Generalized Knowledge Distillation for Topic Models

Kohei Watanabe(✉) and Koji Eguchi(✉)

Graduate School of Advanced Science and Engineering, Hiroshima University,
Higashi-Hiroshima, Japan
{m224406,kxeguchi}@hiroshima-u.ac.jp

Abstract. Topic modeling is used in the analysis of textual data to estimate the underlying topics within the dataset. Knowledge distillation has been attracting attention as a means of transferring knowledge from a large teacher model to a small student model in the field of deep learning. Knowledge distillation can be categorized into three types depending on the type of knowledge to be distilled: response-based, feature-based, and relation-based. To the best of our knowledge, previous studies on knowledge distillation used in topic models have all focused on response and/or feature knowledge, but these methods cannot transfer the structural knowledge of the teacher model to the student model. To solve this problem, we propose a generalized knowledge-distillation method that combines all three types of knowledge distillation, including the relation-based knowledge distillation with contrastive learning, which had not been used for neural topic models. Our experiments show that our neural topic model, trained with the proposed method, improves topic coherence compared to baseline models without knowledge distillation.

Keywords: Topic models · Knowledge distillation · Contrastive learning

1 Introduction

Topic modeling is a common method for estimating latent topics behind data from documents and has been applied to various tasks. A typical topic model, latent Dirichlet allocation (LDA) [2], generates documents probabilistically assuming that there are multiple latent topics behind each document. LDA is typically trained using variational Bayesian methods; however, the challenge is that a new inference process needs to be mathematically derived depending on the purpose of the model. Neural topic models have been proposed to solve this problem. One such model is Srivastava et al.'s PRODLDA [8], which is based on a variational autoencoder (VAE) [6]. It can approximate complex posterior distributions using a flexible inference network that is based on neural networks.

F. Liu et al. (Eds.): PRICAI 2023, LNAI 14326, pp. 355–361, 2024.
https://doi.org/10.1007/978-981-99-7022-3_32

In deep learning, knowledge distillation has attracted attention as a method for transferring knowledge from a large-scale teacher model to a small-scale student model. Knowledge distillation can be classified into three types depending on the type of knowledge to be distilled: response-based, feature-based, and relation-based [4]. In a previous study on knowledge distillation for neural topic models, Hoyle et al. proposed a response-based knowledge-distillation method that trains student neural topic models using the output of BERT, which is pretrained on large corpora, as the teacher model [5]. Adhya et al. also conducted response-based and feature-based knowledge distillation simultaneously using a large neural topic model as the teacher and a small neural topic model as the student [1]. However, these methods focus only on the individual sample representations, which means that they are unable to transfer structural knowledge, the relationships between samples, from the teacher model to the student model.

To solve this problem, we propose a relation-based knowledge-distillation method using contrastive learning for neural topic models. The method uses contrastive loss to distill the structural knowledge of the teacher by learning the latent representations of the student model, while maintaining the relationships in the individual document representations generated by the teacher model. We further propose a generalized knowledge distillation by combining response-based, feature-based, and relation-based knowledge distillation. Through evaluation experiments measuring topic coherence, we show that the neural topic model trained using the proposed method improves on a baseline neural topic model [3] and its variant.

2 Overview of Neural Topic Models

As an earlier neural topic model, PRODLDA [8] was developed using VAE [6]. A generalization of PRODLDA is SCHOLAR [3]. These neural topic models replace the Dirichlet prior used in the original LDA [2] with a logistic normal prior ($\mathcal{LN}$) to facilitate inference. Now suppose $\boldsymbol{w}_i^{\mathrm{BoW}}$ is a V-dimensional vector counting the words in document $\boldsymbol{w}_i$, and $\boldsymbol{z}_i$ is its corresponding topic vector. The VAE-based neural topic model learns to minimize the Kullback-Leibler (KL) divergence between the true posterior distribution $p(\boldsymbol{z}_i)$ and variational distribution $q(\boldsymbol{z}_i|\boldsymbol{w}^{\mathrm{BoW}})$, which cannot be obtained analytically. The evidence lower bound (ELBO) is expressed as

