Tree-Structured Neural Topic Model

Masaru Isonuma[1] **Junichiro Mori**[1,2] **Danushka Bollegala**[3] **Ichiro Sakata**[1]
[1] **The University of Tokyo** [2] **RIKEN** [3] **University of Liverpool**
{isonuma, isakata}@ipr-ctr.t.u-tokyo.ac.jp
mori@mi.u-tokyo.ac.jp danushka@liverpool.ac.uk

Abstract

This paper presents a tree-structured neural topic model, which has a topic distribution over a tree with an infinite number of branches. Our model parameterizes an unbounded ancestral and fraternal topic distribution by applying doubly-recurrent neural networks. With the help of autoencoding variational Bayes, our model improves data scalability and achieves competitive performance when inducing latent topics and tree structures, as compared to a prior tree-structured topic model (Blei et al., 2010). This work extends the tree-structured topic model such that it can be incorporated with neural models for downstream tasks.

1 Introduction

Probabilistic topic models, such as latent Dirichlet allocation (**LDA**; Blei et al., 2003), are applied to numerous tasks including document modeling and information retrieval. Recently, Srivastava and Sutton (2017); Miao et al. (2017) have applied the autoencoding variational Bayes (**AEVB**; Kingma and Welling, 2014; Rezende et al., 2014) framework to basic topic models such as LDA. AEVB improves data scalability in conventional models.

The limitation of the basic topic models is that they induce topics as flat structures, not organizing them into coherent groups or hierarchies. Tree-structured topic models (Griffiths et al., 2004), which detect the latent tree structure of topics, can overcome this limitation. These models induce a tree with an infinite number of nodes and assign a generic topic to the root and more detailed topics to the leaf nodes. In Figure 1, we show an example of topics induced by our model. Such characteristics are preferable for several downstream tasks, such as document retrieval (Weninger et al., 2012), aspect-based sentiment analysis (Kim et al., 2013) and extractive summarization (Celikyilmaz and Hakkani-Tur, 2010), because they provide succinct information from multiple viewpoints. For instance, in the case of document retrieval of product reviews, some users are interested in the general opinions about bag covers, while others pay more attention to specific topics such as the hardness or color of the covers. The tree structure can navigate users to the documents with desirable granularity.

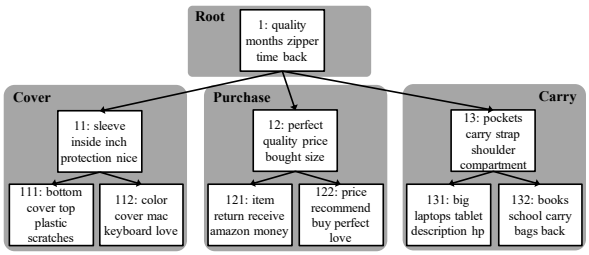

Figure 1: Topics inferred by our tree-structured topic model from Amazon reviews of laptop bags. The five most frequent words are shown and manually labeled.

However, it is difficult to use tree-structured topic models with neural models for downstream tasks. While neural models require a large amount of data for training, conventional inference algorithms, such as collapsed Gibbs sampling (Blei et al., 2010) or mean-field approximation (Wang and Blei, 2009), have data scalability issues. It is also desirable to optimize the tree structure for downstream tasks by jointly updating the neural model parameters and posteriors of a topic model.

To overcome these challenges, we propose a tree-structured neural topic model (**TSNTM**), which is parameterized by neural networks and is trained using AEVB. While prior works have applied AEVB to flat topic models, it is not straightforward to parameterize the unbounded ancestral and fraternal topic distribution. In this paper, we provide a solution to this by applying doubly-recurrent neural networks (**DRNN**; Alvarez-Melis and Jaakkola, 2017), which have two recurrent structures over respectively the ancestors and siblings.

Experimental results show that the TSNTM achieves competitive performance against a prior work (Blei et al., 2010) when inducing latent topics and tree structures. The TSNTM scales to larger datasets and allows for end-to-end training with neural models of several tasks such as aspect-based sentiment analysis (Esmaeili et al., 2019) and abstractive summarization (Wang et al., 2019).

