Unsupervised Hierarchical Topic Modeling via Anchor Word Clustering and Path Guidance

Jiyuan Liu and **Hegang Chen** and **Chunjiang Zhu** and **Yanghui Rao**[*]
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
Department of Computer Science, University of North Carolina at Greensboro, NC, USA
{liujy563,chenhg25}@mail2.sysu.edu.cn, chunjiang.zhu@uncg.edu,
raoyangh@mail.sysu.edu.cn

Abstract

Hierarchical topic models nowadays tend to capture the relationship between words and topics, often ignoring the role of anchor words that guide text generation. For the first time, we detect and add anchor words to the text generation process in an unsupervised way. Firstly, we adopt a clustering algorithm to adaptively detect anchor words that are highly consistent with every topic, which forms the path of *topic→anchor word*. Secondly, we add the causal path of *anchor word→word* to the popular Variational Auto-Encoder (VAE) framework via implicitly using word co-occurrence graphs. We develop the causal path of *topic+anchor word→higher-layer topic* that aids the expression of topic concepts with anchor words to capture a more semantically tight hierarchical topic structure. Finally, we enhance the model's representation of the anchor words through a novel contrastive learning. After jointly training the aforementioned constraint objectives, we can produce more coherent and diverse topics with a better hierarchical structure. Extensive experiments on three datasets show that our model outperforms state-of-the-art methods.

1 Introduction

Topic models, which can automatically discover coherent and meaningful topics from text corpora, have been widely used for text analysis (Rubin et al., 2012; Wang et al., 2018; Jelodar et al., 2020). In such methods, each topic is interpreted as relevant words to represent a semantic concept. Different from traditional flat topic models, Hierarchical Topic Models (HTMs) aim to leverage the hierarchical nature of topics to build a rational topic structure (Zhang et al., 2022). HTMs have 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., 2014).

Existing HTMs can be divided into two categories. The first category is conventional models like hLDA (Griffiths et al., 2003) and its variants (Kim et al., 2012). They infer parameters through Gibbs sampling or Variational Inference, which require high computational costs or complex derivation (Chen et al., 2021b, 2023). The second category is neural hierarchical topic models, including HNTM (Chen et al., 2021a), HyperMiner (Xu et al., 2022), and so forth (Isonuma et al., 2020; Chen et al., 2021b, 2023; Duan et al., 2021; Wu et al., 2024b). These methods generally follow the VAE framework and employ back-propagation for faster parameter inferences (Wu et al., 2024a). However, the generation process of most previous methods overlooks the role of anchor words and directly generates words through topics, resulting in insufficient mining of fine-grained topic-word information. In this study, we adopt the framework of the second category and attempt to alleviate the above issue by exploiting anchor words.

The anchor word guided method (Arora et al., 2012) is based on the separability assumption (Donoho and Stodden, 2003), which assumes that each topic contains at least one highly relevant anchor word that uniquely identifies the topic. For example, while *"resurrection"*, *"pray"*, *"sin"*, and *"christ"* are associated with a topic about *christianity*, only *"christ"* is unambiguous, so it could serve as this topic's anchor word. Based on this assumption, there are two types of anchor word guided topic models. The first type detects anchor words by an interactive process, and uses these words to identify topics in the context of non-negative matrix factorization (Arora et al., 2012; Mimno and Lee, 2014; Arora et al., 2013). These works require intensive matrix calculations, making them unstable and noisy (Arora et al., 2013). The second type obtains anchor words through label information (Jagarlamudi et al., 2012; Gallagher et al., 2017; Lin et al., 2023). These works can be seen as a semi-

[*] The corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7505–7517
November 12-16, 2024 ©2024 Association for Computational Linguistics

supervised anchor word detection fashion, which is unachievable without any external information.

Although aforementioned works demonstrate the important role of anchor words in guiding topic-word relationships, they have their own shortcomings and the application of anchor words in unsupervised hierarchical topic mining is scarce. So there are two key issues here: First, how to **detect anchor words** unsupervisedly on a neural hierarchical topic model; Second, how anchor words can help model **fine-grained topic-word relationships and hierarchical topic structures**.

In our proposed **A**nchor Word Clustering and **P**ath Guided framework for unsupervised **H**ierarchical **T**opic **M**odeling (AP-HTM), we detect anchor words and introduce four causal paths to constrain the text generation process. Fig. 1(a) and Fig. 1(b) show the text generation processes of our AP-HTM and other HTMs (Chen et al., 2021b; Li et al., 2022; Chen et al., 2023), respectively. First, we unsupervisedly detect the anchor words of each topic by a clustering algorithm (Meng et al., 2022; Xie et al., 2016), which forms the path of *topic→anchor word*, as shown in Fig. 1(c). Second, two causal paths related to the anchor words are added to the text generation process. In order to obtain a semantically tight final embedding space, the causal path of *anchor word→word* (Fig. 1(d)) is introduced in each layer. We use the word co-occurrence graph to implicitly capture the relationships between anchor words and other words, e.g., *"christ"→"scriptures", "bible"*. Then, the topic generation structure of *topic+anchor word→higher-layer topic* (Fig. 1(e)) is adopted between layers to capture the topic hierarchical relationships, utilizing anchor words as auxiliary information to enrich the conceptual relevance between topics. Further, for decoupling of anchor words and other words in causal diagram, we employ a novel contrastive learning (Fig. 1(f)), masking anchor words during inference that makes the encoder pay more attention to the vital information of anchor words.

In summary, our contributions are as follows:

- We propose a new framework AP-HTM, which introduces anchor words as important components into text generation. And we adopt a clustering algorithm to adaptively detect anchor words in the anchor space.
- We add four causal path constraints to the VAE framework that can guide to identify high-quality hierarchical topics.
- We introduce a novel anchor-based contrastive learning approach that updates the negative sample during the training process, which enhance model representation of fine-grained anchor words.
- Extensive experiments are conducted on three datasets to evaluate our model. The results show that the performance of AP-HTM is significantly better than the state-of-the-art baselines.

2 Related works

2.1 Hierarchical Topic Models

As an alternative to flat topics models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003), hLDA (Griffiths et al., 2003) was porposed to generate topic hierarchies with a nested Chinese Restaurant Process (nCRP). More variants based on traditional Bayesian probabilistic methods have been explored in early research (Mimno et al., 2007; Blei et al., 2010; Perotte et al., 2011; Kim et al., 2012). Later, CluHTM (Viegas et al., 2020) used Nonnegative Matrix Factorization (NMF) (D. Lee and Seung, 2000) with cluster of words embeddings, HyHTM (Shahid et al., 2023) extended it with hierarchical information from hyperbolic geometrys. But they cannot infer topic distributions of documents.

