SAE-NTM: Sentence-Aware Encoder for Neural Topic Modeling

Hao Liu, Jingsheng Gao, Suncheng Xiang, Ting Liu, Yuzhuo Fu*
School of SEIEE, Shanghai Jiao Tong University, China
{liuh236, gaojingsheng, xiangsuncheng17, louisa_liu, yzfu}@sjtu.edu.cn

Abstract

Incorporating external knowledge, such as pre-trained language models (PLMs), into neural topic modeling has achieved great success in recent years. However, employing PLMs for topic modeling generally ignores the maximum sequence length of PLMs and the interaction between external knowledge and bag-of-words (BOW). To this end, we propose a sentence-aware encoder for neural topic modeling, which adopts fine-grained sentence embeddings as external knowledge to entirely utilize the semantic information of input documents. We introduce sentence-aware attention for document representation, where BOW enables the model to attend on topical sentences that convey topic-related cues. Experiments on three benchmark datasets show that our framework outperforms other state-of-the-art neural topic models in topic coherence. Further, we demonstrate that the proposed approach can yield better latent document-topic features through improvement on the document classification.

1 Introduction

Topic models have been widely used to identify human-interpretable topics and learn text representations, which have been applied for various tasks in Natural Language Processing (NLP) such as information retrieval (Lu et al., 2011), summarization (Nguyen et al., 2021), and semantic similarity detection (Peinelt et al., 2020). A typical topic models is based on the latent Dirichlet allocation (LDA) (Blei et al., 2003) and Bayesian inference. However, to avoid the complex and expensive iterative inference of conventional topic models, topic modeling with the deep neural network has been the leading research direction in this field (Miao et al., 2016; Srivastava and Sutton, 2017; Ding et al., 2018).

Neural topic models (NTMs) usually exploit the BOW representation as input, disregarding the syntactic and semantic relationships among the words in a document, thus leading to relatively inferior quality of topics. Recently, pre-trained language models (PLMs) (Kenton and Toutanova, 2019; Reimers and Gurevych, 2019) demonstrate their strong ability to capture sentential coherence by achieving state-of-the-art performance on many natural language processing tasks. Therefore, several approaches have been proposed to incorporate external knowledge into topic models to address the limitations of BOW. A typical method to take external knowledge as additional features (Bianchi et al., 2021; Jin et al., 2021) concentrates the outputs of PLMs with BOW data. Another way (Hoyle et al., 2020) is to distill the knowledge of the teacher PLMs to generate a smoothed pseudo-document, which guides the training of a student topic model.

However, there are still limitations to the above approaches. Firstly, the document-level sequences are too long to be modeled, since the token-level sequence in the context is usually considered as input to the PLMs. Extracting the document-level semantic embedding with PLMs as external knowledge ignores the restriction on sequence length, which loses massive semantic information from input text. Secondly, the difference in learning objectives between NTMs and PLMs makes it challenging to incorporate external knowledge. The encoder of NTMs is designed to handle the sparse BOW data, unable to take into account the dense contextual document embedding from PLMs.

To address these limitations, we build upon the framework of variational autoencoders (VAE) (Kingma and Welling, 2013) and propose a sentence-aware encoder for incorporating external semantic knowledge into topic models. The proposed approach integrates the advantages of NTMs and PLMs as encoders. Specifically, the encoder of the topic model is responsible for processing document-level BOW data like most NTMs, while the PLMs is used to encode sentence-

*Corresponding Author

Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023), pages 106–111

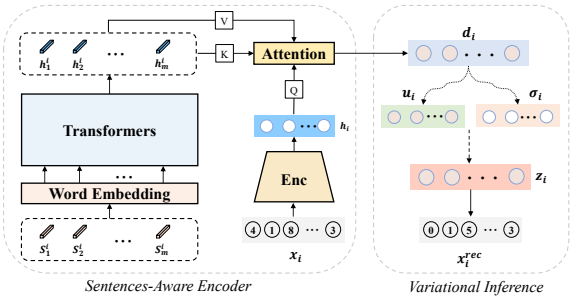

Figure 1: Basic Architecture of SAE-NTM. The sentence-aware encoder deals with the BOW data $\boldsymbol{x}_i$ and sentence sequences $\{\boldsymbol{s}_1^i, \cdots, \boldsymbol{s}_m^i\}$ of the i^{th} document, while variational inference reconstructs the BOW data $\boldsymbol{x}_i^{rec}$ from document representation $\boldsymbol{d}_i$.

level semantic information as its original training objective. Different from previous approaches, our proposed framework considers cross-attention (Vaswani et al., 2017) between the BOW data and sentence embeddings, which leverages fine-grained semantic information for topic discovery.

