Nonparametric Forest-Structured Neural Topic Modeling

Zhihong Zhang[*], **Xuewen Zhang**[*], **Yanghui Rao**[†]
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
{zhangzhh33, zhangxw53}@mail2.sysu.edu.cn, raoyangh@mail.sysu.edu.cn

Abstract

Neural topic models have been widely used in discovering the latent semantics from a corpus. Recently, there are several researches on hierarchical neural topic models since the relationships among topics are valuable for data analysis and exploration. However, the existing hierarchical neural topic models are limited to generate a single topic tree. In this study, we present a nonparametric forest-structured neural topic model by firstly applying the self-attention mechanism to capture parent-child topic relationships, and then build a sparse directed acyclic graph to form a topic forest. Experiments indicate that our model can automatically learn a forest-structured topic hierarchy with indefinite numbers of trees and leaves, and significantly outperforms the baseline models on topic hierarchical rationality and affinity.

1 Introduction

Topic model has been widely used in modeling a collection of documents and encoding the text content to a low dimensional feature space. Traditional topic models can be divided into probabilistic graphical models and matrix factorization based methods. Probabilistic graphical models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003b), rely on approximate approaches (e.g., variational inference and Gibbs sampling) with complex derivation or high computational costs to estimate parameters (Srivastava and Sutton, 2017). Matrix factorization based methods (Lee and Seung, 1999) can effectively decompose the document-word representation into two sub-matrices but are subject to a low stability (Chen et al., 2021b). Recently, neural topic models based on Neural Variational Inference (NVI) (Srivastava and Sutton, 2017; Miao et al., 2017; Chen et al., 2021b) have attracted great attention owing to the advantages of fast parameter inference and flexibility.

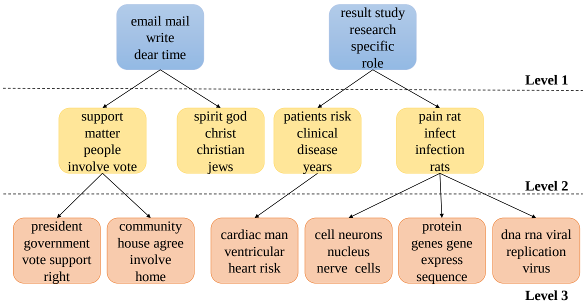

email mail write dear time
result study research specific role
Level 1
support matter people involve vote
spirit god christ christian jews
patients risk clinical disease years
pain rat infect infection rats
Level 2
president government vote support right
community house agree involve home
cardiac man ventricular heart risk
cell neurons nucleus nerve cells
protein genes gene express sequence
dna rna viral replication virus
Level 3

Figure 1: An example of forest-structured topics. Each topic is represented by 5 top words.

Despite some notable successes for neural topic models, most of the existing methods can only extract topics at the same level. This may cause confusions on identifying the hierarchical structure about the relationships among topics, which is valuable for data analysis and exploration in various domains (Paisley et al., 2015). To address this, a few neural topic models have been developed to model the hierarchical structure of topics (Isonuma et al., 2020; Chen et al., 2021b,a). The above methods, however, all assume that the hierarchical structure is a tree with a single root node. This is a significant limitation because for a real-world corpus, topics can be organized into several trees, where the structure of each tree is independent. As an illustration, Figure 1 shows two root topics on email and study. Topics at level 2 include email contents (i.e., politics and religion) and those describing the mode of study (i.e., patients and rats). Some of the topics at level 2 have several children topics which are very specific to distinguish the scope of politics (e.g., country and community) or focus of biology (e.g., cell, protein and genes), and others have only one child topic or none. It indicates that the real-world hierarchical topic structure is more likely to be a forest rather than a tree.

The forest-structured topic models still face challenges since the hierarchical structure of topics

[*] The first two authors contributed equally to this work.
[†] The corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2585–2597
October 12–17, 2022.

needs to be (1) rational—root topics are general and children topics are specific to their corresponding parent topics (Viegas et al., 2020); (2) affinitive—each topic is more similar to their children topics than topics from other parents (Kim et al., 2012); (3) diverse—the topic-word distributions associated with parents and children are distinguishable (Blei et al., 2003a); (4) flexible—children of each topic are automatically assigned (Kim et al., 2012) and topic numbers at each level are unbounded (i.e., nonparametric) (Chen et al., 2021b). Besides, the number of root topics, depth, and width of a forest structure are hard to be pre-defined.

In this work, we propose a nonparametric Forest-structured Neural Topic Model (nFNTM) to tackle these challenges, which firstly captures parent-child topic relationships based on the self-attention mechanism (Vaswani et al., 2017)[1], and then learns a sparse Directed Acyclic Graph (DAG) to build a topic forest by federating document-topic distributions and parent-child relationships based on NVI. To our best knowledge, the current topic models with a DAG structure are based on Bayesian learning (Li and McCallum, 2006; Mimno et al., 2007) or Non-negative Matrix Factorization (NMF) (Liu et al., 2018; Viegas et al., 2020), and there is no work under the NVI framework. To sum up, the main contributions are summarized as follows:

- We are the first to introduce the sparse DAG into neural topic modeling with the aim of learning a forest-structured topic hierarchy.
- We develop a self-attention mechanism to capture the relationships among topics.
- We evaluate nFNTM on three benchmark datasets. Empirical results indicate that our model significantly outperforms baselines.

2 Related Work

Traditional topic models, such as LDA (Blei et al., 2003b), are powerful tools for modeling text in an unsupervised fashion, while they lack the exploration of the relationship among topics. To overcome this issue, a tree-structured topic model named hLDA (Blei et al., 2003a) was first proposed. In hLDA, the nested Chinese Restaurant Process (nCRP) was used to generate a topic tree. To alleviate the single-path constraint assumed by nCRP, a nested Hierarchical Dirichlet Process (HDP), i.e., nHDP (Paisley et al., 2015) was developed, which provided the ability of cross-thematic borrowing while keeping general topic areas in separate subtrees. The nested Chinese Restaurant Franchise (nCRF) process developed in (Ahmed et al., 2013) combined the advantages of HDP (Teh et al., 2004) and nCRP. In (Kim et al., 2012), the recursive Chinese Restaurant Process (rCRP) was proposed to discover a hierarchical topic structure with unbounded depth and width.

These models based on the Chinese restaurant process can be effectively employed to discover the hierarchical topic structure by Bayesian learning, but the posterior inference method requires a high computational cost. The scalability of NMF-based methods (Liu et al., 2018; Viegas et al., 2020) is also quite limited. There is a new direction to build tree-structured topic models based on NVI due to its advantages of fast parameter inference and flexibility. A tree-structured neural topic model (TSNTM) (Isonuma et al., 2020) was proposed, which applied doubly-recurrent neural networks to parameterize topic distributions over a tree. But it lacked the ability of learning appropriate semantic embeddings for topics and relied on heuristic rules to update the tree structure. A nonparametric tree-structured neural topic model (nTSNTM) (Chen et al., 2021b) tackled these weaknesses by directly sampling the leaf topics and generating the paths from bottom up automatically. nTSNTM used a common stick-breaking construction to infer topic distributions from the leaf nodes to the root node and applied dependency matrices to keep track of the affiliations among topics. However, the dependency matrices which determine the topic hierarchy are neural weights between the network layers. It results in the structure (i.e., depth and width) of the tree can only be set in advance.

3 Methodology

3.1 Model Architecture

Our nFNTM consists of an encoder, a topic attention, and a decoder, as shown in Figure 2.

3.1.1 Encoder

Given a collection of documents, each document $\boldsymbol{d} \in \mathbb{R}^V$ is represented by Bag-of-Words (BoW), where V is the vocabulary size. For the encoder,

[1] We use the self-attention mechanism as it is effective to draw the global dependencies between input tokens with little reliance on the external information (Yao et al., 2021). Furthermore, it allows efficient computation by parallelization.

