Dynamic Structured Neural Topic Model with Self-Attention Mechanism

Nozomu Miyamoto[1] **Masaru Isonuma**[1] **Sho Takase**[2]
Junichiro Mori[1,3] **Ichiro Sakata**[1]
[1] The University of Tokyo [2] Tokyo Institute of Technology
[3] RIKEN Center for Advanced Intelligence Project
{nmiyamoto, isonuma, isakata}@ipr-ctr.t.u-tokyo.ac.jp
sho.takase@linecorp.com mori@mi.u-tokyo.ac.jp

Abstract

This study presents a *dynamic structured neural topic model*, which can handle the time-series development of topics while capturing their dependencies. Our model captures the topic branching and merging processes by modeling topic dependencies based on a self-attention mechanism. Additionally, we introduce citation regularization, which induces attention weights to represent citation relations by modeling text and citations jointly. Our model outperforms a prior dynamic embedded topic model (Dieng et al., 2019) regarding perplexity and coherence, while maintaining sufficient diversity across topics. Furthermore, we confirm that our model can potentially predict emerging topics from academic literature.

1 Introduction

Topic models are dominant tools for discovering the underlying semantic structure in a collection of documents. As a part of such topic models that can capture the chronological transition of topics have been intensively studied in recent years.

The dynamic topic model (**DTM**; Blei and Lafferty, 2006) is a pioneering work that captures the time-series evolution of topics. It successfully visualizes the changes in the topic proportion and the word distributions of each topic over time. Recently, neural networks have empowered topic models to handle a significant collection of documents. The dynamic embedded topic model (**D-ETM**; Dieng et al., 2019) introduces word embeddings and amortized variational inference into DTM, which significantly improves topic quality while reducing computational time. D-ETM is widely applied to large-scale time-series documents, such as scientific papers and social media (Churchill and Singh, 2022; Murakami et al., 2021).

However, DTM and D-ETM assume that topics evolve independently without interaction. This assumption is inappropriate, particularly for modeling scientific papers, where documents are dependent on each other through citation relations. For example, the recent *text-to-image* techniques are evolved from multiple topics such as image processing, natural language processing, and deep learning. Conventional dynamic topic models cannot capture how past topics contributed to the emergence of new topics. Furthermore, these models cannot predict emerging topics because the posterior word distribution of topics is parameterized for each time step, as explained later in detail.

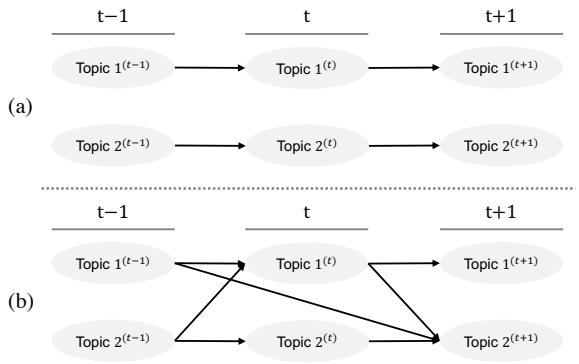

Figure 1: Comparison between (a) DTM/D-ETM and (b) DSNTM (ours)

To overcome these challenges, we propose a dynamic structured neural topic model (**DSNTM**), which captures the dependencies among topics over time (Fig. 1). Specifically, DSNTM models topic dependencies based on a self-attention mechanism (Vaswani et al., 2017; Lin et al., 2017), which reveals how past topics branches or merges into new topics. We can quantitatively evaluate which past topics contributes to the emergence of new topic by observing attention weights. In addition, the self-attention mechanism shares the parameters used for inferring topics at each time step, enabling the model to predict emerging topics.

Additionally, we introduce citation regularization, which induces attention weights to reflect the citation relations among documents. Citation

Findings of the Association for Computational Linguistics: ACL 2023, pages 5916–5930

regularization enables DSNTM to model text and citation jointly, improving the inferred topics' quality and accurately capturing their transitions. Due to the high expressive power of the self-attention mechanism and additional citation information, our model can capture the complex branching and merging processes of topics over time.

In the experiment, we used datasets consisting of over 20,000 scientific papers on computer science and natural language processing retrieved from the Semantic Scholar Open Research Corpus (**S2ORC**; Lo et al., 2020). Experimental results show that DSNTM outperforms recent neural topic models (Dieng et al., 2019, 2020) regarding perplexity and coherence, while maintaining sufficient diversity across topics. We also confirmed that DSNTM accurately captures the topic branching and merging processes and can potentially predict emerging topics in the academic literature.

2 Related Work

Extending the dynamic topic model (Blei and Lafferty, 2006), several variants have been proposed, such as the dependent Dirichlet processes mixture model (Lin et al., 2010), infinite dynamic topic model (Ahmed and Xing, 2010), and D-ETM (Dieng et al., 2019). However, these studies treat time-series changes in topics independently and cannot capture dependencies among topics.

Relating to structured topic models, several models have been proposed such as tree-structured topic model (Griffiths et al., 2003; Isonuma et al., 2020; Chen et al., 2021) and pachinko allocation models (Li and McCallum, 2006; Mimno et al., 2007). In addition, several studies have modeled the structure among documents by jointly modeling citation network and text (Nallapati et al., 2008; Tu et al., 2010; Chang and Blei, 2010; Lim and Buntine, 2015). However, these studies were not intended to track the time series transitions of topics.

The dynamic and static topic model (**DSTM**; Hida et al., 2018) extends the pachinko allocation model to capture the dynamic structure over time and the static structure among topics at each time step. DSTM has several drawbacks against our model. It cannot capture topic dependencies across multiple time steps, and thus, incorporating citation information is challenging. In addition, it cannot predict emerging topics as topic transitions are parameterized for each time step. Moreover, collapsed Gibbs sampling is used to infer posteriors, which is not scalable for large datasets.

The dynamic topic model on networked documents (**NetDTM**; Zhang and Lauw, 2022) models time-series documents and citation networks simultaneously. However, NetDTM does not capture the relations between topics nor predict emerging topics, which significantly differ from ours.

Contrary to the aforementioned studies, our work introduces amortized variational inference using the self-attention mechanism. This inference technique enables us to capture the topic branching and merging process across multiple time steps with significant scalability. Furthermore, our model can predict emerging topics from past topics.