Recently, Neural Hierarchical Topic Models (NHTMs) have emerged in the framework of VAE (Kingma and Welling, 2013). Some works are based on traditional non-parametric models (Zhang et al., 2022). Isonuma et al. (2020) first proposed a tree-structure topic model with two simplified doubly-recurrent neural networks. Chen et al. (2021b) proposed nTSNTM with a stick-breaking process prior. Parameter settings that specify the number of topics at each level of the hierarchy are also gaining attention. SawETM (Duan et al., 2021) exploited a sawtooth connection module to mitigate the problem of posterior collapse. HyperMiner (Xu et al., 2022) modeled topic and word embeddings in the hyperbolic space. NG-HTM (Chen et al., 2023) used a Gaussian mixture prior and nonlinear structural equations to model dependencies. TraCo (Wu et al., 2024b) leveraged a transport plan dependency method to regularize topic hierarchy.

2.2 Anchor Word Guided Topic Models

Introducing anchor words into topic models has been a widely adopted way to improve topic quality and interpretability (Mimno and Lee, 2014). Initially, a series of works based on NMF were

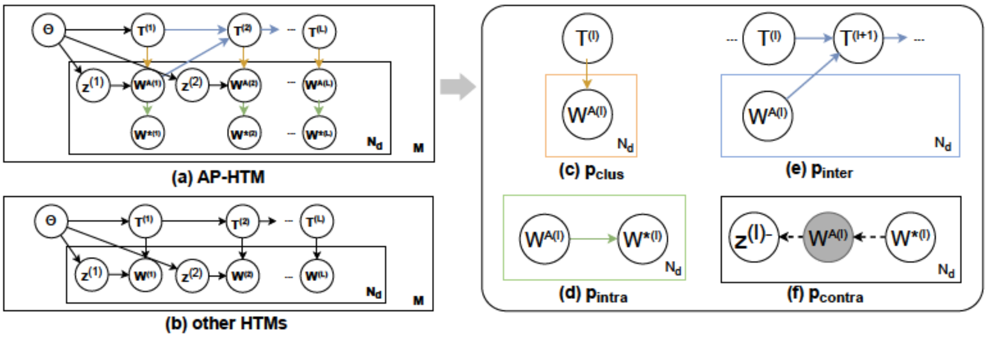


Figure 1: The text generation process of (a) our AP-HTM and (b) other HTMs, where Θ is the set of parameters for the generation process, $T^{(l)}$ is the topic embedding at layer l, and z is document-topic distribution, N_d is the number of words in document d, W^A and W^* are the word embedding in anchor space and final space. (c) p_{clus}, (d) p_{intra}, and (e) p_{inter} show the decomposition of the three causal paths, and (f) p_{contra} demonstrates the effect of applying anchor perturbation (grey variable) to W^A (details in Section 4.3). The solid line represents the generation model, and the dotted line represents the inference process.

proposed, and Arora et al. (2012) proposed the concept of anchor word in topic modeling, distinguishing the different roles of anchor words from other words. There is also a series of work that improved the method of automatically finding anchor words. FastAnchorWords (Arora et al., 2013) is proposed to select anchor words, which provided efficiency and practicality in practical applications. Mimno and Lee (2014) proposed a greedy anchor method to find exact convex hulls in the low dimensional space. Nguyen et al. (2014) proposed a new regularization priori in the anchor method to improve the interpretability and flexibility of the model.

Parallel to this, there is another thread of works that use mutual information between labels and words to semi-supervisedly extract anchor words. SeedLDA (Jagarlamudi et al., 2012) paired each topic with a seed topic and biased documents to topics if they have corresponding anchor words. Anchored CorEx (Gallagher et al., 2017) provided guidance on topic modeling by flexibly integrating word-level domain knowledge into the model via anchor words. SeededNTM (Lin et al., 2023) used label information to extract anchor words for multilevel supervisions.

3 Background

We define the basic hierarchical topic modeling process in the following parts.

Encoder: Given a collection of documents, we process each document d into a Bag-of-Words (BoW) vector $x_{bow} \in \mathbb{R}^V$, where V is the vocabulary size. The Gaussian mixture encoder network can be described as follows:

$$h_e = f(x_{bow}), \tag{1}$$

$$c = \text{Gumbel Softmax}(h_e), \tag{2}$$

$$h_e^{(1)} = \text{Reparameter}(h_e, c), \tag{3}$$

$$h_e^{(l+1)} = \tanh(h_e^{(l)} \Pi^{(l)}), \tag{4}$$

$$z^{(l)} = softmax(h_e^{(l)}), \tag{5}$$

where h_e is the initial hidden representation in encoder, and $h_e^{(l)}$ is the hidden document representation at layer l, $\Pi^{(l)}$ is the topic hierarchy matrix between layers l and $l+1$, L is the total number of topic layers, $z^{(l)}$ is the document-topic distribution at layer l, $f(\cdot)$ stands for multilayer neural networks, and the Gumbel Softmax layer produces a c-dimensional label. Following Dilokthanakul et al. (2016), the number of mixture components c is set to 10.

Decoder: The document decoder can be expressed as follows:

$$\Phi^{(l)} = softmax\left(T^{(l)} \times W^T\right), \tag{6}$$

$$\hat{x} = \sum_{l=1}^{L} \hat{x}^{(l)} = \sum_{l=1}^{L} z^{(l)} \Phi^{(l)}, \tag{7}$$

where $\Phi^{(l)}$ is the topic-word distribution at layer l, and $\hat{x}$ is the document reconstructed from decoder. The symbol description is detailed in Table 1.

4 Method

After introducing the basic hierarchical topic modeling process in Section 3, we briefly describe four paths that extend the basic VAE framework

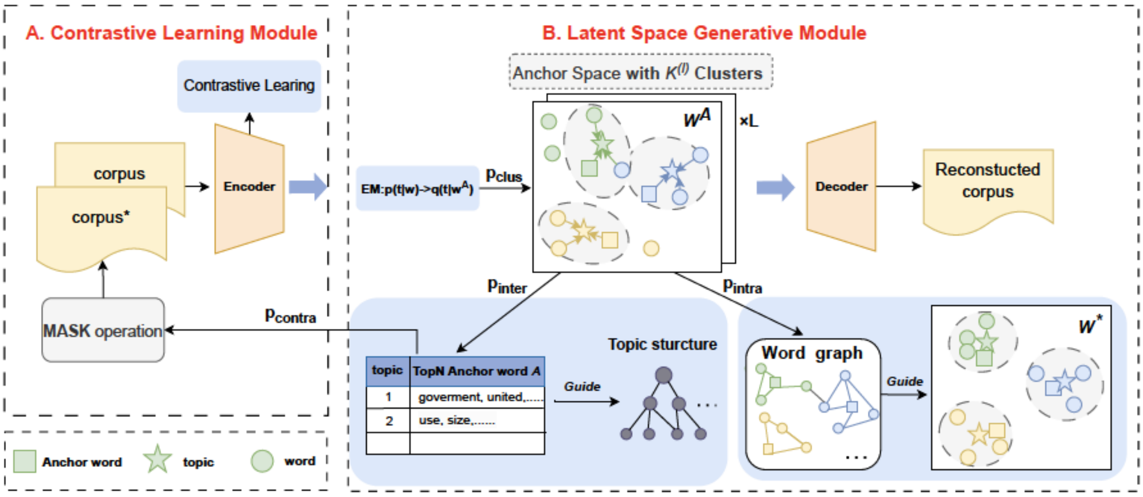


Figure 2: The architecture of the proposed AP-HTM model.