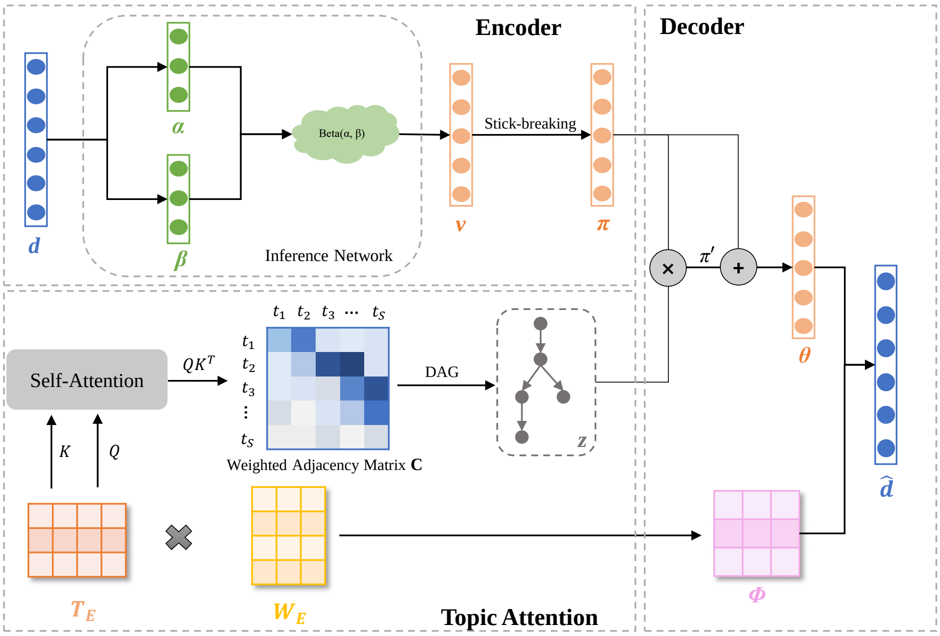


Figure 2: Structure of nFNTM.

we transform $\boldsymbol{d}$ into document-topic distribution $\boldsymbol{\pi}$ by the Stick-Breaking Process (SBP) (Ishwaran and James, 2001), which provides a solution to define atomic measures associated Bayesian nonparametric methods. Any almost sure (a.s.) discrete probability measure $\mathcal{P}$ is an SBP if it can be represented by:

$$\mathcal{P} = \sum_{i=1}^{\infty} \pi_i \delta_{x_i}, \quad \pi_i = \begin{cases} v_1 & i = 1, \\ v_i \prod_{t<i} (1 - v_t) & i > 1, \end{cases} \tag{1}$$

where $x_i \sim \mathcal{H}$, $\mathcal{H}$ is the base probability measure, δ_{x_i} is a discrete measure concentrated at x_i, $\boldsymbol{v} \sim \text{Beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ being the prior parameters, $\{\pi_i\}$ are random weights independent of $\mathcal{H}$ and satisfy $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{\infty} \pi_i = 1$. SBP specifies to the Dirichlet process if $\boldsymbol{v} \sim \text{Beta}\left(1, \hat{\boldsymbol{\beta}}\right)$, then the joint distribution over the infinite sequence of stick-breaking weights with concentration parameter $\hat{\boldsymbol{\beta}}$ is $\{\pi_i\} \sim \text{GEM}(\hat{\boldsymbol{\beta}})$ (Teh et al., 2004).

Note that the existing nTSNTM (Chen et al., 2021b) chooses the Kumaraswamy distribution (Kumaraswamy, 1980) to approximate the Beta distribution since it does not have a differentiable non-centered parametrization. However, the Beta distribution is a one-parameter subfamily of symmetric distributions and has more ways of generating the distribution via physical processes (Jones, 2009). To estimate the Beta distribution unbiasedly, we inference it through computing implicit reparameterization gradients and the details will be introduced in Section 3.3. As shown in Figure 2, we introduce an inference network to build the Beta distribution. We obtain $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by $\boldsymbol{\alpha} = l_\alpha(\boldsymbol{\eta})$ and $\boldsymbol{\beta} = l_\beta(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = \text{MLP}(\boldsymbol{d})$ is the low-dimensional representation of $\boldsymbol{d}$, l_α and l_β are the single layer of linear connection, and MLP denotes a multilayer perceptron.

3.1.2 Topic Attention

We intend the model to attend on relationships among topics in a manner that the resultant attention is distributed according to the topics generated from the corpus. We employ topic embeddings $\mathbf{T_E} \in \mathbb{R}^{S \times d_t}$ to perform attention on topics, where $S \to \infty$ is the breaking number in SBP, i.e., the number of topics learned by SBP, and d_t is the dimension of topic embeddings. Considering that the relevance of each topic should be computed and learned independently, we regard each topic embedding as a subspace and project the input hidden representations to different topic embedding subspaces as follows:

$$\mathbf{Q} = \mathbf{T_E W_Q}, \quad \mathbf{K} = \mathbf{T_E W_K}, \tag{2}$$

where $\mathbf{W_Q} \in \mathbb{R}^{d_t \times d_r}$ and $\mathbf{W_K} \in \mathbb{R}^{d_t \times d_r}$ are trainable parameters, d_r is the dimension of each topic embedding subspace, $\mathbf{Q} \in \mathbb{R}^{S \times d_r}$ and $\mathbf{K} \in \mathbb{R}^{S \times d_r}$ denote the matrices of queries and keys. Based on queries and keys, we calculate a weighted adjacency matrix $\mathbf{C} \in \mathbb{R}^{S \times S}$, where each element

$C_{i,j}$ denotes the parent-child correlation degree of relevance between topics t_i and t_j[2]. To ensure that each element $C_{i,j}$ is discrete, a softmax function with temperature τ is applied (Hinton et al., 2015):

$$C_{i,j} = \frac{exp(\frac{\hat{C}_{i,j}}{\tau})}{\sum_{k=1}^{S} exp(\frac{\hat{C}_{i,k}}{\tau})}, \tag{3}$$

where $\hat{\mathbf{C}} = \mathbf{Q}\mathbf{K}^T$.

To further incorporate the parent-child correlations into document-topic distributions meanwhile maintaining the nonparametric characteristic from SBP, we firstly calculate the parent document-topic distribution by integrating the weighted adjacency matrix, i.e., $\boldsymbol{\pi}' = \boldsymbol{\pi} \times \mathbf{C}$. Particularly, as the weighted adjacency matrix has not captured parent-child correlations at the beginning, the document-topic distribution is initially represented by $\boldsymbol{\pi}$. With the learning of our self-attention module, the quality of the weighted adjacency matrix $\mathbf{C}$ is steadily enhanced, and $\boldsymbol{\pi}'$ becomes a valuable supplement of $\boldsymbol{\pi}$ due to strong parent-child topic relationships. Then, we generate the consolidated document-topic distribution by exploiting both $\boldsymbol{\pi}'$ and $\boldsymbol{\pi}$ as $\boldsymbol{\theta} = (1-\gamma)\boldsymbol{\pi}' + \gamma\boldsymbol{\pi}$, where γ is a decay coefficient.

3.1.3 Decoder

As the BoW document representations lack of the word relatedness information, we incorporate pre-trained word embeddings $\mathbf{W_E}$ (Viegas et al., 2020; Wu et al., 2020) into the network. For the decoder, we firstly obtain the topic-word distribution $\boldsymbol{\Phi}$ = softmax$(\mathbf{T_E} \times \mathbf{W_E})$, where $\mathbf{W_E} \in \mathbb{R}^{d_t \times V}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{S \times V}$. Then, we reconstruct document $\hat{\boldsymbol{d}}$ by combining document-topic distribution $\boldsymbol{\theta}$ with topic-word distribution $\boldsymbol{\Phi}$.