3 Background

We first review the embedded topic model (**ETM**; Dieng et al., 2020), and then explain D-ETM, which combines ETM with DTM before introducing our DSNTM.

3.1 Embedded Topic Model (ETM)

ETM (Dieng et al., 2020) is a topic model that introduces word embeddings into LDA. The generative process of documents is the following:

1. For each document index $d \in \{1, \dots, D\}$:
   Draw topic proportion: $\boldsymbol{\theta}_d \sim \mathcal{LN}(\mathbf{0}, \boldsymbol{I})$ (1)
2. For each word index $n \in \{1, \dots, N_d\}$ in d:
   Draw topic assignment: $z_{d,n} \sim \mathrm{Cat}(\boldsymbol{\theta}_d)$ (2)
   Draw word: $w_{d,n} \sim \mathrm{Cat}(\boldsymbol{\beta}_{z_{d,n}})$ (3)

Here, $\mathrm{Cat}(\cdot)$ and $\mathcal{LN}(\cdot, \cdot)$ denote the categorical distribution and logistic-normal distribution (Atchison and Shen, 1980), respectively. $\boldsymbol{\beta}_k \in \mathbb{R}^V$ represents the word distribution of the k^{th} topic computed as follows:

$$\boldsymbol{\beta}_k = \mathrm{softmax}(\boldsymbol{\rho}^\top \boldsymbol{\alpha}_k). \quad (4)$$

where $\boldsymbol{\rho} \in \mathbb{R}^{L \times V}$ denotes the L-dimensional word embeddings of the entire vocabulary. The $\boldsymbol{\rho}_v \in \mathbb{R}^L$ corresponds to the v^{th} word embedding. $\boldsymbol{\alpha}_k \in \mathbb{R}^L$ denotes the embedding representation of the k^{th} topic in the semantic space of words, which is called topic embedding.

3.2 Dynamic Embedded Topic Model (D-ETM)

D-ETM (Dieng et al., 2019) analyzes time-series documents by changing the topics over time. Contrary to ETM, D-ETM assumes a discrete-time

Markov chain for the topic embedding in Eq. (5) and the topic proportion mean in Eq. (7). The generative process of documents is described as follows:

1. For each time step $t \in \{1, \dots, T\}$:

   Draw word distribution for each topic k:

$$\boldsymbol{\alpha}_k^{(t)} \sim \mathcal{N}(\boldsymbol{\alpha}_k^{(t-1)}, \sigma^2 \boldsymbol{I}) \quad (5)$$

$$\boldsymbol{\beta}_k^{(t)} = \mathrm{softmax}(\boldsymbol{\rho}^\top \boldsymbol{\alpha}_k^{(t)}) \quad (6)$$

   Draw topic proportion mean:

$$\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{\eta}_{t-1}, \delta^2 I) \quad (7)$$

2. For each document index $d \in \{1, \dots, D\}$:

   Draw topic proportion: $\boldsymbol{\theta}_d \sim \mathcal{LN}(\boldsymbol{\eta}_{t_d}, \gamma^2 \boldsymbol{I})$ (8)

3. For each word index $n \in \{1, \dots, N_d\}$ in d:

   Draw topic assignment: $z_{d,n} \sim \mathrm{Cat}(\boldsymbol{\theta}_d)$ (9)

   Draw word: $w_{d,n} \sim \mathrm{Cat}(\boldsymbol{\beta}_{z_{d,n}}^{(t_d)})$ (10)

where $\boldsymbol{\alpha}_k^{(t)}$ and $\boldsymbol{\beta}_k^{(t)}$ are the topic embedding and word distribution assigned to the k^{th} topic in the t^{th} time step, respectively. σ, δ , and γ are model hyperparameters, which control the variance of normal distributions.

Dieng et al. (2019) approximate the posterior distribution of $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ with amortized variational inference (Kingma and Welling, 2014; Rezende et al., 2014). Particularly, for the topic embeddings $\boldsymbol{\alpha}$, the mean-field family is used for the approximation as follows:

$$q(\boldsymbol{\alpha}_k^{(t)}) = \mathcal{N}(\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\sigma}_k^{(t)}) \quad (11)$$

$$q(\boldsymbol{\alpha}) = \prod_k \prod_t q(\boldsymbol{\alpha}_k^{(t)}) \quad (12)$$

where $\boldsymbol{\mu}_k^{(t)} \in \mathbb{R}^L$ and $\boldsymbol{\sigma}_k^{(t)} \in \mathbb{R}^L$ are learnable vectors representing the mean and variance of $\boldsymbol{\alpha}_k^{(t)}$, respectively.

However, this mean-field approximation has two limitations.

Dependencies among topics cannot be modeled Eq. (12) assumes that topics are independent of each other. This assumption is typically inappropriate for time-series documents. For instance, academic topics sometimes emerge from interactions among several past topics. Topic dependencies must be modeled to consider such interactions.

Emerging topics cannot be predicted D-ETM infers a topic by parameterizing $\boldsymbol{\mu}_k^{(t)}$ and $\boldsymbol{\sigma}_k^{(t)}$ for each time step t. As documents in the t^{th} time step are used to infer these parameters, topics cannot be inferred for the time steps that are not contained in the dataset. The parameters must be shared across all time steps to predict emerging topics.

To overcome these limitations, our DSNTM introduces the self-attention mechanism to infer those parameters. The self-attention mechanism enables DSNTM to capture the topic dependencies, while sharing the parameters across all time steps.

4 Dynamic Structured Neural Topic Model (DSNTM)

This section describes the proposed DSNTM. The generative process of documents is the same as that of D-ETM.

4.1 Inference of Topic Embeddings

Contrary to D-ETM, we use structured variational inference to infer the topic embeddings. We compute a topic embedding from all previous topic embeddings using the self-attention mechanism.

$$\tilde{\boldsymbol{\alpha}}_k^{(t)} = \mathrm{self-attention}(\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) \quad (13)$$

$$q(\boldsymbol{\alpha}_k^{(t)} | \tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) = \mathcal{N}(f_\mu(\tilde{\boldsymbol{\alpha}}_k^{(t)}), f_\sigma(\tilde{\boldsymbol{\alpha}}_k^{(t)})) \quad (14)$$

where $\tilde{\boldsymbol{\alpha}}_k^{(t)} \in \mathbb{R}^L$ denotes the transformed topic embedding. f_μ and f_σ are multi-layer perceptrons (**MLP**) that convert $\tilde{\boldsymbol{\alpha}}_k^{(t)}$ to a variational normal distribution.