$$\mathrm{ELBO} = \mathbb{E}_{q(\boldsymbol{z}_i|\cdot)}\left[\mathcal{L}_{RE}\right] - \mathrm{D}_{\mathrm{KL}}\left[q\left(\boldsymbol{z}_i \mid \boldsymbol{w}_i^{\mathrm{BoW}}\right) \| p\left(\boldsymbol{z}_i \mid \alpha\right)\right], \tag{1}$$

where $\mathcal{L}_{RE} = (\boldsymbol{w}_i^{\mathrm{BoW}})^\top \log \sigma(\boldsymbol{\eta}_i)$. The notation $\sigma(\cdot)$ is a softmax function, $\sigma(\boldsymbol{\eta}_i)$ corresponds to the word distribution (multinomial distribution over the vocabulary) of document $\boldsymbol{w}_i$, $\mathcal{L}_{RE}$ is the reconstruction error, and $\mathrm{D}_{\mathrm{KL}}\left[q\left(\boldsymbol{z}_i \mid \boldsymbol{w}_i^{\mathrm{BoW}}\right) \| p\left(\boldsymbol{z}_i \mid \alpha\right)\right]$ is the KL divergence between $q(\boldsymbol{z}_i|\boldsymbol{w}_i^{\mathrm{BoW}})$ and $p(\boldsymbol{z}_i|\alpha)$. As in VAE, the inference process uses a multilayer neural network to generate the variational parameters. Since the logistic normal distribution is assumed for the prior distribution of $\boldsymbol{z}$, the inference network outputs a mean vector $\boldsymbol{\mu}(\cdot)$ and diagonal covariance matrix $\boldsymbol{\sigma}^2(\cdot)$. The variational distribution is $q(\boldsymbol{z}_i \mid \boldsymbol{w}_i^{\mathrm{BoW}}) = \mathcal{LN}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$.

$$\boldsymbol{\mu}_i = \mathbf{W}_\mu \boldsymbol{\pi}_i + \boldsymbol{b}_\mu, \quad \log \boldsymbol{\sigma}_i^2 = \mathbf{W}_\sigma \boldsymbol{\pi}_i + \boldsymbol{b}_\sigma, \quad \boldsymbol{\pi}_i = f\left(\mathbf{W}_w \boldsymbol{w}_i^{\mathrm{BoW}}\right), \tag{2}$$

where f is the multilayer perceptron and the variational parameters are all the weight matrices $\mathbf{W}_w$, $\mathbf{W}_\mu$, and $\mathbf{W}_\sigma$ and biases $\boldsymbol{b}_\mu$ and $\boldsymbol{b}_\sigma$ in Eq. (2).

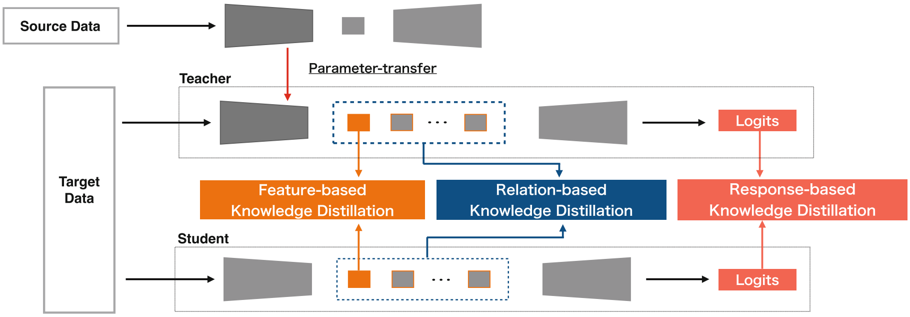
Source Data
Parameter-transfer
Teacher
Target Data
Logits
Feature-based Knowledge Distillation
Relation-based Knowledge Distillation
Response-based Knowledge Distillation
Student
Logits

Fig. 1. Conceptual diagram of generalized knowledge distillation.

