wu, Pan, and Luu 2023). Some follow conventional models (Pham and Le 2021; Zhang, Zhang, and Rao 2022). Isonuma et al. (2020) first propose a tree-structure topic model with two simplified doubly-recurrent neural networks. Chen et al. (2021b) propose nTSNTM with a stick-breaking process prior. Lately parametric settings attract more attention, *i.e.,* specify the number of topics at each level of a hierarchy (Wang et al. 2022, 2023). Chen et al. (2021a) propose a manifold regularization on topic dependencies. Li et al. (2022) use skip-connections for decoding and train with a policy gradient approach. Xu et al. (2022) model topic and word embeddings in hyperbolic space. Chen et al. (2023) use a Gaussian mixture prior and nonlinear structural equations to model dependencies. We follow the popular parametric setting, but differently focus on the low affinity, rationality, and diversity issues of hierarchical topic modeling. To address these issues, we propose the transport plan dependency to regularize topic hierarchy building and the context-aware disentangled decoder to separate semantic granularity.

Methodology

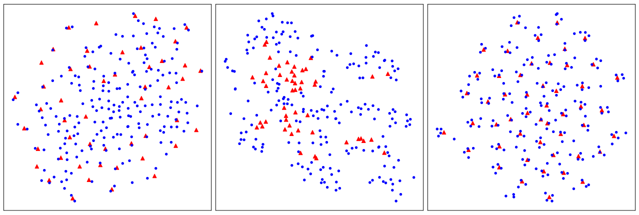
In this section, we recall the problem setting and notations of hierarchical topic modeling. Then we propose our transport plan dependency method and context-aware disentangled decoder. Finally we present our **Tra**nsport Plan and **Co**ntext-aware Hierarchical Topic Model (**TraCo**).

Problem Setting and Notations

Consider a collection of N documents: $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ with V unique words (vocabulary size). Following Chen et al. (2021a); Duan et al. (2021), we aim to discover a topic hierarchy with L levels from this collection, where level ℓ has $K^{(\ell)}$ latent topics. We build this hierarchy with dependency matrices describing the hierarchical relations between topics at two levels. For example, $\boldsymbol{\varphi}^{(\ell)} \in \mathbb{R}^{K^{(\ell+1)} \times K^{(\ell)}}$ denotes the dependency matrix between topics at level ℓ and $\ell+1$, where $\varphi^{(\ell)}_{kk'}$ is the relation between Topic#k at level $\ell+1$ and Topic#k' at level ℓ. Child topics should have high dependencies on their parents and low on others. Following LDA, we define each latent topic as a distribution over words (topic-word distribution), *e.g.,* Topic#k at level ℓ is defined as $\boldsymbol{\beta}^{(\ell)}_k \in \mathbb{R}^V$. Then $\boldsymbol{\beta}^{(\ell)} = (\boldsymbol{\beta}^{(\ell)}_1, \dots, \boldsymbol{\beta}^{(\ell)}_{K^{(\ell)}}) \in \mathbb{R}^{V \times K^{(\ell)}}$ is the topic-word distribution matrix of level ℓ. In addition, we infer doc-topic distributions at each level, *i.e.,* topic proportions in a document. For example, we denote $\boldsymbol{\theta}^{(\ell)} \in \Delta_{K^{(\ell)}}$ as the doc-topic distribution of a document $\mathbf{x}$ at level ℓ, where $\Delta_{K^{(\ell)}}$ is a probability simplex.

Parameterizing Hierarchical Latent Topics

At first we parameterize hierarchical latent topics. Following Miao, Grefenstette, and Blunsom (2017); Dieng, Ruiz, and Blei (2020), we project both words in the vocabulary



(a) HyperMiner (b) NGHTM (c) TraCo

Figure 2: t-SNE visualization (van der Maaten and Hinton 2008) of learned child (●) and parent (▲) topic embeddings of two levels. (a,b): Some child topic embeddings are *not* close enough to their parents; some are excessively gathered together. (c): TraCo pushes each child topic embedding only close to its parent and away from others, and avoids gathering excessive ones together.

and topics at all levels into an embedding space. In detail, we have V word embeddings: $\mathbf{W}=(\mathbf{w}_1,\dots,\mathbf{w}_V)\in\mathbb{R}^{D\times V}$ where D is the dimension. Similarly, we have $K^{(\ell)}$ topic embeddings for level ℓ: $\mathbf{T}^{(\ell)}=(\mathbf{t}^{(\ell)}_1,\dots,\mathbf{t}^{(\ell)}_{K^{(\ell)}})\in\mathbb{R}^{D\times K^{(\ell)}}$. Each topic (word) embedding represents its semantics. To model latent topics at level ℓ, we calculate its topic-word distribution matrix $\boldsymbol{\beta}^{(\ell)}$ following Wu et al. (2023b) as

$$\beta^{(\ell)}_{k,i}=\frac{\exp(-\|\mathbf{t}^{(\ell)}_k-\mathbf{w}_i\|^2/\tau)}{\sum_{k'=1}^{K}\exp(-\|\mathbf{t}^{(\ell)}_{k'}-\mathbf{w}_i\|^2/\tau)} \tag{1}$$

where $\beta^{(\ell)}_{k,i}$ is the correlation between i-th word and Topic#k at level ℓ with τ as a hyperparameter. Here we model the correlation as the Euclidean distance between word and topic embeddings and normalize over all topics at level ℓ.

Transport Plan Dependency

In this section we analyze why topic hierarchies are of low affinity and diversity, and then propose a novel solution called the Transport Plan Dependency (TPD).