To summarize, the main contributions of this paper are as followed: (1) We propose a novel framework **SAE-NTM**: **S**entences-**A**ware **E**ncoder for **N**eural **T**opic **M**odeling which leverages the cross-attention for incorporating external semantic knowledge in a sentence-aware manner. (2) Quantitative and qualitative experiments demonstrate that our proposed approach significantly outperforms the existing state-of-the-art topic models in topic coherence. (3) We show that the BOW-guided attention yields practical latent document-topic features, achieving better performance on the document classification task.

2 Methodology

2.1 SAE: Sentence-Aware Encoder

In this section, we introduce the sentence-level semantic information as external knowledge and propose a method to efficiently combine BOW data with external knowledge for document representations, as shown in Figure 1.

Encoder for bag-of-words and sentence sequences. Neural topic models with variational autoencoders usually take high-dimensional, sparse word counts $\boldsymbol{x}_i$ as input and transform it into a low-dimensional dense feature $\boldsymbol{h}_i$ to fit the variational autoencoders framework as formulated in Eq.1.

$$\boldsymbol{h}_i = Enc\left(\boldsymbol{x}_i\right) \tag{1}$$

Where $Enc : \mathbf{R}^V \rightarrow \mathbf{R}^L$ is usually a multi-layer perceptron (MLP) for the inference of the i^{th} document representation.

Complementary to the orderless BoW, the context of the document carries more affluent and more sophisticated semantic information. And it can be represented as contextual embeddings by pre-trained language models (e.g., BERT (Kenton and Toutanova, 2019)) from large corpora, which have a fine-grained ability to capture aspects of linguistic context. In this paper, we employ sentence-transformers (Reimers and Gurevych, 2019) to encode each sentence in the document as follows:

$$\left\{\boldsymbol{h}_1^i, \cdots, \boldsymbol{h}_m^i\right\} = Trans\left(\left\{\boldsymbol{s}_1^i, \cdots, \boldsymbol{s}_m^i\right\}\right) \tag{2}$$

where s_j^i is a sequence of tokens and h_j^i is the aggregated contextual embedding from the pre-trained sentence-transformers for the j^{th} sentence.

Sentence-aware Attention. The contextual embeddings $\left\{\boldsymbol{h}_1^i, \cdots, \boldsymbol{h}_m^i\right\}$ and BOW representation $\boldsymbol{h}_i$ jointly constitute the input of sentence-aware encoder. Then sentence-aware attention is employed to accomplish the interaction of word counts and semantic embeddings formulated in Eq.3.

$$\begin{aligned} \boldsymbol{d}_i &= \sum_{j=1}^{j=m} \alpha_j^i h_j^i \\ \alpha_j^i &= \frac{\exp(score(h_i, h_j^i))}{\sum_{k=1}^{k=m} \exp(score(h_i, h_k^i))} \end{aligned} \tag{3}$$

Where the representation $\boldsymbol{d}_i$ of the i^{th} document is a weighted sum of contextual embeddings $\left\{\boldsymbol{h}_1^i, \cdots, \boldsymbol{h}_m^i\right\}$ and α_j^i is the normalized attention of the j^{th} sentence. Typically, the scoring function *score* is scaled dot-product attention. Sentence embeddings as external knowledge provide rich textual information, while the BOW data guides topic model in the assignment of attention on topical sentences, which contributes to capturing the co-occurrence patterns of the words.

2.2 Variational inference

Starting with the document representation, variational inference (Kingma and Welling, 2013) consider Logistic-Normal distribution as the posterior distribution $q(\mathbf{z} \mid \mathbf{x})$, whose mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\sigma}_i$ vectors are separately derived from the document representation through a linear layer. Then the reparameterization trick in Eq.4 is used to estimate the gradient.

$$\boldsymbol{z}_i = \operatorname{softmax}\left(\boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \cdot \boldsymbol{\varepsilon}_i\right) \tag{4}$$

