Improving Cross-Lingual Neural Topic Modeling with Document-Level Prototype-based Contrastive Learning

Anonymous ACL submission

Abstract

Cross-lingual topic modeling (CLTM) is an essential task in the field of data mining and natural language processing, aiming to extract aligned and semantically coherent topics from bilingual corpora. Recent advances in cross-lingual neural topic models have widely leveraged bilingual dictionaries to achieve word-level topic alignment. However, two critical challenges remain in cross-lingual topic modeling, the topic mismatch issue and the degeneration of intra-lingual topic interpretability. Due to linguistic diversity, some translated word pairs may not represent semantically coherent topics despite being lexical equivalents, and the objective of cross-lingual topic alignment in CLTM can consequently degrade topic interpretability within intra languages. To address these issues, we propose a novel document-level prototype-based contrastive learning paradigm for cross-lingual topic modeling. Additionally, we design a retrieval-based positive sampling strategy for contrastive learning without data augmentation. Furthermore, we introduce ProtoXTM, a cross-lingual neural topic model based on doucment-level prototype-based contrastive learning. Extensive experiments indicate that our approach achieves state-of-the-art performance on cross-lingual and mono-lingual benchmarks, demonstrating enhanced topic interpretability.

1 Introduction

Cross-lingual topic modeling (CLTM) aims to discover aligned and semantically coherent structures in bilingual corpora. CLTM has been widely applied in various natural language processing (NLP) tasks, including cross-lingual information retrieval (Vulić et al., 2013), entity linking (Zhang et al., 2013), sentiment analysis (Lin et al., 2016), and trend tracking (Tsou et al., 2020). The traditional polylingual topic model (Mimno et al., 2009) discovers aligned topics using tuple-based comparable documents in different languages.

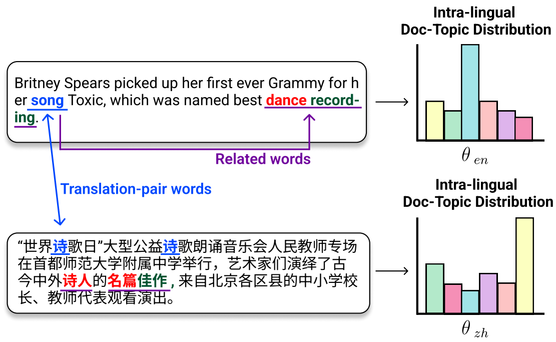

Figure 1: A motivating example of topic mismatch issue in cross-lingual topic modeling.

However, in real-world scenarios, obtaining bilingual parallel corpora is challenging. Previous studies (Shi et al., 2016; Yuan et al., 2018; Yang et al., 2019; Wu et al., 2020, 2023a) have leveraged external information, such as bilingual word dictionaries to achieve topic alignment. Despite the success of these works, cross-lingual topic modeling still faces two critical issues.

Topic Mismatch: *Do translation-based word pairs always guarantee semantically similar and well-aligned topics?* As illustrated by our motivating example in Figure 1, we observe a case where translation word pairs appear in two semantically distinct negative bilingual documents. The english word *"song"* and the chinese word *"诗"*, highlighted in blue, form a translation pair words.

The red and green words are words that are semantically related to blue anchor words within documents of each languages. However, the two documents exhibit divergent topic distributions within their respective intra-lingual corpora. This issue arises due to linguistic diversity and cultural differences.

Degenerating intra-lingual topic interpretability: We investigate the topics generated by a state-of-the-art cross-lingual neural topic model, InfoCTM (Wu et al., 2023a) and a mono-lingual neural topic model, BERTopic (Grootendorst, 2022). Table 1 presents the top-related words for the topic "music" identified by each model. In the topic produced by InfoCTM, several underlined words are aligned translation pairs and English words in parentheses are ground-truth translation of Chinese words. Although these words are correctly aligned across languages, they detract from the intra-lingual topic interpretability. In contrast, the topic generated by BERTopic comprises semantically consistent words that clearly represent the theme. This observation suggests that the objective of alignment in cross-lingual topic models such as InfoCTM can compromise intra-lingual topic interpretability. To address these issues, our proposed approach focuses on two key aspects:

InfoCTM		**BERTopic**
Topic # 13		**Topic # 157**
EN	**ZH**	**EN**
sing	秀(show)	albums
concert	高潮(climax)	chart
exhibit	唱歌(singing)	album
artist	演出(performance)	charts
album	歌(song)	soundtrack
songs	展(exhibition)	band
rap	直播(broadcast)	musicians
broadcast	演艺(performance)	singles
song	游(tour)	dj
travel	艺术家(artist)	songs

Table 1: Comparison of topics generated by InfoCTM (Wu et al., 2023a) and BERTopic (Grootendorst, 2022) on ECNews dataset.

First, We pre-train separate mono-lingual NTMs to cluster documents based on topics in each language. This prevents the deterioration of intra-lingual topic interpretability during CLTM training. Second, unlike word-level alignment, we propose a document-level contrastive learning method to align topics at the document-level. However, document-level contrastive learning remain additional challenges, such as **(1)** depending data augmentation technique for generating positive samples (Nguyen et al., 2024) and **(2)** necessary high computational costs on large-scale datasets. To overcome these challenges, we propose **R**etrieval-based **P**ositive **S**ampling (**RPS**) strategy for document-level contrastive learning without data augmentation. Our RPS method leverages the traditional information retrieval algorithm, BM25 (Robertson and Zaragoza, 2009) to sample positive documents in the target language corpus. In addition, we propose a contrastive learning paradigm for cross-lingual topic modeling, termed **D**ocument-level **P**rototype-based **C**ontrastive **L**earning (**DPCL**). Unlike standard instance-wise contrastive learning, our DPCL performs contrastive learning based on topic cluster prototypes, enabling computational efficiency even with large-scale datasets. Furthermore, we introduce ProtoXTM, a cross-lingual neural topic modeling framework based on document-level prototype-based contrastive learning. ProtoXTM mitigates both the degenerating intra-lingual topic interpretability issue and the topic mismatch issue, thereby enhancing cross-lingual topic alignment while preserving the interpretability of intra-lingual topics.

In a nutshell, our main contributions can be summarized as follows:

- To the best of our knowledge, we are the first to identify two critical issues in cross-lingual topic modeling, the topic mismatch issue and the degeneration of intra-lingual topic interpretability.

- We propose DPCL, a novel document-level prototype-based contrastive learning paradigm tailored for effective cross-lingual topic modeling. Moreover, we design a retrieval-based positive sampling strategy for contrastive learning without data augmentation to support DPCL.