Why Low Affinity and Diversity? As illustrated in Figure 1, previous models struggle with the low affinity and diversity issues. We consider the reason lies in their ways of modeling topic dependencies. Specifically, previous methods model dependencies between topics as the similarities between their topic embeddings. For instance, most studies compute the dot-product of topic embeddings as similarities and normalize them with a softmax function (Chen et al. 2021b; Duan et al. 2021). However, these dependencies are unconstrained and cannot regularize the building of topic hierarchies. As shown in Figures 2a and 2b, this incurs the low affinity and diversity issues: **(i) The dependencies may lack sparsity**, indicating child topic embeddings are *not* close enough to their parents. As a result, child topics are insufficiently associated with their parent topics, which damages the affinity of hierarchies. **(ii) The dependencies could be imbalanced**, indicating excessive child topic embeddings are gathered together close to only a few parents. In consequence, these topics become siblings and contain similar semantics, which impairs the diversity of hierarchies.

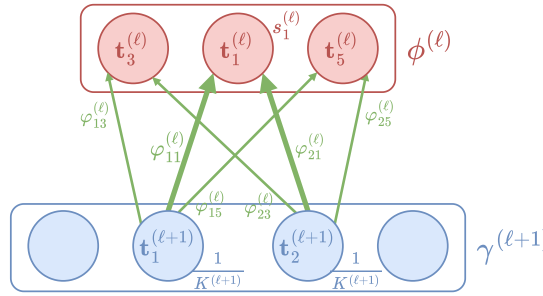


Figure 3: Illustration of TPD. It models the dependency $\varphi^{(\ell)}_{kk'}$ as the transport plan from topic embedding $\mathbf{t}^{(\ell+1)}_k$ to $\mathbf{t}^{(\ell)}_{k'}$ in measures $\gamma^{(\ell+1)}$ and $\phi^{(\ell)}$, constrained by the weight of $\mathbf{t}^{(\ell+1)}_k$ as $1/K^{(\ell+1)}$ and $\mathbf{t}^{(\ell)}_{k'}$ as $s^{(\ell)}_{k'}$. Here TPD pushes $\mathbf{t}^{(\ell+1)}_1$ close to $\mathbf{t}^{(\ell)}_1$ and away from others, similar for $\mathbf{t}^{(\ell+1)}_2$.

Modeling Dependencies as Transport Plans Based on the above analysis, to solve the low affinity and diversity issues, we propose a new Transport Plan Dependency (TPD) method that regularizes topic hierarchy building with sparse and balanced dependencies. Figure 3 illustrates TPD, and Figure 2c shows its effectiveness.

To constrain dependencies, we model them as the transport plan of a particularly defined optimal transport problem. Specifically, we define discrete measures on the topic embeddings at levels $\ell+1$ and ℓ respectively as $\gamma^{(\ell+1)}=\sum_{k=1}^{K^{(\ell+1)}} 1/K^{(\ell+1)}\sigma_{\mathbf{t}^{(\ell+1)}_k}$ and $\phi^{(\ell)}=\sum_{k'=1}^{K^{(\ell)}} s^{(\ell)}_{k'}\sigma_{\mathbf{t}^{(\ell)}_{k'}}$, where σ_x denotes the Dirac unit mass on x. Here the measures specify the weight of each topic embedding at level ℓ+1 as $1/K^{(\ell+1)}$, and each at level ℓ as $s^{(\ell)}_{k'}$ where $\mathbf{s}^{(\ell)}=(s^{(\ell)}_1,\dots,s^{(\ell)}_{K^{(\ell)}})$ is a weight vector and its sum is 1. Then we formulate an entropic regularized optimal transport problem between them:

$$\mathop{\arg\min}_{\boldsymbol{\pi}^{(\ell)}\in\mathbb{R}_+^{K^{(\ell+1)}\times K^{(\ell)}}} \mathcal{L}_{\mathrm{OT}_\varepsilon}(\gamma^{(\ell+1)},\phi^{(\ell)}),\quad \text{where}$$

$$\mathcal{L}_{\mathrm{OT}_\varepsilon}(\gamma^{(\ell+1)},\phi^{(\ell)})=\sum_{k=1}^{K^{(\ell+1)}}\sum_{k'=1}^{K^{(\ell)}}C^{(\ell)}_{kk'}\pi^{(\ell)}_{kk'}+\varepsilon\pi^{(\ell)}_{kk'}(\log\pi^{(\ell)}_{kk'}-1)$$

$$\text{s.t. } \boldsymbol{\pi}^{(\ell)}\mathbb{1}_{K^{(\ell)}}=1/K^{(\ell+1)}\mathbb{1}_{K^{(\ell+1)}},\ (\boldsymbol{\pi}^{(\ell)})^\top\mathbb{1}_{K^{(\ell+1)}}=\mathbf{s}^{(\ell)}. \tag{2}$$

The first term of $\mathcal{L}_{\mathrm{OT}_\varepsilon}$ is the original optimal transport problem, and the second term is the entropic regularization with hyperparameter ε to make this problem tractable (Canas and Rosasco 2012). Eq. (2) is to find a transport plan $\boldsymbol{\pi}^{(\ell)}$ that minimizes the total cost of transporting the weights of topic embeddings at level ℓ+1 to topic embeddings at ℓ and fulfills the two constraints. Here $\pi^{(\ell)}_{kk'}$ indicates the transport weight from $\mathbf{t}^{(\ell+1)}_k$ to $\mathbf{t}^{(\ell)}_{k'}$, and we compute the transport cost between them as Euclidean distance: $C^{(\ell)}_{kk'}=\|\mathbf{t}^{(\ell+1)}_k-\mathbf{t}^{(\ell)}_{k'}\|^2$. We denote $\mathbf{C}^{(\ell)}$ as the transport cost matrix. Eq. (2) has two constraints on $\boldsymbol{\pi}^{(\ell)}$ to balance transport weights where $\mathbb{1}_K$ is a K-dimensional column vector of ones.

