Supplementary Materials for “Context-guided Embedding Adaptation for Effective Topic Modeling in Low-Resource Regimes”

Anonymous Author(s)
Affiliation
Address
email

1 Key Notations

In Table 1, we list the key notations, descriptions and corresponding dimensions used in this paper.

Table 1: Notations used in the paper.

Symbol	Dimensionality	Description
M	-	number of total training tasks
J	-	number of documents in each task
K	-	number of topics in each task
V	-	number of vocabulary terms, shared across tasks
D	-	dimensionality of the word latent space
$\mathcal{T}^{(i)}$	-	the i^{th} training task
$\mathbf{X}^{(i)}$	$\mathbb{R}^{V\times J}$	the BoWs representations for documents in the i^{th} task
$\mathbf{H}^{(i)}$	$\mathbb{R}^{300\times J}$	the deterministic hidden features of BoWs $\mathbf{X}^{(i)}$
$\boldsymbol{c}^{(i)}$	$\mathbb{R}^{K}$	context variable that summarizes the topic proportion information
$\boldsymbol{\theta}_j^{(i)}$	$\mathbb{R}^{K}$	topic proportion of the j^{th} in the i^{th} task
$\boldsymbol{\beta}^{(i)}$	$\mathbb{R}^{V\times K}$	topic-word matrix for the i^{th} task
$\mathbf{A}^{(i)}$	$\mathbb{R}^{V\times V}$	the adjacency matrix of dependency graph for the i^{th} task
$\mathbf{e}_v^{(i)}$	$\mathbb{R}^{D}$	initialized features of the v^{th} word appeared in the i^{th} task
$\mathbf{z}_v^{(i)}$	$\mathbb{R}^{D}$	adaptive embedding of the v^{th} word appeared in the i^{th} task
$\pi_k^{(i)}$	-	coefficient of the k^{th} Gaussian component for the i^{th} task
$\boldsymbol{\mu}_k^{(i)}$	$\mathbb{R}^{D}$	mean of the k^{th} Gaussian component for the i^{th} task
$\boldsymbol{\Sigma}_k^{(i)}$	$\mathbb{R}^{D\times D}$	covariance of the k^{th} Gaussian component for the i^{th} task

2 Algorithms for training and testing

In this section, we present the training and meta-testing procedures of our Meta-CETM in Alg. 1 and Alg. 2, respectively.

Algorithm 1: Training process

Input: A set of training corpora $\{\mathcal{D}_c\}_{c=1}^{C}$; initialized model parameters Ψ

Randomly sample tasks from each training corpus $\mathcal{D}_c$ to obtain $\{\mathcal{T}^{(i)}\}_{i=1}^{M}$;

for *each task* $\mathcal{T}^{(i)}, i = 1, 2, \cdots, M$ **do**

  Build semantic graph $\mathbf{A}^{(i)}$ with established dependency parsing tools;

  Infer adaptive word embeddings $\mathbf{Z}^{(i)}$ according to Eq. 6;

  Initialize parameters of the Gaussian mixture prior: π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$;

  Update to the optimal value $\pi_k^{(i)}$, $\boldsymbol{\mu}_k^{(i)}$ and $\boldsymbol{\Sigma}_k^{(i)}$ using EM based on Eq. 7;

  Compute the topic-word matrix $\boldsymbol{\beta}^{(i)}$ according to Eq. 3;

  Infer the latent context varibale $\boldsymbol{c}^{(i)}$ using Eq. 5;

  for *each document* $\mathbf{x}_j^{(i)}, j = 1, 2, \cdots, J$ **do**

    Infer topic proportion $\boldsymbol{\theta}_j^{(i)}$ with Eq. 4;

    Calculate the log-likelihood $p(\mathbf{x}_j^{(i)} | \boldsymbol{\theta}_j^{(i)}, \boldsymbol{\beta}^{(i)})$;

  Derive the ELBO as Eq. 8 and update Ψ using SGD;