2 Related Works

Following the pioneering work of tree-structured topic models by Griffiths et al. (2004), several extended models have been proposed (Ghahramani et al., 2010; Zavitsanos et al., 2011; Kim et al., 2012; Ahmed et al., 2013; Paisley et al., 2014). Our model is based on the modeling assumption of Wang and Blei (2009); Blei et al. (2010), while parameterizing a topic distribution with AEVB.

In the context of applying AEVB to flat document or topic modeling (Miao et al., 2016; Srivastava and Sutton, 2017; Ding et al., 2018), Miao et al. (2017) proposed a model, which is closely related to ours, by applying recurrent neural networks (**RNN**) to parameterize an unbounded flat topic distribution. Our work infers the topic distributions over an infinite tree with a DRNN, which enables us to induce latent tree structures.

Goyal et al. (2017) used a tree-structured topic model (Wang and Blei, 2009) with a variational autoencoder (**VAE**) to represent video frames as a tree. However, their approach is limited to smaller datasets. In fact, they used only 1,241 videos (corresponding to documents) for training and separately updated the VAE parameters and the posteriors of the topic model by mean-field approximation. This motivates us to propose the TSNTM, which scales to larger datasets and allows for end-to-end training with neural models for downstream tasks.

3 Tree-Structured Neural Topic Model

We present the generative process of documents and the posterior inference by our model. As shown in Figure 2, we draw a path from the root to a leaf node and a level for each word. The word is drawn from the multinomial distribution assigned to the topic specified by the path and level.

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

   Draw a Gaussian vector: $\boldsymbol{x}_d \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2)$ (1)

   Obtain a path distribution: $\boldsymbol{\pi}_d = f_\pi(\boldsymbol{x_d})$ (2)

   Obtain a level distribution: $\boldsymbol{\theta_d} = f_\theta(\boldsymbol{x_d})$ (3)

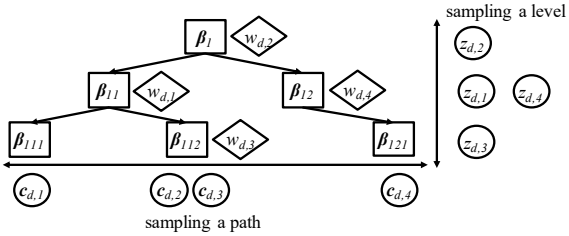

Figure 2: Sampling process of a topic for each word.

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

   Draw a path: $\boldsymbol{c}_{d,n} \sim \mathrm{Mult}(\boldsymbol{\pi}_d)$ (4)

   Draw a level: $z_{d,n} \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$ (5)

   Draw a word: $w_{d,n} \sim \mathrm{Mult}(\boldsymbol{\beta}_{\boldsymbol{c}_{d,n}[z_{d,n}]})$ (6)

where $\boldsymbol{\beta}_{\boldsymbol{c}_{d,n}[z_{d,n}]} \in \Delta^{V-1}$ is the word distribution assigned to a topic, $\boldsymbol{c}_{d,n}[z_{d,n}]$. While Wang and Blei (2009); Blei et al. (2010) draw a path for each document, this constrains a document to be generated from only the topics in the path. Hence, we draw a path for each word, enabling a document to be generated from all topics over a tree.

Wang and Blei (2009) draws a path and a level distribution via the tree-based stick-breaking construction given by (7) and (8):

$$\nu_k \sim \mathrm{Beta}(1, \gamma),\ \pi_k = \pi_{par(k)} \nu_k \prod_{j=1}^{k-1} (1 - \nu_j) \quad (7)$$

$$\eta_l \sim \mathrm{Beta}(1, \alpha),\ \theta_l = \eta_l \prod_{j=1}^{l-1} (1 - \eta_j) \quad (8)$$

Here, $k \in \{1, \ldots, K\}$ and $par(k)$ denote the k-th topic and its parent, respectively. $l \in \{1, \ldots, L\}$ denotes the l-th level. See Appendix A.1 for more details.

In contrast, we introduce neural architectures, f_π and f_θ, to transform a Gaussian sample to a topic distribution, allowing for posterior inference with AEVB. Specifically, we apply a DRNN to parameterize the path distribution over the tree.