Symbols	Descriptions
W	the word embedding from pre-trained model
W^A, W^*	the word embedding in anchor and final space
d	the document in a corpus
Π	the topic hierarchy
T, T^*	the embedding of topic and fusion topic
A	top$N$ anchor words
θ	set of parameters for the generation process
z	document-topic distribution
Φ^A, Φ^*	topic-word distribution in anchor and final space
$\Phi_k^{(l)}$	k^{th} topic-word distribution at layer l
M	the number of documents
N_d	the number of words in document d
L	the number of layers
$K^{(l)}$	the topic number at layer l
r	embedding dimension of the latent space
$\hat{\mathbf{A}}^{\mathbf{adj}}$	modified adjacency matrix
S_A, S_P	anchor word set and the least relevant word set

Table 1: Symbol description.

in Section 4.1, which are introduced to model fine-grained topic-word relationships and the hierarchical structure. In Section 4.2 and Section 4.3, we present the specifics of how our model works in the form of two modules (Fig. 2). Section 4.4 shows the training details of our AP-HTM.

4.1 Enhancing VAE via Four Causal Paths

In response to the two issues, we enhance the basic VAE framework in Section 3 through the following four causal paths.

(I) p_{clus}: In order to obtain a sharpened anchor space $W^A \in \mathbb{R}^{V \times r}$ with $K^{(l)}$ topics in each layer, we conduct clustering on the pre-trained word space $W \in \mathbb{R}^{V \times r}$, where r is the embedding dimension. And we adaptively extract $TopN$ anchor words $A^{(l)}$ in W^A.

(II) p_{intra}: In order to obtain a semantically consistent and diverse word embedding $W^* \in \mathbb{R}^{V \times r}$ in the final space, we use the word co-occurrence graph $\mathcal{G}$ to guide anchor space W^A.

(III) p_{inter}: In order to obtain a reasonable hierarchical structure $\Pi^{(l)}$, we fuse anchor words $A^{(l)}$ and topic embeddings $t^{(l)}$ as new topic embeddings $t^{*(l)}$ to guide the generation of hierarchical relationships between $t^{*(l)}$ and $t^{(l+1)}$.

(IV) p_{contra}: In order to obtain an encoder that accurately captures the relationship between anchor words and topics, we dynamically update the anchor word set $S_A = \bigcup_{l \in L} A^{(l)}$, by masking the original document d as negative samples for contrastive learning.

4.2 Latent Space Generative Module

4.2.1 Anchor Space Clustering

In this part, we aim to obtain anchor words for each topic through a clustering algorithm, e.g., Xie et al. (2016), thereby introducing anchor space W^A with $K^{(l)}$ well separated clusters, which realizes the path of *topic→anchor word* (Fig. 1(c)).

Following a previous work (Meng et al., 2022), we use the expectation–maximization (EM) algorithm to gradually sharpen the posterior topic-word distribution. In the E-Step, we estimate a new soft cluster assignment of each word based on the current parameters; in the M-step, we update the model parameters given the cluster assignments.

E-Step. To estimate the cluster assignment of each word, we compute the posterior topic distribution of the k^{th} topic at l^{th} layer $t_k^{(l)}$ and i^{th} word w_i obtained via the Bayes rule:

$$p\left(t_k^{(l)}|w_i\right) = \frac{p\left(w_i|t_k^{(l)}\right)p\left(t_k^{(l)}\right)}{\sum_{k'=1}^{K^{(l)}} p\left(w_i|t_{k'}^{(l)}\right)p\left(t_{k'}^{(l)}\right)}. \quad (8)$$

According to Eq. (6), $p\left(w_i|t_k^{(l)}\right) = \phi_{k,i}^{(l)} = \exp(t_k^{(l)} \cdot {w_i}^T)$. And we assume that a topic $t_k^{(l)}$ is sampled from a uniform distribution over the $K^{(l)}$ topics, so $p\left(t_k^{(l)}\right) = \frac{1}{K^{(l)}}$. The posterior is simplified as $p\left(t_k^{(l)}|w_i\right) = \frac{\exp\left(t_k^{(l)} \cdot {w_i}^T\right)}{\sum_{k'=1}^{K} \exp\left(t_{k'}^{(l)} \cdot {w_i}^T\right)}$.

Then we compute a new estimate of the cluster assignments $q(t_k|w_i^A)$ to be used for updating the model in the *M-Step* following (Xie et al., 2016):

$$q\left(t_k^{(l)}|{w_i}^A\right) = \frac{p\left(t_k^{(l)}|w_i\right)^2 / s_k^{(l)}}{\sum_{k'=1}^{K^{(l)}} p\left(t_{k'}^{(l)}|w_i\right)^2 / s_{k'}^{(l)}},$$
$$s_k^{(l)} = \sum_{i=1}^{V} p\left(t_k^{(l)}|w_i\right). \quad (9)$$

M-Step. We update the model parameters to maximize the expected log-probability of the current cluster assignment under the new cluster assignment estimate $\mathbb{E}_q[\log p]$, which is equivalent to minimizing the following cross entropy loss:

$$\mathcal{L}_{\text{clus}} = -\sum_{l=1}^{L}\sum_{i=1}^{V}\sum_{k=1}^{K^{(l)}} q\left(t_k^{(l)}|{w_i}^A\right) \times \log p\left(t_k^{(l)}|w_i\right). \quad (10)$$

As shown in Fig. 1(a), we here obtain the anchor space, where the words are embedded as $W^A = [w_1^A, w_2^A, \ldots, w_V^A]$ mentioned in p_{clus}.

4.2.2 Generating Anchor Words

Similar to Eq. (6), we get the topic-word distribution $\Phi^{A^{(l)}}$ at layer l in anchor space W^A by:

$$\Phi^{A^{(l)}} = softmax\left(T^{(l)} \times (W^A)^T\right). \quad (11)$$

We then obtain the anchor words $A^{(l)}$ as follows:

$$A^{(l)} = TopN(\Phi^{A^{(l)}}) \in \mathbb{R}^{K^{(l)} \times N^w}, \quad (12)$$

where $TopN(\cdot)$ returns a vector that retains top N^w words of each row in Φ^A. Here we set N^w as 1 and get top 1 anchor words $A^{(l)}$ of layer l.

4.2.3 Intra-topic Path

After sharpening the pre-train space W via clustering, we obtain the anchor space W^A with good clustering structure, in which each topic maintains the most representative words as anchor words A. Next, we will constrain the *anchor word→word* path (Fig. 1(d)) to get the final space W^*.