3 Methodology

On the basis of the neural topic model SCHOLAR [3], our method unify response-based and feature-based knowledge distillation using transfer learning and relation-based knowledge distillation using contrastive learning. It differs from previous methods in that we apply relation-based knowledge distillation [9] to the neural topic model, which has not been studied previously, and in that we propose to integrate the three types of knowledge distillation in a unified framework. As knowledge distillation require s employing an identical dataset for both student and teacher models, we initialize the teacher model's weight matrix $\mathbf{W}_w$ for the target data by leveraging the weight matrix $\mathbf{W}_w$ pre-trained on a source data. Figure 1 shows a conceptual diagram of generalized knowledge distillation.

For the inference process of neural topic models described in Sect. 2, we use the following objective function instead of $\mathcal{L}_{RE}$ in Eq. (1),

$$\mathcal{L} = (1-\gamma)\mathcal{L}_{RE} + \gamma\mathcal{L}_{ResKD} + \lambda_1\mathcal{L}_{FeaKD} + \lambda_2\mathcal{L}_{RCD}. \tag{3}$$

Here, $\mathcal{L}_{ResKD}$, $\mathcal{L}_{FeaKD}$, and $\mathcal{L}_{RCD}$ corresponds to response-based, feature-based, and relation contrastive distillation, respectively. The details of these terms are explained in the rest of this section. The notations $\gamma, \lambda_1, \lambda_2$ are hyperparameters to adjust the effect of each term.

Response-Based Knowledge Distillation: The generative process of the models trained with our proposed method is the same as that of SCHOLAR. The inference process uses the SCHOLAR inference network but adds a pseudo-document $\boldsymbol{w}_i^t$ to Eq. (2), which is generated from the logit of the teacher model.

$$\boldsymbol{\pi}_i = f\left(\left[\mathbf{W}_w \boldsymbol{w}_i^{\mathrm{BoW}}; \mathbf{W}_{w^t} \boldsymbol{w}_i^t\right]\right), \tag{4}$$

where $\left[\mathbf{W}_w \boldsymbol{w}_i^{\mathrm{BoW}}; \mathbf{W}_{w^t} \boldsymbol{w}_i^t\right]$ denotes the horizontal concatenation of $\mathbf{W}_w \boldsymbol{w}_i^{\mathrm{BoW}}$ and $\mathbf{W}_{w^t} \boldsymbol{w}_i^t$. To apply knowledge distillation to a neural topic model, the following objective function $\mathcal{L}_{ResKD}$ is used

$$\mathcal{L}_{ResKD} = \tau^2 (\boldsymbol{w}_i^t)^\top \log \hat{\boldsymbol{w}}_i, \quad \boldsymbol{w}_i^t = \sigma(\boldsymbol{\eta}_i^t/\tau) N_i, \quad \hat{\boldsymbol{w}}_i = \sigma(\boldsymbol{\eta}_i/\tau), \tag{5}$$

where $\boldsymbol{w}_i^t$ is the probability estimated from the logit $\boldsymbol{\eta}_i^t$ of the teacher model, scaled by the document length N and treated as a smoothed pseudo-document, and τ is the temperature of the softmax function.

Feature-Based Knowledge Distillation: Feature-based knowledge distillation distills the topic multinomial distribution of the documents from the teacher model to the student model as knowledge. The objective function of feature-based knowledge distillation is expressed as

$$\mathcal{L}_{FeaKD} = -\sum (\boldsymbol{z}_i^t - \boldsymbol{z}_i^s)^2 \tag{6}$$

where $\boldsymbol{z}_i^t$ and $\boldsymbol{z}_i^s$ indicate the latent representations (i.e., features or topics) generated by the teacher and student models, respectively, for document $\boldsymbol{w}_i$.