Algorithm 2: Meta-test for a new task

Input: A new corpus $\mathcal{D}_{test}$, trained model parameters Ψ

Output: Adaptive topic-word matrix $\boldsymbol{\beta}$

Randomly sample a task $\mathcal{T}_{new}$ from the given corpus $\mathcal{D}_{test}$;

Get the corresponding BoWs $\mathbf{X}_{new}$ and dependency graph $\mathbf{A}_{new}$ for the current task;

Infer the adaptive word embeddings $\mathbf{Z}_{new}$ with part of the trained model parameters Ψ;

Initialize parameters of the Gaussian mixture prior: π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$;

Compute optimal π_k^*, $\boldsymbol{\mu}_k^*$ and $\boldsymbol{\Sigma}_k^*$ using EM based on Eq. 7;

Derive the adaptive topic-word matrix $\boldsymbol{\beta}_{new}$ by Eq. 3;

3 An Illustration of Our Settings

In the main paper, we mention corpus, task, document, support set and query set to present our framework, which is a bit messy to follow. Here, we provide a clarification of these mechanics following the literature in few-shot learning problems for better understanding.

Considering the 20Newsgroups [1] (20NG) dataset, we refer to a "**corpus**" as a collection of documents belonging to the same class so that 20NG consists of 20 corpora, each of which contains documents from one of the 20 classes.

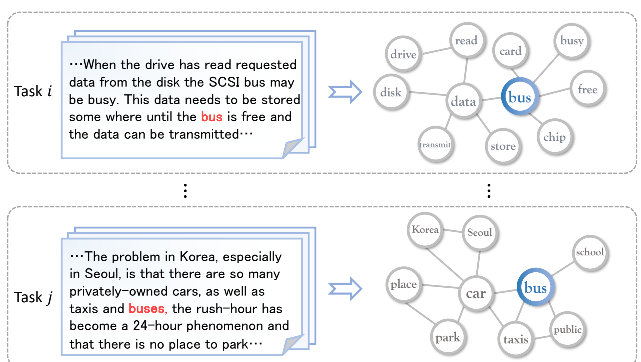

Figure 1: An illustration of word sense variation caused by different contexts. The task i is sampled from a corpus about "hardware", and the task j is sampled from a corpus related to "autos".

Further, a "**task**" is a smaller unit than a "corpus", which only comprises a few (typically 5 or 10) related documents. Consequently, we could sample a number of tasks from each training corpus (we select 12 out of the 20 corpora for training). Then our goal is to utilize these sampled tasks to train a generalizable topic model that can efficiently adapt to a new task from the test corpus (the remaining 8 corpora are used for testing). In addition, for each task at the testing stage, we split its documents into two parts, one for fine-tuning or retraining the topic model, called the **support set**, and the other for evaluating the model's performance, called the **query set**. Note that we do not design different generative processes for the corpus documents versus the task documents. In essence, our proposed

Meta-CETM only characterizes the generative process of the task documents by jointly modeling the syntactic graph $\mathbf{A}$ and the observed BoW $\mathbf{X}$ in each task. In Fig. 1, we visualize the task and the corresponding unweighted dependency graph $\mathbf{A}$.

4 Derivation of Formulas

In this section, we provide the detailed derivation process of variational evidence lower bound (ELBO) in Eq. 8 and the expectation maximization solver process for multivariate Gaussian distribution in Eq. 7 in our main paper.