To ensure the sparsity and balance of dependencies, we

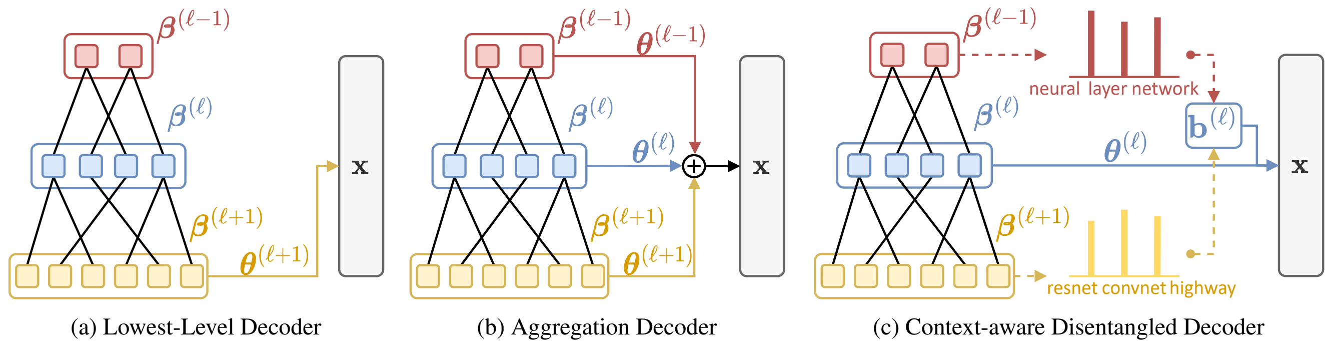


Figure 4: Comparison of decoders for hierarchical topic modeling. Here $\boldsymbol{\beta}^{(\ell)}$ and $\boldsymbol{\theta}^{(\ell)}$ are the topic-word distribution matrix and doc-topic distribution at level ℓ respectively. $\mathbf{x}$ is an input document to be decoded. (a): Decoding only with the lowest level. (b): Decoding with all levels. (c): Decoding with each level individually. For example, here the decoding using level ℓ incorporates the contextual topical bias $\mathbf{b}^{(\ell)}$. The bias includes topical semantics from contextual levels ($\ell-1$ and $\ell+1$), like the top related words "neural layer network" and "resnet convnet highway". This encourages topics at level ℓ ($\boldsymbol{\beta}^{(\ell)}$) to cover semantics different from them, like "deep convolutional cnn" (See this example in case studies). It is similar for other levels.

model them as the optimal transport plan solution of Eq. (2):

$$\boldsymbol{\varphi}^{(\ell)} = \mathrm{sinkhorn}(\mathcal{L}_{\mathrm{OT}_\varepsilon}(\gamma^{(\ell+1)}, \phi^{(\ell)})). \tag{3}$$

We resort to Sinkhorn's algorithm (Sinkhorn 1964; Cuturi 2013) to approximate the optimal transport plan (See details in Appendix A). This makes the obtained $\boldsymbol{\varphi}^{(\ell)}$ a differentiable variable parameterized by transport cost matrix $\mathbf{C}$ (Salimans et al. 2018; Genevay, Peyré, and Cuturi 2018). Here to obtain sparse and balanced dependencies, we model the dependency between Topic#k at level $\ell+1$ and Topic#k' at level ℓ as the transport weight between their topic embeddings $\mathbf{t}_k^{(\ell+1)}$ and $\mathbf{t}_{k'}^{(\ell)}$. Early studies prove that the optimal transport plan becomes sparse under a small ε (Peyré, Cuturi et al. 2019; Genevay, Dulac-Arnold, and Vert 2019). Therefore the modeled dependencies can keep sparsity. Besides, the two constraints in Eq. (2) ensure that the sparse transport plan needs to transport multiple topic embeddings at level $\ell+1$ with a total weight of $s_{k'}^{(\ell)}$ to topic embedding $\mathbf{t}_{k'}^{(\ell)}$ at level ℓ. Thus the modeled dependencies under these constraints can maintain balance.

Objective for TPD To regularize topic hierarchy building, we formulate the objective for TPD with the dependencies:

$$\mathcal{L}_{\mathrm{TPD}}^{(\ell)} = \sum_{k=1}^{K^{(\ell+1)}} \sum_{k'=1}^{K^{(\ell)}} C_{kk'} \varphi_{kk'}^{(\ell)} \tag{4}$$

where we minimize the total distance between topic embeddings at two levels weighted by dependencies. As shown in Figure 3, since dependencies $\boldsymbol{\varphi}^{(\ell)}$ are sparse, Eq. (4) pushes a child topic embedding only close to its parent and away from others. This facilitates the affinity of learned hierarchies. As the dependencies are also balanced, it properly aggregates child topic embeddings and avoids gathering excessive ones together. This improves the diversity of learned hierarchies. We demonstrate these in ablation studies.