- We introduce ProtoXTM, a novel cross-lingual neural topic modeling framework based on document-level prototype-based contrastive learning, which addresses the topic mismatch issue and the degeneration of intra-lingual topic interpretability.

- We conduct extensive experiments on non-parallel bilingual benchmark datasets and show ProtoXTM outperforms state-of-the-art cross-lingual and mono-lingual topic model baselines, generate coherent and aligned topics and transferable document representations.

2 Related Works

Mono-lingual Topic Modeling. Inspired by Auto-Encoding Variational Bayes (Kingma and Welling, 2013) neural variational inference based on Variational AutoEncoder (VAE) has been proposed to approximate the posterior distribution. ProdLDA

(Srivastava and Sutton, 2017) overcomes the limitations of the reparameterization trick in VAE by employing a Laplacian approximation for Dirichlet parameters. Recently, (Wu et al., 2024a; Xu et al., 2023; Bianchi et al., 2021a,b; Akash and Chang, 2024) has demonstrated improved topic quality by integrating contextualized embeddings from large language models.

Cross-lingual Topic Modeling. The traditional polylingual topic model (PLTM) (Mimno et al., 2009) was introduced using a similar approach to mono-lingual probabilistic topic models like LDA (Latent Dirichlet Allocation) (Blei et al., 2003). For cross-lingual topic alignment in non-parallel corpora, privious studies (Jagarlamudi and Daumé, 2010; Shi et al., 2016; Yuan et al., 2018; Hao and Paul, 2018; Yang et al., 2019) proposed word-level topic alignment methods based on bilingual dictionaries. To the best of our knowledge, (Wu et al., 2020) were the first to propose the Neural Multilingual Topic Model (NMTM), which incorporates topic-word distributions across languages using a bilingual dictionary to achieve cross-lingual topic alignment. Subsequently, (Wu et al., 2023a) addressed dictionary limitations and repetitive topic issues by introducing a cross-lingual vocabulary linking method and mutual information maximization to align the topic-word distributions of positive word pairs across languages.

Contrastive Learning. Contrastive learning is a widely used technique in machine learning that focuses on improving data representations by learning similarities and differences between data points (Oord et al., 2018; Wu et al., 2018; Hadsell et al., 2006). In mono-lingual topic modeling, recent studies (Han et al., 2023; Wu et al., 2022; Nguyen and Luu, 2021; Nguyen et al., 2024) have leveraged contrastive learning to generate coherent topics. For cross-lingual topic modeling, contrastive learning has also been explored in aligning topics across different languages (Zosa and Pivovarova, 2022; Wu et al., 2023a). However, M3L-Contrast (Zosa and Pivovarova, 2022) exclusively relies on prealigned bilingual corpora, whereas InfoCTM (Wu et al., 2023a) applies contrastive learning on the word-level (i.e., topic-word distribution). Distinct from this work, our approach focuses on document-level contrastive learning for cross-lingual topic modeling.

3 Proposed Methodology

Problem Setting. We denote non-parallel bilingual corpus as $\boldsymbol{X_1}$,$\boldsymbol{X_2}$ on language l_1 and language l_2, which consists of M_1, M_2 documents $\{\mathbf{x_i}^{l_1}\}_{i=1}^{M_1}$, $\{\mathbf{x_j}^{l_2}\}_{j=1}^{M_2}$. Two primary goal of CLTM are **(1) topic inference and transfer**, inferring the corresponding document-topic distribution $\theta_i^{l_1}$, $\theta_j^{l_2} \in R^K$ where K is the number of topics from $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and CLTM should be a transfer between similar documents on across languages. For **(2) topic discovery and alignment**, k-th topic-word distribution $\beta_k^{l_1} \in R^{V_1}$ and $\beta_k^{l_2} \in R^{V_2}$ are semantically consistent across languages where V_1, V_2 are the vocabulary size. In addition, we mainly aim for topic alignment on across languages by considering a group of documents with similar topics on intra-lingual corpus. For this purpose, we need to integrate the informations of the document-topic distribution on the intra-lingual corpus into the CLTM training objective.

3.1 Overview: Model Architecture

In this subsection, we brief introduce our ProtoXTM architecture. We follow NMTM (Wu et al., 2020) architecture, VAE-based shared encoder network and double decoder network structure for CLTM. Inspired by (Bianchi et al., 2021b; Zosa and Pivovarova, 2022), replace input BoW with pretrained contextualized multilingual embeddings from (Reimers and Gurevych, 2019). The framework is shown in the bottom of Figure 2 and a detailed description of ProtoXTM is described in the following.

Shared Encoder Network. The shared encoder network of ProtoXTM is a multi-layer perceptron (MLP) architecture designed to encode text $\mathbf{x}^{l_1}$ and $\mathbf{x}^{l_2}$ into an unified latent space. Contextualized representation of document as input and processes it through fully connected layers with Softplus activations and dropout for regularization. The shared encoder maps the hidden representation to the μ and Σ of a Gaussian distribution using separate linear layers, followed by batch normalization (BN) to stabilize and regularize the latent space.

Unified Latent Space. Our ProtoXTM uses pretrained contextualized multilingual embeddings and shared encoder to represent texts in different languages in an unified latent space, stabilizing the comparison between semantically consistent documents. The latent representation $\mathbf{z}$ is stochastically sampled using the reparameterization trick (Kingma and Welling, 2013), formulated as $\mathbf{z} = \mu + \Sigma \odot \epsilon$, where $\odot$ denotes the Frobenius inner product and $\epsilon \sim \mathcal{N}(0,1)$. Here, $\mathbf{z}^{l_1}$ and $\mathbf{z}^{l_2}$ represent the latent representations of documents in languages l_1 and l_2, respectively. The topic representation is further normalized into a probability simplex to obtain the document-topic distribution matrix $\boldsymbol{\theta}^{l_1}, \boldsymbol{\theta}^{l_2} \in \Delta^K$ by a softmax function $\boldsymbol{\theta}^{l_1} = \text{softmax}(\mathbf{z}^{l_1})$ and $\boldsymbol{\theta}^{l_2} = \text{softmax}(\mathbf{z}^{l_2})$. Our DPCL method can consider both intra-lingual and cross-lingual topics of a document in an unified latent space.