3.1 Parameterizing Topic Distribution

A DRNN is a neural network decoder for generating tree-structured objects from encoded representations (Alvarez-Melis and Jaakkola, 2017). A DRNN consists of two RNNs over respectively the ancestors and siblings (see Appendix A.2). We assume that their two recurrent structures can parameterize the unbounded ancestral and fraternal path distribution conditioned on a Gaussian sample $\boldsymbol{x}$, using a finite number of parameters.

The hidden state, $\boldsymbol{h}_k$, of the topic k is given by:

$$\boldsymbol{h}_k = \tanh(\boldsymbol{W}_p \boldsymbol{h}_{par(k)} + \boldsymbol{W}_s \boldsymbol{h}_{k-1}) \quad (9)$$

where $\boldsymbol{h}_{par(k)}$ and $\boldsymbol{h}_{k-1}$ are the hidden states of a parent and a previous sibling of the k-th topic, respectively. We alternate the breaking proportions, $\boldsymbol{\nu}$, in (7) and obtain the path distribution, $\boldsymbol{\pi}$, as:

$$\nu_k = \text{sigmoid}(\boldsymbol{h}_k^\top \boldsymbol{x}) \quad (10)$$

Moreover, we parameterize the unbounded level distribution, $\boldsymbol{\theta}$, by passing a Gaussian vector through a RNN and alternating the breaking proportions, $\boldsymbol{\eta}$, in (8) as:

$$\boldsymbol{h}_l = \tanh(\boldsymbol{W} \boldsymbol{h}_{l-1}) \quad (11)$$

$$\eta_l = \text{sigmoid}(\boldsymbol{h}_l^\top \boldsymbol{x}) \quad (12)$$

3.2 Parameterizing Word Distribution

Next, we explain the word distribution assigned to each topic[1]. We introduce the embeddings of the k-th topic, $\boldsymbol{t}_k \in \mathbb{R}^H$, and words, $\boldsymbol{U} \in \mathbb{R}^{V \times H}$, to obtain the word distribution, $\boldsymbol{\beta}_k \in \Delta^{V-1}$, by (13).

$$\boldsymbol{\beta}_k = \text{softmax}(\frac{\boldsymbol{U} \cdot \boldsymbol{t}_k^\top}{\tau^{\frac{1}{l}}}) \quad (13)$$

where $\tau^{\frac{1}{l}}$ is a temperature value and produces more sparse probability distribution over words as the level l gets to be deeper (Hinton et al., 2014).

As the number of topics is unbounded, the word distributions must be generated dynamically. Hence, we introduce another DRNN to generate topic embeddings as $\boldsymbol{t}_k = \text{DRNN}(\boldsymbol{t}_{par(k)}, \boldsymbol{t}_{k-1})$.

Several neural topic models (Xie et al., 2015; Miao et al., 2017; He et al., 2017) have introduced diversity regularizer to eliminate redundancy in the topics. While they force all topics to be orthogonal, this is not suitable for tree-structured topic models, which admit the correlation between a parent and its children. Hence, we introduce a tree-specific diversity regularizer with $\bar{\boldsymbol{t}}_{ki} = \boldsymbol{t}_i - \boldsymbol{t}_k$ as:

$$\sum_{k \notin \text{Leaf}} \sum_{i,j \in \text{Chi}(k): i \neq j} \left(\frac{\bar{\boldsymbol{t}}_{ki}^\top \cdot \bar{\boldsymbol{t}}_{kj}}{\|\bar{\boldsymbol{t}}_{ki}\| \|\bar{\boldsymbol{t}}_{kj}\|} - 1 \right)^2 \quad (14)$$

where Leaf and $\text{Chi}(k)$ denote the set of the topics with no children and the children of the k-th topic, respectively. By adding this regularizer to the variational objective, each child topic becomes orthogonal from the viewpoint of their parent, while allowing parent–children correlations.

[1] $\boldsymbol{\beta}_k$ can be drawn from another distribution, but here we set it as a model parameter following Miao et al. (2017).