Computation of Self-attention We present an outline of the self-attention mechanism in Fig. 2. To calculate Eq. (13), we obtain the *key* $\boldsymbol{K}_{1:K}^{(1:t-1)} \in \mathbb{R}^{K(t-1) \times L}$ and *value* $\boldsymbol{V}_{1:K}^{(1:t-1)} \in \mathbb{R}^{K(t-1) \times L}$ from all previous transformed topic embeddings $\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}$. On the other hand, the *query* $\boldsymbol{q}_k^{(t-1)} \in \mathbb{R}^L$ is obtained from the k^{th} transformed topic embedding $\tilde{\boldsymbol{\alpha}}_k^{(t-1)}$ at time step $t-1$.

$$\boldsymbol{K}_{1:K}^{(1:t-1)} = f_k(\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) \quad (15)$$

$$\boldsymbol{V}_{1:K}^{(1:t-1)} = f_v(\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) \quad (16)$$

$$\boldsymbol{q}_k^{(t-1)} = f_q(\tilde{\boldsymbol{\alpha}}_k^{(t-1)}) \quad (17)$$

where f_k, f_v and f_q denote MLPs. Subsequently, we compute the attention weight $\boldsymbol{a}_k^{(t)} \in \mathbb{R}^{K(t-1)}$ between each past topic and the k^{th} topic at time step t, similar to Vaswani et al. (2017).

$$\boldsymbol{a}_k^{(t)} = \mathrm{softmax}\left(\frac{\boldsymbol{q}_k^{(t-1)} \boldsymbol{K}_{1:K}^{(1:t-1)\top}}{\sqrt{L}}\right) \quad (18)$$

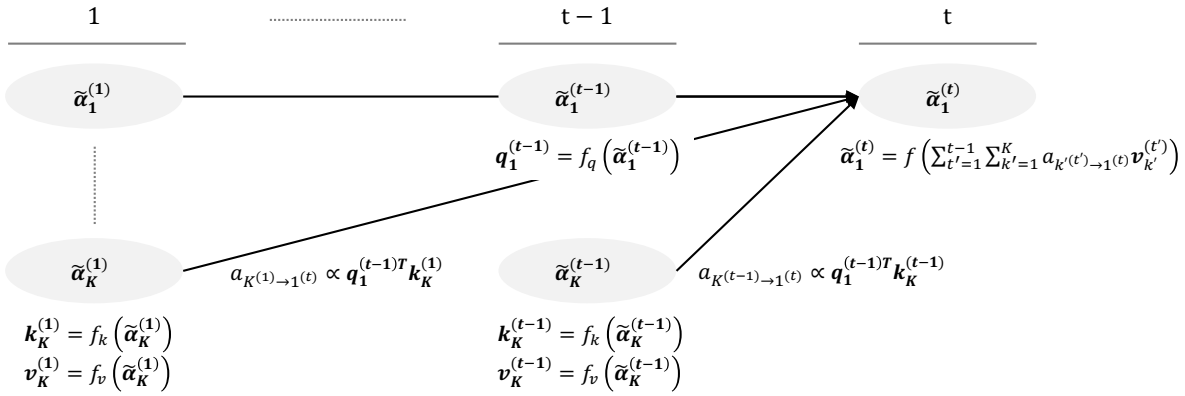

Figure 2: The process of generating the transformed topic embedding $\tilde{\alpha}$ by the self-attention mechanism. The parameters of f_k, f_v, and f_q are shared across all topics and time steps. $a_{k'^{(t')}\to k^{(t)}}$ is the attention weight between the k'^{th} topic at time t' and the k^{th} topic at time t. f denotes the residual connection and layer normalization.

Then, we obtain $\tilde{\boldsymbol{\alpha}}_k^{(t)}$ by calculating the sum of $\boldsymbol{V}_{1:K}^{(1:t-1)}$ weighted by the attention weights. We use a residual connection to obtain $\tilde{\boldsymbol{\alpha}}_k^{(t)}$ for preventing gradient explosion and disappearance (Simonyan and Zisserman, 2014).

$$\Delta\tilde{\boldsymbol{\alpha}}_k^{(t)} = \boldsymbol{a}_k^{(t)} \boldsymbol{V}_{1:K}^{(1:t-1)} \tag{19}$$

$$\tilde{\boldsymbol{\alpha}}_k^{(t)} = \mathrm{LayerNorm}(\Delta\tilde{\boldsymbol{\alpha}}_k^{(t)} + \tilde{\boldsymbol{\alpha}}_k^{(t-1)}) \tag{20}$$

Here, we use the layer normalization (Ba et al., 2016) to compute the transformed topic embeddings. At the time step $t = 1$, we initialize the transformed topic embeddings $\tilde{\boldsymbol{\alpha}}_k^{(1)}$ from a normal distribution.

$$\tilde{\boldsymbol{\alpha}}_k^{(1)} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{21}$$

Note that the prior distribution assumes that $\boldsymbol{\alpha}_k^{(t)}$ is drawn from a normal distribution as follows: $\boldsymbol{\alpha}_k^{(t)} \sim \mathcal{N}(\boldsymbol{\alpha}_k^{(t-1)}, \sigma^2 \boldsymbol{I})$. This assumption regularizes $\boldsymbol{\alpha}_k^{(t)}$ to be close to its previous topic $\boldsymbol{\alpha}_k^{(t-1)}$.

Motivations behind Self-attention We use the self-attention mechanism for the following reasons:

(1) Instead of the self-attention mechanism, we can also capture the dependencies among topics by simply parameterizing the topic embeddings as follows:

$$\tilde{\boldsymbol{\alpha}}_{1:K}^{(t)} = \boldsymbol{W} \tilde{\boldsymbol{\alpha}}_{1:K}^{(t-1)} \tag{22}$$

Here, $\boldsymbol{W} \in \mathbb{R}^{K\times K}$ is a learnable weight matrix, where $w_{i,j}$ represents the dependency between topic i and j. However, this parameterization cannot capture the dependencies across multiple time steps, and a method that allows an arbitrary number of inputs is required. The self-attention mechanism, which handles an arbitrary-length sequence, satisfies this requirement and captures the topic dependencies across multiple time steps.