Double Decoder Network. The double decoder network of ProtoXTM is designed to independently reconstruct BoW representations for different languages while leveraging an unified latent topic space. Each language has a dedicated decoder consisting of topic-word distribution matrix $\boldsymbol{\beta}^{l_1}$, $\boldsymbol{\beta}^{l_2}$ and a corresponding BN layer to stabilize reconstruction documents.

3.2 ProtoXTM Framework

In this subsection, we introduce our ProtoXTM framework. Our framework ProtoXTM consists of the following three stages.

3.2.1 Stage 1: Pre-training and Document Clustering

Our one of the primary goal is to achieve topic alignment across languages while maintaining intra-lingual topic coherence. Recent studies (Sia et al., 2020; Grootendorst, 2022; Han et al., 2023) have demonstrated that clustering-based topic modeling approaches can effectively discover coherent topics. However, document clustering heavily depends on the quality of contextualized embeddings (Zhang et al., 2022). As an alternative, we apply a standard mono-lingual NTM, CTM (Bianchi et al., 2021b) to infer the document-topic distributions for each intra-lingual corpus. Based on the inferred document-topic distributions, we assign each document to the topic with the highest probability as follows:

$$\theta_i^l = [p_1, p_2, \ldots, p_k], \tag{1}$$

$$label(\mathbf{x_i}^l) := \arg \max_{n \in \{1,\ldots,k\}} p_n, \tag{2}$$

where $\sum_{n=1}^{k} p_n = 1, \quad p_n \geq 0 \, \forall n$. Denoted by $label(\mathbf{x_i}^l)$ is a cluster pseudo label of document $\mathbf{x_i}^l$ and θ_i^l is a doc-topic distribution of document $\mathbf{x_i}^l$ on intra-lingual corpus of language l. Our approach serves as a pseudo-labeling mechanism, clustering documents in the intra-lingual corpus according to their most probable topic.

3.2.2 Stage 2: Retrieval-based Positive Sampling

Our positive sampling strategy for document-level contrastive learning consists of two-step process, as described in Figure 2, in the upper right corner. In Figure 2, we illustrate only the scenario in which l_1 serves as the source language and l_2 as the target language.

Topic Translation and Word Replacement. For sample semantically similar documents (i.e., positive samples) across languages for each cluster, we translate the topic representations obtained from Stage 1 using a pre-trained neural machine translation model (M2M) (Fan et al., 2021). Specifically, the top-k words representing each topic are concatenated into a single sentence, which is then translated at the sentence level. The translated sentence is subsequently split back into individual words. If any translated word does not exist in the target vocabulary set, it is replaced with its nearest neighbor in the vocabulary using a pre-trained word embedding (FastText) (Bojanowski et al., 2017).

Retrieval-based Positive Sampling. We utilize BM25, a traditional ranking function in the field of information retrieval. For each query within a topic, BM25 is used to compute the relevance scores between the query and all documents in the target language corpus. The BM25 scores for all queries within the topic are then summed to compute the BM25 score for the each topic, as follows:

$$\text{BM25}(D_j^t, Q_k) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D_j^t) \cdot (m_1 + 1)}{f(q_i, D_j^t) + m_1 \cdot \left(1 - b + b \cdot \frac{|D_j^t|}{\text{avgdl}}\right)}, \tag{3}$$

where $f(q_i, D_j^t)$ is the number of times that the keyword q_i occurs in a document D_j^t, $|D_j^t|$ is the length of the document D_j^t in the words, $avgdl$ is the average document length in the text collection from which documents are drawn. m_1 and b are hyper-parameters for BM25, denoted by $D_j^t \in \boldsymbol{X_t} = \{D_1^t, \ldots, D_N^t\}$, where $\boldsymbol{X_t}$ is target language corpus and Q_k denote the query set of words representing the $k-th$ topic in the source language, defined as $Q_k = \{q_1, q_2, \ldots, q_n\}$, where $q_i \in Q_k$. $\text{IDF}(q_i)$ is the inverse document

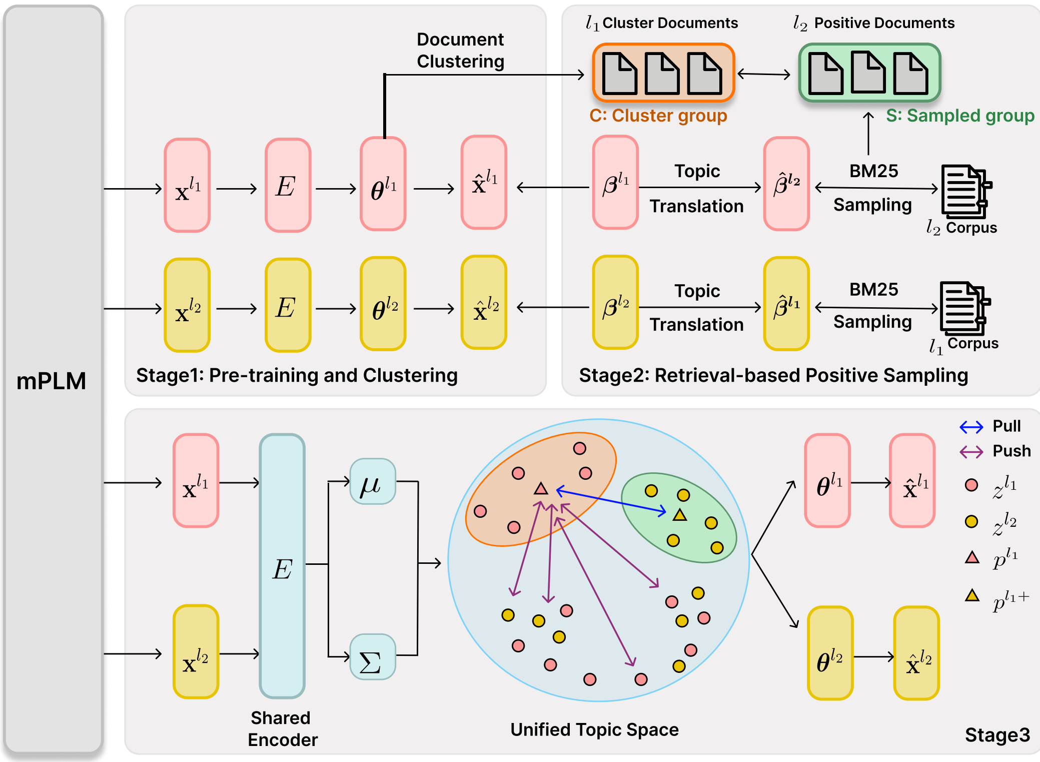

Figure 2: Overall process of our proposed methodology.

frequency (IDF) for the query's keyword q_i, and BM25 takes into account the extent to which that keyword appears in the $\boldsymbol{X_t}$. We define the top-n documents with the highest BM25 scores as the positive samples for the cluster representing the corresponding topic. The our entire positive sampling strategy is performed bidirectionally across different languages.