Method	K=50			K=200		
	20NG	Wiki	IMDb	20NG	Wiki	IMDb
W-LDA (Nan et al., 2019)	0.274 ± 0.012	0.492 ± 0.014	0.134 ± 0.003	0.159 ± 0.002	0.316 ± 0.007	0.090 ± 0.001
SCHOLAR (Chang et al., 2009)	0.322 ± 0.005	0.480 ± 0.009	0.166 ± 0.004	0.262 ± 0.003	0.416 ± 0.005	0.140 ± 0.002
CLNTM (Nguyen and Luu, 2021)	0.327 ± 0.002	0.486 ± 0.013	0.167 ± 0.002	0.267 ± 0.002	0.425 ± 0.003	0.144 ± 0.001
CTM (Bianchi et al., 2021)	0.329 ± 0.003	0.484 ± 0.016	0.176 ± 0.002	0.283 ± 0.005	0.432 ± 0.004	0.163 ± 0.004
SCHOLAR + BAT (Hoyle et al., 2020)	0.343 ± 0.006	0.501 ± 0.007	0.170 ± 0.004	0.301 ± 0.002	0.437 ± 0.003	0.160 ± 0.002
SAE-NTM (Ours)	**0.352 ± 0.006**	**0.511 ± 0.011**	**0.196 ± 0.001**	**0.314 ± 0.002**	**0.472 ± 0.004**	**0.174 ± 0.002**

Table 1: Results of average NPMI scores with 50 and 200 topics on three datasets. For each group of results, we repeat the experiment five times with different random initialization and report the standard deviation.

where $\varepsilon_i \sim \mathcal{N}(0, 1)$ denotes samples from the normal distribution and $\boldsymbol{z}_i$ is the latent document-topic vector. Next, it attempts to reconstruct the original BOW data $\boldsymbol{x}_i$ by modeling the words distributions of topics $\boldsymbol{\phi}$ as follows:

$$\boldsymbol{x}_i^{rec} \sim \text{Multi}\left(\text{softmax}\left(\boldsymbol{z}_i \boldsymbol{\phi}^T\right), \boldsymbol{N}\right) \quad (5)$$

where $\phi \in \mathbf{R}^{V \times K}$ is the word-topic matrix and $\mathbf{N}$ is a vector of document lengths.

Finally, SAE-NTM are trained by maximizing the Evidence Lower Bound (ELBO) of the marginal likelihood of the BoW data:

$$\mathcal{L}(\mathbf{x}) = -\mathbf{E}_q\left[\log p(\mathbf{x} \mid \mathbf{z})\right] + \mathbf{KL}\left[q(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})\right] \quad (6)$$

where $\log p(\mathbf{x} \mid \mathbf{z})$, $q(\mathbf{z} \mid \mathbf{x})$ and $p(\mathbf{z})$ are respectively the reconstructed data likelihood, the posterior distribution and prior Dirichlet distribution.

3 Experiments

In this section, we design empirical experiments to answer the following questions of concern in topic modeling. First, how effectively does SAE-NTM perform quantitatively and qualitatively in terms of topic quality? Second, how does SAE-NTM perform in automated document-topic inference for downstream tasks? Besides, more details about the impact of external knowledge on topic modeling can be found in Appendix A.

3.1 Experimental Settings

Datasets. We evaluate our proposed SAE-NTM on three benchmark datasets, which differ significantly in the domain, vocabulary size, and document length: 20Newsgroups (20NG, Lang, 1995) [1], Wikitext-103 (Wiki, Merity et al., 2016) [2], IMDb movie reviews (IMDb, Maas et al., 2011) [3]. For consistency with prior work, we adopt the same preprocessing steps and train/dev/test split from the original papers for 20NG (i.e., 48/12/40), Wiki (i.e., 70/15/15), IMDb (i.e., 50/25/25).

Baselines. We compare our model with existing state-of-the-art neural topic models: W-LDA (Nan et al., 2019) is a neural model with wassestein autoencoder, which approximates the Dirichlet prior by minimizing Maximum Mean Discrepancy. SCHOLAR (Card et al., 2018) is a VAE-based neural topic model with a logistic normal prior to facilitate approximate Bayesian inference and provide a flexible way to incorporate document metadata. SCHOLAR+BAT (Hoyle et al., 2020) is a knowledge-distilled neural topic model where a BERT-based autoencoder as a teacher provides contextual knowledge for the student model. CTM (Bianchi et al., 2021) is a combined topic model with the incorporation of contextualized document embeddings in neural topic models. CLNTM (Nguyen and Luu, 2021) is a contrastive learning version of the neural topic model through a word-based sampling strategy.