Relation Contrastive Distillation: Now, we describe the method for achieving relation-based knowledge distillation by maximizing the mutual information of the relation Y^t between the latent representations of the teacher model and that $Y^{t,s}$ between the latent representations of the teacher model and student model. The idea is inspired by [9]; however, we employ it in the context of inference of neural topic models. Let $p(W)$ be the empirical distribution of the document set $W = \{\boldsymbol{w}_i : i = 1, ..., D\}$ of the training data and model the conditional marginal distributions of topic relations $p(Y^t|W)$ and $p(Y^{t,s}|W)$ as follows.

$$\boldsymbol{w}_i, \boldsymbol{w}_j, \boldsymbol{w}_m, \boldsymbol{w}_n \sim p(W), \quad \boldsymbol{y}_{i,j}^t = g^t(\boldsymbol{z}_i^t, \boldsymbol{z}_j^t), \quad \boldsymbol{y}_{m,n}^{t,s} = g^{t,s}(\boldsymbol{z}_m^t, \boldsymbol{z}_n^s), \tag{7}$$

where $\boldsymbol{z}_i^t$ is the latent representation generated by the decoder of the teacher neural topic model for document $\boldsymbol{w}_i$, and $\boldsymbol{z}_n^s$ is that generated by the student neural topic model for document $\boldsymbol{w}_n$. The g^t is a network that computes the relation between the latent representations of the teacher model and $g^{t,s}$ is a network that computes the relation between the latent representations of the teacher model and student model. We also model $p(Y^t, Y^{t,s}|W)$ as follows.

$$\boldsymbol{w}_i, \boldsymbol{w}_j \sim p(W), \quad \boldsymbol{y}_{i,j}^t = g^t(\boldsymbol{z}_i^t, \boldsymbol{z}_j^t), \quad \boldsymbol{y}_{i,j}^{t,s} = g^{t,s}(\boldsymbol{z}_i^t, \boldsymbol{z}_j^s). \tag{8}$$

The mutual information of $p(Y^t|W)$ and $p(Y^{t,s}|W)$ is expressed as follows.

$$I(Y^t, Y^{t,s}) = \mathbb{E}_{p(Y^t,Y^{t,s}|W)} \log \frac{p(Y^t, Y^{t,s}|W)}{p(Y^t|W)p(Y^{t,s}|W)}. \tag{9}$$

To derive the objective function, we define a latent variable δ that indicates whether the relation pairs $(\boldsymbol{y}^t, \boldsymbol{y}^{t,s})$ are generated from the joint distribution or product of marginal distributions. When $\delta = 1$, it means that $\boldsymbol{y}^t$ and $\boldsymbol{y}^{t,s}$ are computed by the same input pair, as in Eq. (8), and when $\delta = 0$, it means that $\boldsymbol{y}^t$ and $\boldsymbol{y}^{t,s}$ are computed by independently selected input pairs, as in Eq. (7). Maximizing the mutual information is equivalent to maximizing the following objective function $\mathcal{L}_{RCD}$ of relation contrastive distillation [9].

$$\mathcal{L}_{RCD} = \sum_{q(\delta=1)} \log h(\boldsymbol{y}^t, \boldsymbol{y}^{t,s}) + N \sum_{q(\delta=0)} \log[1 - h(\boldsymbol{y}^t, \boldsymbol{y}^{t,s})], \tag{10}$$

where$\{(\boldsymbol{y}^t, \boldsymbol{y}^{t,s})|\delta = 1\}$ is a positive pair and $\{(\boldsymbol{y}^t, \boldsymbol{y}^{t,s})|\delta = 0\}$ is a negative pair, and N is the number of negative pairs for a positive pair. h is a model for approximating true distribution $q(\delta = 1|Y^t, Y^{t,s})$, where $h : \{Y^t, Y^{t,s}\} \rightarrow [0, 1]$. Not only h, but also the student network and subnetworks are optimized when $\mathcal{L}_{RCD}$ is minimized.

Table 1. Datasets that differ in total number of documents D and vocabulary size. V

	Wiki (Source)	IMDb (Target)	20NG (Target)	BBC (Target)
D	6,078,287	50,000	18,745	2,225
V	50,000	5,000	1,995	9,635

Table 2. NPMI and sample standard deviation.

