Nonlinear Structural Equation Model Guided Gaussian Mixture Hierarchical Topic Modeling

Hegang Chen and **Pengbo Mao** and **Yuyin Lu** and **Yanghui Rao***
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
{chenhg25,maopb,luyy37}@mail2.sysu.edu.cn, raoyangh@mail.sysu.edu.cn

Abstract

Hierarchical topic models, which can extract semantically meaningful topics from a text corpus in an unsupervised manner and automatically organise them into a topic hierarchy, have been widely used to discover the underlying semantic structure of documents. However, the existing models often assume in the prior that the topic hierarchy is a tree structure, ignoring symmetrical dependencies between topics at the same level. Moreover, the sparsity of text data often complicate the analysis. To address these issues, we propose NSEM-GMHTM as a deep topic model, with a Gaussian mixture prior distribution to improve the model's ability to adapt to sparse data, which explicitly models hierarchical and symmetric relations between topics through the dependency matrices and nonlinear structural equations. Experiments on widely used datasets show that our NSEM-GMHTM generates more coherent topics and a more rational topic structure when compared to state-of-the-art baselines. Our code is available at https://github.com/nbnbhwyy/NSEM-GMHTM.

1 Introduction

Topic models, which can uncover the hidden semantic structure in a text corpus, have been widely applied to text analysis. Specifically, a topic model aims to discover a set of semantically meaningful topics from a document set. Each topic captures a common pattern of word co-occurrences in the document and is often interpreted semantically as a coherent set of words representing a common concept. Although traditional topic models like the Latent Dirichlet Allocation (LDA) (Blei et al., 2003b) and the Embedded Topic Model (ETM) (Dieng et al., 2020) are able to achieve this goal, they assume that topics are independent, which limits the ability of these models to explore the topic structure. To remedy the defect, a series of hierarchical extensions, such as the hierarchical LDA (hLDA) (Blei et al., 2003a), the recursive Chinese Restaurant Process (rCRP) (Kim et al., 2012), and the nested Hierarchical Dirichlet Process (nHDP) (Paisley et al., 2015), have been proposed. Commonly, these models learn hierarchical topics in tree structures, which assume that the topics in the upper layers are more general/abstract than those in the lower layers. Consequently, revealing hierarchical relations between topics provides the user an intuitive way to better understand text data. However, these methods rely on approximate approaches (e.g., variational inference and Gibbs sampling) and require complex derivation or high computational costs to estimate parameters.

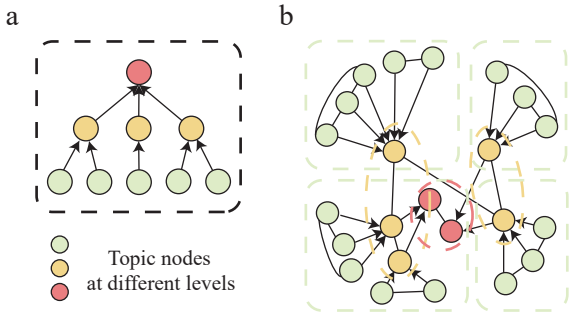

Figure 1: Illustration of (a) topics discovered by previous hierarchical topic models, and (b) topics found by our hierarchical topic model guided by nonlinear structural equations.

With the development of deep neural networks and the proposal of Neural Variational Inference (NVI), there is a growing interest in developing Neural Hierarchical Topic Models (NHTMs) due to their fast parameter inference and flexibility (Isonuma et al., 2020; Chen et al., 2021; Duan et al., 2021; Xu et al., 2022). Generally, NHTMs are based on Variational Auto-Encoder (VAE) and model topic hierarchy as the relationship between neurons at different levels in the encoder or decoder, such as the Tree-Structured Neural Topic

*The corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10377–10390

Model (TSNTM) (Isonuma et al., 2020) and the nonparametric TSNTM (nTSNTM) (Chen et al., 2021). However, most NHTMs rely on a single isotropic multivariable Gaussian prior distribution, which often fails to well approximate the posterior distributions of sparse data (Xiong et al., 2019). A tighter estimation of the posterior distribution could greatly improve the power of VAE in fitting and analyzing sparse data. Moving beyond topic mining, the application of Gaussian Mixture Model (GMM) as a priori for latent variables has recently shown promising performance in the fields of image generation and bioinformatics (Xiong et al., 2019; Yang et al., 2019).

We further note that previous NHTMs have focused only on relationships between topics at different levels. For example, Chen et al. (2021) built a topic tree bottom-up through a dependency matrix, where the parent topic can be considered as a generalization of its child topics. However, the generation of high level topics may also be influenced by the structure between topics at lower levels. For example, in the case of modules in biochemical networks or communities in social networks, information cross-talk between nodes at the same level plays a crucial role in the extraction of higher level abstraction modules (Clauset et al., 2008). Furthermore, returning to the nature of topics, they can be thought of as words with highly generalised semantics. Intuitively, as defined by Speer et al. (2017), not only are there hierarchical relations between topics with different levels of generalisation like Chicago and city (ISA), but there should also be symmetrical relations that belong to the same level, such as cut and knife (CapableOf) and learned and learned (RelatedTo). Unfortunately, existing NHTMs tend to predefine topics as tree structures, focusing only on modelling topic hierarchy relationships and neglecting symmetrical relationships between topics that may also help researchers better understand and process textual information. Furthermore, the use of topic symmetric structures to help models better capture document semantics has not been much explored. In addition, some works (Liu et al., 2018; Viegas et al., 2020) generate a document via Directed Acyclic Graph (DAG) structured topics, but the structure of their generated topics is often unclear.

To overcome these limitations, in this paper, we propose the Nonlinear Structural Equation Model guided Gaussian Mixture Hierarchical Topic Model (NSEM-GMHTM), a deep generative model of documents. As shown in Figure 1, in contrast to the previous hierarchical topic models, the core idea is to apply a Nonlinear Structural Equation Model (NSEM) to explicitly construct the symmetric dependencies between topics to facilitate the extraction of a more comprehensive and clear topic structure. In particular, we introduce gaussian mixture distribution as a prior for latent variables, enabling the network to learn more complex distributions, and further improving the power of VAE in fitting and analyzing sparse data. Experiments show that our model outperforms state-of-the-art baselines on several widely adopted metrics, validating the rationality of topic structure generated by our model. Futhermore, ablation studies and extensive qualitative evaluations have shown that NSEM guided NHTM results in a better topic structure, which further demonstrates the validity of our method.

