LLM-guided Preference Learning for Neural Topic Models

Abstract

1 Introduction

2 Related Work

3 Background

3.1 Notations

Denote $\mathbf{X} = \{x^d\}_{d=1}^D$ as a collection of Bag-of-Words (BoW) representations of D documents with the vocabulary of V words. Topic models aim to discover K hidden topics in this corpus. The pre-trained language model embedding of document d is x^d_{PLM}. The clustering algorithm applied to x^d_{PLM} produces G clusters. We have $\beta \in \mathbb{R}^{V \times K}$ = $(\beta_1, \ldots, \beta_K)$, where each $\beta_k \in \mathbb{R}^{V \times 1}$, as the topic-word distributions of K desired topics.

With L as the word embedding dimension, we set $\mathbf{w}_v \in \mathbb{R}^L, v \in \{1, 2, \ldots, V\}$ and $\mathbf{t}_k \in \mathbb{R}^L, k \in \{1, 2, \ldots, K\}$ to be the word embeddings of word v and topic embeddings of topic k, respectively. Each document x^d has the topic proportion $\theta_d \in \mathbb{R}^K$ indicating what topic it includes. $\mathbb{1}_N$ denotes a vector of length N, where each entry is set to 1.

3.2 VAE-based Topic Model

Similar to many recent neural topic models (Dieng et al., 2020; Wu et al., 2023), our approach is built on a VAE framework, which consists of two primary components: (i) an inference encoder that produces document-topic distributions; and (ii) a generative decoder that reconstructs the original text using the encoder's output and the topic-word proportions. For the encoder, the Bag-of-Words (BoW) representation of a document x^d is processed through neural networks to obtain the parameters of a normal distribution, where the mean $\mu = h_\mu(x^d)$ and the diagonal covariance matrix $\Sigma = \text{diag}(h_\Sigma(x^d))$ are computed. The reparameterization trick (Kingma and Welling, 2013) is then employed to sample a latent variable α from the posterior distribution $q(\alpha|x) = \mathcal{N}(\alpha|\mu, \Sigma)$, while the prior distribution of α is $p(\alpha) = \mathcal{N}(\alpha|\mu_0, \Sigma_0)$. Afterwards, the softmax function is applied to α, producing the topic proportion $\theta = \text{softmax}(\alpha)$.

Regarding the second component, VAE-based neural topic models aim to construct an effective representation for the topic-word distributions $\beta \in \mathbb{R}^{V \times K}$. There are several approaches to modeling β, such as directly inferring it through an optimization process (Srivastava and Sutton, 2017) or decomposing β into the product of word embeddings $\mathcal{W}$ and topic embeddings $\mathcal{T}$. Alternatively, (Wu et al., 2023) propose another form of β that effectively addresses the issue of topic collapse as follows:

$$\beta_{ij} = \frac{\exp\left(-\|\mathbf{w}_i - \mathbf{t}_j\|^2/\tau\right)}{\sum_{j'=1}^K \exp\left(-\|\mathbf{w}_i - \mathbf{t}_{j'}\|^2/\tau\right)}, \quad (1)$$

where τ is a temperature hyperparameter. The word embeddings $\mathcal{T}$ are typically initialized using pre-trained embeddings such as GloVe (Pennington et al., 2014).

VAE-based models aim to reconstruct the BoW representations of documents using the topic-word distribution matrix β and document-topic proportion θ_d as $\hat{x}^d \sim \text{Multinomial}(\text{softmax}(\beta\theta_d))$. The topic modeling loss consists of a reconstruction term and a regularization term, as follows:

$$\mathcal{L}_{\text{TM}} = \frac{1}{D}\sum_{d=1}^{D}\Big[-(x^d)^\top \log(\text{softmax}(\beta\theta_d)) + \text{KL}(q(\alpha|x^d)\|p(\alpha))\Big]$$

4 Methodology

4.1 Collecting Preferences with an LLM

For each topic k extract the top-M words by $\beta^{(0)}_{k,w}$ (or $p(w \mid k)$). Use an LLM to label each word

*Equal contribution

†Corresponding author: linhnv@soict.hust.edu.vn

as good (+) if it is related to the topic or bad (-) if it is not. To reduce noise, we use a voting technique: repeat the LLM query multiple times (e.g. 10) and take a majority vote for each word. We then partition the top-M list into a *good* list L^+ and a *bad* list L^-. To construct preference pairs (w^+, w^-), we randomly match words from the two lists. In practice we handle three cases:

- **Case 1:** $|L^+| > |L^-|$. We randomly pair each word in L^- with a word in L^+ (removing used words), then match the remaining words in L^+ to bad words (allowing reuse) to complete the pairing.
- **Case 2:** $|L^+| < |L^-|$. Symmetric to Case 1, swapping the roles of L^+ and L^-.
- **Case 3:** $|L^+| = |L^-|$. We simply pair each word in L^+ with one word in L^- and vice versa.

Store the resulting preference dataset as $\mathcal{D}_{pref} = \{(k, w^+, w^-)\}$.

4.2 LLM-guide preference learning in Neural Topic Model

In topic-word distribution, each top word will represent the relationship between it and the topics. Therefore, enhance the preference of top words in topic can improve the semantic quality of the topics and the unrelated words can be pushed to the back. To achieve that, we implement preference learning in the concept of neural topic models.

To be specific, after training neural topic model, for each topic k, we attain a base or reference policy $\pi_\phi^{ref}(w|x,k) = softmax(\beta_k^{ref})_w$, with ϕ is the parameters of neural topic models. The goal is to optimize the policy $\pi_\phi(w|x,k)$ such that it assigns higher probability to outputs (words for a topic) that are preferred according to the reward signal $r(k,w)$. The preference dataset used is achieved from Section 4.1 that $\mathcal{D}_{pref} = \{(k, w^+, w^-)\}$.