3.2 Topic Forest Hierarchy

To construct the topic forest, we build a forest hierarchy of topics from the weighted adjacency matrix $\mathbf{C}$ which contains the parent-child relationships among topics. However, the structure of weighted adjacency matrix $\mathbf{C}$ may not be a reasonable hierarchical structure, i.e., it may have loops if without constraints. To tackle this challenge, we propose to construct the structure of topics as a sparse DAG. Zheng et al. (2018) have proved that a weighted adjacency matrix $\mathbf{C}$ is a DAG if and only if $h(\mathbf{C}) = \mathrm{tr}\left(e^{\mathbf{C} \circ \mathbf{C}}\right) - d = 0$, where $\circ$ is the Hadamard product and $e^{\mathbf{C}}$ is the matrix exponential of $\mathbf{C}$. We employ $h(\mathbf{C}) = 0$ with an augmented Lagrangian method to ensure the acyclicity of the weighted adjacency matrix.

Considering that the topic hierarchy is reasonable and children topics can be represented by their parent topics, we assume that the sum of children topics' document-topic distributions is similar to their parent topics' document-topic distributions. Accordingly, we learn the forest-structured topic hierarchy by minimizing the difference between documents reconstructed by $\boldsymbol{\pi} \times \boldsymbol{\Phi}$ and those reconstructed by their parent document-topic distributions (i.e., $\boldsymbol{\pi}' \times \boldsymbol{\Phi}$) under the constraint of $h(\mathbf{C}) = 0$ using the augmented Lagrangian method, as follows:

$$\begin{aligned}\min_{\mathbf{C}\in\mathbb{R}^{S\times S}} \quad \mathcal{L}_C = \frac{1}{2}\|(\boldsymbol{\pi} - \boldsymbol{\pi}') \times \boldsymbol{\Phi}\|_F^2 \\ + \frac{\rho}{2}|h(\mathbf{C})|^2 + \epsilon h(\mathbf{C}),\end{aligned} \tag{4}$$

where ρ is a penalty parameter, and ϵ is the Lagrange multiplier. We update parameters ρ and ϵ by following (Zheng et al., 2018), as follows: $\begin{cases}\rho_i = & 2\rho_{i-1} \\ \epsilon_i = & \epsilon_{i-1} + \rho h_{i-1}\end{cases}$, where $\rho_0 = 1$, $\epsilon_0 = 0$, i is a training epoch, and h is a constraint value.

The generative process of nFNTM is described as follows:

1. For each document $\boldsymbol{d}$:
   - (a) Draw a topic proportion $\boldsymbol{\pi} \sim \mathrm{GEM}(\hat{\boldsymbol{\beta}})$;
   - (b) Get the weighted adjacency matrix $\mathbf{C}$;
   - (c) Get the correlational topic distribution $\boldsymbol{\theta}$.
2. For each word $w_{d,n} \in \boldsymbol{d}$:
   - (a) Draw a topic $z_{d,n} \sim \mathrm{Mult}(\boldsymbol{\theta})$;
   - (b) Obtain the topic-word distribution $\boldsymbol{\Phi}$;
   - (c) Draw a word $w_{d,n} \sim \mathrm{Mult}(\boldsymbol{\Phi}_{z_{d,n}})$.

3.3 Parameter Inference

We apply NVI to inference network parameters, which is proven to be efficient and flexible (Srivastava and Sutton, 2017; Miao et al., 2017; Isonuma et al., 2020; Chen et al., 2021b). The likelihood of each reconstructed document $\hat{\boldsymbol{d}}$ is estimated by $p\left(\hat{\boldsymbol{d}} \mid \boldsymbol{\theta}, \boldsymbol{\Phi}\right) = \sum_{\boldsymbol{z}} p\left(\hat{\boldsymbol{d}} \mid \boldsymbol{\Phi}_{\boldsymbol{z}}\right) p\left(\boldsymbol{z} \mid \boldsymbol{\theta}\right)$, where $\boldsymbol{z}$ is the topic assigned for each word in $\hat{\boldsymbol{d}}$. To maximize the log-likelihood, we derive the lower bound

[2] Different from (Vaswani et al., 2017), we capture the relationships among topics without using the value matrix. This is because we found experimentally that topics trained with value matrix tend to be the same, which might result from concentrating on several important topics.

as follows:

$$\begin{aligned}\mathcal{L}_B = &\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\Phi}|\boldsymbol{d})}\left[\log p\left(\hat{\boldsymbol{d}} \mid \boldsymbol{\theta}, \boldsymbol{\Phi}\right)\right] \\ &- D_{KL}\left[q\left(\boldsymbol{\theta} \mid \boldsymbol{d}\right) \| p\left(\boldsymbol{\theta}\right)\right] - \mathcal{L}_C,\end{aligned} \tag{5}$$

where $\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\Phi}|\boldsymbol{d})}\left[\log p\left(\hat{\boldsymbol{d}} \mid \boldsymbol{\theta}, \boldsymbol{\Phi}\right)\right]$ is the reconstruction loss, $D_{KL}\left[q\left(\boldsymbol{\theta} \mid \boldsymbol{d}\right) \| p\left(\boldsymbol{\theta}\right)\right]$ is the Kullback-Leibler (KL) divergence between the prior Beta distribution $p\left(\boldsymbol{\theta}\right)$ and the posterior Beta distribution $q\left(\boldsymbol{\theta} \mid \boldsymbol{d}\right)$. The KL divergence of two Beta distributions is given below:

$$\begin{aligned}&KL(Beta(\alpha_1, \beta_1)||Beta(\alpha_2, \beta_2)) = \\ &\quad \ln\Gamma(\alpha_2) + \ln\Gamma(\beta_2) + \ln\Gamma(\alpha_1 + \beta_1) \\ &\quad - (\ln\Gamma(\alpha_1) + \ln\Gamma(\beta_1) + \ln\Gamma(\alpha_2 + \beta_2)) \\ &\quad + (\alpha_1 - \alpha_2)\digamma(\alpha_1) \\ &\quad + (\beta_1 - \beta_2)\digamma(\beta_1) \\ &\quad + (\alpha_2 + \beta_2 - \alpha_1 - \beta_1)\digamma(\alpha_1 + \beta_1),\end{aligned} \tag{6}$$

where $\digamma(x) = \frac{d}{dx}\ln\digamma(x) = \frac{\digamma'(x)}{\digamma(x)}$.

Note that we transform $\boldsymbol{d}$ to the variational Beta distribution, thus $q\left(\boldsymbol{\theta} \mid \boldsymbol{d}\right)$ is derived by:

$$\begin{aligned}q\left(\boldsymbol{\theta} \mid \boldsymbol{d}\right) &= \text{Beta}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})) \\ &= \text{Beta}(\boldsymbol{\theta} \mid l_\alpha\left(\text{MLP}(\boldsymbol{d})\right), l_\beta\left(\text{MLP}(\boldsymbol{d})\right)).\end{aligned} \tag{7}$$

To compute reparameterization gradients, the Beta samples are obtained from Gamma samples since the latter do not require inverting the standardization function (Figurnov et al., 2018): for $z_1 \sim \text{Gamma}(\boldsymbol{\alpha}, 1)$ and $z_2 \sim \text{Gamma}(\boldsymbol{\beta}, 1)$, it has $\frac{z_1}{z_1+z_2} \sim \text{Beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The lower bound $\mathcal{L}_B$ is used to calculate gradients and parameters are updated by Adam (Kingma and Ba, 2015).

4 Experiment

4.1 Experimental Setup

Datasets To evaluate our nFNTM[3] comprehensively, we conduct experiments on three datasets which are observed to be hierarchical (Chen et al., 2021b): 20News (Miao et al., 2017), Reuters (Wu et al., 2020), and Wikitext-103 (Nan et al., 2019). According to (Liu et al., 2018; Viegas et al., 2020), a topic often comprises of some sub-topics and owns a hierarchical structure in Web directory of news (e.g., 20News and Reuters) and encyclopedia (e.g., Wikitext-103 which is extracted from Wikipedia). For these datasets, each document is associated with a manually-curated hierarchy of categories. Take 20News as an example, "rec.autos" and "rec.motor.cycles" are sub categories of "rec". Instead of relying on the prior coarse-grained categories, we build the hierarchical structure from a corpus at the fine-grained topic level, by automatically mining a set of representative words for each topic in a forest structure to help a user comprehend her/his interested topics.