(2) The self-attention mechanism parameterizes MLP f_q, f_k, and f_v to compute the embeddings of subsequent topics from previous topics. As the parameters of MLPs are shared over time, the self-attention mechanism allows DSNTM to predict the emerging topic embeddings.

4.2 Overall Inference and ELBO

Under our proposed probabilistic model, the likelihood of documents is given by

$$\begin{aligned}
&p(\boldsymbol{w}_{1:D}|\sigma,\delta,\gamma) \\
&= \int \Big\{\prod_d \prod_n \sum_{z_{d,n}} p(w_{d,n}|\boldsymbol{\beta}_{z_{d,n}}^{(t_d)}) p(z_{d,n}|\boldsymbol{\theta}_d) p(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d})\Big\} \\
&\quad \Big\{\prod_t \prod_k p(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}) p(\boldsymbol{\alpha}_k^t|\boldsymbol{\alpha}_k^{t-1})\Big\} d\boldsymbol{\theta} d\boldsymbol{\eta} d\boldsymbol{\alpha} \\
&= \int \Big\{\prod_d \prod_n (\boldsymbol{\beta}^{(t_d)} \cdot \boldsymbol{\theta}_d)_{w_{d,n}} p(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d})\Big\} \\
&\quad \Big\{\prod_t \prod_k p(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}) p(\boldsymbol{\alpha}_k^t|\boldsymbol{\alpha}_k^{t-1})\Big\} d\boldsymbol{\theta} d\boldsymbol{\eta} d\boldsymbol{\alpha}
\end{aligned} \tag{23}$$

Subsequently, let $q(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\alpha})$ be the variational distribution of the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\alpha}|\boldsymbol{w}_{1:D})$. Following D-ETM, $q(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\alpha})$ is computed as follows:

$$\begin{aligned}
&q(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\alpha}) \\
&= \prod_d q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d}, \boldsymbol{w}_d) \times \prod_t q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \tilde{\boldsymbol{w}}_t) \\
&\times \prod_t \prod_k q(\boldsymbol{\alpha}_k^{(t)}|\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)})
\end{aligned} \tag{24}$$

$$q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d}, \boldsymbol{w}_d) = \mathcal{LN}(f_\mu(\tilde{\boldsymbol{\theta}}_{t_d}), f_\sigma(\tilde{\boldsymbol{\theta}}_{t_d}))$$
$$\tilde{\boldsymbol{\theta}}_{t_d} = f_\theta([\boldsymbol{\eta}_{t_d}; \boldsymbol{w}_d]) \quad (25)$$

$$q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \tilde{\boldsymbol{w}}_t) = \mathcal{N}(f_\mu(\tilde{\boldsymbol{\eta}}_t), f_\sigma(\tilde{\boldsymbol{\eta}}_t))$$
$$\tilde{\boldsymbol{\eta}}_t = [\boldsymbol{h}_t; \boldsymbol{\eta}_{t-1}] \quad (26)$$

where $\boldsymbol{w}_d$ is the bag-of-words (BoW) representation of document d. $\boldsymbol{h}_t$ is the hidden state of a long-short term memory network (**LSTM**; Hochreiter and Schmidhuber, 1997) that uses the normalized BoW representation $\tilde{\boldsymbol{w}}_t$ of all documents at time t as input. $[\cdot;\cdot]$ denotes the concatenation of vectors.

The evidence lower bound (**ELBO**) for the document log-likelihood is derived as follows:

$$\begin{aligned} L_{doc} = & \sum_d \mathrm{E}_{q(\boldsymbol{\theta}_d)q(\boldsymbol{\eta}_t)q(\boldsymbol{\alpha}_k^{(t)})}\left[\boldsymbol{w}_d^\top \log(\boldsymbol{\beta}^{(t_d)} \cdot \boldsymbol{\theta}_d)\right] \\ & - \sum_t \sum_k \mathrm{D_{KL}}\left[q(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_{1:K}^{(1:t-1)})||p(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_k^{(t-1)})\right] \\ & - \sum_d \mathrm{D_{KL}}\left[q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d}, \boldsymbol{w}_d)||p(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d})\right] \\ & - \sum_t \mathrm{D_{KL}}\left[q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \tilde{\boldsymbol{w}}_t)||p(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1})\right] \end{aligned} \quad (27)$$

5 Citation Regularization

DSNTM has difficulty interpreting the attention weights among topics. Ideally, the attention should model the dependency among topics, representing citation relations between documents. Therefore, we let the attention weights to be interpretable by regularizing them to correspond with citation relations. Citation regularization also improves the quality of the inferred topics by jointly modeling the text and citations.

To regularize the attention weights, we model the citations between documents based on the topic proportion $\boldsymbol{\theta}$, the attention weights $\boldsymbol{a}$, and the paper proportion $\boldsymbol{\phi}$ as shown in Fig. 3. Formally, for each document pair $(i, j) \in \{1, \ldots, D\} \times \{1, \ldots, D\}$, the citation is modeled as follows:

1. Draw citing topic assignment: $z_i \sim \mathrm{Cat}(\boldsymbol{\theta}_i)$ (28)
2. Draw cited topic assignment: $z_j \sim \mathrm{Cat}(\boldsymbol{a}_{z_i}^{(t_i)})$ (29)
3. Draw cited document: $d_j \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_j})$ (30)

where $\boldsymbol{a}_k^{(t_i)} \in \mathbb{R}^{K(t_i-1)}$ is the attention weight, which denotes the probability distribution across all previous topics $z_j \in \{1, \ldots, K\} \times \{1, \ldots, t_i - 1\}$. The paper proportion $\boldsymbol{\phi}_k \in \mathbb{R}^D$ denotes the probability distribution of cited documents where a topic k is assigned, as explained in next section.

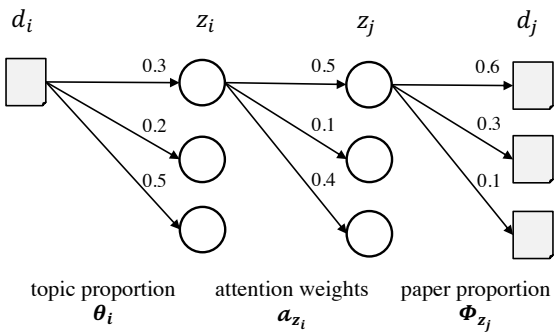

Figure 3: Citation modeled by topic proportion, attention weights, and paper proportion.