4.1 Variational ELBO

$$
\begin{aligned}
\log p(X^{(i)}, A^{(i)}) &= \log \iiint p(X^{(i)}, A^{(i)}, \Theta^{(i)}, c^{(i)}, Z^{(i)}) d\Theta^{(i)} dc^{(i)} dZ^{(i)} \\
&= \log \iiint p(X^{(i)} \mid \Theta^{(i)}, Z^{(i)}) p(\Theta^{(i)} \mid c^{(i)}) p(c^{(i)}) p(A^{(i)} \mid Z^{(i)}) p(Z^{(i)}) d\Theta^{(i)} dc^{(i)} dZ^{(i)} \\
&= \log \mathbb{E}_Q \left[\frac{p(X^{(i)} \mid \Theta^{(i)}, Z^{(i)}) p(\Theta^{(i)} \mid c^{(i)}) p(c^{(i)}) p(A^{(i)} \mid Z^{(i)}) p(Z^{(i)})}{q(\Theta^{(i)} \mid X^{(i)}, c^{(i)}) q(c^{(i)} \mid X^{(i)}) q(Z^{(i)} \mid A^{(i)}, E^{(i)})} \right] \\
&\geq \mathbb{E}_Q \left[\log \frac{p(X^{(i)} \mid \Theta^{(i)}, Z^{(i)}) p(\Theta^{(i)} \mid c^{(i)}) p(c^{(i)}) p(A^{(i)} \mid Z^{(i)}) p(Z^{(i)})}{q(\Theta^{(i)} \mid X^{(i)}, c^{(i)}) q(c^{(i)} \mid X^{(i)}) q(Z^{(i)} \mid A^{(i)}, E^{(i)})} \right] \\
&= \mathbb{E}_Q \left[\log \prod_{j=1}^{J} p(x_j^{(i)} \mid \theta_j^{(i)}, Z^{(i)}) \right] + \mathbb{E}_Q \left[\log \prod_{j=1}^{J} \frac{p(\theta_j^{(i)} \mid c^{(i)})}{q(\theta_j^{(i)} \mid x_j^{(i)}, c^{(i)})} \right] \\
&+ \mathbb{E}_Q \left[\log \frac{p(c^{(i)})}{q(c^{(i)} \mid X^{(i)})} \right] + \mathbb{E}_Q \left[\log p(A^{(i)} \mid Z^{(i)}) \right] + \mathbb{E}_Q \left[\log \frac{p(Z^{(i)})}{q(Z^{(i)} \mid A^{(i)}, E^{(i)})} \right] \\
&= \sum_{j=1}^{J} \mathbb{E}_Q \left[\log p(x_j^{(i)} \mid \theta_j^{(i)}, Z^{(i)}) \right] + \sum_{j=1}^{J} \mathbb{E}_Q \left[\log \frac{p(\theta_j^{(i)} \mid c^{(i)})}{q(\theta_j^{(i)} \mid x_j^{(i)}, c^{(i)})} \right] \\
&+ \mathbb{E}_Q \left[\log \frac{p(c^{(i)})}{q(c^{(i)} \mid X^{(i)})} \right] + \mathbb{E}_Q \left[\log p(A^{(i)} \mid Z^{(i)}) \right] + \mathbb{E}_Q \left[\log \frac{p(Z^{(i)})}{q(Z^{(i)} \mid A^{(i)}, E^{(i)})} \right] \\
&= \mathcal{L}_{ELBO}
\end{aligned}
\tag{1}
$$

4.2 Solving topic parameters $\{\pi_k^{(i)}, \mu_k^{(i)}, \Sigma_k^{(i)}\}_{k=1}^{K}$ with Expectation Maximization

The log likelihood function is given by

$$
\ln p(Z^{(i)} \mid \pi^{(i)}, \mu^{(i)}, \Sigma^{(i)}) = \sum_{v=1}^{V} \ln \left[\sum_{k=1}^{K} \pi_k^{(i)} \mathcal{N}(z_v^{(i)} \mid \mu_k^{(i)}, \Sigma_k^{(i)}) \right]. \tag{2}
$$

1. Deriving $\mu_k^{(i)}$

Setting the derivatives of $\ln p(Z^{(i)} \mid \pi^{(i)}, \mu^{(i)}, \Sigma^{(i)})$ w.r.t the means $\mu_k^{(i)}$ to zero, we have

$$
-\sum_{v=1}^{V} \frac{\pi_k^{(i)} \mathcal{N}(z_v^{(i)} \mid \mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{s=1}^{K} \pi_s^{(i)} \mathcal{N}(z_v^{(i)} \mid \mu_s^{(i)}, \Sigma_s^{(i)})} \Sigma_k^{(i)} (z_v^{(i)} - \mu_k^{(i)}) = 0. \tag{3}
$$