All datasets have undergone a preprocessing of removing stop words and deleting low-frequency words. Table 1 shows the numbers of training and test documents, as well as the vocabulary size.

Dataset	#Docs(Train)	#Docs(Test)	Vocabulary size
20News	11,314	7,531	1,995
Reuters	7,769	3,019	2,000
Wikitext-103	28,472	120	20,000

Table 1: Statistics of datasets.

Baselines We employ four tree-structured topic models, including hLDA[4] (Blei et al., 2003a), rCRP[5] (Kim et al., 2012), TSNTM[6] (Isonuma et al., 2020), and nTSNTM[7] (Chen et al., 2021b), and three DAG-structured topic models, including hPAM (Mimno et al., 2007), HSOC (Liu et al., 2018), and CluHTM (Viegas et al., 2020) as baselines. For tree-structured baselines, the max-depth of topic tree is set to 3 by following (Isonuma et al., 2020). In addition, we adopt seven flat topic models for comparison, including parametric models of GSM[8] (Miao et al., 2017), GSB (Miao et al., 2017), NB-NTM[9] (Wu et al., 2020) and GNB-NTM[9] (Wu et al., 2020), and nonparametric models of HDP[10] (Teh et al., 2004), iTM-VAE[11] (Ning et al., 2020), and HiTM-VAE[11] (Ning et al., 2020). Except for GSB, we use the open source codes for all other baseline models.

Hyper-parameter Settings To ensure fair comparisons, we follow (Chen et al., 2021b) to set topic numbers to 50 and 200 for all parametric models. For non-parametric models based on SBP (i.e., iTM-VAE and nTSNTM), the maximum number of topics is set to 200, and the concentration

[3] https://github.com/Angr4Mainyu/nFNTM

[4] https://github.com/joewandy/hlda

[5] https://github.com/uilab-github/rCRP

[6] https://github.com/misonuma/tsntm

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

[8] https://github.com/linkstrife/NVDM-GSM

[9] https://github.com/mxiny/NB-NTM

[10] https://github.com/arnim/HDP

[11] https://github.com/walkerning/itmvae_public

parameter $\hat{\beta}$ of the GEM distribution is set to 20. According to (Chen et al., 2021b), we select the topic with a total probability exceeding 95% as an active topic. Besides, the penalty parameter ρ is updated by $\rho = 2^x$, where x denotes the number of training epochs. The temperature τ changes exponentially from 5 to 1×10^{-4} to ensure the adjacency matrix to be sparse. The hidden layer size of the encoder is set to 256, which is consistent with other models. In the construction of the weighted adjacency matrix, we use an exponential change strategy to dynamically adjust the decay coefficient. This is because the quality of such a weighted adjacency matrix is gradually enhanced when learning the self-attention module. Finally, the decay coefficient γ changes exponentially from 1 to 0.5. We implement our model by pytorch and run it on a computer with NVIDIA 1080Ti and 128GB RAM.

4.2 Topic Hierarchy Analysis

To evaluate the rationality, affinity, and diversity of the topic hierarchy generated by different models, we adopt four metrics: topic specialization (Kim et al., 2012), Cross-Level Normalized Point-wise Mutual Information (CLNPMI) (Chen et al., 2021b), hierarchical affinity (Kim et al., 2012), and Topic Uniqueness (TU) (Nan et al., 2019). Key words of each topic are ranked by the topic-word matrix $\mathbf{\Phi}$ (Blei et al., 2003b).

Topic Hierarchical Rationality For the tree-structured topics, the topics closer to the root node will be more general, while topics closer to the leaf node will be more specific. Topic specialization score is to quantify this feature by computing the cosine similarity of the word distribution between each topic and the entire corpus. Since our forest-structured model generates several trees, we pick out all three-layer topic trees to obtain the average score for comparison. Figure 3 shows the topic specialization results, from which we can observe that our model outperforms baselines except for HSOC and CluHTM. Although the topic specializations of HSOC and CluHTM at different levels are close to 1, they still lack rationality since the root topic in a tree should be more general than others.

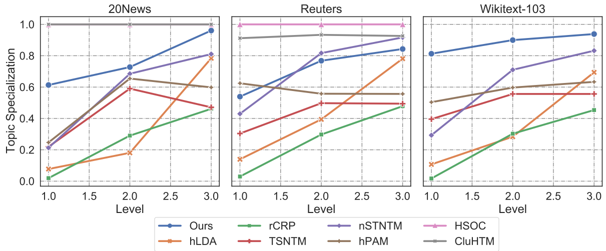


Figure 3: Topic specialization of different topics trees generated from the three datasets. A higher score with a growing trend means better performance.

Models	20News		Reuters		Wikitext-103	
	CLNPMI	*TU*	*CLNPMI*	*TU*	*CLNPMI*	*TU*
hLDA	0.065	0.051	0.050	0.447	0.063	0.597
rCRP	0.098	0.285	0.072	0.227	0.088	0.355
TSNTM	0.086	0.430	0.027	0.370	0.065	0.615
nTSNTM	0.109	0.745	0.102	0.708	0.113	0.730
hPAM	0.046	0.606	0.011	0.470	0.047	0.713
HSOC	0.128	0.231	0.047	0.211	-	-
CluHTM	0.123	0.116	0.016	0.117	-	-
Ours	**0.152**	**0.757**	**0.125**	**0.798**	**0.118**	**0.766**

Table 2: The average CLNPMI and TU scores of hierarchical topic models with top 5, 10, and 15 words for each topic. A higher score means better performance and the best scores are highlighted by boldface.

To measure the relationship between two connected topics, a metric of CLNPMI (Chen et al., 2021b) was proposed by calculating the average Normalized Point-wise Mutual Information (NPMI) score of every parent and its children topics, as follows: $CLNPMI\,(W_p, W_c) = \sum_{w_i \in W_p'} \sum_{w_j \in W_c'} \frac{\mathrm{NPMI}(w_i, w_j)}{|W_p'||W_c'|}$, where $W_p' = W_p - W_c$ and $W_c' = W_c - W_p$, in which W_p and W_c denote the top N words of a parent topic its child topic, respectively. A higher CLNPMI indicates that children topics are more coherent with their corresponding parent topics. We compare our nFNTM with hierarchical topic models mentioned above, and the CLNPMI results of all models are shown in Table 2. Note that the two NMF-based models (i.e., HSOC and CluHTM) had not converged after running for more than 48 hours on Wikitext-103, thus we did not include their results in the table. As shown in Table 2, our nFNTM achieves the best performance on all datasets.

Topic Uniqueness To evaluate the diversity of hierarchical topics, we calculate the topic uniqueness by: $TU = \frac{Count(Set(W_{topN}))}{N \times S}$, where $Count(Set(W_{topN}))$ is the number of distinct words in top N words of all topics. A higher TU means that the generated topics are more diverse. Table 2 shows the TU results for all models, from which we can observe that our model generates more diverse topics than others. The baselines of HSOC and CluHTM perform quite poor since they need to preset up to 600 topics for convergence.

Topic Hierarchical Affinity A reasonable assumption for topic hierarchy is that topics with parent-child relationships show larger similarities in their topic-word distributions than topics without any parent-child relationship (Kim et al., 2012). According to (Kim et al., 2012), we firstly evaluate the similarity between parent-child topics and non-parent-child topics by computing the cosine similarity between their topic-word distributions. Then, the topic hierarchical affinity is measured according to the difference of those similarities.

Let $\mathbf{\Phi}(k)$ be a topic at level k, $\lambda(k)$ be children topics of $\mathbf{\Phi}(k)$, and $\bar{\lambda}(k)$ be non-children topics of $\mathbf{\Phi}(k)$. The topic hierarchical affinity metric compares the average cosine similarity $\mathrm{S}_{\lambda}(k)$ between $\mathbf{\Phi}(k)$ and all topics in $\lambda(k)$ against the average cosine similarity $\mathrm{S}_{\bar{\lambda}}(k)$ between $\mathbf{\Phi}(k)$ and all topics in $\bar{\lambda}(k)$. A large difference between $\mathrm{S}_{\lambda}(k)$ and $\mathrm{S}_{\bar{\lambda}}(k)$ indicates a good topic hierarchical affinity.