5.1 Obtaining paper proportion

From Bayes' theorem, we calculate the probability where a paper $d_j \in \{1, \ldots, D\}$ is cited according to a topic z_j as follows:

$$\begin{aligned} p(d_j|z_j) &= \frac{p(z_j|d_j)p(d_j)}{p(z_j)} \\ &\propto p(z_j|d_j) \end{aligned} \quad (31)$$

Here, we assume that the prior $p(d_j)$ is uniformly distributed, and $p(z_j)$ can be ignored because it is constant regardless of d_j. As $p(d_j|z_j) = \phi_{z_j}^{(d_j)}$ and $p(z_j|d_j) = \theta_{d_j}^{(z_j)}$, $\phi_{z_j}^{(d_j)}$ can be simply computed by normalizing $\theta_{d_j}^{(z_j)}$:

$$\phi_{z_j}^{(d_j)} = \frac{\theta_{d_j}^{(z_j)}}{\sum_{d_j} \theta_{d_j}^{(z_j)}} \quad (32)$$

5.2 Overall Inference and ELBO

Hence, under our modeling assumption, the likelihood of a citation $c_{i,j} \in \{0, 1\}$ is given by

$$\begin{aligned} & p(c_{i,j} = 1|\boldsymbol{\theta}, \boldsymbol{\alpha}) \\ & = \sum_k^K \sum_{k'}^K p(d_j|\boldsymbol{\phi}_{k'}) p(z_j = k'|\boldsymbol{a}_k^{(t_i)}) p(z_i = k|\boldsymbol{\theta}_i) \end{aligned} \quad (33)$$

where $c_{i,j} = 1$ indicates that the document d_i cites the document d_j. Finally, the likelihood of both documents and citations can be described as fol-

lows:

$$
\begin{aligned}
&p(\boldsymbol{w}_{1:D}, c_{1,1}, \ldots, c_{D,D} | \sigma, \delta, \gamma) \\
&= \int \Big\{ \prod_i \prod_n (\boldsymbol{\beta}^{(t_i)} \cdot \boldsymbol{\theta}_i)_{w_{i,n}} p(\boldsymbol{\theta}_i | \boldsymbol{\eta}_{t_i}) \Big\} \\
&\quad \Big\{ \prod_t \prod_k p(\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1}) p(\boldsymbol{\alpha}_k^t | \boldsymbol{\alpha}_k^{t-1}) \Big\} \\
&\quad \Big\{ \prod_i \prod_j p(c_{i,j} | \boldsymbol{\theta}, \boldsymbol{\alpha}) \Big\} d\boldsymbol{\theta} d\boldsymbol{\eta} d\boldsymbol{\alpha} \qquad (34)
\end{aligned}
$$

The ELBO for both document and citation log-likelihood is derived as follows:

$$
\begin{aligned}
L =& \sum_i \mathrm{E}_{q(\boldsymbol{\theta}_i)} \Big[\boldsymbol{w}_i^\top \log(\boldsymbol{\beta}^{(t_i)} \cdot \boldsymbol{\theta}_i) \Big] \\
&- \sum_t \sum_k \mathrm{D}_{\mathrm{KL}} \Big[q(\boldsymbol{\alpha}_k^{(t)} | \boldsymbol{\alpha}_{1:K}^{(1:t-1)}) || p(\boldsymbol{\alpha}_k^{(t)} | \boldsymbol{\alpha}_k^{(t-1)}) \Big] \\
&- \sum_i \mathrm{D}_{\mathrm{KL}} \Big[q(\boldsymbol{\theta}_i | \boldsymbol{\eta}_{t_i}, \boldsymbol{w}_i) || p(\boldsymbol{\theta}_i | \boldsymbol{\eta}_{t_i}) \Big] \\
&- \sum_t \mathrm{D}_{\mathrm{KL}} \Big[q(\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1}, \tilde{\boldsymbol{w}}_t) || p(\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1}) \Big] \\
&+ \sum_i \sum_j \mathrm{E}_{q(\boldsymbol{\theta}_i) q(\boldsymbol{\eta}_t) q(\boldsymbol{\alpha}_k^{(t)})} \Big[\log p(c_{i,j} | \boldsymbol{\theta}, \boldsymbol{\alpha}) \Big] \\
=& \, L_{doc} + L_{cit} \qquad (35)
\end{aligned}
$$

Here, L_{doc} is defined in Eq. (27), and L_{cit} is defined as the following equation:

$$
L_{cit} = \sum_i^D \sum_j^D \mathrm{BCE} \left[p(c_{i,j} | \boldsymbol{\theta}, \boldsymbol{\alpha}), c_{i,j} \right] \qquad (36)
$$

where BCE denotes the binary cross entropy.

6 Experiment

6.1 Experimental Setup[1]

Dataset In our experiments, we used *ACL* and *CS* dataset, which were based on the Semantic Scholar Open Research Corpus (S2ORC; Lo et al., 2020)[2]. S2ORC contains over 136 million academic papers, each of which contains publication year, abstract text, cited paper's data, ACL ID, and field of study. We used the abstracts of papers that are published at *ACL conferences (ACL ID is not "None") from 2006 to 2019 for *ACL* dataset. For *CS* dataset, we used the abstracts of papers where the field of study includes "Computer Science" and are published from 2006 to 2019. We used the top 40,000 papers w.r.t. the number of citations for *CS* dataset.

[1] https://github.com/miyamotononno/DSNTM

[2] https://github.com/allenai/s2orc

Dataset	ACL	CS
# of time steps	7	7
# of words in vocabulary	5,540	10,449
# of docs for training	14,110	23,991
# of docs for validation	4,704	7,997
# of docs for evaluation	4,704	7,998

Table 1: Summary statistics of the datasets.

The papers were randomly splitted into 3:1:1 ratio for training, validation, and evaluation. We also filtered out stop words, i.e., words with a document frequency of 70% or above, words appearing in less than ten documents, numbers, punctuation marks, and stop words used in Dieng et al. (2019). The papers published in two consecutive years were grouped into a single time step so that each time step contained a sufficient number of papers. For example, papers published between 2006 and 2007 were grouped into $t=1$. The statistics of the datasets are summarized in Table 1.