Define the posterior probabilities as

$$
\gamma_{vk} = p(y_v^{(i)} = k \mid z_v^{(i)}) = \frac{\pi_k^{(i)} \mathcal{N}(z_v^{(i)} \mid \mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{s=1}^{K} \pi_s^{(i)} \mathcal{N}(z_v^{(i)} \mid \mu_s^{(i)}, \Sigma_s^{(i)})}. \tag{4}
$$

Multiplying by ${\Sigma_k^{(i)}}^{-1}$ and rearranging, we can obtain the updating formula for $\mu_k^{(i)}$ as

$$
\mu_k^{(i)} = \frac{\sum_v \gamma_{vk} \cdot z_v^{(i)}}{\sum_v \gamma_{vk}}. \tag{5}
$$

2 Deriving $\Sigma_k^{(i)}$

Similarly, we set the derivatives of $\ln p(Z^{(i)} \mid \pi^{(i)}, \mu^{(i)}, \Sigma^{(i)})$ w.r.t $\Sigma_k^{(i)}$ to zero, then we have

$$-\frac{1}{2}\sum_{v=1}^{V}\frac{\pi_k^{(i)}\mathcal{N}(z_v^{(i)} \mid \mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{s=1}^{K}\pi_s^{(i)}\mathcal{N}(z_v^{(i)} \mid \mu_s^{(i)}, \Sigma_s^{(i)})}{\Sigma_k^{(i)}}^{-1}\left[1 + (z_v^{(i)} - \mu_k^{(i)})^T {\Sigma_k^{(i)}}^{-1}(z_v^{(i)} - \mu_k^{(i)})\right] = 0. \quad (6)$$

Using γ_{vk} in Eq. 4 and rearranging, we get the updating formula for $\Sigma_k^{(i)}$ as

$$\Sigma_k^{(i)} = \frac{\sum_v \gamma_{vk} \cdot (z_v^{(i)} - \mu_k^{(i)})(z_v^{(i)} - \mu_k^{(i)})^T}{\sum_v \gamma_{vk}}. \quad (7)$$

3 Deriving $\pi_k^{(i)}$

Finally, using Lagrange multiplier algorithm, our goal is to maximize the following formula:

$$\sum_{v=1}^{V}\ln\left[\sum_{k=1}^{K}\pi_k^{(i)}\mathcal{N}(z_v^{(i)} \mid \mu_k^{(i)}, \Sigma_k^{(i)})\right] + \lambda(\sum_{k=1}^{K}\pi_k^{(i)} - 1), \quad (8)$$

where $\sum_{k=1}^{K}\pi_k^{(i)} = 1$.

Then setting the derivatives of the above equation w.r.t $\pi_k^{(i)}$ to zero, we have

$$\sum_{v=1}^{V}\frac{\pi_k^{(i)}\mathcal{N}(z_v^{(i)} \mid \mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{s=1}^{K}\pi_s^{(i)}\mathcal{N}(z_v^{(i)} \mid \mu_s^{(i)}, \Sigma_s^{(i)})} + \lambda = 0. \quad (9)$$

Multiplying $\pi_k^{(i)}$ and rearranging, we obtain

$$\pi_k^{(i)} = -\frac{\sum_{v=1}^{V}\frac{\pi_k^{(i)}\mathcal{N}(z_v^{(i)}|\mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{s=1}^{K}\pi_s^{(i)}\mathcal{N}(z_v^{(i)}|\mu_s^{(i)}, \Sigma_s^{(i)})}}{\lambda} = -\frac{\sum_v \gamma_{vk}}{\lambda}. \quad (10)$$

Considering $\sum_{k=1}^{K}\pi_k^{(i)} = 1$, then $\sum_k -\frac{\sum_v \gamma_{vk}}{\lambda} = 1$, and $\lambda = \sum_v \sum_k \gamma_{vk}$.