Inspired by a previous work (Arora et al., 2012), we construct a word co-occurrence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ (words as nodes) and $\mathcal{E}$ (counts of corresponding biterms) are sets of nodes and edges, respectively. we compute the modified adjacency matrix $\hat{\mathbf{A}}^{\mathbf{adj}}$ by:

$$\hat{\mathbf{A}}^{\mathbf{adj}} = (\mathbf{D}+\mathbf{I}_N)^{-1/2}(\mathbf{A}^{adj}+\mathbf{I}_N)(\mathbf{D}+\mathbf{I}_N)^{-1/2}. \quad (13)$$

In the above, $\mathbf{A}^{\mathbf{adj}} \in \mathbb{R}^{V \times V}$ is the adjacency matrix, D is the degree matrix of $\mathbf{A}^{\mathbf{adj}}$, and I_N is the unit matrix. Unlike a previous work using GCNs (Zhu et al., 2018) for learning the graph structure data, we achieve graph-guided embeddings through a simple and effective traction formulation as follows:

$$W^* = W^A + b\hat{\mathbf{A}}^{\mathbf{adj}} \times W^A, \quad (14)$$

where b is a hyperparameter that controls the degree of graph guidance. Intuitively, two word nodes that are similar in $\mathcal{G}$ are considered semantically related, which are therefore drawn closer to each other.

We then adjust the model parameters by link prediction, as follows:

$$\mathcal{L}_{\text{intra}} = \frac{1}{V^2}\sum_{i=1}^{V}\sum_{j=1}^{V}\left[\cos(w_i^*, w_j^*) - \hat{\mathbf{A}}^{\mathbf{adj}}\right]^2. \quad (15)$$

This path implicitly utilizes W^A and $\mathcal{G}$ to form the final space $W^* = [w_1^*, w_2^*, \ldots, w_V^*]$ mentioned in p_{intra}.

4.2.4 Inter-topic Path

In order to leverage the guiding role of anchor words in the hierarchical structure between topics, we fuse each topic embedding and its anchor word embedding in a simple way to form a new fused topic embedding, which is used to constrain the path of *topic+anchor word→higher-layer topic* (Fig. 1(e)). The fusion topic embedding $t_i^{*(l)}$ of i^{th} topic at layer l can be achieved by:

$$t_i^{*(l)} = \sigma(\mathbf{w}[t_i^{(l)}; A_i^{(l)}] + b), \quad (16)$$

where $\mathbf{w} \in \mathbb{R}^{2r \times r}$ and $b \in \mathbb{R}^r$ are the weights vector and bias, respectively, and $[\cdot;\cdot]$ represents the concatenation. The size of embedding vector remains unchanged after the fusion.

Similar to the Section 4.2.1, we use the EM algorithm to constrain hierarchical relationships between topics.

E-Step. We have the relationship between i^{th} topic of layer $l+1$ and j^{th} topic of layer l as:

$$p(t_j^{(l+1)}|t_i^{(l)}) = \Pi_{i,j}^{(l)}, i \in K^{(l)}, j \in K^{(l+1)}, \quad (17)$$

where $\Pi^{(l)} \in \mathbb{R}^{K^{(l)} \times K^{(l+1)}}$ is the matrix of learnable parameters that represents the topic hierarchy between layers l and $l+1$. Then we compute a posteriori estimation under p_{inter} as:

$$q(t_j^{(l+1)}|t_i^{*(l)}) = \frac{\exp\left(t_i^{*(l)} \cdot t_j^{(l+1)T}\right)}{\sum_{j'}^{K^{(l+1)}} \exp\left(t_i^{*(l)} \cdot t_{j'}^{(l+1)T}\right)}. (18)$$

M-Step. We update the model parameters to encourage the topic hierarchy to utilize fine-grained anchor word information, which is equivalent to minimizing the following cross-entropy loss:

$$\mathcal{L}_{\text{inter}} = -\sum_{l=1}^{L-1} \sum_{j=1}^{K^{(l+1)}} \sum_{i=1}^{K^{(l)}} q\left(t_j^{(l+1)} \middle| t_i^{*(l)}\right) \times \log p\left(t_j^{(l+1)} \middle| t_i{}^{(l)}\right). \quad (19)$$

In the same way as p_{inter}, we can obtain a reasonable hierarchy Π through the EM algorithm.

4.3 Anchor-based Contrastive Learning

In this section, in order to accurately detect anchor words, we perform mask operations on anchor words in W^A during inference as negative sampling, as shown in Fig. 1(f). We introduce this anchor-based contrastive learning to improve the embedding capability of the encoder.

4.3.1 Sampling Strategy

Negative Sampling We first obtain anchor words $A^{(l)}$ in Section 4.2.2, and then update the anchor word set by $S_A = \bigcup_{l \in L} A^{(l)}$. We believe that the anchor words are highly relevant to the semantics of the topic. Then, the mask operation is defined as removing words from S_A in the original document d to obtain a negative sample d^-. We enforce d^- to decouple the anchor word factor from d.

Positive Sampling Similar to the negative sampling strategy mentioned above, we obtain a set S_P of words with minimal relevance to all topics based on Φ^A. By removing the words in this set S_P from the original document d, we obtain d^+ as a positive sample of d. We believe that d^+ retains the salient topics from d.

4.3.2 Contrastive Learning

Let the document-topic distribution in layer l obtained from d, d^- and d^+ be $z^{(l)}, z^{(l)-}$ and $z^{(l)+}$. We then calculate the contrastive loss as follows:

$$\mathcal{L}_{cl} = \sum_{l=1}^{L} \log(1 + \frac{\beta \cos(z^{(l)}, z^{(l)-})}{\cos(z^{(l)}, z^{(l)+})}), \quad (20)$$

where β controls the weight of negative samples.

It is worth noting that under our generative framework, we adaptively update the negative and positive sampling during training, which is different from previous static contrastive learning.

4.4 Joint Training

By introducing the constraints of the aforementioned four paths, the path loss function is:

$$\mathcal{L}_{\text{path}} = \mathcal{L}_{\text{clus}} + \gamma \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{cl}}, \quad (21)$$

where γ is a hyperparameter that controls the connectivity about the nodes in the graph.

Our framework can be viewed as the extensions of VAE, thus we use $\mathcal{L}_{\text{ELBO}}$ to maximize the Evidence Lower BOund (ELBO). The overall loss function of our model is:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{path}} + \mathcal{L}_{\text{ELBO}}, \quad (22)$$

where λ is a hyperparameter that controls the weighting of the original VAE and the paths we introduce. More details of the inference of the model parameters can be found in Appendix A.

5 Experiments

5.1 Experimental Settings

Datasets: Our experiments are conducted on three widely-used benchmark text datasets, varying in different sizes, including 20News (Miao et al., 2017), NIPS (Tan et al., 2017), and Wikitext-103 (Nan et al., 2019). All datasets have been processed to remove stop words and filter low frequency words by following Chen et al. (2023). The statistics of datasets are shown in Appendix B.