2 Related Work

Following the pioneering work on topic models (Blei et al., 2003b), several extension models, such as hLDA (Blei et al., 2003a) and rCRP (Kim et al., 2012) have been proposed to explore the relationships between topics. Although these models showed clear competitiveness in hierarchical topic modeling, they are limited by the expensive iterative inference step, which is not conducive to further model expansion (Ranganath et al., 2014).

NVI-based topic models (Miao et al., 2017; Ding et al., 2018; Srivastava and Sutton, 2017) commonly converted a document to a Bag-of-Words (BoW) representation determined on the frequency count of each vocabulary token in the document. The BoW input was processed through an MLP followed by variational inference which sampled a latent document-topic vector. A decoder network then reconstructed the original BoW using the latent document-topic vector via a topic-word distribution. Building hierarchical topic models based on NVI is a promising direction due to the fast parameter inference and flexibility. Isonuma et al. (2020) proposed a tree-structured neural topic model, which applied doubly-recurrent neural networks to parameterize topic distributions over a tree. Chen et al. (2021) developed a tree-structured topic model by using nonparametric NVI, which first learned the potential components of the stick-breaking process for each document and then modelled the affiliation of components through depen-

dency matrices between network layers.

Besides tree-structured topic models, several works proposed to generate a document by a DAG structured topic hierarchy. For instance, Mimno et al. (2007) proposed the hierarchical Pachinko Allocation Model (PAM) by connecting the root topic to lower-level topics through multinomial distributions. Liu et al. (2018) and Viegas et al. (2020) applied Nonnegative Matrix Factorization (NMF) to generate hierarchical topics in a DAG structure. Although the aforementioned DAG-based approaches captured all the relations between topics, the generated topic structures of them were not clear compared to those of the tree-structured topic models. In turn, the tree-structured topic models ignored the symmetric relations between topics at the same level. Unlike previous approaches, our approach uses NSEM and dependency matrices to capture both symmetric and hierarchical dependencies between topics, which helps to further clarify the structure of topics.

Recently, some works attempted to use other prior distributions. For neural topic models, Wu et al. (2020) combined the mixed counting models and variational inference to develop the Negative Binomial Neural Topic Model (NB-NTM) and the Gamma Negative Binomial Neural Topic Model (GNB-NTM). For HNTMs, Duan et al. (2021) proposed SawETM, which used a Weibull prior to model sparse and nonnegative documents, and mitigated the problem of posterior collapse to some extent with a Sawtooth Connection module. Xu et al. (2022) built on an existing method (Duan et al., 2021) by proposing to embed topics and words into a hyperbolic space, which enhanced the model's ability to mine the implicit semantic hierarchy. For a more comprehensive comparison of the models, we used these models as baselines for our work.

3 The Proposed Model

In this section, we propose NSEM-GMHTM for text analysis, which aims at exploring a topic structure. The motivation for designing NSEM-GMHTM focuses on tackling two main challenges: (i) How to clearly construct topic symmetric and hierarchical dependencies; (ii) How to design expressive neural networks to improve the ability of models to adapt and analyse sparse data. Below, we firstly introduce the details of the related technology, and then describe the decoder and encoder of NSEM-GMHTM as shown in Figure 2. Finally, we provide details of model inference.

3.1 Gaussian Mixture VAE

Variational inference has the potential to transform intractable inference problems into solvable optimization problems (Wainwright et al., 2008), and thus expands the set of available tools for inference to include optimization techniques as well. Despite this, a key limitation of classical variational inference is the need for the likelihood and the prior to be conjugate in order for most problems to be tractably optimized, which in turn limits the applicability of such algorithms.

VAE is the result of a combination of variational inference with the flexibility and scalability offered by neural networks (Kingma and Welling, 2014; Rezende et al., 2014), which uses neural networks to output the conditional posterior and thus allows the variational inference objective to be tractably optimized via stochastic gradient descent. Such a framework learns the distribution of input data well, enabling it to combine with the traditional probabilistic graphical models (e.g., LDA) and infer model parameters quickly (Srivastava and Sutton, 2017). However, the standard VAE uses a single isotropic multivariable Gaussian prior distribution over the latent variables and often underfits sparse data (Xiong et al., 2019).

Applying GMM as the prior over the latent variables has been used in unsupervised learning for generating more disentangled and interpretable latent representations. Following Dilokthanakul et al. (2016), it can be modeled with a joint distribution $p(\mathbf{x}, \mathbf{z}, c)$, and the joint probability can be factorized as follows:

$$p(\mathbf{x}, \mathbf{z}, c) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z} \mid c)p(c) \quad (1)$$

$$c \sim Mult(\boldsymbol{\pi}) \quad (2)$$

$$\mathbf{z} \mid c \sim \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_{c_k}, \boldsymbol{\sigma}_{c_k}^2 \mathbf{I}\right)^{c_k} \quad (3)$$

$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{z}), \boldsymbol{\sigma}^2(\boldsymbol{z})\right) \text{ or } \mathcal{B}\left(\boldsymbol{\mu}(\boldsymbol{z})\right) \quad (4)$$

where K is a predefined number of components in the mixture, $\mathbf{x}$ is the input variable, $\mathbf{z}$ is the latent variable, and the one-hot vector c is sampled from the mixing probability $\boldsymbol{\pi}$, which chooses one component from the Gaussian mixture.