3.2.3 Stage 3: Topic Alignment by DPCL

In this subsection, we propose a novel contrastive learning method named DPCL, which employs document-level prototype-based contrastive learning instead of standard instance-wise contrastive learning. We use InfoNCE (Oord et al., 2018) to compute the loss functions for both directions. From the stage1 and stage2, we obtain $i-th$ cluster group c_i of In the intra-lingual corpus of language l_1, denoted as $c_i = \{z_1^{l_1}, \dots, z_m^{l_1}\}$ and $i-th$ sampled group s_i of language l_2 corpus, denoted as $s_i = \{z_1^{l_2}, \dots, z_n^{l_2}\}$, where $C = \{c_1, \dots, c_k\}$, $S = \{s_1, \dots, s_k\}$. Denoted by C and S are entire cluster group set and entire sampled group set, respectively. The entire group of documents belonging to the same cluster is treated as the anchor. The anchor feature is defined as the prototypes of all documents in the mini-batch that belong to the each cluster. Similarly, the contrastive feature is defined as the prototypes of all positive samples from the other language that are associated with the anchor cluster in the mini-batch. For a given source language l_1 and target language l_2, we compute the anchor prototype $p_i^{l_1}$ and the positive prototype $p_i^{l_1+}$ as follows:

$$p_i^{l_1} = \frac{1}{m} \sum_{k=1}^{m} z_k^{l_1}, \quad z_k^{l_1} \in c_i \tag{4}$$

$$p_i^{l_1+} = \frac{1}{n} \sum_{k=1}^{n} z_k^{l_2}, \quad z_k^{l_2} \in s_i \tag{5}$$

For the anchor prototype, negative samples include all documents in the mini-batch except for those belonging to the anchor cluster and its positive samples in the other language. Since documents in the same language but belonging to different clusters are expected to represent different topics, our negative sampling strategy considers intra-lingual topic distributions while enabling alignment with other language documents that share similar topics. $\mathcal{L}_{DPCL-l_{12}}$ is defined for the case where the source language is l_1 and the target language is l_2. Based on the above description, we formulate

$\mathcal{L}_{DPCL-l_{12}}$ as follow:

$$\mathcal{L}_{DPCL-l_{12}} = -\frac{1}{K}\sum_{i=1}^{K}\Bigg[(p_i^{l_1} \cdot p_i^{l_1+}/\tau) - \log\left(\sum_{j=0}^{r}\exp(p_i^{l_1} \cdot z_j^{l_1-}/\tau) + \sum_{j=0}^{r}\exp(p_i^{l_1} \cdot z_j^{l_2-}/\tau)\right)\Bigg], \quad \text{where } z_j^{l_1-} \in \{\mathbf{z}^{l_1} \setminus c_i\}, \quad z_j^{l_2-} \in \{\mathbf{z}^{l_2} \setminus s_i\} \tag{6}$$

Overall loss function $\mathcal{L}_{DPCL}$ include $\mathcal{L}_{DPCL-l_{12}}$, $\mathcal{L}_{DPCL-l_{21}}$ and τ is a temperature hyper-parameter, $\mathcal{L}_{DPCL}$ as follows:

$$\mathcal{L}_{DPCL} = \mathcal{L}_{DPCL-l_{12}} + \mathcal{L}_{DPCL-l_{21}} \tag{7}$$

3.2.4 Overall Training Objective

We follow (Bianchi et al., 2021b), the generative objective function for ProtoXTM is the same as ELBO of VAE (Kingma and Welling, 2013) which needs to be maximized in order to maximize the log-likelihood of the input pre-trained multi-lingual document embeddings. Our topic modeling objective function of language l_1 as follows:

$$\mathcal{L}^{l_1} = \frac{1}{M_1}\sum_{i=1}^{M_1}\Big[-(\mathbf{x}_i^{l_1})^\top \log\left(\text{softmax}(\boldsymbol{\beta}^{l_1}\theta_i^{l_1})\right) + \text{KL}\left(q(\mathbf{z}^{l_1} \mid \mathbf{x}_i^{l_1}) \parallel p(\mathbf{z}^{l_1})\right)\Big] \tag{8}$$

The first term represents the reconstruction error, quantified by the cross-entropy between the reconstructed document and the input document. On the other hand, the second term is the KL divergence of the learned an unified latent space distribution. In language l_2, the topic modeling objective function operates in the same manner as in l_1. The overall objective function for ProtoXTM is formulated as follows:

$$\mathcal{L} = \mathcal{L}^{l_1} + \mathcal{L}^{l_2} + \lambda * \mathcal{L}_{DPCL}, \tag{9}$$

where λ control hyperparameter the relative significance of $\mathcal{L}_{DPCL}$. Denoted by $\mathcal{L}^{l_1}$ and $\mathcal{L}^{l_2}$ are the topic modeling objective function of language l_1 and languege l_2, respectively. Please refer to the detailed training process of Stage 3 in our ProtoXTM framework in Algorithm 1 in Appendix B.

4 Experiments

4.1 Experimental Setup

We have conducted the experiments using TopMost (Wu et al., 2024b), a comprehensive toolkit for comparing and optimizing topic modeling in various scenarios.[1]

Datasets. We conduct experiments on two benchmark English-Chinese bilingual datasets: EC-News and Amazon Review. Datasets were already included in TopMost in pre-processed formats. The statistics of the processed datasets are shown in Table 7 in Appendix A.

Baselines. We compare our ProtoXTM with the following cross-lingual and mono-lingual topic models. Following **cross-lingual topic models,** (1) NMTM (Wu et al., 2020), the first cross-lingual neural topic model based on VAE, and (2) InfoCTM (Wu et al., 2023a), a state-of-the-art cross-lingual neural topic model using mutual information maximization. Following **mono-lingual topic models,** (3) ProdLDA (Srivastava and Sutton, 2017), a VAE-based standard neural topic model, (4) ETM (Dieng et al., 2020), which incorporates word embedding to model topics, (5) ZeroshotTM (Bianchi et al., 2021b), a neural topic model replacing input BoW with contextualized embeddings. (6) BERTopic (Grootendorst, 2022), a clustering-based method, apply pre-trained document embeddings, and (7) ECRTM (Wu et al., 2023b), which topic embedding clustering regularization to improve topic coherence.