Baseline Methods As baseline methods, we measured the performance of ETM (Dieng et al., 2020) and D-ETM (Dieng et al., 2019) by using the published code of ETM[3] and D-ETM[4]. To evaluate the effectiveness of the self-attention mechanism, we also compared DSNTM that adopts a linear layer instead of the self-attention as shown in Eq. (22). We denote it by "DSNTM w/o self-attention."

Implementation Details The number of topics was set to 20 for all models and kept constant over time for fair comparison with the baseline models. Hyperparameters of each model were tuned based on the validation perplexity in *ACL* dataset. Further details are provided in Appendix A.

6.2 Experimental Results

We quantitatively evaluated the performance of topic models using the following three criteria. We run each model eight times and show the average performance and its 95% confidence interval in Table 2 and 3. Lower is better for perplexity, while higher is better for coherence and diversity.

Perplexity We used perplexity (Rosen-Zvi et al., 2004) to evaluate the generalization ability of topic models as a generative model. It measures the

[3] https://github.com/adjidieng/ETM

[4] https://github.com/adjidieng/DETM

	Perplexity	Coherence	Diversity
ETM (Dieng et al., 2020)	1,590.6± 2.4	0.023±0.003	**0.911±0.010**
D-ETM (Dieng et al., 2019)	1,187.8± 7.4	0.091±0.003	0.788±0.016
DSNTM w/o self-attention	1,260.5±37.8	0.054±0.005	0.631±0.043
DSNTM	1,079.9± 8.9	0.084±0.006	0.851±0.009
DSNTM w/ citation regularization	**1,054.0± 7.4**	**0.101±0.006**	0.895±0.014

Table 2: Evaluation of each model for *ACL* dataset.

	Perplexity	Coherence	Diversity
ETM (Dieng et al., 2020)	3,011.8± 4.4	0.022±0.002	**0.956±0.006**
D-ETM (Dieng et al., 2019)	2,519.0±41.8	0.078±0.004	0.954±0.010
DSNTM w/o self-attention	2,195.2±25.2	0.078±0.004	0.904±0.019
DSNTM	2,185.0±19.2	0.079±0.009	0.929±0.009
DSNTM w/ citation regularization	**2,156.8±30.5**	**0.105±0.005**	0.948±0.006

Table 3: Evaluation of each model for *CS* dataset.

ability to predict words in unseen documents. Perplexity is computed as follows:

$$\text{Perplexity} = \exp\Big(-\frac{\sum_{d=1}^{D} \log\, p(\boldsymbol{w}_d)}{\sum_{d=1}^{D} N_d}\Big) \quad (37)$$

where N_d is the number of words in the test document d. As computing $p(\boldsymbol{w}_d)$ is intractable, we calculated perplexity using ELBO following Miao et al. (2017); Srivastava and Sutton (2017).

Across the two datasets, our DSNTM achieved a lower perplexity than the baseline models. DSNTM outperformed D-ETM by a large margin specifically for *CS* dataset. In addition, DSNTM with citation regularization outperformed DSNTM, indicating that citation information contributed to the generalization ability of the topic model.

Coherence We measured topic coherence by calculating the average pointwise mutual information (Mimno et al., 2011) to assess the interpretability of topics. Specifically, we used the normalized pointwise mutual information (**NPMI**; Lau et al., 2014) of the two words included in the top 10 most likely words of the topic k.

$$\text{Coherence} = \frac{1}{K}\sum_{k}^{K}\frac{1}{45}\sum_{i=1}^{10}\sum_{j=i+1}^{10}\text{NPMI}(w_i^{(k)}, w_j^{(k)}) \quad (38)$$

NPMI is calculated using the following formula:

$$\text{NPMI}(w_i, w_j) = \frac{\log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{\log P(w_i, w_j)} \quad (39)$$

where $P(w_i, w_j)$ is the probability where words w_i and w_j co-occurs in a document, and $P(w_i)$ is the marginal probability of word w_i.

Our DSNTM significantly outperformed ETM and DSNTM without self-attention, while achieving a slightly lower score than D-ETM. However, the citation regularization let DSNTM outperform D-ETM. This result demonstrates that the topic interpretability is sufficiently ensured by the citation regularization.

Diversity We calculated the percentage of unique words in the top 25 frequent words of all topics to measure the diversity of topics following Dieng et al. (2020, 2019).

$$\text{Diversity} = \frac{N_u}{25K} \quad (40)$$

where N_u denotes the number of unique words that appear in all topics. Both models achieved competitive scores with the baseline models. This result indicates that our models improves the topic quality, while ensuring sufficient topic diversity.

7 Discussion

7.1 Visualization of Topic Transition

We discuss that the attention weights capture academic topics merging and branching processes.

Fig. 4 presents an example of topic merging on *CS* dataset. We show a topic about *motion tracking* in 2018-2019 (i.e., citing topic) and the two most influential topics on its emergence with respect to the attention weights in 2016-2017 (i.e.,

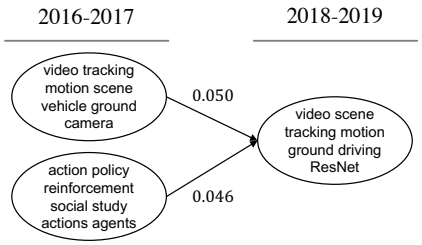

Figure 4: Example of topic merging process. We show the seven most frequent words in each topic. The edge values indicate the attention weight.

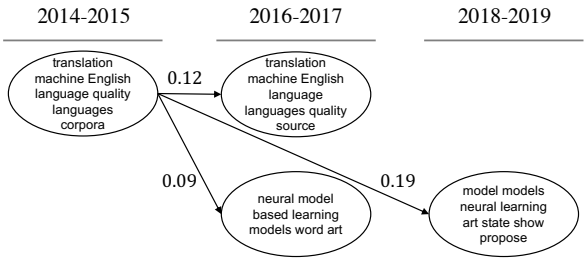

Figure 5: Example of topic branching process. We show the seven most frequent words in each topic. The edge values represent the influence on each topic, which is calculated in Appendix B.