Inferring Doc-Topic Distributions of Levels

We infer doc-topic distributions over each level for document decoding. We first infer $\boldsymbol{\theta}^{(L)}$, the doc-topic distributions over topics at the lowest level L following normal topic models (Srivastava and Sutton 2017; Wu et al. 2020a; Wu, Li, and Miao 2021). In detail, we define a random variable $\mathbf{r} \in \mathbb{R}^{K^{(L)}}$ with a logistic normal prior $\mathcal{LN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the mean and diagonal covariance matrix. We model its variational distribution as $q_\Theta(\mathbf{r}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To model parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, we use a neural network encoder f_Θ parameterized by Θ with the Bag-of-Words of document $\mathbf{x}$ as inputs. Then we sample $\mathbf{r}$ via the reparameterization trick as $\mathbf{r} = \boldsymbol{\mu} + (\boldsymbol{\Sigma})^{1/2}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We compute $\boldsymbol{\theta}^{(L)}$ with a softmax function as $\boldsymbol{\theta}^{(L)} = \mathrm{softmax}(\mathbf{r})$. Thereafter, we infer doc-topic distributions of a higher level ℓ as

$$\boldsymbol{\theta}^{(\ell)} = \big(\prod_{\ell'=\ell}^{L-1} (K^{(\ell'+1)} \boldsymbol{\varphi}^{(\ell')})^\top\big) \boldsymbol{\theta}^{(L)} \quad \text{where} \quad l < L. \tag{5}$$

Here we transform $\boldsymbol{\theta}^{(L)}$ via the dependencies of each level, and the multiplication of $K^{(\ell'+1)}$ rescales $\boldsymbol{\varphi}^{(\ell')}$ to produce normalized doc-topic distribution $\boldsymbol{\theta}^{(\ell)}$.

Context-aware Disentangled Decoder

In this section we explore why the low rationality issue happens. Then we propose a novel Context-aware Disentangled Decoder (CDD) to address this issue.

Why Low Rationality? As exemplified in Figure 1, early methods suffer from low rationality, *i.e.,* child topics have the same granularity as parent topics instead of being specific to them. We conceive the underlying reason lies in their decoders. As shown in Figure 4, previous decoders can be classified into two types. The first type is **lowest-level decoders** (Duan et al. 2021; Xu et al. 2022). Their decoding only engages the lowest-level topics. Higher-level topics are

Model	NeurIPS		ACL		NYT		20NG	
	TC	TD	TC	TD	TC	TD	TC	TD
nTSNTM	0.389	0.050	0.369	0.077	0.348	0.064	0.361	0.118
HNTM	0.356	0.118	0.368	0.191	0.345	0.134	0.376	0.151
NGHTM	0.362	0.235	0.371	0.235	0.359	0.145	0.358	0.166
SawETM	0.367	0.586	0.362	0.483	0.374	0.523	0.366	0.430
DCETM	0.385	0.173	0.398	0.090	0.371	0.652	0.380	0.561
ProGBN	0.377	0.436	0.372	0.462	0.372	0.502	0.370	0.495
HyperMiner	0.374	0.632	0.364	0.636	0.365	0.580	0.359	0.456
TraCo	**0.438**	**0.824**	**0.421**	**0.823**	**0.401**	**0.782**	**0.394**	**0.718**

Table 1: Topic quality results of Topic Coherence (TC) and Diversity (TD). The best are in bold. The superscript ‡ means the gain of is statistically significant at 0.05 level.

the linear combinations of these lowest-level topics via dependency matrices. In consequence, this entangles topics at all levels to cover the same semantic granularity, causing low rationality. The second type is **aggregation decoders** (Chen et al. 2021b,a; Li et al. 2022; Chen et al. 2023). Their decoding involves all levels, which still entangles topics at all levels. This endows the same semantics to these topics, so they become relevant but have similar granularity. As a result, learned hierarchies tend to have low rationality even with high affinity. Recently Duan et al. (2023) craft documents with more related words for the decoding of higher levels, but their granularity cannot be separated, still experiencing low rationality. See supports in the experiment section.

Contextual Topical Bias Motivated by the above, we aim to separate semantic granularity for each level to address the low rationality issue. Unfortunately, it is ***non-trivial*** since semantic granularity is unknown and varies in each domain. Some studies borrow external knowledge graphs (Wang et al. 2022; Duan et al. 2023), but such auxiliary information cannot fit various domains and mostly are unavailable. To overcome this challenge, we propose a new Context-aware Disentangled Decoder (CDD). Figure 4c illustrates CDD.

To separate semantic granularity, we propose to introduce a contextual topical bias to the decoding of each level. We denote this bias as a learnable variable $\mathbf{b}^{(\ell)} \in \mathbb{R}^V$ for level ℓ. We expect it to contain the topical semantics from the contextual levels of level ℓ in a hierarchy, so that level ℓ turns to cover other different semantics. Let $\mathbf{p}^{(\ell)}$ denote such topical semantics of level ℓ, and we model it as

$$\mathbf{p}^{(\ell)} = \sum_{\ell' \in \{\ell-1, \ell+1\}} \sum_{k=1}^{K^{(\ell')}} \mathrm{topK}(\boldsymbol{\beta}_k^{(\ell')}, N_{\text{top}}). \tag{6}$$

Here $\mathrm{topK}(\cdot,\cdot)$ returns a vector that retains the top N_{top} elements of $\boldsymbol{\beta}_k^{(\ell')}$ and fills the rest with 0. As such, $\mathbf{p}^{(\ell)}$ represents the contextual topical semantics as it includes the top related words of all topics at level ℓ–1 and ℓ+1 (only involves level ℓ+1 (ℓ–1) if level ℓ is the top-level (lowest-level)). Then we assign these contextual topical semantics to the bias $\mathbf{b}^{(\ell)}$:

$$b_i^{(\ell)} = p_i^{(\ell)} \quad \text{where} \quad p_i^{(\ell)} \neq 0. \tag{7}$$

So $\mathbf{b}^{(\ell)}$ contains the topical semantics from the contextual levels and also allows flexible bias learning on the semantics *not* covered by these levels. See an example in Figure 4c.