3.2 Nonlinear Structural Equation Model

Structural Equation Model (SEM) is a multivariate statistical model to analyze structural relationships

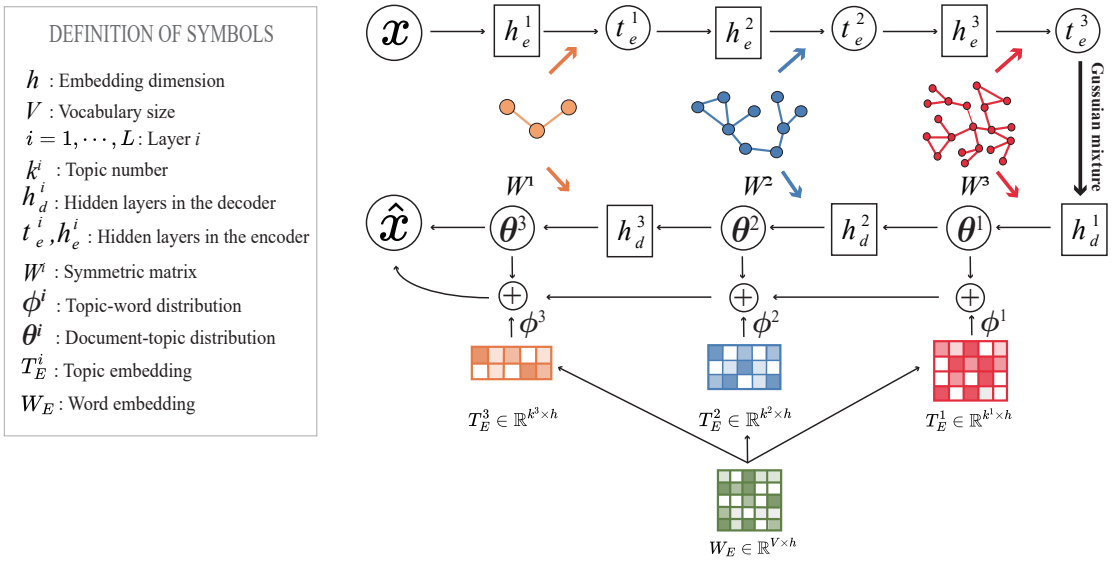

Figure 2: The workflow of NSEM-GMHTM with 3 layers.

among different random variables. The basic SEM was first developed to model the covariance matrix for random variables (Bollen, 1989). Later, SEM was found to be very powerful in modeling the relationship between observed features and hidden latent variables and was widely used in econometrics and sociology for causal inference (Goldberger, 1972; Luo et al., 2020). More importantly, SEM can be adopted to detect the conditional dependency among random variables and therefore also used to predict the graph structure of Bayesian networks and Markov random fields (Yu et al., 2019). Let $W \in \mathbb{R}^{D \times D}$ be the weighted adjacency matrix of D variables (nodes) and $X \in \mathbb{R}^{D \times h}$ be a sample of a joint distribution of D variables over h features, where each row corresponds to one variable. The linear SEM model reads:

$$X = W^T X + Z \tag{5}$$

where $Z \in \mathbb{R}^{D \times h}$ stands for a noise matrix following a Gaussian distribution. By combining traditional linear SEM with deep learning capable of capturing complex nonlinear mappings, a nonlinear version of SEM (i.e., NSEM) was proposed by Yu et al. (2019). It can be defined as follows:

$$X = f_1\left(\left(\mathbf{I} - W^T\right)^{-1} Z\right) \tag{6}$$

$$Z = \left(\mathbf{I} - W^T\right) f_2(X) \tag{7}$$

where $\mathbf{I}$ denotes the identity matrix. $f_1(\cdot)$ and $f_2(\cdot)$ stand for multilayer neural networks. By extending NSEM to bioinformatics, Shu et al. (2021) successfully predicted regulatory relationships between genes, which proved that it could help models capture symmetric dependencies between topics.

3.3 Modeling Process

Inspired by previous works, we introduce GMM as a prior for latent variables, enabling the network to learn more complex distributions while improving the model's ability to fit and analyze sparse data. Additionally, to explore a more comprehensive topic structure from a collection of documents, NSEM-GMHTM extends the NSEM proposed by Yu et al. (2019) to capture symmetric dependencies between topics at the same level. The details of our model are described in the following.

Document encoder: Given a collection of documents, we process each document into a Bag-of-Words (BoW) vector $\boldsymbol{x}_{bow} \in \mathbb{R}^V$, where V is the vocabulary size. Following the definition of Dilokthanakul et al. (2016), the Gaussian mixture encoder network can be described as follows:

$$h_e^1 = f_1(\boldsymbol{x}_{bow}) \tag{8}$$

$$t_e^i = \left(\left(\mathbf{I} - |W^i|^T\right)\left(h_e^i\right)^T\right)^T \tag{9}$$

$$h_e^{i+1} = \tanh\left(f_2\left(t_e^i\right)\right) \tag{10}$$

$$c = \text{Gumbel Softmax}(t_e^L) \tag{11}$$

$$t_d^1 = \text{Reparameter}\left(t_e^L, c\right) \tag{12}$$

where the Gumbel Softmax layer produces a K-dimensional label. Its i_{th} dimension contains the probability that the input vector belonging to the i_{th} Gaussian mixture component. During training, this set of probabilities is gradually enforced to be

concentrated on one component (Jang et al., 2017). Following Dilokthanakul et al. (2016), the number of mixture components K is set to 10. For each layer of topics, we train both $(\mathbf{I} - |W|^T)$ and $\left(\mathbf{I} - |W|^T\right)^{-1}$ in the encoder and decoder to capture the symmetric relations of topics, helping the model better understand the implicit semantic structure of the corpus. It is worth noting that h_e^i and t_e^i denote hidden features without and with the integration of the symmetric relation, respectively.