Evaluation Metrics. To evaluate topic coherence quality, we adopt two complementary perspectives. **(1) Cross-lingual topic coherence**, measured by CNPMI (Cross-lingual Normalized Point-wise Mutual Information) (Hao et al., 2018), is a widely used metric for assessing both the coherence and alignment of cross-lingual topics. CNPMI evaluates the degree to which semantically similar words appear across languages within a topic, thereby capturing cross-lingual consistency. **(2) Intra-lingual topic coherence** is assessed using NPMI (Normalized Point-wise Mutual Information) (Lau et al., 2014), which assigns higher scores to topics where the top-related word pairs exhibit high co-occurrence probability relative to their marginal probabilities. Additionally, Cv (Coherence Value) (Röder et al., 2015) is employed as another coherence metric. Based on Fitelson's confirmation measure and computed via a sliding window approach over the reference corpus, Cv has been shown to correlate well with human judgments of topic quality. Furthermore, to evaluate the quality of the document-topic distributions, we employ

[1]https://github.com/BobXWu/TopMost

	ECNews					Amazon Review				
	CNPMI	NPMI – EN	NPMI – ZH	Cv – EN	Cv – ZH	CNPMI	NPMI – EN	NPMI – ZH	Cv – EN	Cv – ZH
ProdLDA		-0.2084	-0.2393	0.3881	0.3646		-0.2121	-0.2303	0.4199	0.3879
ETM		-0.1974	-0.1566	0.3695	0.3658		-0.2219	-0.2160	0.4310	0.3338
ZeroshotTM		-0.1548	**-0.0628**	0.4101	0.4486		-0.0970	**-0.1518**	0.4451	0.3973
BERTopic		**-0.0699**	-0.0949	0.4027	**0.5214**		**-0.0268**	-0.1933	0.4075	**0.4116**
ECRTM		-0.2909	-0.2888	**0.4922**	0.3722		-0.0818	-0.1852	**0.4652**	0.3639
NMTM	0.0253	-0.1757	-0.1607	0.3941	0.3620	0.0455	-0.1526	-0.2062	0.4153	**0.4152**
InfoCTM	0.0370	-0.2409	-0.2601	0.4301	0.4055	0.0275	-0.2305	-0.2699	0.4117	0.3362
ProtoXTM (ours)	**0.0717**	**-0.0847**	**-0.0076**	**0.4456**	**0.5334**	**0.0564**	**-0.0979**	**-0.1635**	**0.4570**	0.4130

Table 2: Cross-lingual and intra-lingual topic coherence measures, for models containing 10 topics. The best-performing method is highlighted in **bold**.

	ECNews		Amazon Review	
	Purity	NMI	Purity	NMI
NMTM	0.5832	0.2574	0.5820	0.0245
InfoCTM	0.5768	0.2227	0.6287	0.0264
ProtoXTM (ours)	**0.6204**	**0.2752**	**0.6292**	**0.0298**

Table 3: Performance comparison on document-topic distribution transferability. The best-performing method is highlighted in **bold**.

	CNPMI	NPMI – EN	NPMI – ZH	Cv – EN	Cv – ZH
w/o DPCL	0.0420	-0.0950	**-0.0656**	0.4131	0.4520
DPCL-EN only	0.0442	-0.0989	-0.0830	0.4130	0.4328
DPCL-ZH only	0.0529	-0.0896	-0.0788	0.4264	0.4478
ProtoXTM	**0.0621**	**-0.0838**	-0.0731	**0.4413**	**0.4566**

Table 4: Ablation studies on the ECNews dataset.

a document clustering task using two evaluation metrics, Purity and NMI (Normalized Mutual Information) (Manning et al., 2008). NMI quantifies the mutual information between the predicted topic assignments and the ground-truth labels, normalized to fall within the range [0, 1]. Purity measures the extent to which each cluster contains data points from a single class.

4.2 Experimental Results

Topic Quality. For a given dataset, we have reported the mean value over 5 random runs. Tables 2 and 9 present the results of three topic coherence measures for 10 and 20 topics, respectively. We compute all topic coherence measures using the top 15 related words for each topic. From the results, ProtoXTM improves CNPMI performance by up to 93.8% and outperforms other cross-lingual topic model baselines in every settings by solving the problem of topic mismatch between translated words across languages through document-level topic alignment. ProtoXTM demonstrated competitive performance in intra-lingual topic coherence compared to various mono-lingual neural topic models. While NMTM and InfoCTM exhibited lower intra-lingual topic coherence than other mono-lingual topic models, ProtoXTM achieved high topic coherence even within an intra-lingual language while performing cross-lingual topic alignment. This result indicates that ProtoXTM enables topic alignment while preserving intra-lingual topic coherence across different languages and our topic-based clustering approach using mono-lingual topic models can mitigate the issue of degenerated intra-lingual topic coherence in cross-lingual topic models.

Doc-Topic Distribution Quality. To evaluate the language transferability of document-topic distributions in cross-lingual topic models, we concatenated the infered document-topic distributions from two different languages. Following (Adhya and Sanyal, 2024), each document was assigned to the topic with the highest probability in document-topic distribution. Intuitively, an integrated cluster contains documents from both languages, meaning that the quality of these clusters reflects the degree of transferability across languages. Table 3 present the results of clustering performance for 20 clusters, respectively. From the results, we could find that our ProtoXTM outperforms clustering performances with the other baselines. These results indicate that ProtoXTM facilitates language transfer across different languages by enabling semantically similar documents to share topics through the inferred document-topic distributions of the other language.

4.3 Ablation Study

We conduct an ablation study on the ECNews for 20 topics, Table 4 presents the comparison of different variations of the our ProtoXTM framework. The **w/o DPCL** variant removes the overall DPCL loss function from the ProtoXTM frame-

	CNPMI	NPMI – EN	NPMI – ZH	Cv – EN	Cv – ZH
ProtoXTM (I)	0.0648	-0.0851	-0.0245	**0.4497**	0.5253
ProtoXTM (P)	**0.0717**	**-0.0847**	**-0.0076**	0.4456	**0.5334**

Table 5: Comparison of contrastive learning strategy using topic coherence metrics.