3.3 Variational Inference with AEVB

Under our proposed probabilistic model, the likelihood of a document is given by (15):

$$\begin{aligned} & p(\boldsymbol{w}_d | \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \boldsymbol{\beta}) \\ &= \int_{\boldsymbol{\pi}, \boldsymbol{\theta}} \Big\{ \prod_n \sum_{\boldsymbol{c}_n, z_n} p(w_n | \boldsymbol{\beta}_{\boldsymbol{c}_n [z_n]}) p(\boldsymbol{c}_n | \boldsymbol{\pi}) p(z_n | \boldsymbol{\theta}) \Big\} \\ &\qquad p(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0) \, d\boldsymbol{\pi} d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\pi}, \boldsymbol{\theta}} \Big\{ \prod_n (\boldsymbol{\beta} \cdot \boldsymbol{\phi})_{w_n} \Big\} p(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0) d\boldsymbol{\pi} d\boldsymbol{\theta} \end{aligned} \quad (15)$$

where $\boldsymbol{\phi} \in \Delta^{K-1}$ is the topic distribution and is derived as $\phi_k = \sum_{l=1}^{L} \theta_l (\sum_{\boldsymbol{c}: c_l = k} \pi_{\boldsymbol{c}})$.

From (15), by integrating out the latent variables $\boldsymbol{c}_n$ and z_n, the evidence lower bound for the document log-likelihood is derived as:

$$\begin{aligned} \mathcal{L}_d = & \mathbf{E}_{q(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{w}_d)} \Big[\sum_n \log (\boldsymbol{\beta} \cdot \boldsymbol{\phi})_{w_n} \Big] \\ & - \text{KL} \Big[q(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{w}_d) || p(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0) \Big] \end{aligned} \quad (16)$$

where $q(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{w}_d)$ is the variational distribution approximating posteriors.

Following the AEVB framework, we introduce multi-layer perceptrons (**MLP**) f_μ and f_{σ^2} for transforming bag-of-words vector $\boldsymbol{w}_d$ to the variational Gaussian distribution. The variational distribution of the posteriors is re-written as:

$$\begin{aligned} q(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{w}_d) &= q(f_\pi(\boldsymbol{x}), f_\theta(\boldsymbol{x}) | \boldsymbol{w}_d) \\ &= \mathcal{N}(\boldsymbol{x} | f_\mu(\boldsymbol{w}_d), f_{\sigma^2}(\boldsymbol{w}_d)) \end{aligned} \quad (17)$$

We sample $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\theta}}$ from $q(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{w}_d)$ by sampling $\hat{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and computing $\hat{\boldsymbol{x}} = f_\mu(\boldsymbol{w}_d) + \hat{\epsilon} \cdot f_{\sigma^2}(\boldsymbol{w}_d)$. The priors, $p(\boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2)$, is also re-written as $\mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2)$.

To sum up, the evidence lower bound is approximated with sampled topic distribution $\hat{\boldsymbol{\phi}}$ as:

$$\begin{aligned} \mathcal{L}_d \approx & \sum_n \log (\boldsymbol{\beta} \cdot \hat{\boldsymbol{\phi}})_{w_n} - \\ & \text{KL} \big[\mathcal{N}(\boldsymbol{x} | f_\mu(\boldsymbol{w}_d), f_{\sigma^2}(\boldsymbol{w}_d)) || \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2) \big] \end{aligned} \quad (18)$$

3.4 Dynamically Updating the Tree Structure

To allow an unbounded tree structure, we introduce two heuristic rules for adding and pruning the branches. We compute the proportion of the words in topic k: $p_k = (\sum_{d=1}^{D} N_d \, \hat{\phi}_{d,k}) / (\sum_{d=1}^{D} N_d)$. For each non-leaf topic k, if p_k is more than a threshold, a child is added to refine the topic. For each topic k, if the cumulative proportion of topics over descendants, $\sum_{j \in \text{Des}(k)} p_j$, is less than a threshold, the k-th topic and its descendants are removed ($\text{Des}(k)$ denotes the set of topic k and its descendants). We also remove topics with no children at the bottom.

4 Experiments

4.1 Datasets

In our experiments, we use the *20NewsGroups* and the *Amazon product reviews*. The *20NewsGroups* is a collection of 20 different news groups containing $11,258$ training and $7,487$ testing documents[2]. For the *Amazon product reviews*, we use the domain of *Laptop Bags* provided by Angelidis and Lapata (2018), with $31,943$ training, 385 validation and 416 testing documents[3]. We use the provided test documents in our evaluations, while randomly splitting the remainder of the documents into training and validation sets.