Document decoder: Considering the generative model of NSEM-GMHTM with L layers, from bottom to top, the document decoder can be expressed as follows:

$$h_d^i = \left(\left(\mathbf{I} - |W^i|^T\right)^{-1}\left(t_d^i\right)^T\right)^T \quad (13)$$

$$t_d^{i+1} = \tanh\left(h_d^i |M^i|\right) \quad (14)$$

$$\boldsymbol{\theta}^i = softmax\left(h_d^i\right) \quad (15)$$

$$\boldsymbol{\phi}^i = softmax\left(T_E^i \times W_E\right) \quad (16)$$

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{L} \hat{\boldsymbol{x}}^i = \sum_{i=1}^{L} \boldsymbol{\theta}^i \boldsymbol{\phi}^i \quad (17)$$

where $W^i \in \mathbb{R}^{k^i \times k^i}$ is a symmetric matrix, $M^i \in \mathbb{R}^{k^i \times k^{i+1}}$ is a dependency matrix to capture the hierarchical relationships between topics at different levels, and k^i denotes the topic number at layer i. It is worth noting that, the weights of W^i and M^i are constrained to be nonnegative to maintain interpretability as to the directionality of topic structure. We calculate topic-word distribution $\boldsymbol{\phi}^i$ by Equation (16) with topic embeddings T_E^i and word embeddings W_E. Then we reconstruct document $\hat{\boldsymbol{x}}^i$ by combining document-topic distribution $\boldsymbol{\theta}^i$ with topic-word distribution $\boldsymbol{\phi}^i$. To allow each layer to be useful by itself, we make the decoder reconstruct each layer back to an $\hat{\boldsymbol{x}}^i$. More details of the inference of the model parameters can be found in Appendix A.

4 Experiments

4.1 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 undergone data preprocessing of removing stop words and deleting low-frequency words. The statistics of datasets are listed in Table 1.

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

Table 1: The statistics of datasets.

Baselines and parameter settings: For hierarchical topic models, we adopt TSNTM (Isonuma et al., 2020)[1], CluHTM (Viegas et al., 2020)[2], SawETM (Duan et al., 2021)[3], HyperMiner[4] (Xu et al., 2022), and nTSNTM (Chen et al., 2021)[5] as our baselines. For all these models, the max-depth of topic hierarchy is set to 3 by following Isonuma et al. (2020). For nonparametric or flat topic models, we adopt HDP (Teh et al., 2004)[6] , ETM (Dieng et al., 2020)[7], NB-NTM & GNB-NTM (Wu et al., 2020)[8], and iTM-VAE & HiTM-VAE (Ning et al., 2020)[9] as baselines. HDP is a classical nonparametric topic model that allows potentially an infinite number of topics. ETM is a document generative model that combines LDA (Blei et al., 2003b) with word embeddings. It assumes that topics and words exist in the same embedding space, thus learning interpretable word and topic embeddings. For iTM-VAE & HiTM-VAE, they extended the method in Nalisnick and Smyth (2017) to introduce nonparametric processes into the NVI framework by extracting potential infinite topics.

To better compare parametric and nonparametric topic models, we follow Chen et al. (2021) to set topic numbers to 50 and 200 for all flat parametric models. For nonparametric models (i.e., HDP, iTM-VAE & HiTM-VAE, CluHTM, and nTSNTM), we use the best hyperparameters reported in the original papers. For the parametric hierarchical topic models (i.e., SawETM, HyperMiner, and NSEM-GMHTM), the topic numbers of different layers are set as 128, 32 and 8. It is worth mentioning that for all the indicators below except topic specialization (Kim et al., 2012), we calculate the average score for the 5, 10, and 15 top words. More training details of methods can be found in Appendix B.

[1] http://github.com/misonuma/tsntm
[2] http://github.com/feliperviegas/cluhtm
[3] http://github.com/BoChenGroup/SawETM
[4] https://github.com/NoviceStone/HyperMiner
[5] http://github.com/hostnlp/nTSNTM
[6] http://github.com/arnim/HDP
[7] http://github.com/adjidieng/ETM
[8] http://github.com/mxiny/NB-NTM
[9] http://github.com/walkerning/itmvae_public

Dataset	20News		NIPS		Wikitext-103	
Model	50	200	50	200	50	200
ETM	0.263	0.248	0.098	0.068	0.214	0.217
NB-NTM	0.265	0.281	0.107	0.103	0.127	0.125
GNB-NTM	0.292	0.278	0.101	0.126	0.127	0.093
HDP	0.273		0.131		0.157	
iTM-VAE	0.278		0.098		0.184	
HiTM-VAE	0.294		0.135		0.233	
SawETM	0.264		0.133		0.154	
nTSNTM	0.262		0.101		0.169	
TSNTM	0.282		0.116		0.237	
HyperMiner	0.263		0.135		0.225	
CluHTM	0.219		0.122		-	
NSEM-GMHTM	**0.307**		**0.147**		**0.255**	

Table 2: The average NPMI scores between top 5, 10, and 15 words in each topic. The higher score means better performance and the best scores are in boldface. We do not report the results of CluHTM on Wikitext-103 since it failed to achieve convergence in 48 hours.

4.2 Evaluation on Topic Interpretability

In this part, we use the widely adopted NPMI (Chen et al., 2021; Isonuma et al., 2020; Viegas et al., 2020; Bouma, 2009) to evaluate topic interpretability. As mentioned by Lau et al. (2014), NPMI is a measurement of topic coherence that is closely consistent with the ranking of topic interpretability by human annotators. 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 8.9% on 20News, 26.7% on NIPS, and 7.6% on Wikitext-103, where percentage improvements are determined over the second best NHTMs. Compared to SawETM, our NSEM-GMHTM's NPMI is on average 30.8% higher across the three datasets, presumably because SawETM only constructs topic hierarchies, which are not entirely accurate, whereas NSEM-GMHTM novelly models symmetric relationships of topics and is therefore able to capture the structural properties between topics at the same level. In addition, our method shows competitive performance compared to the best flat baselines. In particular, the NPMI of NSEM-GMHTM is improved by 8.9% compared to HiTM-VAE on the NIPS dataset.

4.3 Topic Structure Analysis

In this section, we use the evaluation metrics proposed by prior works, including topic specialization (Kim et al., 2012), cross-level normalized pointwise mutual information (CLNPMI) (Chen et al., 2021), topic uniqueness (TU) (Nan et al., 2019), and overlap rate (OR) (Chen et al., 2021) to comprehensively assess the topic hierarchy generated by NSEM-GMHTM from different perspectives and to compare it with state-of-the-art approaches. Key words of topics are ranked from the topic-word matrix ϕ (Blei et al., 2003a).