Model	IMDb	20NG	BBC
SCHOLAR	0.164 (0.006)	0.316 (0.005)	0.279 (0.011)
SCH.+Wiki	0.162 (0.003)	0.321 (0.003)	0.280 (0.006)
SCH.+ResKD+FeaKD+RCD	**0.167 (0.002)**	**0.349 (0.010)**	**0.321 (0.012)**

4 Experiments and Results

We used the English Wikipedia dataset (Wiki)[1] as the source data for pre-training SCHOLAR, and the IMDb dataset of movie reviews (IMDb)[2], 20News-

[1] https://huggingface.co/datasets/wikipedia.

[2] http://ai.stanford.edu/~amaas/data/sentiment/.

groups dataset (20NG)[3], and BBC dataset (BBC)[4] as the target data to be analyzed. We split the datasets into training, development, and test sets (train/dev/test) in the following proportions: 20NG: 48/12/40, IMDb: 50/25/25, BBC: 70/15/15. The vocabulary of the Wiki dataset used for the pre-training was formed by keeping the top 50,000 words that occurred in most documents. Details of the datasets are listed in Table 1. We set the number of topics to 50 in the evaluation experiment. We used Optuna[5] to tune the hyperparameters τ, γ, λ_1, and λ_2.

The models trained with the proposed method were evaluated using normalized pointwise mutual information (NPMI) [7], a measure of topic coherence based on the co-occurrence of words in a corpus, using a test set of the top 10 words for each topic in the same corpus. Table 2 lists the experimental results. The NPMI in the table is the average of five runs with different random initialization. The baseline models are SCHOLAR [3] and SCH.+Wiki, which was trained by transferring parameters from the SCHOLAR pre-trained on the large dataset, i.e., Wiki, and used as a teacher model in the knowledge distillation. The model (SCH.+ResKD+FeaKD+RCD) trained using the proposed method, which combines the three types of knowledge distillation (response-based, feature-based and relation-based), achieved the best NPMI on all three datasets compared with the two baselines: SCHOLAR [3] and SCH.+Wiki. We found that the SCH.+Wiki achieved better NPMI than the original SCHOLAR on the 20NG and BBC datasets, but slightly worse on the IMDb dataset.

5 Conclusions

We proposed a generalized knowledge distillation for training neural topic models, by unifying three types of knowledge distillation: response-based, feature-based, and relation-based. The response-based and feature-based knowledge-distillation are based on parameter transfer from a teacher model trained with a larger dataset. The relation-based knowledge distillation is based on contrastive learning that transfers topic relationships of a teacher model into a student model. This is the first work on relation-based knowledge distillation for neural topic models, to our knowledge. Evaluation experiments indicated that all three types of knowledge distillation improved the performance of the neural topic models trained with our method in several datasets. For future work, we plan to investigate which type of teacher is best suited for each of dataset to be analyzed. The use of large language models as teacher models is also a possible extension of our work.

Acknowledgements. This work was supported in part by the Grant-in-Aid for Scientific Research (#23K11231) from JSPS, Japan, and in part by ROIS NII Open Collaborative Research 2023 (#23FS02).

[3] https://github.com/akashgit/autoencoding_vi_for_topic_models.

[4] http://mlg.ucd.ie/datasets/bbc.html.

[5] https://optuna.org/.

References

1. Adhya, S., Sanyal, D.K.: Improving neural topic models with Wasserstein knowledge distillation. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13981, pp. 321–330. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28238-6_21
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR **3**, 993–1022 (2003)
3. Card, D., Tan, C., Smith, N.A.: Neural models for documents with metadata. In: ACL 2018, pp. 2031–2040 (2018)
4. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. IJCV **129**, 1789–1819 (2021)
5. Hoyle, A.M., Goel, P., Resnik, P.: Improving neural topic models using knowledge distillation. In: EMNLP 2020, pp. 1752–1771 (2020)
6. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR 2014 (2014)
7. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: EACL 2014, pp. 530–539 (2014)
8. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. In: ICLR 2017 (2017)
9. Zhu, J., et al.: Complementary relation contrastive distillation. In: CVPR 2021, pp. 9260–9269 (2021)