Since all topics between paired levels in hPAM are fully connected, it is impossible to clearly distinguish topics with parent-child and non-parent-child relationships. Thus, we exclude hPAM in this part. As shown in Figure 4, our model achieves high similarities between parent-child topics and low similarities between non-parent-child topics, indicating a good hierarchical affinity. Note that both HSOC and CluHTM are based on NMF, which rely on a predefined number of topic trees. To avoid missing potential topic sub-structures, these models need to set a large number of topic trees (e.g., 100) according to their default settings. Given superabundant topic trees, the similarity between non-parent-child topics will be underestimated due to the huge number of non-parent-child topics, resulting in a competitive hierarchical affinity. However, it leads to a very poor TU score for HSOC or CluHTM, as already shown in Table 2.

4.3 Topic Interpretability

We employ the NPMI score (Lau et al., 2014) to evaluate the interpretability of topics since this metric is shown to be close to human judgments (Lau et al., 2014). We extract top 5, 10, and 15 words for each topic and compute the average NPMI scores of all models over the three datasets. The higher the value of NPMI score, the more interpretable the generated topics is. The corresponding numbers of topics automatically determined by our nonparametric model on 20News, Reuters, and Wikitext-103 are 61, 76, and 121, respectively.

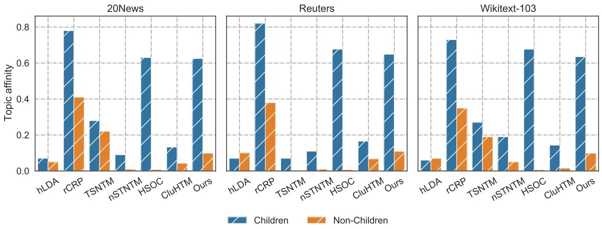


Figure 4: Topic hierarchical affinity results. For each associated topic, a larger difference of word distributions between children and non-children topics is better.

Table 3 shows the NPMI scores of topics generated by different models. Although our nFNTM outperforms hierarchical topic models by a large margin, it performs slightly worse than the top-performing flat topic model. The reason is that NPMI is the average score of all topics, which may be unfair for hierarchical topic models. Flat topic models treat topics independently and result in specific topics, while higher-level topics in hierarchical topic models are general with a lower NPMI score. Figure 5 shows two examples of topic trees generated by nFNTM on Wikitext-103 (left) and 20News (right), which focus on different topics, i.e., *politics* and *war* in the left tree, and *play*, *religion*, and *computer* in the right tree. Take the left tree as an illustration, the NPMI score of the root topic [*day hour people home left*] is obviously lower than that of the leaf topic [*force army battle attack war*], which affects the overall result. We can also observe that the topics closer to root are more general and those closer to leaves are more specific. Besides, children topics are related to parent topics, e.g., encryption is a child of computer, and law is a child of governance. These examples of topics validate the advantage of a forest structure when compared with a tree structure: the former can not only learn reasonable parent-child relationships among topics, but also generate a flexible topic hierarchy with unbounded depth and width.

4.4 Ablation Study

We perform ablation experiments on our model to validate the effectiveness of each module. Table 4 shows the ablation results of our nFNTM without three modules, where "Ours w/o SBP" denotes a parametric model with Gaussian distribution instead of SBP, "Ours w/o self-attention" means that the weighted adjacency matrix is randomly set rather than learned by the proposed self-attention mechanism, and "Ours w/o self-attention + DAG"

Datasets	20News		Reuters		Wikitext-103	
Model	50	200	50	200	50	200
GSM	0.211	0.165	0.198	0.155	0.214	0.217
GSB	0.231	0.191	0.152	0.136	0.229	0.131
NB-NTM	0.188	0.223	0.248	0.245	0.127	0.125
GNB-NTM	**0.240**	0.228	0.237	0.255	0.127	0.093
HDP	0.192		0.266		0.157	
iTM-VAE	0.195		0.201		0.184	
HiTM-VAE	0.237		**0.269**		0.233	
rCRP	0.186		0.206		0.201	
hLDA	0.221		0.185		0.186	
TSNTM	0.212		0.206		0.213	
nTSNTM	0.219		0.234		0.237	
hPAM	0.213		0.229		0.223	
HSOC	0.223		0.210		-	
CluHTM	0.219		0.161		-	
Ours	0.235		0.251		**0.240**	

Table 3: The average NPMI scores of different models using top 5, 10, and 15 words for each topic. A higher score means better performance and the best scores are highlighted by boldface.

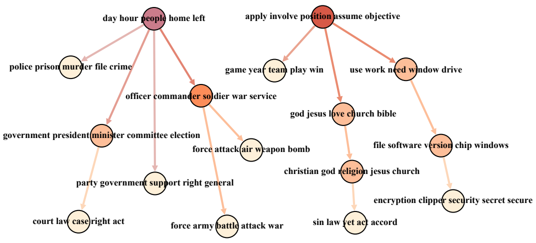

day hour people home left
police prison murder file crime
officer commander soldier war service
government president minister committee election
force attack air weapon bomb
party government support right general
court law case right act
force army battle attack war
apply involve position assume objective
game year team play win
use work need window drive
god jesus love church bible
file software version chip windows
christian god religion jesus church
encryption clipper security secret secure
sin law yet act accord

Figure 5: Two examples of topic trees generated by nFNTM on Wikitext-103 (left) and 20News (right). Each node represents a topic with top 5 words, and the arrow direction is from parent to children.

denotes reconstructing documents by their topic distributions generated from SBP. Note that there is a flat topic hierarchy without DAG, thus the CLNPMI metric can not be calculated. As shown in Table 4, nonparametric and self-attention modules are beneficial to generate a good topic hierarchy and achieve improvements in topic interpretability.

4.5 Concentration Parameter Evaluation

Here, we vary the values of concentration parameter $\hat{\beta}$ to validate the nonparametric property of our model on 20News. Figure 6 shows that the topic numbers of all nonparametric models are promoted by increasing the value of $\hat{\beta}$, which is reasonable since a larger $\hat{\beta}$ leads to a smoother distribution of SBP, and the smoother distribution generates more topic numbers than a denser distribution. It also demonstrates that our model generates more topics for a larger $\hat{\beta}$ since these topics are dispersed (Wu et al., 2020), and our model performs better than other models on approximating the nonparametric property of HDP.

Model	#Topics	NPMI	CLNPMI
Ours	61	**0.235**	**0.152**
Ours w/o SBP	50	0.223	0.135
Ours w/o self-attention	57	0.209	0.097
Ours w/o self-attention+DAG	48	0.203	-

Table 4: Ablation evaluation on 20News.

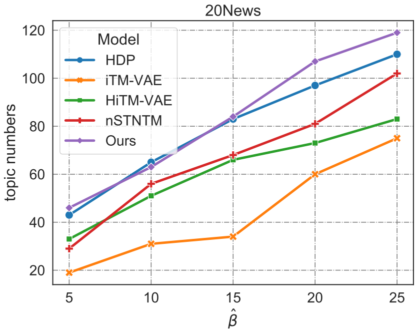
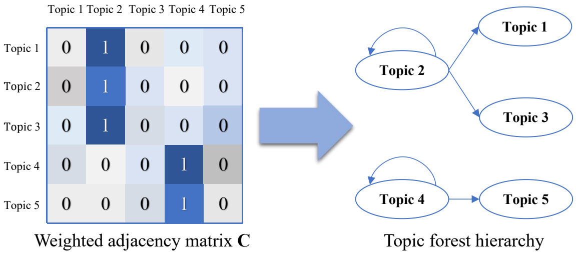


Figure 6: Topic numbers derived by different nonparametric models with various values of $\hat{\beta}$.