Semantic rationality of topic hierarchy: In the real world, higher-level topics are generalized representations of their lower-level counterparts, for example, *basketball* and *football* can be subsumed within the larger topic of *sport*. In other words, the semantics of topics at higher levels should be more general, while the ones close to the bottom should be more specific. Topic specialization (Kim et al., 2012) quantifies this feature by calculating the cosine distance of the word distribution between each topic and the entire corpus. A higher specialization score implies that the topic is more specialized. Therefore, we adopt topic specialization as an indicator for evaluating semantic rationality of topic hierarchy. Figure 3 illustrates the topic specialization scores of all hierarchical topic models at each level. The results show that NSEM-GMHTM achieves a reasonable pattern of topic specialisation across different datasets, i.e., the scores get lower while the levels get deeper. Oppositely, CluHTM gets topic specialization scores close to 1 at all levels on 20News and NIPS datasets, which indicates unreasonable topic hierarchies.

Furthermore, when the model is insufficient to capture the complex underlying topic structure within the corpus, it tends to undergo mode collapse, which generates topics that are particularly similar. We, therefore, measure the semantic redundancy of the topic hierarchy with the widely-used topic uniqueness (TU) (Nan et al., 2019), which is calculated as follows:

$$\mathrm{TU}(k) = \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{1}{\mathrm{cnt}(n,k)} \qquad (18)$$

where K is the number of topics and $\mathrm{cnt}(n,k)$ is the total number of times the n_{th} top word in the k_{th} topic appears in the top N words across all topics. The results in Table 3 indicate that our model significantly outperforms the baselines. In summary, these results show a clear hierarchy and low redundancy in the semantics of topics generated by NSEM-GMHTM, which demonstrates the semantic rationalization of topic hierarchy.

Structural rationality of topic hierarchy: As mentioned by Viegas et al. (2020), a reasonable

Dataset	Metric	SawETM	CluHTM	TSNTM	nTSNTM	HyperMiner	NSEM-GMHTM
Wikitext-103	CLNPMI↑	0.060	-	0.086	**0.113**	0.079	0.090
	TU↑	0.221	-	0.615	0.730	0.520	**0.797**
	OR↓	0.064	-	0.078	0.080	0.162	**0.017**
20News	CLNPMI↑	0.138	0.123	0.109	0.144	0.143	**0.146**
	TU↑	0.716	0.577	0.430	0.683	0.388	**0.811**
	OR↓	0.064	0.332	0.052	0.030	0.143	**0.011**
NIPS	CLNPMI↑	0.034	0.098	0.113	0.022	**0.048**	0.028
	TU↑	0.431	0.285	0.116	0.373	0.662	**0.719**
	OR↓	0.071	0.447	0.078	0.063	0.135	**0.025**

Table 3: The CLNPMI, TU, and OR scores of all hierarchical topic models, where - indicates that the results could not be obtained within 48 hours.

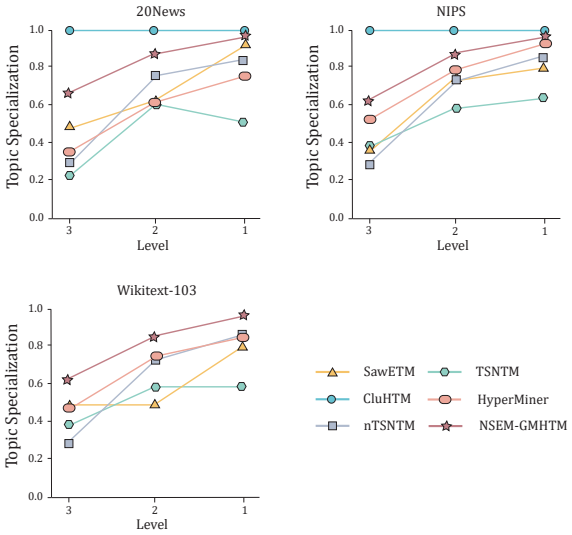

Figure 3: Topic specialization of different hierarchical topic models at each level.

topic structure also indicates that child topics are coherent with their corresponding parent topics. However, it is also inconsistent with the assumption of a topic hierarchy if the parent and child topics are too similar. Therefore, to measure the relationship between parent and a child topics, we use CLNPML and OR to quantify coherent and redundancy between topics respectively. CLNPMI is proposed by Chen et al. (2021) to calculate the average NPMI value of every parent and its children topics by $\text{CLNPMI}(W_p, W_c) = \sum_{w_i \in W_p'} \sum_{w_j \in W_c'} \frac{\text{NPMI}(w_i, w_j)}{|W_p'||W_c'|}$, where $W_p' = W_p - W_c$ and $W_c' = W_c - W_p$, in which W_p and W_p denote the top N words of a parent topic and a child topic respectively. OR measures the averaged repetition ratio of top N words between parent topics and their children, which is defined as: $\frac{|W_p \cap W_c|}{N}$. Following Duan et al. (2021), we treat the 2 most relevant lower-level topics of each upper-level topic as parent-child topics. Table 3 shows the performance of different models on multiple datasets, which demonstrates that our topic structure ensures the most diversity while remains

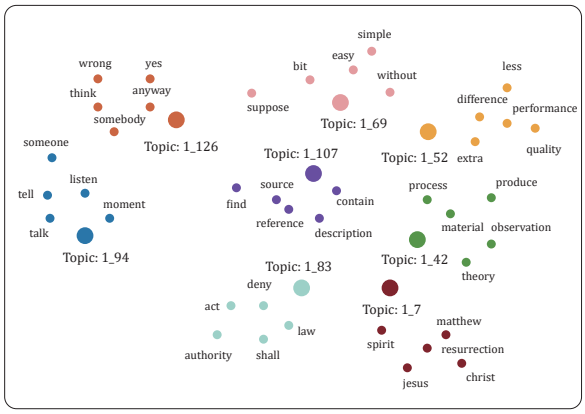

Figure 4: Visualization of word and topic embeddings, where Topic: l_i denotes the i_{th} topic at the l_{th} layer.

good coherences between parent and child topics, proving the structural rationality of topic hierarchy.