4.2 Baseline Methods

As baselines, we use a tree-structured topic model based on the nested Chinese restaurant process (**nCRP**) with collapsed Gibbs sampling (Blei et al., 2010). In addition, we use a flat neural topic model, i.e. the recurrent stick-breaking process (**RSB**), which constructs the unbounded flat topic distribution via an RNN (Miao et al., 2017).

4.3 Implementation Details

For the TSNTM and RSB, we use 256-dimensional word embeddings, a one-hidden-layer MLP with 256 hidden units, and a one-layer RNN with 256 hidden units to construct variational parameters. We set the hyper-parameters of Gaussian prior distribution $\boldsymbol{\mu}_0$ and $\boldsymbol{\sigma}_0^2$ as a zero mean vector and a unit variance vector with 32 dimensions, respectively. We train the model using AdaGrad (Duchi et al., 2011) with a learning rate of 10^{-2}, an initial accumulator value of 10^{-1}, and a batch size of 64. We grow and prune a tree with a threshold of 0.05 in Section 3.4 and set a temperature as $\tau = 10$ in Section 3.2 [4].

Regarding the nCRP-based model, we set the nCRP parameter as $\gamma = 0.01$, the GEM parameter as $\pi = 10, m = 0.5$, and the Dirichlet parameter as $\eta = 5$.

The hyperparameters of each model are tuned based on the perplexity on the validation set in the *Amazon product reviews*. We fix the number of levels in the tree as 3 with an initial number of branches 3 for both the second and third levels.

[2]For direct comparison against Miao et al. (2017), we use the training/testing splits and the vocabulary provided at https://github.com/akashgit/autoencoding_vi_for_topic_models.

[3]https://github.com/stangelid/oposum

[4]The code to reproduce the results is available at: https://github.com/misonuma/tsntm.

NPMI	20News	Amazon
RSB (Miao et al., 2017)	0.201	0.102
nCRP (Blei et al., 2010)	0.198	0.112
TSNTM (Our Model)	0.220	0.121

Table 1: Average NPMI of the induced topics.

Perplexity	20News	Amazon
RSB (Miao et al., 2017)	931	472
nCRP (Blei et al., 2010)	681	303
TSNTM (Our Model)	886	460

Table 2: Average perplexity of each model.

4.4 Evaluating Topic Interpretability

Several works (Chang et al., 2009; Newman et al., 2010) pointed out that perplexity is not suitable for evaluating topic interpretability. Meanwhile, Lau et al. (2014) showed that the normalized pointwise mutual information (**NPMI**) between all pairs of words in each topic closely corresponds to the ranking of topic interpretability by human annotators. Thus, we use NPMI instead of perplexity as the primary evaluation measure following Srivastava and Sutton (2017); Ding et al. (2018).

Table 1 shows the average NPMI of the topics induced by each model. Our model is competitive with the nCRP-based model and the RSB for each dataset. This indicates that our model can induce interpretable topics similar to the other models.

As a note, we also show the average perplexity over the documents of each model in Table 2. For the AEVB-based models (RSB and TSNTM), we calculate the upper bound of the perplexity using ELBO following Miao et al. (2017); Srivastava and Sutton (2017). In contrast, we estimate it by sampling the posteriors in the nCRP-based model with collapsed Gibbs sampling.

Even though it is difficult to compare them directly, the perplexity of the nCRP-based model is lower than that of the AEVB-based models. This tendency corresponds to the result of Srivastava and Sutton (2017); Ding et al. (2018), which report that the model with collapsed Gibbs sampling achieves the lowest perplexity in comparison with the AEVB-based models. In addition, Ding et al. (2018) also reports that there is a trade-off between perplexity and NPMI. Therefore, it is natural that our model is competitive with the other models regarding to NPMI, while there is a significant difference in achieved perplexity.