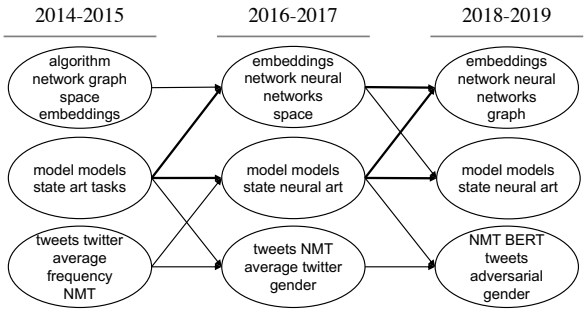

Figure 6: Example of topic transition. We present the top five most frequent words in each topic. The thickness of edges represent the attention weight (0.05 for thick edges and $0.02 - 0.05$ for thin edges).

cited topic). Each cited topic represents *motion tracking* and *reinforcement learning*. To investigate the validity of this topic merging, we checked the citation relations between the top 50 papers w.r.t. the citing topic's paper proportion and the top 50 papers w.r.t. the cited topic's paper proportion using the test dataset. While many papers on the citing topic refer to papers on motion tracking in previous years, some papers refer to papers on reinforcement learning. As reinforcement learning is used as the learning method of a tracker to achieve a light computation and satisfactory tracking accuracy for object tracking, the topic of reinforcement learning greatly influences the topic of object tracking. The attention weights reveal the merging process of academic topics.

Subsequently, we present an example of topic branching using *ACL* dataset (Fig. 5). We show a topic about *machine translation* in 2014-2015 (i.e., cited topic) and three subsequent topics that are most highly influenced by the cited topic (i.e., citing topic). Each citing topic describes *machine translation* (2016-2017) and *neural network* (2016-2017 and 2018-2019). We assessed the validity of this branching in the same manner as topic merging. As of 2014-2015, statistical machine translation (SMT) was predominant, whereas neural machine translation (NMT) was a nascent area in machine translation research. After 2016, NMT was intensively studied by incorporating SMT knowledge of SMT, while NMT models were imported into other text generation tasks (e.g., summarization). This trend induced the topics on neural network in 2016-2019. DSNTM successfully captures such topic branching processes in the academic literature.

Finally, we present an overview of the topic transition process using *ACL* dataset (Fig. 6). The topics in the first, second, and third rows represent *graph*, *neural network*, and *social media*, respectively. We can follow the prevalence of the neural network techniques to other research areas, such as graphs and social media, by observing the frequent words in the topic and attention weights. DSNTM enables us to grasp the current trends in the research area without following the citations of articles.

7.2 Prediction of Emerging Topics

In this section, we discuss the predictive performance of emerging topics using DSNTM on *CS* dataset. We trained DSNTM on the papers in 2006-2017 and predicted topics in 2018-2019 by computing the posterior word distribution for each topic using the self-attention mechanism (Eq. (6), (13), and (14)). We then prepared another model trained on the papers from 2006-2019. The topics inferred by this model are regarded as a proxy for the ground truth of predicted topics. We discuss the quality of the predicted topics by comparing the inferred topics in 2018-2019 across two models.

To measure the prediction performance, we calculated the minimum KL divergence between the word distribution of the predicted topics and ground truth topics: $\sum_{i=1}^{K} \min_j \mathrm{D}_{\mathrm{KL}}[\boldsymbol{\beta}_i^{pred}, \boldsymbol{\beta}_j^{truth}]/K$. This value measures the difference between the predicted topics and their nearest ground truth topics.

We computed this value for topics in 2018-2019 and compared it with the average value computed for topics in 2006-2017, which provides the baseline of the predictive performance.

We show that the average KL divergence for 2018-2019 (i.e., predictive performance) is 5.33, whereas that for 2006-2017 (i.e., baseline) is 5.24. This result indicates that the predicted topics are sufficiently close to the ground truth topics. Although further studies are needed, this result suggests that DSNTM can potentially predict emerging topics in the academic literature.

8 Conclusion

In this study, we proposed a novel dynamic-structured neural topic model, DSNTM, which captures dependencies among topics using the self-attention mechanism. We also introduced a citation regularizer, which induces the attention weights to correspond to citation relations.

Experimental results demonstrated that DNSTM outperforms previous dynamic topic models regarding perplexity and coherence while maintaining sufficient diversity across topics. In addition, DSNTM can identify the process of topic merging and branching while showing the potential to predict emerging topics. We expect that DSNTM will make it easier for non-specialists to keep track of the evolution of topics in a given research area without retracing the citations of copious articles and assist their search for a novel topic.

Limitations

As a limitation of the modeling assumption, DSNTM assumes that the number of topics is constant over time; however, this assumption is inappropriate for some time-series documents, such as scientific papers. As the number of scientific papers is increasing annually, increasing the number of topics over time would be appropriate for modeling the time-series evolution of academic literature.

We used the abstracts of the papers as text, and the attention was computed using textual information. However, citations mainly appear in the body text when a paper cites other papers. Therefore, there might be a discrepancy between the attention among topics and the citation relation among papers because the attention cannot not consider information in the body text. In future work, it would be desirable to evaluate our model using a corpus containing the body text of the papers.

Generally, topic models sometimes infer the incorrect information about topics, such as the frequent words appearing in topics, the topic proportion in each document, and the dependencies among topics. It would be the potential risk to induce the misunderstanding of users.

Ethics Statement

Our study complies with the ACL Ethics Policy. We used S2ORC (Lo et al., 2020, CC BY-NC 4.0), PyTorch (Paszke et al., 2019, BSD-style license) as scientific artifacts. Our study is conducted under the licenses and terms of the scientific artifacts. S2ORC is a collection of academic papers and generally does not contain any information that uniquely identifies individual people or offensive content. We did not use the author's information in our experiments.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. This work was supported by NEDO JPNP20006, JST ACT-X JPMJAX1904, JST CREST JPMJCR21D1, Japan.

References

Amr Ahmed and Eric P Xing. 2010. Timeline: a dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 20–29.

J Atchison and Sheng M Shen. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67:261–272.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Jonathan Chang and David M Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.

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

Rob Churchill and Lisa Singh. 2022. Dynamic topic-noise models for social media. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 429–443.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

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.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16.

Rem Hida, Naoya Takeishi, Takehisa Yairi, and Koichi Hori. 2018. Dynamic and static topic model for analyzing time-series document collections. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 516–520. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, pages 1–15.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584.

Kar Wai Lim and Wray Buntine. 2015. Bibliographic analysis with the citation network topic model. In *Proceedings of the Sixth Asian Conference on Machine Learning*, volume 39 of *Proceedings of Machine Learning Research*, pages 142–158. PMLR.