Disentangled Decoding with Contextual Topical Bias Instead of entangled decoding as early, we disentangle the decoding for each level with contextual topical biases. To be specific, we decode the document $\mathbf{x}$ with topics at level ℓ by sampling word x from a Multinomial distribution:

$$x \sim \mathrm{Multi}(\mathrm{softmax}(\boldsymbol{\beta}^{(\ell)}\boldsymbol{\theta}^{(\ell)} + \lambda_{\text{b}}\mathbf{b}^{(\ell)})) \tag{8}$$

Here $\boldsymbol{\beta}^{(\ell)}\boldsymbol{\theta}^{(\ell)}$ is the unnormalized generation probabilities following Srivastava and Sutton (2017). Recall that $\boldsymbol{\beta}^{(\ell)}$ is the topic-word distribution matrix, and $\boldsymbol{\theta}^{(\ell)}$ is the doc-topic distribution of $\mathbf{x}$ at level ℓ. The decoding incorporates the contextual topical bias $\mathbf{b}^{(\ell)}$ with a weight hyperparameter λ_{b}, *i.e.,* it knows the topical semantics of contextual levels. Thus the decoding turns to assign $\boldsymbol{\beta}^{(\ell)}$, topics at level ℓ, with semantics different from contextual levels. This explicitly separates different semantic granularity and properly distributes them to topics at different levels. As a result, we can effectively improve the rationality of hierarchies See evidence in ablation studies.

Transport Plan and Context-aware Hierarchical Topic Model

Finally we formulate the objective for our Transport Plan and Context-aware Hierarchical Topic Model (TraCo).

Objective for Topic Modeling Following the ELBO of VAE (Kingma and Welling 2014), we write the topic modeling objective with Eq. (8) as

$$\begin{aligned}\mathcal{L}_{\text{TM}}(\mathbf{x}) = &\frac{1}{L}\sum_{\ell=1}^{L} -\mathbf{x}^\top \log\big(\mathrm{softmax}(\boldsymbol{\beta}^{(\ell)}\boldsymbol{\theta}^{(\ell)} + \lambda_{\text{b}}\mathbf{b}^{(\ell)})\big) \\ &+ \mathrm{KL}\Big[q(\mathbf{r}|\mathbf{x}) \| p(\mathbf{r})\Big]\end{aligned} \tag{9}$$

The first term measures the average reconstruction error over all levels; the second term is the KL divergence between the prior and variational distributions.

Model	NeurIPS				ACL				NYT				20NG			
	PCC	PCD	SD	PnCD	PCC	PCD	SD	PnCD	PCC	PCD	SD	PnCD	PCC	PCD	SD	PnCD
nTSNTM	-0.348	0.603	0.195	0.566	-0.214	0.674	0.268	0.653	-0.450	0.501	0.193	0.479	-0.089	0.745	0.323	0.765
HNTM	-0.214	0.719	0.410	0.775	-0.095	0.867	0.568	0.887	-0.137	0.757	0.380	0.723	-0.332	0.832	0.425	0.796
NGHTM	0.014	0.905	0.635	0.954	0.055	0.902	0.633	0.947	-0.026	0.816	0.351	0.887	-0.011	0.831	0.446	0.863
SawETM	-0.093	0.785	0.816	0.986	-0.095	0.772	0.782	0.977	-0.234	0.641	0.680	0.970	-0.332	0.563	0.543	0.945
DCETM	-0.361	0.605	0.485	0.858	-0.353	0.584	0.387	0.804	-0.041	0.802	0.756	0.978	-0.085	0.742	0.644	0.900
ProGBN	-0.119	0.746	0.576	0.976	-0.058	0.781	0.611	0.976	-0.049	0.753	0.614	0.983	-0.009	0.780	0.626	0.981
HyperMiner	-0.084	0.771	0.808	0.991	-0.063	0.757	0.824	0.990	-0.229	0.638	0.713	0.984	-0.256	0.604	0.584	0.959
TraCo	**0.077**	**0.958**	**0.972**	**0.999**	**0.081**	**0.932**	**0.967**	**0.999**	**-0.021**	**0.946**	**0.946**	**0.998**	**0.037**	**0.895**	**0.894**	**0.997**

Table 2: Topic hierarchy quality results. PCC and PCD refer to the coherence and diversity between parent and child topics respectively; PnCD is the diversity between parent and non-child topics; SD is the diversity between sibling topics. The best are in bold. The superscript ‡ means the gain of TraCo is statistically significant at 0.05 level.

Model	NeurIPS				ACL				NYT				20NG			
	PCC	PCD	SD	PnCD	PCC	PCD	SD	PnCD	PCC	PCD	SD	PnCD	PCC	PCD	SD	PnCD
w/o TPD	-0.033	0.920	0.710	0.991	-0.023	0.898	0.606	0.968	-0.162	0.930	0.821	0.993	-0.084	**0.900**	0.723	0.985
w/o CDD	-0.083	0.772	0.907	0.996	-0.034	0.795	0.905	0.996	-0.140	0.731	0.828	0.991	-0.145	0.719	0.780	0.990
TraCo	**0.077**	**0.958**	**0.972**	**0.999**	**0.081**	**0.932**	**0.967**	**0.999**	**-0.021**	**0.946**	**0.946**	**0.998**	**0.037**	0.895	**0.894**	**0.997**

Table 3: Ablation study: without Transport Plan Dependency (w/o TDP); without Context-aware Disentangled Decoder (w/o CDD). The best are in bold. The superscript ‡ means the gain of TraCo is statistically significant at 0.05 level.