Baseline models: We compared our AP-HTM with the following baselines: 1) **SawETM** (Duan et al., 2021): The hierarchical topic model which introduces a sawtooth connection module to mitigate the problem of posterior collapse. 2) **HyperMiner** (Xu et al., 2022): The hierarchical topic model which exploits hyperbolic embeddings for topic and word representations. 3) **nTSNTM** (Chen

Dataset	Metric	SawETM	HyperMiner	nTSNTM	nFNTM	CluHTM	TraCo	NG-HTM	AP-HTM
NIPS	NPMI↑	0.133	0.135	0.100	0.113	0.137	0.112	0.147	**0.162**
	CLNPMI↑	0.034	0.048	0.022	0.025	0.027	**0.054**	0.028	0.053
	TU↑	0.431	0.662	0.373	0.765	0.554	0.672	0.719	**0.786**
	TQ↑	0.057	0.089	0.037	0.086	0.076	0.076	0.106	**0.128**
20News	NPMI↑	0.264	0.263	0.284	0.246	0.219	0.241	0.307	**0.329**
	CLNPMI↑	0.138	0.153	0.156	0.150	0.164	0.129	0.146	**0.174**
	TU↑	0.716	0.486	0.757	0.844	0.577	0.617	0.811	**0.857**
	TQ↑	0.189	0.128	0.215	0.208	0.126	0.149	0.249	**0.281**
Wikitext-103	NPMI↑	0.154	0.225	0.225	0.228	-	0.134	0.255	**0.295**
	CLNPMI↑	0.060	0.079	0.121	**0.147**	-	0.042	0.090	0.145
	TU↑	0.221	0.520	0.662	0.739	-	0.729	0.797	**0.825**
	TQ↑	0.034	0.117	0.149	0.168	-	0.098	0.203	**0.243**

Table 2: The performance of all hierarchical topic models, where - indicates that the model has not converged after 48 hours of training. The best results are in bold and the second best are underlined.

et al., 2021b): The tree-like topic model that introduces non-parameterization in the determination of topic numbers. 4) **nFNTM** (Zhang et al., 2022): The forest topic model which employs the self-attention mechanism to capture parent-child topic relations. 5) **CluHTM** (Viegas et al., 2020): The Directed Acyclic Graph (DAG)-structured topic model based on non-negative matrix factorization. 6) **NG-HTM** (Chen et al., 2023): A deep topic model with a Gaussian mixture prior distribution and nonlinear structural equations to capture topic relations. 7) **TraCo** (Wu et al., 2024b): A hierarchical topic model with transport plan dependency method and context-aware disentangled decoder. The training details of all methods can be found in Appendix C.

5.2 Quantitative Analysis

Interpretability of Topics: The topic hierarchy generated by the model should have the following properties: 1) a high degree of semantic consistency of individual topics, and 2) a certain degree of semantic similarity between parent and child topics. Therefore, we adopt the widely adopted Normalized Pointwise Mutual Information (NPMI) (Isonuma et al., 2020) to evaluate the interpretability of the intra-topic, and Cross-Level Normalized Point-wise Mutual Information (CLNPMI) (Chen et al., 2021b) to evaluate the subordination between parent and child topics.

As shown in Table 2, the proposed model performs significantly better than previous NHTMs on all datasets, achieving a better NPMI by a margin of 13.8%, 7.0% and 15.8% on three datasets. On the other hand, our model achieves the best CLNPMI score on Wikitext-103 as well as sub-optimal results on the other two datasets. And overall for the CLNPMI metric, our model is much better than the current optimal baseline model. It demonstrates that the hierarchical generation process of adding anchor words ensures the coherence between the parent and child topics, proving the structural rationality of topic hierarchy.

Topic Diversity: The diversity of hierarchical topics reflects the model's ability to mine the richness of the corpus for information. We adopt topic uniqueness (TU) (Nan et al., 2019) to evaluate the diversity of hierarchical topics generated. As shown in Table 2, it's evident that our AP-HTM performs the best for TU on all the datasets, which can be attributed to p_{clus} that clusters the anchor space W^A, while p_{contra} makes the encoder capture the anchor word more accurately.

Topic Quality: Intuitively, higher NPMI scores imply better correlation within topics, which may lead to increased redundancy between topics and thus lower TU scores. Conversely, most of topics with higher TU scores are marginal topics (Wu et al., 2020), which lead to lower NPMI. Therefore, in order to provide a more comprehensive insightful into overall topic quality, we use topic quality (TQ) (Dieng et al., 2020) for evaluaion. As shown in Table 2, our model is significantly higher in TQ than all baselines, indicating high quality hierarchical topics are detected.

Topic Structure Rationality: For a reasonable hierarchy, the semantics of topics at higher levels should be general, while the semantics of topics near the bottom should be more specific. Topic specialization (TS) (Kim et al., 2012) quantifies this feature by the following formula: $\mathrm{TS}(\Phi) = 1 - cos(\Phi, \Phi_{\mathrm{Norm}})$, where Φ and Φ_{Norm} denote a topic-word distribution and the word distribution of the entire corpus, respectively. A higher TS score implies that the topic is more specialized.

As shown in Fig. 6, AP-HTM achieves a reasonable pattern of topic specialisation across different datasets. Meanwhile, the relatively high TS values per layer illustrate the ability of our model to capture more unique topics at each layer.

5.3 Qualitative Analysis

Examples of Anchor Word Guidance: Our model introduces anchor words into text generation for fine-grained word information. To show the effect of anchor words in p_{intra}, we show 5 examples of top 5 topic words in Table 3 , where first and second rows are in anchor space W^A and the final embedding $W*$. We also calculated the relative increase in NPMI scores for each topic after guidance.

NPMI Increase	Label	Top 5 Words				
103%	topic:1_70	chastity christ	scripture jesus	pray bible	sin god	resurrection scriptures
86%	topic:1_65	yesterday sunday	sunday saturday	Canada thursday	friday yesterday	Canadian friday
59%	topic:1_14	interface interface	toolkit microsoft	unix unix	compiler amiga	platforms linux
50%	topic:1_111	father father	son son	woman mother	pitt woman	wife daughter
36%	topic:1_28	ordered ordered	recommended announced	indicated plans	initially announcement	plans recommended

Table 3: Comparison of topic words before and after anchor word guidance. We manually italicize and underline words that are clearly unrelated to the topic.

From both quantitative and qualitative perspectives, it can be explained that anchor word guidance leads to more consistent topics. First, the NPMI score significantly increases. Second, it can be manually observed that some words with lower relevance are replaced by more semantically consistent words, for example, *"Canada"* and *"Canadian"* in Topic:1_65, we believe that the topic of this word cluster is *time*, which obviously does not match the semantics of these two words. This may be due to that p_{clus} captures the high-frequency co-occurrence of *"Canada"* and words about *time*. However, with the guidance of p_{intra}, we adopt the synergy of anchor words and graph to ensure semantic consistency.