4.6 Forest-Structured Topic Visualization

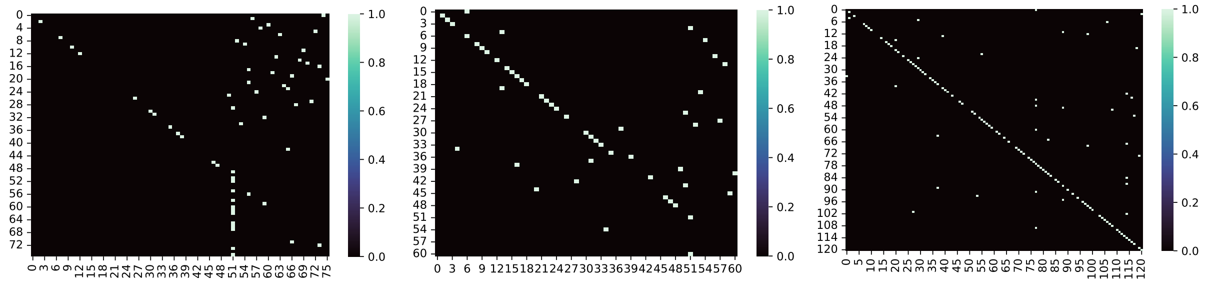
In this part, we qualitatively analyze the rationality of the topic forest generated by our model. We determine parent-child topic relationships by the value of the weighted adjacency matrix $\mathbf{C}$. $C_{i,j} \approx 1$ means that child topic t_i connects to parent topic t_j. For clarity, we show an example of $\mathbf{C}$ and how to build a topic forest from it in Figure 7. Such a topic hierarchy is based on the assumption that a document can be decomposed into a weighted sum of multiple topics, where a topic can also be decomposed into a weighted sum of multiple subtopics (Chen et al., 2021b).



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	0	1	0	0	0
Topic 2	0	1	0	0	0
Topic 3	0	1	0	0	0
Topic 4	0	0	0	1	0
Topic 5	0	0	0	1	0

Figure 7: An example of building a topic forest from a weighted adjacency matrix.

Figure 8(a) shows the visualization of the weighted adjacency matrix generated from Reuters. Except for the dots on the diagonal, the rest of



(a) Reuters

(b) 20News

(c) Wikitex-103

Figure 8: Visualization of the weighted adjacency matrix generated from (a) Reuters, (b) 20News, and (c) Wikitext-103, respectively. The white dot means that there is a correlation between two topics indexed in the row and column.

dots present a discrete distribution to meet the constraints of DAG. In addition, we observe that topic 51 is connected to various topics and becomes the parent topic of multiple topics, which is a general root topic of the whole corpus. Figures 8(b) and 8(c) present the weighted adjacency matrices generated by our model on 20News and Wikitext-103, respectively. The results also indicate that except for the dots on the diagonal, the rest of dots present a discrete distribution on both datasets, i.e., they meet the constraints of DAG.

Besides, we use Gephi[12] to visualize the topic forests generated by our model over 20News, Reuters, and Wikitext-103 in Figures 9, 10, and 11, where each node represents a topic with top 5 words and the arrow direction is from parent to children. The color of each node is determined by indegree. It indicates that the topic forest generated from Reuters contains a master tree whose root topic is about interviews. Furthermore, the topic forests generated from 20News and Wikitext-103 both consist of multiple trees, where the root topics mainly involve news and daily, respectively. All results indicate a rational hierarchical structure of topics obtained by our model, i.e., root topics are general and children topics are specific to their corresponding parents.

4.7 Speed Comparison

For speed comparison, we record the running time by taking the average cost of training each model for 5 times over 10k documents sampled from 20News. The running time for baseline methods are: hLDA - 16.74s , HSOC - 50.97s, CluHTM - 49.72s, and nTSNTM - 1.57s. Our model takes 1.35s, which is significantly faster than models based on Bayesian learning and NMF. Although nTSNTM achieves a competitive result, it applies Kumaraswamy distribution in parameter inference which slows down the model speed. Besides, such a tree-structured method is limited to local information and data sparsity along the hierarchy of topics (Viegas et al., 2020). In particular, the deeper the topic tree, the less coincident data remains. This deteriorates the quality of tree-structured methods inevitably. Compared with a single deep tree, the forest structure with several shallower trees could circumvent the above problem theoretically. Furthermore, with varied numbers of trees, a topic forest has a more flexible hierarchical structure and better adaptability than a fixed tree structure.

5 Conclusion

In this work, we present a forest-structured neural topic model named nFNTM. Particularly, the self-attention mechanism is applied to capture topic correlation and an augmented Lagrangian method is employed to build the topic forest under DAG constraints. By computing reparameterization of Gamma distribution, the SBP is adopted to obtain a nonparametric model. We empirically demonstrate that our nFNTM generates more rational, affinitive, and diverse topics than state-of-the-art hierarchical topic models, meanwhile achieves better topic interpretability than various baselines. Although flat topic models can generate distinguishable topics with larger NPMI scores than hierarchical ones, they treat all topics independently without explaining the corpus well to a user.

Acknowledgements

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 (61972426).

[12] http://gephi.org/

References

Amr Ahmed, Liangjie Hong, and Alexander Smola. 2013. Nested chinese restaurant franchise process: Applications to user tracking and document modeling. In *ICML*, pages 1426–1434.

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.

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 *ACL/IJCNLP*, pages 2343–2353.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. 2018. Implicit reparameterization gradients. In *NeurIPS*, pages 439–450.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*.

Hemant Ishwaran and Lancelot F. James. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173.

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

M.C. Jones. 2009. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6(1):70–81.

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 Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

P. Kumaraswamy. 1980. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1):79–88.

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.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, pages 577–584.

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.

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

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

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.

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

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.

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

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.

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.

Shunyu Yao, Binghui Peng, Christos H. Papadimitriou, and Karthik Narasimhan. 2021. Self-attention networks can process bounded hierarchical languages. In *ACL/IJCNLP*, pages 3770–3785.