Objective for TraCo Based on the above, we write the overall objective for TraCo by combining Eq. (4) and (9):

$$\min_{\Theta,\mathbf{W},\{\mathbf{T}^{(\ell)}\}_{\ell=1}^{L}} \lambda_{\text{TPD}} \frac{1}{L-1}\sum_{\ell=1}^{L-1}\mathcal{L}_{\text{TPD}}^{(\ell)} + \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{TM}}(\mathbf{x}^{(i)}) \quad (10)$$

where λ_{TPD} is a weight hyperparameter. Here $\mathcal{L}_{\text{TPD}}^{(\ell)}$ regularizes topic hierarchy building with sparse and balanced dependencies; $\mathcal{L}_{\text{TM}}$ assigns topics at each level with different semantic granularity and infers doc-topic distributions.

Experiment

In this section we conduct experiments to show the effectiveness of our method.

Experiment Setup

Datasets We experiment with the following benchmark datasets: (i) **NeurIPS** contains the publications at the NeurIPS conference from 1987 to 2017. (ii) **ACL** (Bird et al. 2008) is a paper collection from the ACL anthology from 1970 to 2015. (iii) **NYT** contains news articles of the New York Times with 12 categories. (iv) **20NG** (Lang 1995) includes news articles with 20 labels.

Baseline Models We consider the following state-of-the-art baseline models: (i) **nTSNTM** (Chen et al. 2021b) uses a stick-breaking process prior. (ii) **HNTM** (Chen et al. 2021a) introduces manifold regularization on topic dependencies. (iii) **SawETM** (Duan et al. 2021) proposes a Sawtooth Connection to model topic dependencies. (iv) **DCETM** (Li et al. 2022) uses skip-connections into document decoding and a policy gradient training approach. (v) **HyperMiner** (Xu et al. 2022) projects topic and word embeddings into hyperbolic space. (vi) **NGHTM** (Chen et al. 2023) models dependencies via non-linear equations. (vii) **ProGBN** (Duan et al. 2023) crafts documents with more related words for the decoding of higher levels. We report average results of 5 runs. See more implementation details in the Appendix.

Topic Quality

Evaluation Metrics We adopt the below metrics following normal topic quality evaluation: (i) **T**opic **C**oherence (**TC**) measures the coherence between top words of topics. We evaluate with the widely-used metric C_V , outperforming earlier ones (Newman et al. 2010; Röder, Both, and Hinneburg 2015). (ii) **T**opic **D**iversity (**TD**) refers to differences between topics. Following Dieng, Ruiz, and Blei (2020), we measure TD as the uniqueness of top related words in topics.

Result Analysis Table 1 shows the average TC and TD scores over all levels. We see our TraCo consistently outperforms baselines concerning both TC and TD. Especially TraCo achieves significantly higher TD scores. For example, TraCo reaches a TD score of 0.824 on NeurIPS while the runner-up only has 0.632. These results demonstrate that our model can generate high-quality topics for different levels with better coherence and diversity.

Topic Hierarchy Quality

Evaluation Metrics We consider the following metrics to evaluate topic hierarchy: (i) **P**arent and **C**hild Topic

Coherence (**PCC**) indicates the coherence between parent and child topics. We use CLNPMI (Chen et al. 2021b) to measure it. CLNPMI computes the NPMI (Lau, Newman, and Baldwin 2014) of every two words from a parent topic and its child topic. (ii) **P**arent and **C**hild Topic **D**iversity (**PCD**) measures the diversity between a parent topic and its child (Chen et al. 2021b). PCC and PCD together verify if parent and child topics are relevant and cover different semantic granularity. This evaluates the rationality of a topic hierarchy. (iii) **P**arent and **n**on-**C**hild Topic **D**iversity (**PnCD**) measures the diversity between a parent topic and its non-child (Isonuma et al. 2020; Chen et al. 2021a). It verifies whether a child topic only has a high affinity to its parent topic. (iv) **S**ibling Topic **D**iversity (**SD**) measures the diversity between sibling topics. Note that PCD cannot replace SD since a parent topic may have repeating children. We follow the TD metric (Dieng, Ruiz, and Blei 2020) to compute the above PCD, PnCD, and SD.

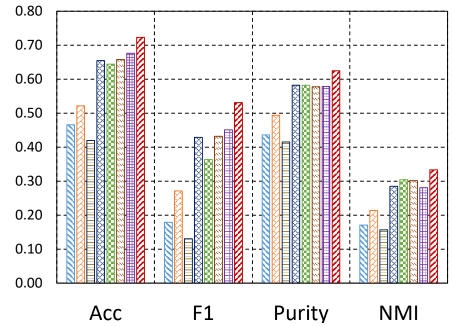
Result Analysis Table 2 reports the topic hierarchy quality results. We have the following observations: **(i) Our model shows higher affinity.** We see that our TraCo significantly surpasses all baselines concerning PCC and PnCD. This signifies that parent topics more relate to their children and differ from non-children in the hierarchies of TraCo, manifesting its enhanced affinity. **(ii) Our model attains better rationality.** Besides the best PCC, our TraCo reaches the best PCD compared to all baselines. For example, TraCo has PCC of 0.077 and PCD of 0.958 on NeurIPS while the runner-up has 0.014 and 0.905. This evidences that parent and child topics contain not only related semantics but also different granularity, which shows higher rationality of our method. **(iii) Our model achieves higher diversity.** Table 2 shows our TraCo outperforms baselines in terms of SD. For example, NGHTM has a close PCC score on NYT, but TraCo reaches much higher SD (0.946 vs. 0.351). This demonstrates our model produces more diverse sibling topics instead of repetitive ones.