Dahua Lin, Eric Grimson, and John Fisher. 2010. Construction of dependent dirichlet processes based on poisson processes. *Advances in neural information processing systems*, 23.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2410–2419.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Riki Murakami, Basabi Chakraborty, and Yukari Shirota. 2021. Dynamic topic tracking and visualization using covid-19 related tweets in multiple languages. In *2021 International Conference on Artificial Intelligence and Big Data Analytics*, pages 16–21.

Ramesh M Nallapati, Amr Ahmed, Eric P Xing, and William W Cohen. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Danilo J Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, page 487–494.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations*.

Yuancheng Tu, Nikhil Johri, Dan Roth, and Julia Hockenmaier. 2010. Citation author topic model in expert search. In *The 23rd International Conference on Computational Linguistics: Posters*, pages 1265–1273.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Delvin Ce Zhang and Hady Lauw. 2022. Dynamic topic models for temporal document networks. In *International Conference on Machine Learning*, pages 26281–26292. PMLR.

A Implementation Details

The hyperparameters of each model were tuned based on the perplexity of the validation set in the *ACL*. The model was trained using Adam (Kingma and Ba, 2015) with a batch size of 512. DSNTM and ETM were trained for 200 epochs with a learning rate of 6.0×10^{-4}. As D-ETM was slow to converge, D-ETM was trained for 600 epochs with a learning rate of 8.0×10^{-4}. We applied the learning rate decay for each model.

To infer the topic embedding $q(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_{1:K}^{(1:t-1)})$, we used a linear layer for $f_q, f_k, f_v, f_\mu, f_\sigma$ to compute the self-attention and the variational distribution in Eq. (14). The dimension of the topic embedding was set $L = 300$. We used the multi-head attention (Vaswani et al., 2017) for the self-attention mechanism, where the number of parallel attention heads is 10.

Regarding the following hyperparameters, we set the same hyperparameters as those used in D-ETM.

To infer the topic proportion $q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d}, \boldsymbol{w}_d)$, We used one-hidden-layer MLPs with 800 hidden units and ReLU activation for f_θ and a linear layer for f_μ and f_σ to compute the variational distribution in Eq. (25).

To construct the inference of the topic proportion mean $q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \tilde{\boldsymbol{w}}_t)$, we first applied a linear layer to the BoW representation of documents at the time step t and obtain 200-dimensional input vector for LSTM. Then, we applied LSTM with three layers of 200 hidden units to the input, and obtain the hidden states of each time step $\boldsymbol{h}_t$. We used a linear layer for f_μ and f_σ to compute the variational distribution in Eq. (26).

The variances of the priors were set to $\delta^2 = \sigma^2 = 0.005$ and $\gamma^2 = 1$. We used 300-dimensional word embeddings pretrained with a skip-gram (Mikolov et al., 2013) used in ETM and D-ETM.

We ran experiments with a single NVIDIA GeForce RTX 2080 Ti for each model. The computational cost and parameters of each model are reported in Table 4. Our DSNTM converged faster than D-ETM regardless citation regularization. The training time was longer when using the citation regularization as it calculates the loss in Eq. (36) with time complexity $\boldsymbol{O}(D^2)$.

Our code is implemented with Python v3.9.13, PyTorch v1.9.0 (Paszke et al., 2019). We use the pretrained word embeddings published by Dieng et al. (2019)[5]. NPMI is computed using the code distributed by Lau et al. (2014)[6].

B Computing the Influence Among Topics

In Fig. 5, we do not directly use the attention weights to represent how much a past topic influences the emergence of new topics. This section describes its reason and how to calculate the influence of a topic on emerging topics.

We assume that the influence of a topic z_j on the emergence of topic z_i can be represented by the probability where z_i emerges given z_j. From Bayes' theorem, we can calculate its probability as follows:

$$\begin{aligned} p(z_i|z_j) &= \frac{p(z_j|z_i)p(z_i)}{p(z_j)} \\ &\propto p(z_j|z_i)p(z_i) \end{aligned} \tag{41}$$

where $p(z_j)$ can be ignored because it is constant regardless of z_i. $p(z_j|z_i)$ is represented by the attention weight from z_i to z_j, denoted as $a_{z_i \to z_j}$. $p(z_i)$ indicates the marginal probability where topic z_i appears across all documents, which is calculated by the sum of its topic proportions across all documents.

$$\begin{aligned} p(z_i) &= \sum_d p(z_i|d)p(d) \\ &= \sum_d \theta_d^{(z_i)} \end{aligned} \tag{42}$$

where we assume that $p(d)$ is uniformly distributed. Thus, we can obtain probability $p(z_i|z_j)$ as follows:

$$v_{i,j} = a_{z_i \to z_j} \sum_d \boldsymbol{\theta}_d^{(z_i)} \tag{43}$$

$$p(z_i|z_j) = \frac{v_{i,j}}{\sum_j v_{i,j}} \tag{44}$$

Therefore, we calculate the influence of a topic z_j on the emergence of topic z_i by considering the marginal probability of topic z_i.

[5] https://github.com/adjidieng/ETM

[6] https://github.com/jhlau/topic_interpretability

Dataset	ACL			CS		
	Parameters	Time	Memory	Parameters	Time	Memory
ETM (Dieng et al., 2020)	5,111,640	6	44	9,038,840	25	84
D-ETM (Dieng et al., 2019)	7,287,480	33	22	12,196,480	81	42
DSNTM w/o self-attention	7,480,380	9	22	12,389,380	23	42
DSNTM	8,022,780	9	23	12,931,780	29	43
DSNTM w/ citation regularization	8,022,780	14	23	12,931,780	55	44

Table 4: Computational cost of each model. Parameters, Time, and Memory denote the total number of model parameters, the total training time (minute) and the peak amount of the memory usage (MB), respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☒ A2. Did you discuss any potential risks of your work?
Section "Limitations"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1

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

B ☑ Did you use or create scientific artifacts?

Section 6,7, Appendix A

☑ B1. Did you cite the creators of artifacts you used?
Section 6.1, Appendix A

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☑ 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)?
Section "Ethics Statement"

☑ 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?
Section "Ethical Statement"

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 6.1

☑ 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.
Section 6.1

C ☑ Did you run computational experiments?

Section 6, 7

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

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?
Section 6.1, Appendix A

☑ 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?
Appendix A

☑ 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.)?
Appendix A

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.