Visualisation of Embedding Space: The top 5 words of the 5 topics in Table 3 are visualized in Fig. 3 via t-SNE visualization (Van der Maaten and Hinton, 2008). We can see that the topics are embedded in the middle of related words, expressing certain semantic information. Besides, words under the same topic are closer, while words under different topics are farther apart. Additionally, related topics are closer in the embedding space, such as Topic: 1_111 *christianity* and Topic:1_70 *family*. It is also worth noting that words with similar semantics under different topics will approach each other, such as *"god", "jesus"* in Topic: 1_111 and *"father", "son"* in Topic:1_70. The comparison of embedding spaces generated by AP-HTM and baselines are shown in Appendix D.

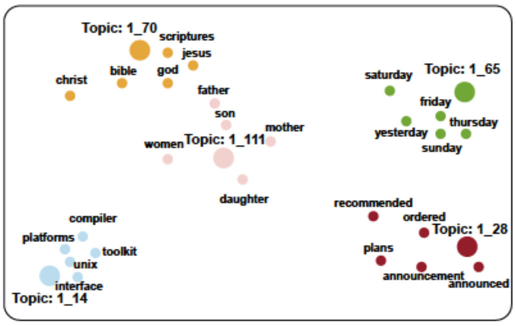


Figure 3: Visualization of word and topic embeddings, where Topic: l_i denotes the i^{th} topic at layer l.

Visualisation of Topic Structure: To demonstrate intuitively the capability of our model in generating reasonable topic structures, we visualized several topic structures extracted by AP-HTM from 20News. As shown in Fig. 4, each rectangle represents a topic and its top 10 words, with arrows from sub-topics to the most relevant topics. Consistent with the results of TS in Fig. 6, topics from root to leaves show a gradual semantic change process from general to specific. In addition, child topics are related to parent topics, e.g., *cancer* is the child of *health*, while *disk* is the child of *use*. These results mean that the semantic meaning of each topic and the connections between the topics of adjacent layers are highly interpretable, indicating that our AP-HTM can learn a reasonable topic hierarchy.

5.4 Ablation Study

We perform ablation experiments on our model to validate the effectiveness of each path. Table 4 shows the ablation results of our AP-HTM, where "Ours w/o p_{clus}, p_{intra} and p_{inter}" denotes removing the corresponding path constraints. "Ours w/o p_{contra}" means that we use the sampling strategy similar to CLNTM (Nguyen and Luu, 2021) rather than our anchor-based method.

We can see from Table 4 that each path contributes to solving two issues mentioned in Section 1. **Firstly, detecting the anchor space via clustering** is the first backbone in our framework. After removing p_{clus}, all metrics degrade largely.

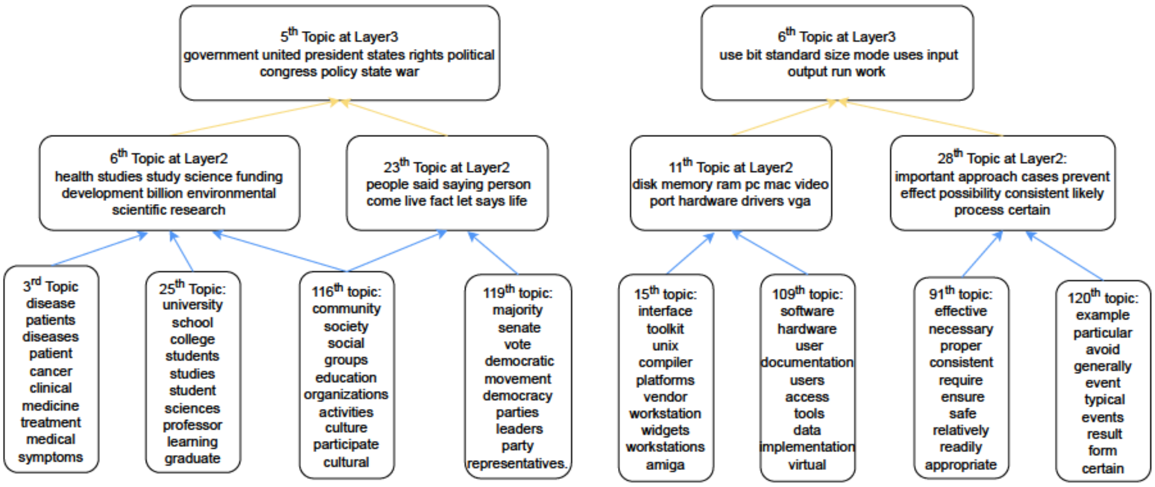

5th Topic at Layer3
government united president states rights political congress policy state war
6th Topic at Layer2
health studies study science funding development billion environmental scientific research
23th Topic at Layer2
people said saying person come live fact let says life
3rd Topic
disease patients diseases patient cancer clinical medicine treatment medical symptoms
25th Topic:
university school college students studies student sciences professor learning graduate
116th topic:
community society social groups education organizations activities culture participate cultural
119th topic:
majority senate vote democratic movement democracy parties leaders party representatives.
6th Topic at Layer3
use bit standard size mode uses input output run work
11th Topic at Layer2
disk memory ram pc mac video port hardware drivers vga
28th Topic at Layer2:
important approach cases prevent effect possibility consistent likely process certain
15th topic:
interface toolkit unix compiler platforms vendor workstation widgets workstations amiga
109th topic:
software hardware user documentation users access tools data implementation virtual
91th topic:
effective necessary proper consistent require ensure safe relatively readily appropriate
120th topic:
example particular avoid generally event typical events result form certain

Figure 4: An example of hierarchical topics learned from 20News by AP-HTM.

Secondly, the intra-topic and inter-topic path constraints also lead to semantically consistent topic words and reasonable hierarchical structures. After removing p_{intra}, although TU increases on 20News and NIPS, it will lead to a significant decline in NPMI, thereby affecting the overall topic quality TQ. Besides, removing p_{inter} leads to a decrease in all metrics on 20News and NIPS. While TU and TQ increase slightly on Wikitext-103, both CLNPMI and NPMI are affected and decrease. Finally, all metrics for "Ours w/o p_{contra}" declined to some extent, indicating that our **anchor-based contrastive learning method** can effectively enhance the embedding ability of encoder and fully utilize the fine-grained information brought by anchor words. In summary, all causal paths of the AP-HTM framework are reasonable and effective.

Datasets	Model	NPMI↑	TU↑	CLNPMI↑	TQ↑
NIPS	Ours	**0.162**	0.786	**0.053**	**0.128**
	Ours w/o p_{clus}	0.160	0.733	0.048	0.117
	Ours w/o p_{intra}	0.139	**0.803**	0.045	0.104
	Ours w/o p_{inter}	0.158	0.773	0.049	0.122
	Ours w/o p_{contra}	0.150	0.738	0.050	0.110
20News	Ours	**0.329**	0.858	**0.175**	**0.281**
	Ours w/o p_{clus}	0.320	0.781	0.155	0.250
	Ours w/o p_{intra}	0.278	**0.904**	0.159	0.251
	Ours w/o p_{inter}	0.302	0.873	0.165	0.263
	Ours w/o p_{contra}	0.325	0.848	0.158	0.276
Wikitext-103	Ours	**0.295**	0.825	**0.145**	0.243
	Ours w/o p_{clus}	0.293	0.817	0.118	0.239
	Ours w/o p_{intra}	0.244	0.824	0.116	0.201
	Ours w/o p_{inter}	0.289	**0.849**	0.139	**0.245**
	Ours w/o p_{contra}	0.294	0.819	0.134	0.241

Table 4: Results of ablation evaluation on all datasets.