Ablation Study

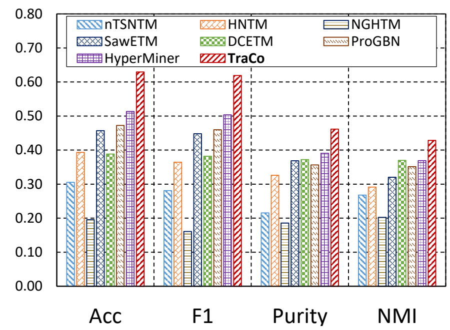
We conduct ablation studies to show the necessity of our TPD and CDD methods. From Table 3, we see that **TPD effectively mitigates the low affinity and diversity issues.** PCC and SD scores degrade largely if without TPD (w/o TPD). For example, PCC decreases from 0.077 to -0.033 and SD from 0.972 to 0.710 on NeurIPS. This implies less related parent and child topics and repetitive siblings. These results verify that our TPD facilitates the affinity and diversity of topic hierarchies. Besides, we notice that **CDD can alleviate the low rationality issue.** PCC and PCD decline significantly if without CDD (w/o CDD), like from 0.081 to -0.034 and from 0.932 to 0.795 on ACL, indicating less distinguishable parent and child topics. This demonstrates that our CDD improves the rationality of topic hierarchies.

Text Classification and Clustering

Apart from the above comparisons, we evaluate inferred doc-topic distributions through downstream tasks: text classification and clustering. Specifically, we train SVM classifiers with learned doc-topic distributions as features and predict document labels, evaluated by Accuracy (Acc) and F1. For clustering, we use the most significant topics in doc-topic distributions as clustering assignments, evaluated by Purity and NMI following Zhao et al. (2021). We take the average classification and clustering results over all hierarchy levels on the NYT and 20NG datasets.



(a) NYT



(b) 20NG

Figure 5: Text classification (Acc and F1) and clustering results (Purity and NMI). The gains of our TraCo are all statistically significant at 0.05 level.

Figure 5 shows our TraCo consistently outperforms baseline methods in terms of both text classification and clustering. These demonstrate that our model can infer higher-quality doc-topic distributions for different hierarchy levels, which can benefit downstream applications. As we infer higher-level doc-topic distributions via dependencies (Eq. (5)), these manifest that the learned dependencies of our model are accurate as well.

Conclusion

In this paper we propose TraCo for hierarchical topic modeling. Our TraCo uses a transport plan dependency method to address the low affinity and diversity issues, and leverages a context-aware disentangled decoder to mitigate the low rationality issue. Experiments demonstrate that TraCo can consistently outperform baselines, producing higher-quality topic hierarchies with significantly improved affinity, diversity, and rationality. Especially TraCo shows better performance on downstream tasks with more accurate topic distributions of documents.

Acknowledgements

We thank all anonymous reviewers for their helpful comments. This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme, AISG Award No: AISG2-TC-2022-005.

References

Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B. R.; Joseph, M. T.; Kan, M.-Y.; Lee, D.; Powley, B.; Radev, D. R.; Tan, Y. F.; et al. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc. of LREC*.

Blei, D. M.; Griffiths, T. L.; and Jordan, M. I. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Canas, G.; and Rosasco, L. 2012. Learning probability measures with respect to optimal transport metrics. *Proc. of NeurIPS*.

Celikyilmaz, A.; and Hakkani-Tur, D. 2010. A hybrid hierarchical model for multi-document summarization. In *Proc. of ACL*.

Chen, H.; Mao, P.; Lu, Y.; and Rao, Y. 2023. Nonlinear Structural Equation Model Guided Gaussian Mixture Hierarchical Topic Modeling. In *Proc. of ACL*.

Chen, Z.; Ding, C.; Rao, Y.; Xie, H.; Tao, X.; Cheng, G.; and Wang, F. L. 2021a. Hierarchical neural topic modeling with manifold regularization. *World Wide Web*.

Chen, Z.; Ding, C.; Zhang, Z.; Rao, Y.; and Xie, H. 2021b. Tree-structured topic modeling with nonparametric neural variational inference. In *Proc. of ACL*.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Proc. of NeurIPS*.

Dieng, A. B.; Ruiz, F. J.; and Blei, D. M. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*.

Duan, Z.; Liu, X.; Su, Y.; Xu, Y.; Chen, B.; and Zhou, M. 2023. Bayesian Progressive Deep Topic Model with Knowledge Informed Textual Data Coarsening Process. In *Proc. of ICML*.

Duan, Z.; Wang, D.; Chen, B.; Wang, C.; Chen, W.; Li, Y.; Ren, J.; and Zhou, M. 2021. Sawtooth factorial topic embeddings guided gamma belief network. In *Proc. of ICML*.

Genevay, A.; Dulac-Arnold, G.; and Vert, J. 2019. Differentiable Deep Clustering with Cluster Size Constraints. *CoRR*.

Genevay, A.; Peyré, G.; and Cuturi, M. 2018. Learning generative models with sinkhorn divergences. In *Proc. of AISTATS*.

Griffiths, T.; Jordan, M.; Tenenbaum, J.; and Blei, D. 2003. Hierarchical topic models and the nested Chinese restaurant process. *Proc. of NeurIPS*.

Guo, D.; Chen, B.; Lu, R.; and Zhou, M. 2020. Recurrent hierarchical topic-guided RNN for language generation. In *Proc. of ICML*.

Isonuma, M.; Mori, J.; Bollegala, D.; and Sakata, I. 2020. Tree-structured neural topic model. In *Proc. of ACL*.