4.4 Qualitative Analysis

Visualisation of embedding space: The top 5 words from eight topics generated by NSEM-GMHTM over 20News are visualized in Figure 4 via UMAP visualization (McInnes et al., 2018). We can observe that the topics are highly interpretable in the word embedding space, where each topic is close to semantically related words. Besides, while words under the same topic are closer together, words under different topics are far apart. Additionally, the related topics are also closer in the embedding space, such as Topic: 1_94 and Topic: 1_126.

Hierarchical structure of topics: To intuitively demonstrate the ability of our model in generating hierarchical topic structures (i.e., relationships between topics at different levels), we visualize several topics extracted by our NSEM-GMHTM from 20News. As shown in Figure 5, each rectangle represents a topic and its top 10 words, and there are arrows from sub-topics to the most related topics. Consistent with the claim of topic specialization, the topics closer to root are more general and those closer to leaves are more specific. Besides, child topics are related to parent topics, e.g., *boston* is a child of *states*, and *authority* is a child of *law*. These results show that the semantic meaning of each topic and the connections between the topics of adjacent layers are highly interpretable, which demonstrates that our method can learn a reasonable topic hierarchy.

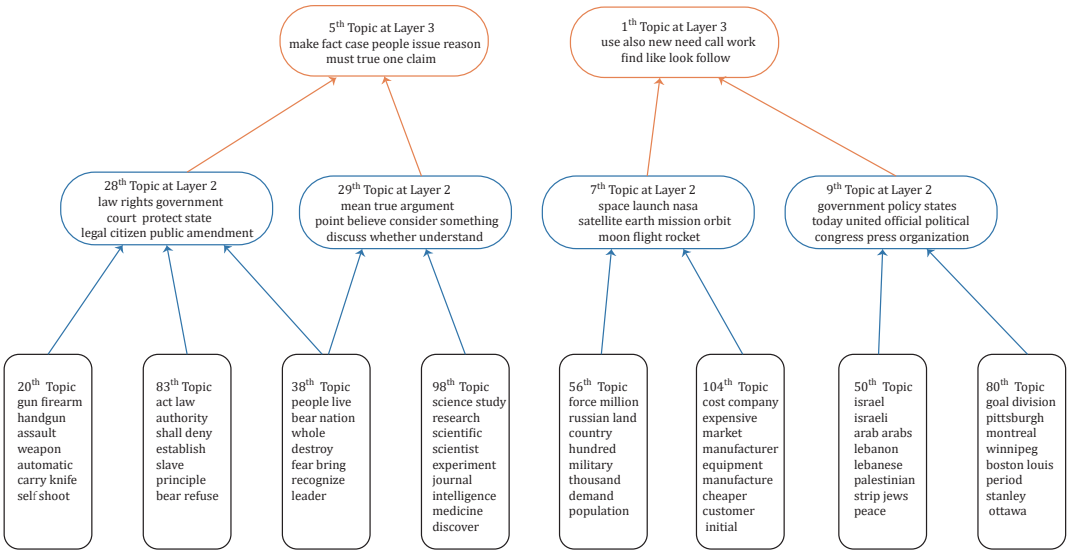

Figure 5: An example of hierarchical topics learned from 20News by NSEM-GMHTM.

Symmetric structure of topics: Apart from excelling at topic hierarchy learning, another appealing characteristic of NSEM-GMHTM is that it can discover an interpretable topic symmetric structure. In this part, we perform the topic symmetric relations discovery experiment on 20News. We query the top ranked same-level topic associations and some examples are shown in Table 4. The results show that our model can capture symmetric dependencies between topics, such as *nsa* for Topic: 1_76 and *secure* for Topic: 1_99, as well as *israel* and *nazi* for the 7_{th} ranked association. Furthermore,

Rank	Label					
2	Topic: 1_76 Topic: 1_99	clipper encryption	des privacy	nsa secure	escrow rsa	encrypted cryptography
6	Topic: 1_57 Topic: 1_77	jesus sin	christ lord	matthew spirit	scripture heaven	resurrection scripture
7	Topic: 1_79 Topic: 1_106	israel nazi	israeli muslim	arab german	palestinian genocide	lebanon nazis
9	Topic: 1_43 Topic: 1_53	gun crime	guns fbi	firearms batf	weapon waco	handgun defense
14	Topic: 1_57 Topic: 1_71	jesus truth	christ believe	matthew interpretation	scripture belief	resurrection follow

Table 4: Top topic relationships ranked by NSEM-GMHTM.

we extract topic symmetric dependencies from the first layer and construct a topic-topic network by selecting 100 topic associations with the greatest weight as edges. To better analyze the topic-topic network, we use Gephi[10] to visualize the topics and identify communities via a community detection algorithm (Blondel et al., 2008). As shown in Figure 6, topics with symmetric associations tend to form clusters and have tighter semantics within clusters (Table 5 and Table 6), suggesting that exploring symmetric associations between topics may be useful in further mining the semantic structure of a text corpus.

Label					
Topic: 1_82	cost	market	costs	cheaper	expensive
Topic: 1_70	armenian	armenians	armenia	azerbaijan	genocide
Topic: 1_100	jews	jewish	greek	adam	jew
Topic: 1_126	surrender	banks	pitt	gordon	intellect
Topic: 1_106	nazi	muslim	german	genocide	nazis
Topic: 1_25	religion	atheism	morality	atheists	religious
Topic: 1_79	israel	israeli	arab	palestinian	lebanon

Table 5: Top 5 words of topics in the green box of Figure 6.

Label					
Topic: 1_76	clipper	des	nsa	escrow	encrypted
Topic: 1_99	encryption	privacy	secure	rsa	cryptography
Topic: 1_64	low	high	rate	higher	rates
Topic: 1_91	state	rights	constitution	political	civil
Topic: 1_32	anti	population	armed	murder	crime
Topic: 1_53	crime	fbi	batf	waco	defense
Topic: 1_43	gun	guns	firearms	weapon	handgun
Topic: 1_40	crime	fbi	batf	waco	defense
Topic: 1_125	key	keys	blocks	pgp	scheme

Table 6: Top 5 words of topics in the blue box of Figure 6.