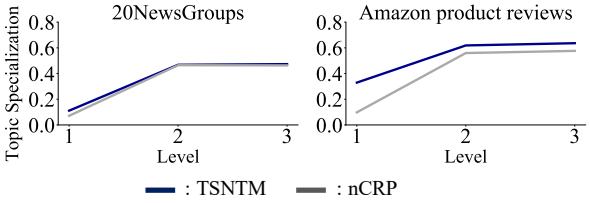

Figure 3: Topic specialization scores for each level.

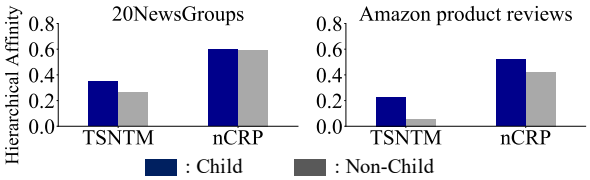

Figure 4: Hierarchical affinity scores.

4.5 Evaluating Tree-Structure

For evaluating the characteristic of the tree structure, we adopt two metrics: topic specialization and hierarchical affinity following Kim et al. (2012).

Topic specialization: An important characteristic of the tree-structure is that the most general topic is assigned to the root, while the topics become more specific toward the leaves. To quantify this characteristic, we measure the specialization score as the cosine similarity of the word distribution between each topic and the entire corpus. As the entire corpus is regarded as the most general topic, more specific topics have lower similarity scores. Figure 3 presents the average topic specialization scores for each level. While the root of the nCRP is more general than that of our model, the tendency is roughly similar for both models.

Hierarchical Affinity: It is preferable that a parent topic is more similar to its children than the topics descended from the other parents. To verify this property, for each parent in the second level, we calculate the average cosine similarity of the word distribution to children and non-children respectively. Figure 4 shows the average cosine similarity over the topics. While the nCRP-based model induces child topics slightly similar to their parents, our model infers child topics with more similarity to their parent topics. Moreover, lower scores of the TSNTM also indicate that it induces more diverse topics than the nCRP-based model.

Example: In Section 1, an example of the induced topics and the latent tree for the laptop bag reviews is shown in Figure 1.

4.6 Evaluating Data Scalability

To evaluate how our model scales with the size of the datasets, we measure the training time until the convergence for various numbers of documents.

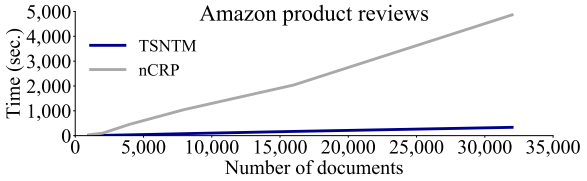

Figure 5: Training time for various number of docs.

We randomly sample several number of documents (1,000, 2,000, 4,000, 8,000, 16,000 and all) from the training set of the *Amazon product reviews* and measure the training time for each number of documents. The training is stopped when the perplexity of the validation set is not improved for 10 consecutive iterations over the entire batches. We measure the time to sample the posteriors or update the model parameters, except for the time to compute the perplexity [5].

As shown in Figure 5, as the number of documents increases, the training time of our model does not change considerably, whereas that of the nCRP increases significantly. Our model can be trained approximately 15 times faster than the nCRP-based model with 32,000 documents.

5 Conclusion

We proposed a novel tree-structured topic model, the TSNTM, which parameterizes the topic distribution over an infinite tree by a DRNN.

Experimental results demonstrated that the TSNTM achieves competitive performance when inducing latent topics and their tree structures, as compared to a prior tree-structured topic model (Blei et al., 2010). With the help of AEVB, the TSNTM can be trained approximately 15 times faster and scales to larger datasets than the nCRP-based model.

This allows the tree-structured topic model to be incorporated with recent neural models for downstream tasks, such as aspect-based sentiment analysis (Esmaeili et al., 2019) and abstractive summarization (Wang et al., 2019). By incorporating our model instead of flat topic models, they can provide multiple information with desirable granularity.

Acknowledgments

We would like to thank anonymous reviewers for their valuable feedback. This work was supported by JST ACT-X Grant Number JPMJAX1904 and CREST Grant Number JPMJCR1513, Japan.

[5] All computational times are measures on the same machine with a Xeon E5-2683-v4 (2.1 GHz, 16 cores) CPU and a single GeForce GTX 1080 (8GB) GPU.