Kim, J. H.; Kim, D.; Kim, S.; and Oh, A. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proc. of CIKM*.

Kim, S.; Zhang, J.; Chen, Z.; Oh, A.; and Liu, S. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proc. of AAAI*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *Proc. of ICLR*.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Proc. of ICML*.

Lau, J. H.; Newman, D.; and Baldwin, T. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proc. of EACL*.

Li, Y.; Wang, C.; Duan, Z.; Wang, D.; Chen, B.; An, B.; and Zhou, M. 2022. Alleviating" Posterior Collapse"in Deep Topic Models via Policy Gradient. *Proc. of NeurIPS*.

Liu, R.; Wang, X.; Wang, D.; Zuo, Y.; Zhang, H.; and Zheng, X. 2018. Topic splitting: a hierarchical topic model based on non-negative matrix factorization. *Journal of Systems Science and Systems Engineering*.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *Proc. of ICML*.

Mimno, D.; Li, W.; and McCallum, A. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proc. of ICML*.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Proc. of NAACL*.

Nguyen, T.; and Luu, A. T. 2021. Contrastive Learning for Neural Topic Model. *Proc. of NeurIPS*.

Paisley, J.; Wang, C.; Blei, D.; and Jordan, M. I. 2013. A nested hdp for hierarchical topic models. *arXiv preprint arXiv:1301.3570*.

Perotte, A.; Wood, F.; Elhadad, N.; and Bartlett, N. 2011. Hierarchically supervised latent Dirichlet allocation. *Proc. of NeurIPS*.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*.

Pham, D.; and Le, T. M. 2021. Neural topic models for hierarchical topic detection and visualization. In *Proc. of KDD*.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. *In Proceedings ofthe 31th International Conference on Machine Learning*.

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proc. of WSDM*.

Salimans, T.; Zhang, H.; Radford, A.; and Metaxas, D. 2018. Improving GANs using optimal transport. *arXiv preprint arXiv:1803.05573*.

Shahid, S.; Anand, T.; Srikanth, N.; Bhatia, S.; Krishnamurthy, B.; and Puri, N. 2023. HyHTM: Hyperbolic Geometry-based Hierarchical Topic Model. In *Proc. of ACL Findings*.

Sinkhorn, R. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*.

Srivastava, A.; and Sutton, C. 2017. Autoencoding Variational Inference For Topic Models. In *Proc. of ICLR*.

Teh, Y.; Jordan, M.; Beal, M.; and Blei, D. 2004. Sharing clusters among related groups: Hierarchical Dirichlet processes. *Proc. of NeurIPS*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*.

Viegas, F.; Cunha, W.; Gomes, C.; Pereira, A.; Rocha, L.; and Goncalves, M. 2020. CluHTM-semantic hierarchical topic modeling based on CluWords. In *Proc. of ACL*.

Wang, D.; Xu, Y.; Li, M.; Duan, Z.; Wang, C.; Chen, B.; Zhou, M.; et al. 2022. Knowledge-aware Bayesian deep topic model. *Proc. of NeurIPS*.

Wang, N.; Wang, D.; Jiang, T.; Du, C.; Fang, C.; and Zhuang, F. 2023. Hierarchical Neural Topic Model with Embedding Cluster and Neural Variational Inference. In *Proc. of SDM*.

Weninger, T.; Bisk, Y.; and Han, J. 2012. Document-topic hierarchies from document graphs. In *Proc. of CIKM*.

Wu, X.; Dong, X.; Nguyen, T.; Liu, C.; Pan, L.; and Luu, A. T. 2023a. InfoCTM: A Mutual Information Maximization Perspective of Cross-Lingual Topic Modeling. *arXiv preprint arXiv:2304.03544*.

Wu, X.; Dong, X.; Nguyen, T.; and Luu, A. T. 2023b. Effective neural topic modeling with embedding clustering regularization. In *Proc. of ICML*.

Wu, X.; and Li, C. 2019. Short Text Topic Modeling with Flexible Word Patterns. In *Proc. of IJCNN*.

Wu, X.; Li, C.; and Miao, Y. 2021. Discovering Topics in Long-tailed Corpora with Causal Intervention. In *Proc. of ACL Findings*.

Wu, X.; Li, C.; Zhu, Y.; and Miao, Y. 2020a. Learning Multilingual Topics with Neural Variational Inference. In *Proc. of NLPCC*.

Wu, X.; Li, C.; Zhu, Y.; and Miao, Y. 2020b. Short Text Topic Modeling with Topic Distribution Quantization and Negative Sampling Decoder. In *Proc. of EMNLP*.

Wu, X.; Luu, A. T.; and Dong, X. 2022. Mitigating Data Sparsity for Short Text Topic Modeling by Topic-Semantic Contrastive Learning. In *Proc. of EMNLP*.

Wu, X.; Nguyen, T.; and Luu, A. T. 2023. A Survey on Neural Topic Models: Methods, Applications, and Challenges. *Research Square*.

Wu, X.; Pan, F.; and Luu, A. T. 2023. Towards the TopMost: A Topic Modeling System Toolkit. *arXiv preprint arXiv:2309.06908*.

Xu, Y.; Wang, D.; Chen, B.; Lu, R.; Duan, Z.; and Zhou, M. 2022. HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding. In *Proc. of NeurIPS*.

Zhang, Z.; Zhang, X.; and Rao, Y. 2022. Nonparametric Forest-Structured Neural Topic Modeling. In *Proc. of COLING*.

Zhao, H.; Phung, D.; Huynh, V.; Le, T.; and Buntine, W. L. 2021. Neural Topic Model via Optimal Transport. In *Proc. of ICLR*.