Hence the updating formula for $\pi_k^{(i)}$ as

$$\pi_k^{(i)} = \frac{\sum_v \gamma_{vk}}{\sum_v \sum_k \gamma_{vk}}. \quad (11)$$

5 More Results

5.1 Topic quality results

In Sec. 3.2.2 in the main paper, we display the topic interpretability results including topic diversity (TD) and topic coherence (TC) of six compared methods. Except for CombinedTM [2] and ZeroShotTM [3], we carry on experiments applying another contextual topic model (CTM) CETopicTM [4] with SimCSE pretrained word embeddings[1] [5] on four datasets. The results are exhibited in Fig. 2. It can be notably noticed CETopicTM [4] achieves much competitive results on both TD and TC scores, even compared with CombinedTM [2] and ZeroShotTM [3]. Such superiority is owed to the fact that CETopicTM utilizes word embeddings learned from large-scale BERT data and it performs clustering on sentence embeddings to generate topics. In our settings, the aim is to provide a framework for training a sufficiently generalized topic model in low-resource regimes, while equipped with BERT embeddings, CETopicTM is highly likely to obtain context-related meanings in advance under most situations. But in some cases where the words or the word meanings have not been encountered or learned by BERT, such as some specialized occasions, CETopicTM may fail to extract interpretable topics.

[1]https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased

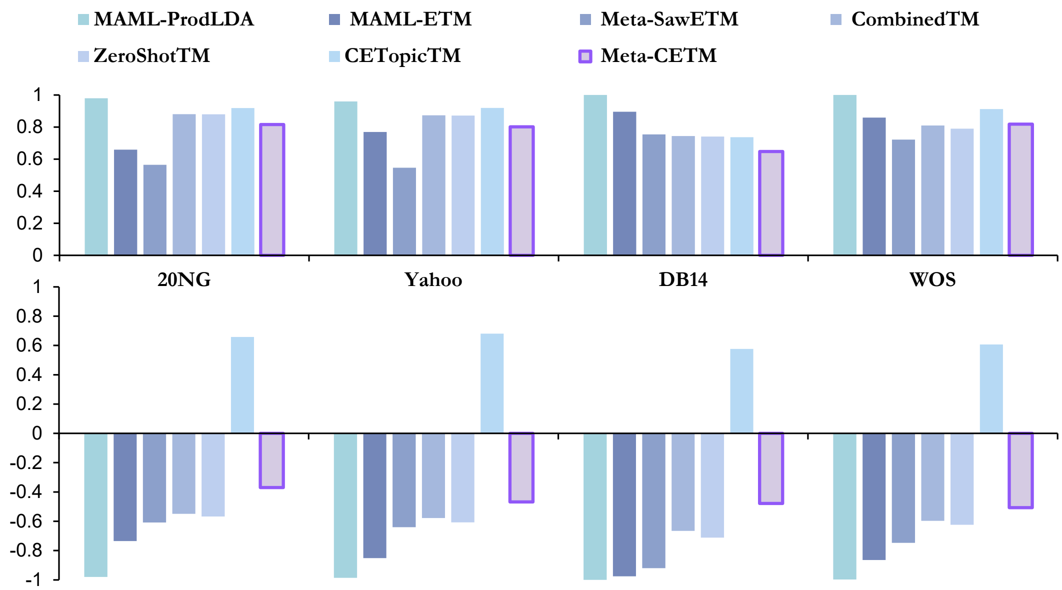

Figure 2: Topic diversity results (top row) and topic coherence results (bottom row) of seven compared methods on four datasets. Compared with Fig. 2 in main paper, we add the results of CETopicTM [4] in this figure.

5.2 Topic visualization results

In Fig. 3 in our the main paper, we visualize the adapted embedding space of different methods to demonstrate our Meta-CETM's successful fast adaption. Further, to better characterize meaningful and coherent topics learned by our model given a few number of documents, we display the text and topics extracted by Meta-SawETM [6], CombinedTM [2] and our Meta-CETM.