References

Amr Ahmed, Liangjie Hong, and Alexander J Smola. 2013. The nested chinese restaurant franchise process: User tracking and document modeling. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1426–1434.

David Alvarez-Melis and Tommi S Jaakkola. 2017. Tree-structured decoding with doubly-recurrent neural networks. In *Proceedings of the 5th International Conference on Learning Representations*.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686.

David M Blei, Thomas L Griffiths, and Michael I Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7.

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

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Babak Esmaeili, Hongyi Huang, Byron Wallace, and Jan-Willem van de Meent. 2019. Structured neural topic models for reviews. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 3429–3439.

Zoubin Ghahramani, Michael I Jordan, and Ryan P Adams. 2010. Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems*, pages 19–27.

Prasoon Goyal, Zhiting Hu, Xiaodan Liang, Chenyu Wang, and Eric P Xing. 2017. Nonparametric variational auto-encoders for hierarchical representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5094–5102.

Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, pages 17–24.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 388–397.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2014. Distilling the knowledge in a neural network. In *the NIPS 2014 Deep Learning and Representation Learning Workshop*.

Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 783–792.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 526–533.

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.

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.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1727–1736.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 100–108.

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

Danilo Jimenez 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.

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

Chong Wang and David M Blei. 2009. Variational inference for the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, pages 1990–1998.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 166–177.

Tim Weninger, Yonatan Bisk, and Jiawei Han. 2012. Document-topic hierarchies from document graphs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 635–644.

Pengtao Xie, Yuntian Deng, and Eric Xing. 2015. Diversifying restricted boltzmann machine for document modeling. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1315–1324.

Elias Zavitsanos, Georgios Paliouras, and George A Vouros. 2011. Non-parametric estimation of topic hierarchies from texts with hierarchical dirichlet processes. *Journal of Machine Learning Research*, 12:2749–2775.

A Appendices

A.1 Tree-Based Stick-Breaking Construction

Figure 6 describes the process of the tree-based stick-breaking construction (Wang and Blei, 2009). At the first level, the stick length is $\pi_1 = 1$. Then, the stick-breaking construction is applied to the first level stick to obtain the path distribution over the second level. For instance, if the second level contains $K = 3$ topics, the probability of each path is obtained as $\pi_{11} = \pi_1 \nu_{11}$, $\pi_{12} = \pi_1 \nu_{12}(1-\nu_{11})$ and the remaining stick $\pi_{13} = \pi_1(1-\nu_{12})(1-\nu_{11})$. Generally, for any values of K, it satisfies $\sum_{k=1}^{K} \pi_{1k} = \pi_1$. The same process is applied to each stick proportion of the second level and continues until it reaches to the bottom level.

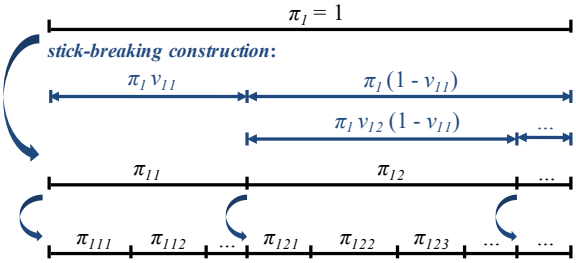

Figure 6: Tree-based stick-breaking construction.

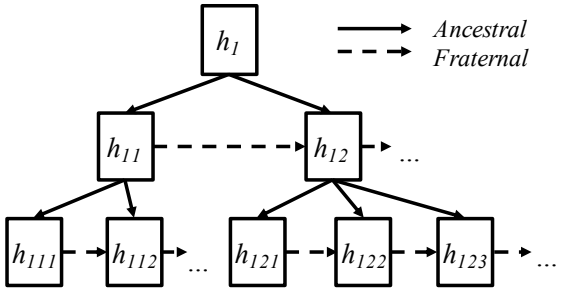

Figure 7: Doubly-recurrent neural networks.

A.2 Doubly-Recurrent Neural Networks

Figure 7 shows the architecture of doubly-recurrent neural networks (Alvarez-Melis and Jaakkola, 2017). It consists of two recurrent neural networks over respectively the ancestors and siblings that are combined in each cell as described in (9).