Batch size	500	1000	5000	10000	20000	30000
ProtoXTM (I)	2.33s	2.58s	4.27s	6.71s	14.96s	44.29s
ProtoXTM (P)	2.65s	2.70s	2.77s	3.25s	3.34s	4.02s

Table 6: Comparison of runtime performance on contrastive learning perspective.

work, relying solely on pre-trained multilingual embeddings without our topic alignment mechanism. **w/o DPCL** achieves competitive CNPMI performance compared to InfoCTM. These results indicate that document-level alignment induced by pre-trained multilingual document embeddings contributes positively to topic alignment. The **DPCL-EN only** variant uses English documents as anchor samples while incorporating only sampled chinese documents, meaning it does not consider topic structures within the chinese corpus itself. Likewise, **DPCL-ZH only** does not consider topic structures within the english corpus itself. The experimental results indicate that both **DPCL-EN only** and **DPCL-ZH only** achieve improved CNPMI scores compared to **w/o DPCL**, reflecting enhanced cross-lingual topic alignment. However, intra-lingual topic coherence does not show substantial improvement in these settings, suggesting that unidirectional DPCL may lead to a loss of intra-lingual topic information within each monolingual corpus. In contrast, ProtoXTM demonstrates improved performance across all topic coherence measures except NPMI-ZH. By incorporating bidirectional topic information between the two languages, ProtoXTM enables mutual enhancement and reinforcement of the topic structures in each language. Our approach simultaneously improves both intra-lingual topic interpretability and cross-lingual topic alignment.

4.4 Learning Strategy Analysis

In this subsection, we explore two different document-level contrastive learning strategies in our ProtoXTM framework. We compare standard instance-wise contrastive learning with our DPCL method in terms of topic coherence quality and runtime performance on ECNews dataset. Denoted by **ProtoXTM (I)** is the standard instance-wise contrasitve learning method and **ProtoXTM (P)** is our DPCL method. As shown in Table 5, our DPCL method outperforms the standard instance-wise contrastive learning in CNPMI and intra-lingual topic coherence, except for Cv-EN. These results suggest that, in contrastive learning, comparing prototypes representing clusters rather than each documents is more effective in topic alignment and coherence. Generally, contrastive learning methods that utilize negative samples within a mini-batch suffer from degraded representation quality as batch size decreases (Grill et al., 2020). Depending on the data scale, performance can be improved through a large batch size (Chen et al., 2020; Tian et al., 2020). As shown in Table 6, standard instance-wise contrastive learning encounters training speed degradation with large batch sizes. In contrast, our DPCL method demonstrates robust training speed performance even under large batch size conditions. Please refer to Appendix E for more detailed our findings.

5 Conclusion

In this paper, we identify two critical issues in cross-lingual topic modeling, the topic mismatch issue and the degeneration of intra-lingual topic interpretability. Furthermore, we propose a novel cross-lingual neural topic modeling framework, ProtoXTM, effectively mitigates topic mismatch issue and intra-lingual topic degradation by retrieval-based positive sampling strategy and document-level prototype-based contrastive learning. Extensive experimental results demonstrate that ProtoXTM outperforms the baseline methods in both cross-lingual and intra-lingual topic coherence, and can infer document-topic distributions with high transferability.

Limitations

Our proposed methodology has achieved promising enhancements by mitigating the topic mismatch and intra-lingual topic degradation issues in cross-lingual topic modeling. However, we consider the following remaining several limitations as future work. First, while we employ traditional retrieval algorithms such as BM25 for positive sampling in document-level contrastive learning, we anticipate that more powerful information retrieval methods based on large language models (LLMs) could further enhance contrastive learning performance. Second, although we utilize an open-source Neural Machine Translation (NMT) model for cross-lingual topic alignment. However, we leave a comprehensive investigation of this sensitivity for future work. Third, the experiments in our work are limited to the English–Chinese benchmark. While previous work (Wu et al., 2023a) demonstrates promising results for Japanese language with limited translation resources cross-lingual topic alignment in truly low-resource languages, where bilingual dictionaries are entirely unavailable, remains an open challenge. Lastly, determining the optimal number of topics is still an unresolved problem in topic modeling (Stammbach et al., 2023). Since the number of topics is a critical hyperparameter that significantly affects model performance, identifying an optimal topic number that balances both cross-lingual topic alignment and topic interpretability in CLTM is an important research direction for future work.

References

Suman Adhya and Debarshi Kumar Sanyal. 2024. GINopic: Topic modeling with graph isomorphism network. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6171–6183, Mexico City, Mexico. Association for Computational Linguistics.

Pritom Saha Akash and Kevin Chen-Chuan Chang. 2024. Enhancing short-text topic modeling with LLM-driven context expansion and prefix-tuned VAEs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15635–15646, Miami, Florida, USA. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. 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. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Sungwon Han, Mingi Shin, Sungkyu Park, Changwook Jung, and Meeyoung Cha. 2023. Unified neural topic model via contrastive learning and term weighting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1802–1817, Dubrovnik, Croatia. Association for Computational Linguistics.

Shudong Hao, Jordan Boyd-Graber, and Michael J. Paul. 2018. Lessons from the Bible on modern topics: Low-resource multilingual topic model evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1090–1100, New Orleans, Louisiana. Association for Computational Linguistics.

Shudong Hao and Michael J. Paul. 2018. Learning multilingual topics from incomparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2595–2609, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jagadeesh Jagarlamudi and Hal Daumé. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval*, pages 444–456, Berlin, Heidelberg. Springer Berlin Heidelberg.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

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, Gothenburg, Sweden. Association for Computational Linguistics.

Zheng Lin, Xiaolong Jin, Xueke Xu, Yuanzhuo Wang, Xueqi Cheng, Weiping Wang, and Dan Meng. 2016. An unsupervised cross-lingual topic model framework for sentiment classification. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 24(3):432–444.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore. Association for Computational Linguistics.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *Advances in neural information processing systems*, 34:11974–11986.

Thong Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Topic modeling as multi-objective contrastive optimization. *arXiv preprint arXiv:2402.07577*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Bei Shi, Wai Lam, Lidong Bing, and Yinqing Xu. 2016. Detecting common discussion topics across culture from news reader comments. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 676–685, Berlin, Germany. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

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

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839.

Yu Tsou, Deng-Neng Chen, and Chia-Yu Lai. 2020. A cross-lingual patent topics model for trend analysis. In *2020 International Computer Symposium (ICS)*, pages 525–528.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2013. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Inf. Retr.*, 16(3):331–368.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liang-Ming Pan, and Anh Tuan Luu. 2023a. Infoctm: a mutual information maximization perspective of cross-lingual topic modeling. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023b. Effective neural topic modeling with embedding clustering regularization.