3.2 Evaluation in topic coherence

Since topic models aim to discover a set of latent topics that are meaningful and useful for humans (Chang et al., 2009), we evaluate topic coherence using the Normalized Mutual Pointwise Information (NPMI) which is significantly correlated with human judgments on topic quality (Aletras and Stevenson, 2013; Lau et al., 2014). Specifically, we first select the top 10 words under each topic generated by topic models, and then estimate NPMI scores with reference co-occurrence counts from the held-out corpus, e.g. the dev or test split.

As shown in Table 1, we report the results of the average NPMI over 5 runs with different random seeds for initialization for robustness. It can be observed that our model yields the most coher-

[1] qwone.com/~jason/20Newsgroups

[2] s3.amazonaws.com/research.metamind.io/wikitext/wikitext-103-v1

[3] ai.stanford.edu/Ëœamaas/data/sentiment

Dataset	Model	NPMI	Topic Words
20NG	SCHOLAR	0.234	encryption enforcement privacy conversation *industry manufacturer* protect *administration* device
	Our Model	0.434	encryption enforcement clipper agency wiretap privacy escrow protect security secure
Wiki	SCHOLAR	0.379	opera composer repertory theatre conductor *libretto* operatic *painting* orchestral *painter*
	Our Model	0.632	cantata bach recitative oboe continuo soloist chorale viola soprano violin
IMDb	SCHOLAR	0.161	religious beliefs christian *society* christ *views* portray *racist issues* jesus
	Our Model	0.333	christ christian religion church jesus religious bible faith god beliefs

Table 2: Some example topics on three datasets, where the italic words are less relevant to the topic.

ent topics across all baselines for three benchmark datasets in NPMI scores. This demonstrates that our method promotes the overall quality of generated topics. More importantly, our model not only significantly outperforms the baseline without external knowledge such as SCHOLAR, but also surpasses other state-of-the-art neural topic models that incorporate external knowledge, such as CTM, SCHOLAR+BAT. It suggests that our approach is more efficient than others for incorporating external knowledge into neural topic models.

In addition to the quantitative evaluation, we also randomly extract sample topics from three datasets to gain an intuitive view on the quality of generated topics, as shown in Table 2. Obviously, the topic words generated by our model capture the concept of topics in the document rather than the baseline model. For example, it can be noticed that in the 20NG dataset our words are closely related to encryption (*agency, wiretap, etc.*), rather than some common words (*industry, manufacturer, etc.*) from SCHOLAR. The words generated by our model in Wiki are more focused on *cantata* and *opera*, while SCHOLAR drifts gradually away from the music topic to *paintings*. Similarly in the IMDb dataset, the topic words generated by our model reflect religion-related themes, which is different from SCHOLAR including off-topic words such as *views, racist, etc*.

3.3 Document Classification

Since the latent vectors inferred by neural topic models can be applied as text features (Nan et al., 2019), we employ the downstream task of document classification to compare the predictive performance of the models in addition to the evaluation of topic coherence. Specifically, we collect latent document-topic features from the trained neural topic models setting number of topics to 50 and use these vectors as inputs to train a Random Forest classifier on the training split separately.

Model	20NG	IMDb
W-LDA	52.3	80.3
SCHOLAR	62.8	82.7
CLNTM	58.4	79.5
CTM	62.4	84.5
SCHOLAR + BAT	65.2	83.1
SAE-NTM (Ours)	**66.1**	**85.9**

Table 3: Test Accuracy between different topic models on document classification.

We report classification accuracy on the test split of 20NG and IMDb in Table 3. It is worth noting that we aim to evaluate the predictive capability of topic models by the performance in document classification, rather than training the model to obtain higher accuracy. The document-topic features provided by our proposed model achieve best accuracy for all the datasets with a significant improvement. It demonstrates that the proposed sentence-aware encoder not only discovers topics that are more meaningful to humans, but also learns better latent document features.

4 Conclusions

In this paper, we propose a Sentence-Aware Encoder for Neural Topic Modeling framework: SAE-NTM to incorporate external knowledge into neural topic models. The proposed method can capture document information by performing attention on sequential sentences in a bag-of-words guided manner. Extensive experiments have shown that our framework can achieve state-of-the-art performance in topic coherence and encode better latent document-topic features. In the future, we would like to explore the possibility of integrating our approach with neural topic models built on other frameworks, such as generative adversarial training (Nan et al., 2019; Wang et al., 2020).

Limitations

The proposed model with sentence-aware encoder aims to efficiently incorporate external knowledge and bag-of-words for topic modeling, which means that in this work we are mainly interested in how documents should be encoded for topic inference. However, the decoder of topic models can also be coupled with word embeddings through factorization, such as embedded topic models (Dieng et al., 2020). It is worth exploring how hierarchical semantic embeddings can be employed for topic modeling with our model.