Given preference pairs, we want to achieve a reward model $r(k,w)$ so that it ranks preferred words higher, the Bradley-Terry stipulates that the human preference distribution P^* can be written as:

$$P^*(w^+ \succ w^- \mid k) \tag{2}$$

$$= \frac{\exp(r(k,w^+))}{\exp(r(k,w^+)) + \exp(r(k,w^-))}. \tag{3}$$

Using the sigmoid function σ, the equation becomes:

$$P(w^+ \succ w^- \mid k) = \sigma\big(r(k,w^+) - r(k,w^-)\big) \tag{4}$$

The cross-entropy (negative log-likelihood) loss for the reward model is:

$$\mathcal{L}_{\mathrm{RM}} = -\mathbb{E}_{(k,w^+,w^-)} \log \sigma\big(r(k,w^+) - r(k,w^-)\big) \tag{5}$$

The next step consists of optimizing the policy $p_\phi(w|x,k)$ through reinforcement learning in order to maximize the reward. However, directly maximizing the reward may cause the policy to deviate excessively from the base policy $\pi_\phi^{ref}(w|x,k)$, resulting in unnatural or overly optimized behavior. To mitigate this issue, a penalty term is introduced to constrain the policy:

$$\max_\phi \mathbb{E}_{x\sim \mathrm{X},\, w\sim\pi_\phi(w|x,k)} \left[r(k,w)\right] \tag{6}$$

$$- \alpha\, D_{\mathrm{KL}}[\pi_\phi(w|x,k) \,\|\, \pi_\phi^{ref}(w|x,k)] \tag{7}$$

Following prior works (Rafailov et al., 2023), it is straightforward to show that the optimal solution to the KL-constrained reward maximization objective in Eq. 7 takes the form:

$$\pi^*(w|x,k) = \frac{1}{C(x,k)}\, \pi^{ref}(w|x,k) \exp\left(\frac{r(k,w)}{\alpha}\right), \tag{8}$$

$$C(x,k) = \sum_{v\in V} \pi^{ref}(v|x,k) \exp\left(\frac{r(k,w)}{\alpha}\right) \tag{9}$$

From the box above, taking the logarithm of both sides and then with some algebra we obtain:

$$r(k,w) = \alpha \log \frac{\pi^*(w|x,k)}{\pi^{ref}(w|x,k)} + \alpha \log C(x,k). \tag{10}$$

Substituting the reparameterization in Eq. 10 for $r(k,w)$ into the preference model Eq. 4, the partition function cancels, and we can express the human preference probability in terms of only the optimal policy π^* and reference policy π^{ref}. Thus, the optimal RLHF policy π^* under the Bradley-Terry model satisfies the preference model:

$$P(w^+ \succ w^- \mid k) = \sigma\big(\alpha \log \frac{\pi^*(w^+|x,k)}{\pi^{ref}(w^+|x,k)} \tag{11}$$

$$- \alpha \log \frac{\pi^*(w^-|x,k)}{\pi^{ref}(w^-|x,k)}\big) \tag{12}$$

Now that we have the probability of human preference data in terms of the optimal policy rather than the reward model, we can formulate a maximum likelihood objective for a parametrized policy π_ϕ:

$$\mathcal{L}^k_{DPO} = -\mathbb{E}_{(k,w^+,w^-)} \log\sigma\big(\alpha \log \frac{\pi^*(w^+|x,k)}{\pi^{ref}(w^+|x,k)} \tag{13}$$

$$- \alpha \log \frac{\pi^*(w^-|x,k)}{\pi^{ref}(w^-|x,k)}\big) \tag{14}$$

Recall that $\pi^{ref}_\phi(w|x,k) = softmax(\beta^{ref}_k)_w$, we have $\pi^{ref}_\phi(w|x,k) = \frac{e^{\beta^{ref}_{kw}}}{Z_k}$ with $Z^{ref}_k = \sum_{v \in V} e^{\beta^{ref}_{kv}}$. So we have:

$$\log \frac{\pi^*(w|x,k)}{\pi^{ref}(w|x,k)} = (\beta_{kw} - \log Z_k) \tag{15}$$

$$- (\beta^{ref}_{kw} - \log Z^{ref}_k) \tag{16}$$

$$= (\beta_{kw} - \beta^{ref}_{kw}) \tag{17}$$

$$- (\log Z_k - \log Z^{ref}_k) \tag{18}$$

Finally, we have:

$$\mathcal{L}^k_{DPO} = -\mathbb{E}_{(k,w^+,w^-)} \log \sigma\big[\alpha((\beta_{k,w^+} - \beta_{k,w^-}) \tag{19}$$

$$- (\beta^{ref}_{k,w^+} - \beta^{ref}_{k,w^-}))\big] \tag{20}$$

In gradient descent process, reference β^{ref} will be frozen. Finally, we aggregate over all topics preference loss that: $\mathcal{L}_{DPO} = \sum_{k=1}^{K} \mathcal{L}^k_{DPO}$

The overall loss for LLM-guided preference neural topic model is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{TM} + \lambda \mathcal{L}_{DPO} \tag{21}$$

4.3 Preference Learning with Plackett-Luce Model

: Instead of $\mathcal{D}_{pref} = \{(k, w^+, w^-)\}$, we use $\mathcal{D}_{pref} = \{(k, w^1, w^2, ..., w^N)\}$ with N is the number of top words, $(w^1, w^2, ..., w^N)$ is the rankings of N top words proposed by LLMs. We will apply Preference Learning with Plackett-Luce Model to refine the topics.

5 Experiments

6 Conclusion

7 Limitations

Acknowledgements

References

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, pages 439–453.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, pages 37335–37357.