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

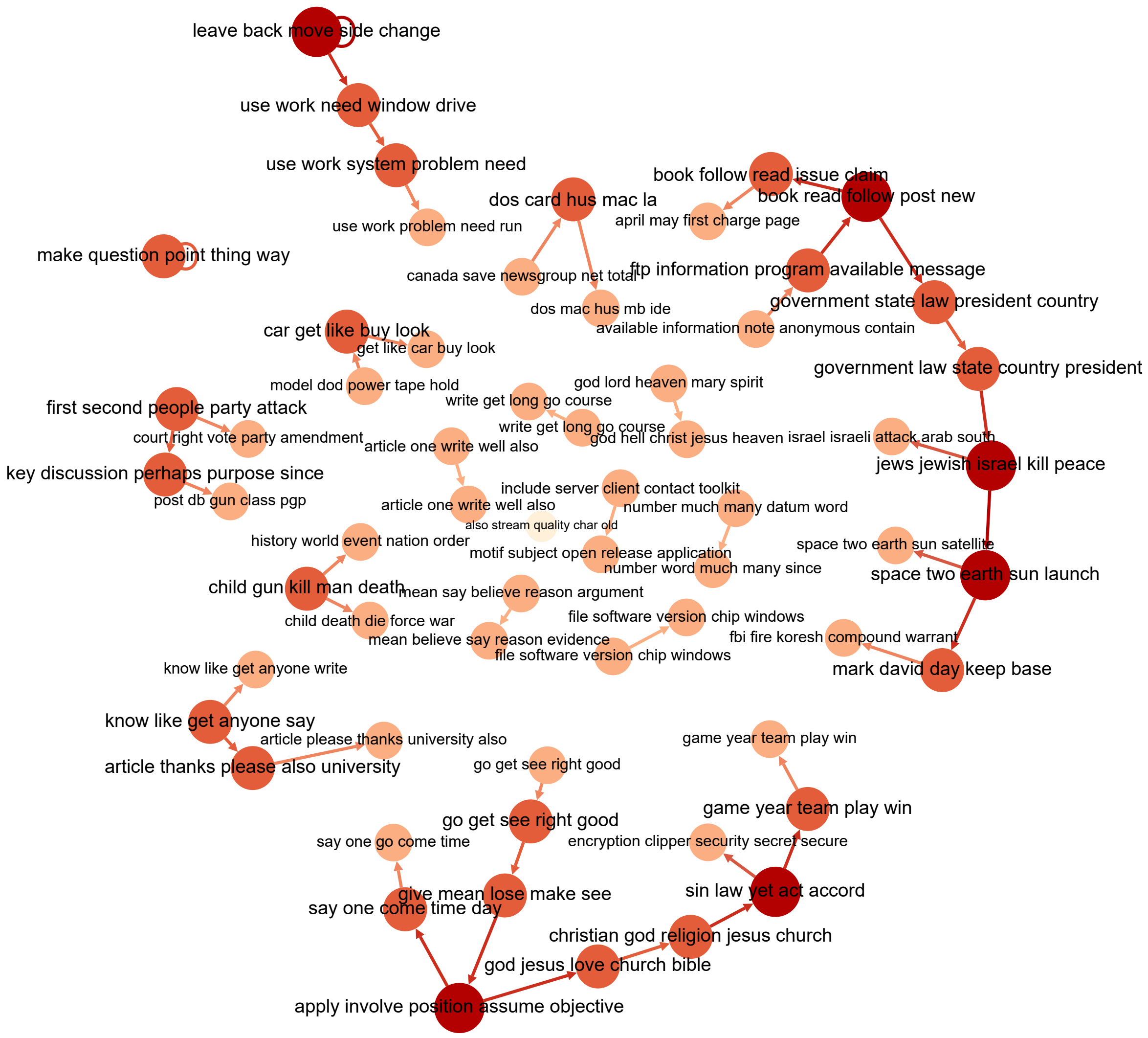

leave back move side change
use work need window drive
use work system problem need
use work problem need run
make question point thing way
dos card hus mac la
canada save newsgroup net total
dos mac hus mb ide
book follow read issue claim
book read follow post new
april may first charge page
ftp information program available message
government state law president country
available information note anonymous contain
government law state country president
car get like buy look
get like car buy look
model dod power tape hold
first second people party attack
court right vote party amendment
key discussion perhaps purpose since
post db gun class pgp
god lord heaven mary spirit
write get long go course
write get long go course
god hell christ jesus heaven israel israeli attack arab south
article one write well also
jews jewish israel kill peace
include server client contact toolkit
article one write well also
number much many datum word
also stream quality char old
history world event nation order
motif subject open release application
number word much many since
space two earth sun satellite
space two earth sun launch
child gun kill man death
mean say believe reason argument
file software version chip windows
child death die force war
mean believe say reason evidence
file software version chip windows
fbi fire koresh compound warrant
mark david day keep base
know like get anyone write
know like get anyone say
article please thanks university also
article thanks please also university
game year team play win
go get see right good
game year team play win
go get see right good
say one go come time
encryption clipper security secret secure
give mean lose make see
say one come time day
sin law yet act accord
christian god religion jesus church
god jesus love church bible
apply involve position assume objective

Figure 9: Visualization of the topic forest generated from 20News. Each node represents a topic and the arrow direction is from parent to children. There are about 8 trees in the forest, two of which have complex levels and structures, and two trees have only one root node. Take three of these trees as an illustration, the top 5 words of root topics and their children topics are (1) [know, like, get, anyone, say]⟶[article, thanks, please, also, university]; (2) [apply, involve, position, assume, objective]⟶[god, jesus, love, church, bible] and [game, year, team, play, win]; (3) [available, information, note, anonymous, contain]⟶[ftp, information, program, available, message]⟶[book, read, follow, post, new]. The root topics are general while children topics are specific.