In *International Conference on Machine Learning*, pages 37335–37357. PMLR.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Learning multilingual topics with neural variational inference. In *Natural Language Processing and Chinese Computing*, pages 840–851, Cham. Springer International Publishing.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2748–2760, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024a. Fastopic: A fast, adaptive, stable, and transferable topic modeling paradigm. *arXiv preprint arXiv:2405.17978*.

Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2024b. Towards the TopMost: A topic modeling system toolkit. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Bangkok, Thailand. Association for Computational Linguistics.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Weijie Xu, Wenxiang Hu, Fanyou Wu, and Srinivasan Sengamedu. 2023. DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9040–9057, Singapore. Association for Computational Linguistics.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2019. A multilingual topic model for learning weighted topic links across corpora with low comparability. 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 1243–1248, Hong Kong, China. Association for Computational Linguistics.

Michelle Yuan, Benjamin Van Durme, and Jordan L Ying. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tao Zhang, Kang Liu, and Jun Zhao. 2013. Cross lingual entity linking with bilingual topic model. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 2218–2224. AAAI Press.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Elaine Zosa and Lidia Pivovarova. 2022. Multilingual and multimodal topic modelling with pretrained embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4037–4048, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Dataset

In this section, we provide detailed description of the bilingual benchmark datasets used in our experiments. ECNews is a bilingual news dataset in English and Chinese, consisting of six categories: business, education, entertainment, sports, technology, and fashion. Amazon Review is a bilingual review dataset collected from the Amazon website in both English and Chinese. For both datasets, we use the preprocessed versions provided by the TopMost toolkit (Wu et al., 2024b). The statistics of the preprocessed datasets are presented in Table 7.

B Training Algorithm

In this section, we provide detailed training procedure of Stage3 in our ProtoXTM framework. Before training the our model, the cluster labels $\mathbf{y_c}^{l_1}, \mathbf{y_c}^{l_2}$ and sampled labels $\mathbf{y_s}^{l_1}, \mathbf{y_s}^{l_2}$ are precomputed during Stage 1 and Stage 2, respectively. The detailed training algorithm for Stage 3 of ProtoXTM is presented in Algorithm 1.

C Implementation Details

In this section, we describe the training environment and model architecture details. All models were implemented using PyTorch 2.1.0 and Python 3.10, and experiments were conducted on a machine equipped with a GeForce RTX 3090 GPU. The encoder network is a 3-layer multilayer perceptron (MLP) with a hidden layer dimension of 128, and model parameters were optimized using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2e-2. For pretrained multilingual document embeddings, we used the *paraphrase-multilingual-MiniLM-L12-v2*

Dataset	Language	#Train Docs	#Vocabulary	labels
Amazon Review	English	25,000	5,000	2
	Chinese	25,000	5,000	2
ECNews	English	46,870	5,000	6
	Chinese	50,000	5,000	6

Table 7: Statistics of the preprocessed datasets

model from Sentence-Transformers[2]. Additionally, we employed 200-dimensional FastText embeddings for both English and Chinese as the pre-trained word embeddings.

D Hyperparameter Setting

In this section, we describe all hyperparameter settings used in our experiments with the ProtoXTM framework. In Stage 1, the number of topics for the pre-training of the separated mono-lingual neural topic models is set to 50. All other settings follow the configuration of (Bianchi et al., 2021b). In Stage 2, the number of query words in each query set (i.e., top-related words) is set to 10, the word replacement threshold is 0.4, and 30 documents are sampled as positives within each cluster group. The BM25 ranking function is used with its default configuration of (Robertson and Zaragoza, 2009). In Stage 3, we set the temperature τ to 0.3 and the $\mathcal{L}_{DPCL}$ weight λ to 1.2 and the batch size B to 1024. We use grid search to determine the value of the above hyperparameter and all hyperparameter settings are kept fixed across our experiments on all datasets.

E Contrastive Learning Strategy Analysis

In this section, we explain the details of our contrastive learning strategy analysis in subsection 4.4. The objective function of **ProtoXTM (I)**, which employs the standard instance-wise contrastive learning is as follows:

$$\mathcal{L}_{ICL-l_{12}} = -\frac{1}{M_1}\sum_{i=1}^{M_1}\Bigg[\sum_{j=0}^{n}(z_i^{l_1}\cdot z_j^{l_1+}/\tau) - \log\left(\sum_{j=0}^{r}\exp(z_i^{l_1}\cdot z_j^{l_1-}/\tau) + \sum_{j=0}^{r}\exp(z_i^{l_1}\cdot z_j^{l_2-}/\tau)\right)\Bigg],$$
$$\text{where } z_j^{l_1-}\in\{\mathbf{z}^{l_1}\setminus c_i\},\quad z_j^{l_2-}\in\{\mathbf{z}^{l_2}\setminus s_i\} \tag{10}$$

Denoted as $\mathcal{L}_{ICL-l_{12}}$, this variant refers to the strandard instance-wise contrastive learning where the source language is l_1 and the target language is l_2, and z^{l_1+} represents documents sampled from the corresponding group. All hyperparameters are set identically to those used in **ProtoXTM (P)**, and the overall objective function of **ProtoXTM (I)** is as follows:

$$\mathcal{L} = \mathcal{L}^{l_1} + \mathcal{L}^{l_2} + \lambda * \mathcal{L}_{ICL}, \tag{11}$$

where $\mathcal{L}_{ICL} = \mathcal{L}_{ICL-l_{12}} + \mathcal{L}_{ICL-l_{21}}$. We analyze the standard instance-wise contrastive learning and our DPCL method in terms of both topic quality and runtime performance.

Topic Quality: As shown in the experimental results in Table 5, our DPCL method outperforms the standard instance-wise contrastive learning approach in both cross-lingual and intra-lingual topic coherence. Previous studies (Han et al., 2023; Nguyen and Luu, 2021; Nguyen et al., 2024) have demonstrated the effectiveness of contrastive learning for topic modeling, but conventional contrastive learning methods are primarily designed for sentence embedding learning (Xu et al., 2023). In contrast, our DPCL method is tailored toward effective topic alignment and inference for cross-lingual topic modeling, rather than learning representations of each documents.