6 Conclusion

In this paper, we propose a neural hierarchical topic model AP-HTM based on anchor words. Unlike the popular prior models, our approach adds the concept of anchor words to the text generation process. We obtain the anchor space by clustering and propose three causal paths guided by anchor words to extend the VAE framework. In addition, we introduce a novel anchor-based contrastive learning to decouple the roles of anchor words in the paths, thus endowing the model with stronger anchor word and topic representation. Extensive experiments show that our framework outperforms state-of-the-art baselines in extracting coherent, unique, and rationally structured topics.

Acknowledgement

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) and NSF grant CNS-2349369.

Limitations

Our framework is only a small step towards mining a comprehensive and high-quality topic hierarchy, and there are two limitations to be explored for future works: 1) After obtaining anchor words about the topic, some external prior information, such as ConceptNet (Speer et al., 2017), can be introduced to further guide the topic model incorporating human knowledge. 2) Metadata (e.g., author, tags, and sentiment) from documents in a corpus can be combined with anchor words of topics to assist in document relations modeling and analyze the relationships between topics in the document.

References

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2): 1–30.

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

Hegang Chen, Pengbo Mao, Yuyin Lu, and Yanghui Rao. 2023. Nonlinear structural equation model guided gaussian mixture hierarchical topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 10377–10390.

Ziye Chen, Cheng Ding, Yanghui Rao, Haoran Xie, Xiaohui Tao, Gary Cheng, and Fu Lee Wang. 2021a. Hierarchical neural topic modeling with manifold regularization. *World Wide Web*, 24: 2139–2160.

Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. 2021b. Tree-structured topic modeling with nonparametric neural variational inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2343–2353.

Daniel. D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, pages 556-562.

Adji B. Dieng, Francisco JR. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8: 439–453.

Nat Dilokthanakul, Pedro A.M. Mediano, Marta Garnelo, Matthew C.H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

David Donoho and Victoria Stodden. 2003. When does non-negative matrix factorization give a correct decomposition into parts? *In Advances in Neural Information Processing Systems*, pages 1141-1148.

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

Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5: 529–542.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. *In Advances in Neural Information Processing Systems*, pages 17-24.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-structured neural topic model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 800–806.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

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 Journal of Biomedical and 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 *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 783–792.

Diederik P. Kingma and Max Welling. 2013. Autoencoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Yewen Li, Chaojie Wang, Zhibin Duan, Dongsheng Wang, Bo Chen, Bo An, and Mingyuan Zhou. 2022. Alleviating "posterior collapse" in deep topic models via policy gradient. *Advances in Neural Information Processing Systems*, pages 22562–22575.

Yang Lin, Xin Gao, Xu Chu, Yasha Wang, Junfeng Zhao, and Chao Chen. 2023. Enhancing neural topic model with multi-level supervisions from seed words. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13361–13377.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, pages 3143–3152.

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

David Mimno and Moontae Lee. 2014. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1328.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 633–640.

Zhao-Yan Ming, Kai Wang, and Tat-Seng Chua. 2010. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–9.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*.

Thang Nguyen, Yuening Hu, and Jordan Boyd-Graber. 2014. Anchors regularized: Adding robustness and extensibility to scalable topic-modeling algorithms. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 359–369.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *In Advances in Neural Information Processing Systems*, pages 11974–11986.

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2014. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 256–270.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Adler Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent dirichlet allocation. *Advances in Neural Information Processing Systems*, pages 2609-2617.

Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine Learning*, 88: 157–208.

Simra Shahid, Tanay Anand, Nikitha Srikanth, Sumit Bhatia, Balaji Krishnamurthy, and Nikaash Puri. 2023. Hyhtm: Hyperbolic geometry based hierarchical topic models. *arXiv preprint arXiv:2305.09258*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Chenhao Tan, Dallas Card, and Noah A. Smith. 2017. Friendships, rivalries, and trysts: Characterizing relations between ideas in texts. *arXiv preprint arXiv:1704.07828*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11): 2579–2605.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves. 2020. Cluhtm-semantic hierarchical topic modeling based on cluwords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8138–8150.

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

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1772–1782.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, 57(2): 1–30.

Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024b. On the affinity, rationality, and diversity of hierarchical topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17, pages 19261–19269.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487.

Yi Xu, Dongsheng Wang, Bo Chen, Ruiying Lu, Zhibin Duan, Mingyuan Zhou, et al. 2022. Hyperminer: Topic taxonomy mining with hyperbolic embedding. *In Advances in Neural Information Processing Systems*, pages 31557–31570.

Zhihong Zhang, Xuewen Zhang, and Yanghui Rao. 2022. Nonparametric forest-structured neural topic modeling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2585–2597.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical methods in Natural Language Processing*, pages 4663–4672.

A Parameter Inference Algorithm

We apply NVI to network parameters, which is efficient and flexibility (Srivastava and Sutton, 2017). Similar to VAEs, one of the training obective of our model is to maximize the ELBO, and the corresponding loss $\mathcal{L}_{ELBO}$ is given below:

$$\mathcal{L}_{ELBO} = \sum_{i=1}^{L} \mathbb{E}_{q(z^i, \Phi^i, c|x)} \left[\log p \left(\hat{x} | z^i, \Phi^i \right) \right] + \\ - D_{KL} \left[q \left(z^L, c \mid x \right) \| p \left(z^L, c \right) \right]. \quad (23)$$

```
**Algorithm 1:** Parameter Inference Algorithm
**Input:** The word embedding $W$ from a pre-trained model;
**Output:** Topic-word distribution $\Phi^*$, topic hierarchy $\Pi$.
Initialize $\Pi$ and topic embeddings $T$;
Construct the word co-occurrence graph $\mathcal{G}$ and compute $\hat{\mathbf{A}}^{\mathbf{adj}}$ by Eq. (13);
**repeat**
    Estimate $z$ using $Enc$;
    $\Phi^A, W^A, \mathcal{L}_{clus} \leftarrow W, T$ by Eqs. (10) and (11);
    Obtain $A$ by Eq. (12);
    $\Phi^*, W^*, \mathcal{L}_{intra} \leftarrow W^A, \hat{R}$ by Eq. (15);
    $\Pi, T^*, \mathcal{L}_{inter} \leftarrow \Pi, A, T$ by Eq. (19);
    Mask words in $S_P, S_A$ and estimate $z^+$, $z^-$ using $Enc$;
    $\mathcal{L}_{cl} \leftarrow z, z^+, z^-$ by Eq. (20) ;
    $\hat{x} \leftarrow z, \Phi^A$ using $Dec$;
    Compute ELBO by Eq. (23);
    $\Phi^* \leftarrow W^*, T$ by Eq. (6);
    Update $\Pi, T, T^*, A, S_P, S_A, W_A, W^*, \Phi^A, \Phi^*$ and $Enc, Dec$ by Eq. (22) ;
**until** Convergence
```

The parameter inference method for AP-HTM is presented in Algorithm 1. We use the variational lower-bound to calculate gradients and apply RMSprop to update parameters.