4.5 Ablation Study

For analyzing the effect of each component of our model, we ablate different components in three cases: 1) Without replacing the Gaussian prior distribution with a Gaussian mixture distribution (w/o GMM). 2) Without using pre-trained word embeddings (w/o PWE). 3) Without introducing NSEM to

[10] https://gephi.org/

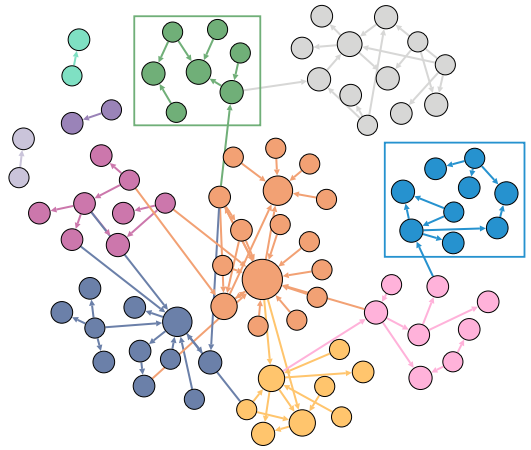

Figure 6: Topic symmetric network's first layer learned from 20News by NSEM-GMHTM. Each community is symbolized by a specific color and the details of green and blue boxes are shown in Table 5 and Table 6, respectively.

capture the symmetric topic relations (w/o NSEM). Table 7 tabulates all metrics on the three datasets for the three cases with NSEM-GMHTM. Results suggested that by introducing PWE and GMM components, the model can better capture the underlying topic hierarchy and achieve topic interpretability improvements. Moreover, the introduction of NSEM helps to enhance the semantic coherence of the topics but reduces the uniqueness of topics. Furthermore, NSEM has a significant impact on TU, NPMI, and OR, indicating that exploring symmetrical relationships between topics can help to generate a more rational topic structure and improve the interpretability of the model. In summary, all components of the NSEM-GMHTM method are reasonable and effective.

Datasets	Model	NPMI↑	TU↑	OR↓	CLNPMI↑
Wikitext-103	Ours	0.255	**0.791**	**0.017**	0.090
	Ours w/o GMM	0.255	0.641	0.021	0.092
	Ours w/o PWE	0.252	0.787	0.025	0.011
	Ours w/o NSEM	**0.261**	0.641	0.045	**0.147**
20News	Ours	**0.307**	**0.811**	**0.011**	0.146
	Ours w/o GMM	0.271	0.436	0.016	0.131
	Ours w/o PWE	0.277	0.698	0.019	0.127
	Ours w/o NSEM	0.284	0.807	0.038	**0.171**
NIPS	Ours	**0.147**	**0.719**	**0.028**	0.025
	Ours w/o GMM	0.129	0.642	0.031	0.025
	Ours w/o PWE	0.126	0.681	0.037	0.031
	Ours w/o NSEM	0.141	0.689	0.042	**0.057**

Table 7: Results of ablation evaluation on all datasets.

Metric	SawETM	TSNTM	nTSNTM	HyperMiner	NSEM-GMHTM
Speed	5.2S	11.3S	38.6S	4.4S	3.8S
#Params	1.9M	1.3M	0.5M	2.2M	1.5M

Table 8: Speed and number of parameters for NHTMs on the 20News dataset.

4.6 Analysis of Model Complexity

Here, we compare the complexity of our model and all benchmarks of NHTMs. Specifically, we average the cost of 10 training epochs for each model on 20News to record the running time. In addition, the number of parameters for models is recorded, as shown in Table 8. It is worth noting that, although CluHTM is excluded due to its unique training strategy, it is clear from Table 3 that its running time is far greater than that of the other NHTMs. We can find that NSEM-GMHTM achieves competitive performance, demonstrating that explicitly modelling hierarchical and symmetric dependencies does not significantly increase the complexity of the model, and further demonstrating the scalability of our model.

5 Conclusion

In this paper, we propose a novel neural topic model named NSEM-GMHTM. Our method explicitly constructs symmetric and hierarchical dependencies between topics through NSEM and dependency matrices. In addition, we introduce GMM as a prior for latent variables to improve the ability of NSEM-GMHTM to fit and analyze sparse data. Extensive experiments have shown that our method outperforms state-of-the-art baselines in extracting coherent and reasonably structured topics. Furthermore, with learned word and topic embeddings, and different types of topic relationships (hierarchical and symmetric), NSEM-GMHTM can discover a clearly interpretable topic structure. Eventually, the topic structures mined by NSEM-GMHTM show defined topic associations beyond the hierarchy, which are more consistent with the semantic relations of generic knowledge graphs such as WordNet (Miller, 1995) and ConceptNet (Speer et al., 2017) compared to other NHTMs, suggesting that our model may be able to exploit knowledge more fully. In the future, we will attempt to further incorporate prior information to guide the discovery of topic structures. In summary, our findings suggest that the discovery of topic structure can benefit from the construction of topic symmetric relations, which may contribute to a better understanding of text data.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (61972426).

Limitations

Our approach is only a small step towards mining more comprehensive, high-quality topic structures, and there are many more issues that need to be addressed in the future. For example, there are still limitations in the current assessment of the structure of topics mined by different models. Examples include assessing the validity of topic hierarchical indicators by topic specialization and the validity of the symmetric structure of topics through clustering as we have demonstrated. All these assessment methods are only a sideways demonstration of the interpretability of the topic structure. Besides, there is still a lot of a prior information available in the field of topic modelling, e.g. WordNet, and it may help researchers to explore further in the field of topic modelling if they can combine prior human knowledge and information on topic-words obtained from models to define quantitative metrics that are more consistent with human understanding.

References

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

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

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Kenneth A Bollen. 1989. *Structural equations with latent variables*. John Wiley & Sons.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *GSCL*, pages 31–40.

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

Aaron Clauset, Cristopher Moore, and Mark EJ Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

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 AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*, abs/1611.02648.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *EMNLP*, pages 830–836.

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 *ICML*, pages 2903–2913.

Arthur S Goldberger. 1972. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, 40(6):979–1001.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-structured neural topic model. In *ACL*, pages 800–806.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *ICLR*.

Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice H. Oh. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *CIKM*, pages 783–792.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539.

Rui Liu, Xingguang Wang, Deqing Wang, Yuan Zuo, He Zhang, and Xianzhu Zheng. 2018. Topic splitting: A hierarchical topic model based on non-negative matrix factorization. *Journal of Systems Science and Systems Engineering*, 27(4):479–496.

Yunan Luo, Jian Peng, and Jianzhu Ma. 2020. When causal inference meets deep learning. *Nature Machine Intelligence*, 2(8):426–427.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426.

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

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

David M. Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *ICML*, pages 633–640.

Eric Nalisnick and Padhraic Smyth. 2017. Stick-breaking variational autoencoders. In *ICLR*.

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

Xuefei Ning, Yin Zheng, Zhuxi Jiang, Yu Wang, Huazhong Yang, Junzhou Huang, and Peilin Zhao. 2020. Nonparametric topic modeling with neural inference. *Neurocomputing*, 399:296–306.

John W. Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2015. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270.

Rajesh Ranganath, Sean Gerrish, and David Blei. 2014. Black box variational inference. In *AISTATS*, pages 814–822.

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

Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. 2021. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1(7):491–501.

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

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

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

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, pages 1385–1392.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos André Gonçalves. 2020. Cluhtm - semantic hierarchical topic modeling based on cluwords. In *ACL*, pages 8138–8150.

Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

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

Lei Xiong, Kui Xu, Kang Tian, Yanqiu Shao, Lei Tang, Ge Gao, Michael Zhang, Tao Jiang, and Qiangfeng Cliff Zhang. 2019. Scale method for single-cell atac-seq analysis via latent feature extraction. *Nature Communications*, 10(1):1–10.

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

Linxiao Yang, Ngai-Man Cheung, Jiaying Li, and Jun Fang. 2019. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *ICCV*, pages 6440–6449.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. Dag-gnn: Dag structure learning with graph neural networks. In *ICML*, pages 7154–7163.

A Parameter Inference Algorithm

We apply NVI to inference network parameters, which is efficient and flexibility (Srivastava and Sutton, 2017). Similar to VAEs, the training objective of our model is to maximize the following Evidence Lower Bound (ELBO):

$$\mathcal{L}_{ELBO} = \sum_{i=1}^{L} \mathbb{E}_{q(\boldsymbol{\theta}^i, \boldsymbol{\phi}^i, c|\boldsymbol{x})} \left[\log p\left(\hat{\boldsymbol{x}} \mid \boldsymbol{\theta}^i, \boldsymbol{\phi}^i\right)\right] + - D_{KL}\left[q\left(\boldsymbol{\theta}^L, c \mid \boldsymbol{x}\right) \| p\left(\boldsymbol{\theta}^L, c\right)\right] \tag{19}$$

Algorithm 1: Parameter Inference Algorithm

Input : The embedding of words W_E and documents $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D\}$;

Output : Topic-word distribution $\boldsymbol{\phi}$, topic hierarchy $\boldsymbol{T_h}$, and topic symmetry $\boldsymbol{T_s}$.

Randomly initialize dependency matrices M, symmetric matrices W, and topic embeddings T_E.
repeat
  for *documents* $\boldsymbol{x}_d \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D\}$ **do**
    Estimate $\{\boldsymbol{\theta}_d\}$ by Eqs. (8-15);
    Infer $\{\boldsymbol{\phi}\}$ by Eq. (16);
    Reconstruction $\hat{\boldsymbol{x}_d} \leftarrow \{\boldsymbol{\theta}_d\}, \{\boldsymbol{\phi}\}$;
    Compute $\mathcal{L}_{ELBO}$ by Eq. (19);
    Update $f(\cdot)$, W, M and T_E;
until *convergence*;
$\boldsymbol{T_h}$ and $\boldsymbol{T_s}$ are built from M, W, and $\boldsymbol{\phi}$.

where the first term is the reconstruction error for the different levels of topics with an additional L1

norm to regularize the symmetric dependency matrix matrix W^i, while the second term is the Kullback–Leibler (KL) divergence that constrains posterior $q\left(\boldsymbol{\theta}^L, c \mid \boldsymbol{x}\right)$ to be close to its prior $p\left(\boldsymbol{\theta}^L, c\right)$ in the generative model. The parameter inference method for NSEM-GMHTM is presented in Algorithm 1. We use the variational lower-bound to calculate gradients and apply RMSprop to update parameters.

B Training Details

NSEM-GMHTM is implemented via PyTorch. To keep simplicity, for the multilayer neural network $f(\cdot)$ in the encoder, we use a fully-connected neural network with $Tanh$ as the activation function. For the embedding-based topic models including ETM, SawETM, nTSNTM, CluHTM, HyperMiner, and NSEM-GMHTM, we incorporate pre-trained word embeddings (Viegas et al., 2020)[11] into them. All experiments were conducted with public model codes, trained for a single run, and on a workstation equipped with an Nvidia RTX 1080-Ti GPU and a Python environment with 128G memory.

[11] https://nlp.stanford.edu/projects/glove

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the "Limitations" section.

☒ A2. Did you discuss any potential risks of your work?
To the best of our knowledge, we haven't identified any potential risks of our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In the "Abstract" section and Section 1 "Introduction".

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In Section 3 "The Proposed Model" and Section 4 "Experiments" .

☑ B1. Did you cite the creators of artifacts you used?
In Section 4 "Experiments" and Appendix B "Training Details".

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will include the license or terms in the README file of our code repository.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We will specify intended use of existing artifacts and the created artifact in the README file of our code repository.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We have adopted widely-used corpora without sensitive information for our experiments.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We report the language and basic information about the artifacts in Section 4.1.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We report relevant statistics in detail in section 4.1.

C ☑ Did you run computational experiments?

In Section 4 "Experiments".

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We describe the model complexity and the equipment used in Section 4.6.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss the experiment settings in Section 4.1.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
To ensure reproducible results, all experiments for the models are run with a fixed random seed. In Section 4 "Experiments" and Appendix B "Training Details".

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We report the used existing packages for preprocessing and evaluation in Section 4.1.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.