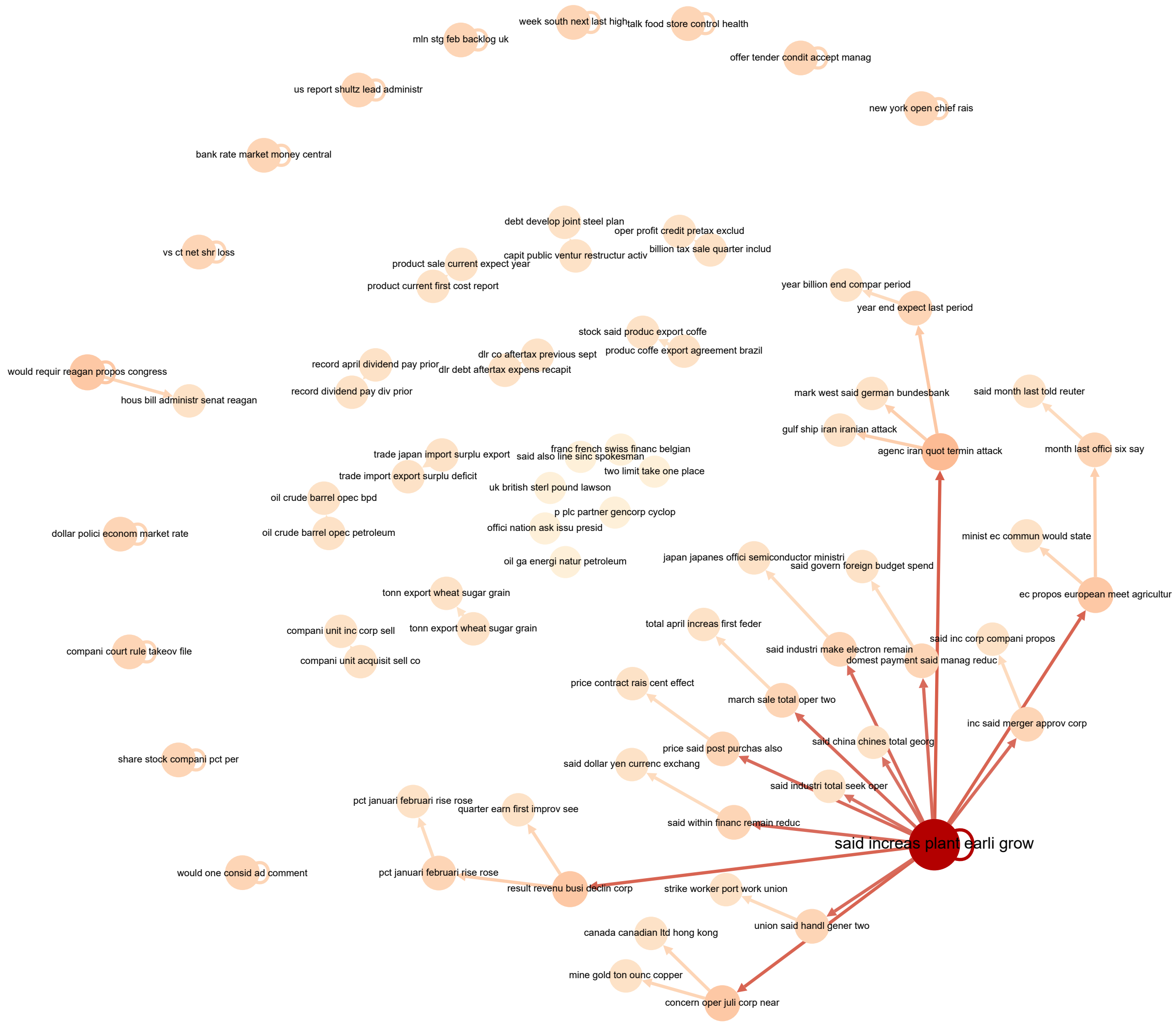

week south next last high
talk food store control health
mln stg feb backlog uk
offer tender condit accept manag
us report shultz lead administr
new york open chief rais
bank rate market money central
debt develop joint steel plan
oper profit credit pretax exclud
vs ct net shr loss
capit public ventur restructur activ
billion tax sale quarter includ
product sale current expect year
product current first cost report
year billion end compar period
year end expect last period
stock said produc export coffe
produc coffe export agreement brazil
dlr co aftertax previous sept
record april dividend pay prior
dlr debt aftertax expens recapit
would requir reagan propos congress
hous bill administr senat reagan
record dividend pay div prior
mark west said german bundesbank
said month last told reuter
gulf ship iran iranian attack
agenc iran quot termin attack
month last offici six say
trade japan import surplu export
franc french swiss financ belgian
said also line sinc spokesman
two limit take one place
trade import export surplu deficit
uk british sterl pound lawson
oil crude barrel opec bpd
p plc partner gencorp cyclop
offici nation ask issu presid
dollar polici econom market rate
oil crude barrel opec petroleum
minist ec commun would state
oil ga energi natur petroleum
japan japanes offici semiconductor ministri
said govern foreign budget spend
ec propos european meet agricultur
tonn export wheat sugar grain
total april increas first feder
compani unit inc corp sell
tonn export wheat sugar grain
said inc corp compani propos
said industri make electron remain
domest payment said manag reduc
compani court rule takeov file
compani unit acquisit sell co
price contract rais cent effect
march sale total oper two
inc said merger approv corp
said china chines total georg
price said post purchas also
share stock compani pct per
said dollar yen currenc exchang
said industri total seek oper
pct januari februari rise rose
quarter earn first improv see
said within financ remain reduc
said increas plant earli grow
would one consid ad comment
pct januari februari rise rose
result revenu busi declin corp
strike worker port work union
union said handl gener two
canada canadian ltd hong kong
mine gold ton ounc copper
concern oper juli corp near

Figure 10: Visualization of the topic forest generated from Reuters, where a huge tree is surrounded by several trees with only two levels. The root topic of the largest tree contains words: [said, increase, plan, earli, grow], which are commonly used during interviews. The surrounding areas are related to the topics of various categories, including [mine, gold, ton, ounce, copper], [strike, worker, port, work, union], and [year, end, expect, last, period].

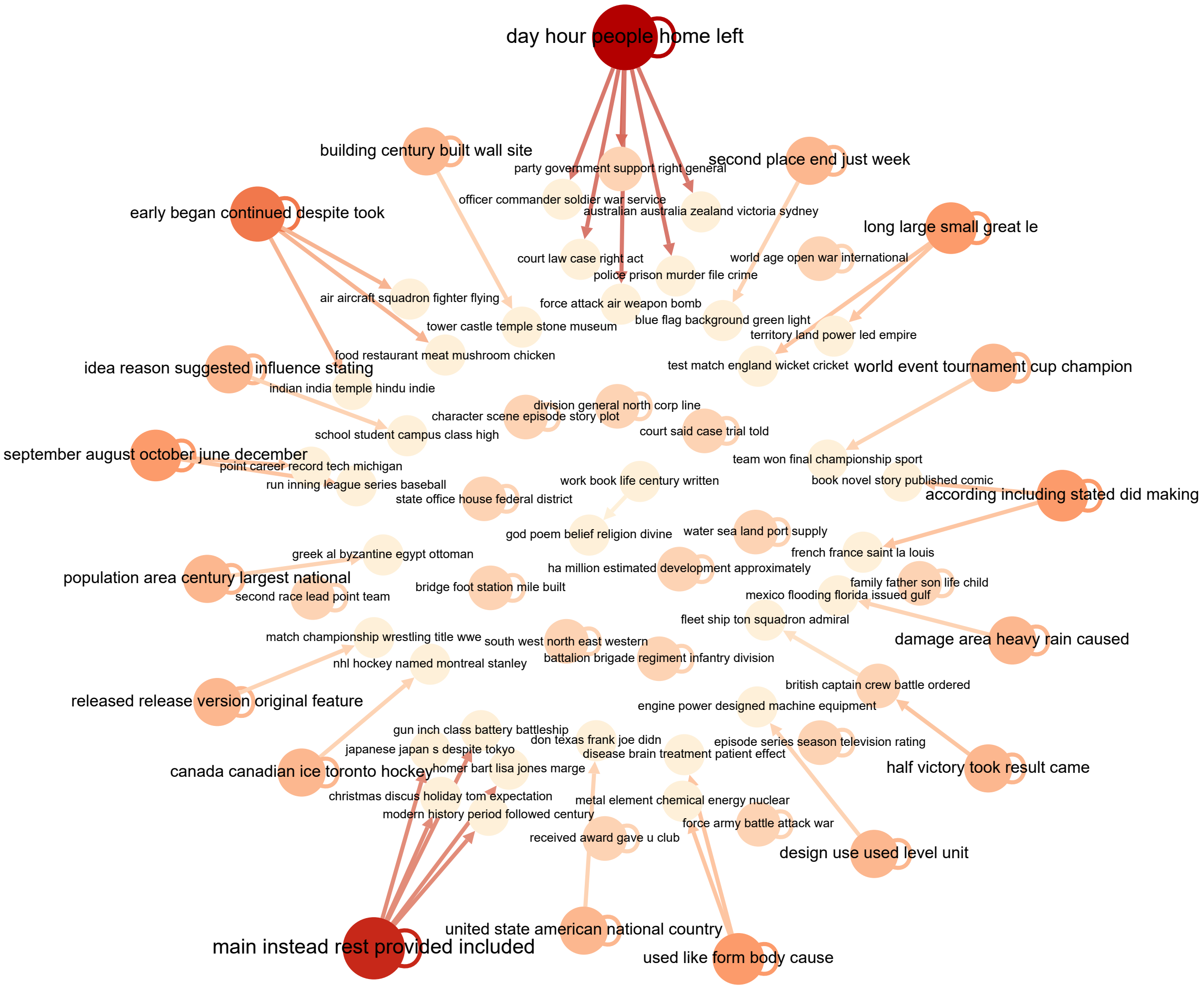

day hour people home left
building century built wall site
party government support right general
second place end just week
early began continued despite took
officer commander soldier war service
australian australia zealand victoria sydney
long large small great le
court law case right act
world age open war international
police prison murder file crime
air aircraft squadron fighter flying
force attack air weapon bomb
blue flag background green light
tower castle temple stone museum
territory land power led empire
food restaurant meat mushroom chicken
idea reason suggested influence stating
test match england wicket cricket
world event tournament cup champion
indian india temple hindu indie
division general north corp line
character scene episode story plot
court said case trial told
school student campus class high
september august october june december
team won final championship sport
point career record tech michigan
work book life century written
book novel story published comic
run inning league series baseball
according including stated did making
state office house federal district
god poem belief religion divine
water sea land port supply
french france saint la louis
greek al byzantine egypt ottoman
ha million estimated development approximately
population area century largest national
bridge foot station mile built
family father son life child
second race lead point team
mexico flooding florida issued gulf
fleet ship ton squadron admiral
match championship wrestling title wwe
south west north east western
damage area heavy rain caused
nhl hockey named montreal stanley
battalion brigade regiment infantry division
british captain crew battle ordered
released release version original feature
engine power designed machine equipment
gun inch class battery battleship
don texas frank joe didn
episode series season television rating
japanese japan s despite tokyo
disease brain treatment patient effect
canada canadian ice toronto hockey
homer bart lisa jones marge
half victory took result came
christmas discus holiday tom expectation
metal element chemical energy nuclear
modern history period followed century
force army battle attack war
received award gave u club
design use used level unit
united state american national country
main instead rest provided included
used like form body cause

Figure 11: Visualization of the topic forest generated from Wikitext-103. Possibly due to a large variety of topics in this corpus, the result presents a more dispersed and shallower structure than others. The top words of two root topics and their children topics are (1) [day, hour, people, home, left]⟶[court, law, case, right, act] and [party, government, support, right, general]; (2) [main, instead, rest, provides, included]⟶[christmas, discus, holiday, tom, expectation] and [gun, inch, class, battery, battleship].