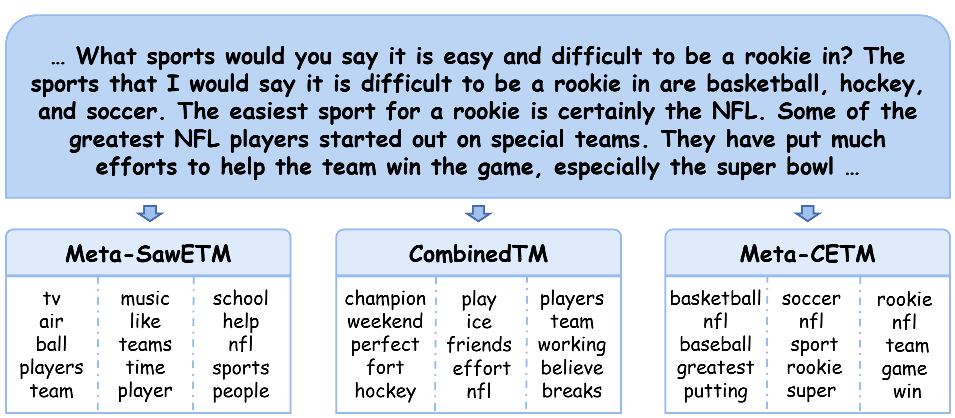

Figure 3: A paragraph of text and top five words of three topics from Meta-SawETM, CombinedTM and our Meta-CETM. It can be clearly found that Meta-CETM learns the most relevant topics among the three models.

5.3 Few-shot document classification results

In main paper, we list the classification results without intervals in Table.2 in terms of the space limit. In this section, we provide the complete results of different compared methods with confidence intervals.

Table 2: 5-way 5-shot and 5-way 10-shot few-shot text classification results with intervals. * denotes all parameters of the model are fine-tuned.

Methods		20NG		DB14	
Rep.	Alg.	5 shot	10 shot	5 shot	10 shot
MLP	MAML	32.01 ± 0.53	36.20 ± 0.21	50.20 ± 1.28	60.30 ± 0.85
	PROTO	35.20 ± 0.66	38.30 ± 0.45	54.13 ± 0.89	57.16 ± 0.72
	FT	29.70 ± 0.75	33.04 ± 0.57	51.11 ± 1.82	53.83 ± 1.74
	FT*	38.87 ± 0.51	48.52 ± 0.34	71.12 ± 1.04	77.94 ± 0.76
CNN	MAML	34.08 ± 0.41	45.40 ± 1.51	66.28 ± 1.07	75.96 ± 0.98
	PROTO	39.86 ± 0.79	49.71 ± 0.62	**78.58** ± 0.90	**81.01** ± 0.65
	FT	<u>45.70</u> ± 0.47	<u>53.63</u> ± 0.29	74.68 ± 1.58	<u>80.75</u> ± 0.96
	FT*	44.53 ± 0.71	51.92 ± 0.39	72.49 ± 1.64	80.07 ± 1.29
HNS-SawETM		39.37 ± 0.78	43.78 ± 0.93	65.93 ± 1.15	71.08 ± 0.67
Meta-SawETM		39.19 ± 0.95	45.83 ± 0.75	67.20 ± 1.53	72.31 ± 1.33
CombinedTM		46.17 ± 0.94	52.73 ± 0.69	68.42 ± 1.19	73.26 ± 1.03
ZeroShotTM		46.65 ± 0.59	52.08 ± 0.53	71.93 ± 1.74	76.09 ± 1.23
Meta-CETM		**50.57** ± 0.27	**58.47** ± 0.14	<u>76.85</u> ± 1.37	79.34 ± 1.18

References

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

[2] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*, 2020.

[3] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online, April 2021. Association for Computational Linguistics.

[4] Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. *arXiv preprint arXiv:2204.09874*, 2022.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[6] Zhibin Duan, Yishi Xu, Jianqiao Sun, Bo Chen, Wenchao Chen, Chaojie Wang, and Mingyuan Zhou. Bayesian deep embedding topic meta-learner. In *International Conference on Machine Learning*, pages 5659–5670. PMLR, 2022.