In this paper, we do not conduct any fine-tuning for the pre-trained language model. Our approach reveals how the frozen pre-trained language model can be effectively used to improve the performance of the topic model with limited computational overhead, given that the parameter size of the pre-trained language model is much larger than that of the topic model. Moreover, fine-tuning pre-trained language models for topic modeling as an unsupervised learning task (Mueller and Dredze, 2021) is challenging.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *ACL*, pages 759–766.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *ACL*, pages 2031–2040.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

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.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. *arXiv preprint arXiv:1809.02687*.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving neural topic models using knowledge distillation. In *EMNLP*, pages 1752–1771.

Yuan Jin, He Zhao, Ming Liu, Lan Du, and Wray Buntine. 2021. Neural attention-aware hierarchical topic model. In *EMNLP*, pages 1042–1052.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. 2011. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Aaron Mueller and Mark Dredze. 2021. Fine-tuning encoders for improved monolingual and zero-shot polylingual neural topic modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3054–3068.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. In *ACL*, pages 6345–6381.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *Advances in Neural Information Processing Systems*, 34:11974–11986.

Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *arXiv preprint arXiv:2109.10616*.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tbert: Topic models and bert joining forces for semantic similarity detection. In *ACL*, pages 7047–7055.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

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.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020. Neural topic modeling with bidirectional adversarial training. *arXiv preprint arXiv:2004.12331*.

Andrew KC Wong and Manlai You. 1985. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, (5):599–609.

A Analysis on individual topics

To evaluate whether our improvements are meaningful on individual topics, we directly compare each of the aligned topics generated by the baseline SCHOLAR without external knowledge and our model. Follow previous works (Hoyle et al., 2020; Nguyen and Luu, 2021), we align the topics by using a variation of competitive linking to greedily approximate the optimal weight of the bipartite graph matching. And the weight of each link is calculated based on the similarity between their word distributions as measured Jenson-Shannon (JS) divergence (Wong and You, 1985; Lin, 1991). We iteratively select the topic pair with the lowest score based on JS divergence, separate the two topics from the topic list, and repeat until the rest JS score exceeds a certain threshold.

Figure 2 shows the JS-divergences for aligned topic pairs for three benchmark corpora. Based on visual inspection, we choose the most aligned 44 topic pairs to conduct the comparison, since there is no conceptual relationship between topic pairs

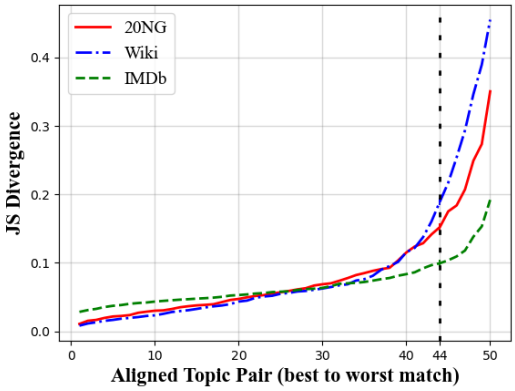

Figure 2: Jensen-Shannon divergence for aligned topic pairs of SCHOLAR and our model.

beyond this point and employ the same threshold across all three datasets for simplicity. Considering these aligned topic pairs conceptually related, we explore the impact of external knowledge on the baseline topic model on a topic-by-topic basis as shown in Figure 3. It can be observed that the number of topics with high NPMI scores from our model is apparently more than that of the baseline model. This means that the overall promotion achieved by our approach can be interpreted as identifying the topic space generated by the baseline models and in most cases, improving the coherence of individual topics.

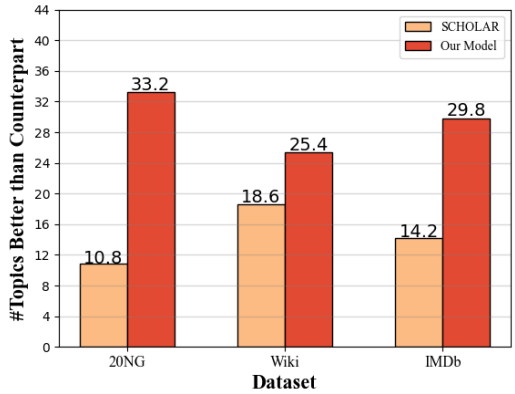

Figure 3: The number of aligned topic pairs which our model improves upon SCHOLAR model.