B Datasets

Statistics about the datasets employed in this paper are shown in Table 5.

Dataset	Docs(Train)	Docs(Test)	Vocabulary size
20News	11314	7531	3997
NIPS	1350	149	3531
Wikitext-103	28472	120	20000

Table 5: Basic dataset statistics.

C Training Details & Hyperparameters

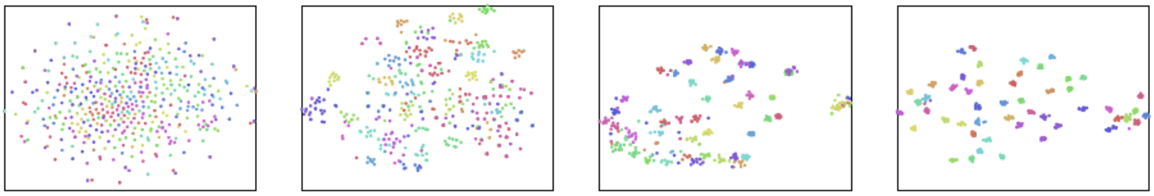
AP-HTM is implemented via PyTorch. For the embedding-based topic models including SawETM, nTSNTM, nFNTM, CluHTM, HyperMiner, NG-HTM, TraCo, and AP-HTM, we leverage the pre-trained GloVe model (Pennington et al., 2014) to obtain the initialization for word embeddings W. All experiments were conducted with model codes available in public, trained for a single run, and on a workstation equipped with an Nvidia RTX 1080-Ti GPU and a Python environment with 128G memory.

For all these models, the max-depth of topic hierarchy is set to 3 by following (Isonuma et al., 2020). To better compare parametric and nonparametric topic models, we follow (Chen et al., 2021b) to use the best hyperparameters reported in the original papers. For nonparametric models (i.e. CluHTM, nTSNTM, and nFNTM), we set the number of topics to 200. For the parametric hierarchical topic models (i.e., SawETM, HyperMiner, NG-HTM, TraCo, and AP-HTM), the topic numbers of different layers $k^{(1)}$, $k^{(2)}$, and $k^{(3)}$ are set as 8, 32, and 128. For AP-HTM, we set the weight parameter b to 0.02, 0.05, 0.006 for NIPS, 20News and Wikitext-103, λ and γ are set to 10 and 0.1, 10 and 1, 10 and 1 for NIPS, 20News, and Wikitext-103. β is set to 0.5. The optimisation of AP-HTM is achieved by RMSprop with a learning rate of 5e-3 and batch size of 512. It is worth mentioning that for all the metrics except topic specialization (Kim et al., 2012), we calculate the average score for the 5, 10, and 15 top words.

D Comparison of Embedding Spaces

By comparing the strong baselines of TraCo, hyperMiner, NG-HTM, and AP-HTM, we can clearly conclude the superiority of AP-HTM in obtaining a reasonable word embedding space. The above four models are embedded by different types of embedding assumptions. TraCo learns the relationship between topics and words through Euclidean space, HyperMiner embeds in hyperbolic space, NG-HTM and our model obtain topic-word distribution Φ through inner product (The visual distance metric obtained by inner product s set to the cosine distance).

As shown in Fig. 5, it is evident that in the space obtained by our AP-HTM, words are more tightly embedded within the topic, and the distance between topics is also more reasonable. In summary,



(a) TraCo (b) HyperMiner (c) NG-HTM (d) AP-HTM

Figure 5: Visualization of the embedding space for (a) TraCo, (b) HyperMiner, (c) NG-HTM, and (d) AP-HTM. We randomly select 50 topics from 3_{rd} layer from 20News, each consisting of top 10 words assigned different colors.

it is well illustrated that the final space W^* supports coherent and diverse topics.

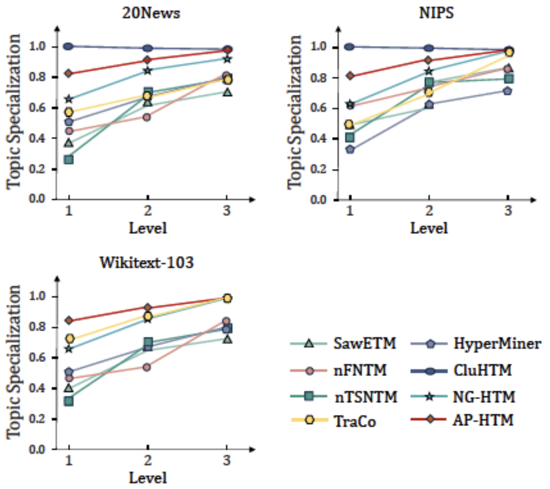


Figure 6: Topic specialization of different topic structures generated on all datasets.

E Intrusion Task Evaluation

We manually conduct an intrusion task for topic word coherency on 20News for three strong baselines and our AP-HTM, and the results are shown in Table 6.

Metric	HyperMiner	TraCo	NGHTM	AP-HTM
coherency	56.9%	58.7%	65.5%	**79.1%**

Table 6: The intrusion task on 20News.

The experimental results on the topic word coherency appear similar to those on the NPMI metric, i.e., our model is much stronger than the most recent strong baselines and 20.7% higher than the suboptimal model (i.e., NGHTM). This experiment with manual metric combined with the experiment with automated metric lead to the conclusion that our model outperforms the existing baseline in terms of extracting topics that are coherent, distinctive, and rationally structured.

Details of our intrusion task are given below:

We randomly select 10 students as volunteers to participate in the experiment, where the model generates top k words for all topics in 20News as two parts. The first part is top 5 words and the second part is top 1 words. Besides, the second part is randomly mixed into the first part as intrusive words. We randomly select the topic words after intrusion in each layer according to the ratio of 1/8 (the number of each evaluation is (8+32+128)/8=21). We then ask the human evaluators to identify the least relevant words as intrusion words and calculate the correct recognition rate of the intrusion instances as the final intrusion metric.

F Runtime and Parameter Size

The runtime and parameter size of different models on 20News are shown in Table 7.

Metric	SawETM	nFNTM	nTSNTM	HyperMiner	TraCo	NGHTM	AP-HTM
Runtime	5.2s	3.3s	38.6s	4.4s	87.3s	3.8s	15.4s
#Params	1.9M	1.2M	0.5M	2.2M	2.2M	1.5M	2.8M

Table 7: Runtime and parameter size on 20News.

Due to the addition of several causal paths on anchor words, the number of parameters and runtime increased. However, the overall computation time and parameter size are still acceptable.