Efficiency: Table 6 presents the runtime performance of **ProtoXTM (I)** and **ProtoXTM (P)** across varying batch sizes, ranging from 500 to 30,000. In the instance-wise contrastive learning setting, all documents participate in contrastive learning, leading to increased computational cost as the batch size grows. However, the DPCL method maintains a fixed number of prototypes representing topics, regardless of batch size, with only the number of negative samples increasing within the mini-batch. As a result, our DPCL method remains robust even with large batch sizes, indicating its potential for effective topic alignment and inference on large-scale datasets.

[2] https://huggingface.co/sentence-transformers

F Case Study

In this section, for qualitative analysis of topic quality, we report the topic word examples yielded by different baseline methods and our ProtoXTM model on the ECNews dataset in Table 8. In our case study, we set the number of topics to 20 and conducted qualitative analysis on two representative topics: “fashion” and “study”. For Chinese terms, the corresponding ground-truth English translations are provided in parentheses, and words with underlines indicate those that lack topical consistency. As shown in Table 8, NMTM and InfoCTM exhibit reduced interpretability by either presenting different topics across the two languages or including inconsistent words within topics. In contrast, we observe that the topics generated by ProtoXTM contain semantically coherent words and consistently express similar topic across languages.

G Quantitative Experimental Results

In this section, we report our quantitative experimental results for topic quality analysis. Table 9 present the results of three topic coherence measures for 20 topics.

Methods	Top-related word examples
NMTM	**EN-Topic#13**: fashionably youtube videos runway facetime **ZH-Topic#13**: 时装(fashion) 设计师(designer) 嘉宾(guest) 评选(selection) 时尚(fad) **EN-Topic#18**: education school loans charter college **ZH-Topic#18**: 录取(admit) 本科(undergraduate course) 分数线(cutline) 批次(group) 院校(college)
InfoCTM	**EN-Topic#6**: designers math speed models fashion **ZH-Topic#6**: 流行(trend) 时装(fashion) 模特(model) 传播(spread) 周末(weekend) **EN-Topic#3**: students pilot education pleasure college **ZH-Topic#3**: 学子(student) 教室(classroom) 教学(teaching) 测试(test) 教师(teacher)
ProtoXTM	**EN-Topic#15**: fashion style dress clothing vintage **ZH-Topic#15**: 时尚(fad) 穿(wear) 设计(design) 造型(styling) 外套(overcoat) **EN-Topic#13**: college education students university campus **ZH-Topic#13**: 考试(exam) 学生(student) 学校(school) 大学(university) 教育(education)

Table 8: Top-related word examples generated by different baseline methods.

	ECNews					Amazon Review				
	CNPMI	NPMI – EN	NPMI – ZH	Cv – EN	Cv – ZH	CNPMI	NPMI – EN	NPMI – ZH	Cv – EN	Cv – ZH
ProdLDA		-0.2602	-0.2469	**0.4660**	0.4081		-0.2189	-0.2567	0.4135	**0.4112**
ETM		-0.2044	-0.1531	0.4101	0.3915		-0.1988	-0.1926	0.3932	0.3409
ZeroshotTM		-0.1330	**-0.0749**	0.4251	0.4494		-0.0928	**-0.1795**	0.4424	0.3830
BERTopic		**-0.0679**	-0.1165	0.4256	**0.4969**		**-0.0414**	-0.1952	0.4055	0.3960
ECRTM		-0.2375	-0.2669	0.4519	0.4111		-0.1048	-0.1818	**0.4978**	0.3621
NMTM	0.0279	-0.1829	-0.1390	0.4142	0.3967	0.0251	-0.1823	-0.2051	0.4200	0.3610
InfoCTM	0.0419	-0.2274	-0.2413	0.4224	0.3922	0.0397	-0.2301	-0.2333	**0.4479**	0.3471
ProtoXTM (ours)	**0.0621**	**-0.0838**	**-0.0731**	**0.4413**	**0.4566**	**0.0645**	**-0.0830**	**-0.1692**	0.4456	**0.3826**

Table 9: Cross-lingual and intra-lingual topic coherence measures, for models containing 20 topics. The best-performing method is highlighted in **bold**.

Algorithm 1 Training Procedure of Stage3 in ProtoXTM framework

Input: mini-batch size B, pre-trained document embeddings $\mathbf{x}^{l_1}, \mathbf{x}^{l_2}$, cluster labels ${\mathbf{y_c}}^{l_1}, {\mathbf{y_c}}^{l_2}$, sampled labels ${\mathbf{y_s}}^{l_1}, {\mathbf{y_s}}^{l_2}$, topic number K, temperature τ, $\mathcal{L}_{DPCL}$ weight λ

Output: learned shared encoder f, encoder parameter W_{enc}, decoder parameter $W_{dec}^{l_1}$, $W_{dec}^{l_2}$, topic-word distributions matrix $\boldsymbol{\beta}^{l_1}, \boldsymbol{\beta}^{l_2}$, document-topic distribution matrix $\boldsymbol{\theta}^{l_1}, \boldsymbol{\theta}^{l_2}$

Initialize parameters W_{enc}, $W_{dec}^{l_1}$, $W_{dec}^{l_2}$
for each training epoch $t = 1$ to T **do**
  for batch of B documents $(\mathbf{x}^{l_1}, \mathbf{x}^{l_2})$ **do**
    Encode documents with f:
      $\mathbf{z}^{l_1} \leftarrow f(\mathbf{x}^{l_1}), \quad \mathbf{z}^{l_2} \leftarrow f(\mathbf{x}^{l_2})$
    Compute anchor prototypes using cluster labels ${\mathbf{y_c}}^{l_1}, {\mathbf{y_c}}^{l_2}$ by Eq. 4
    Compute positive prototypes using sampled labels ${\mathbf{y_s}}^{l_1}, {\mathbf{y_s}}^{l_2}$ by Eq. 5
    Compute $\mathcal{L}_{DPCL}$ by Eq. 6,7.
    Compute document-topic distributions:
      $\boldsymbol{\theta}^{l_1} \leftarrow \text{softmax}(\mathbf{z}^{l_1})$,
      $\boldsymbol{\theta}^{l_2} \leftarrow \text{softmax}(\mathbf{z}^{l_2})$
    Compute reconstructed documents:
      $\hat{\mathbf{x}}^{l_1} \leftarrow \text{softmax}(\boldsymbol{\beta}^{l_1}\boldsymbol{\theta}^{l_1})$,
      $\hat{\mathbf{x}}^{l_2} \leftarrow \text{softmax}(\boldsymbol{\beta}^{l_2}\boldsymbol{\theta}^{l_2})$
    Compute $\mathcal{L}^{l_1}$ and $\mathcal{L}^{l_2}$ by Eq. 8
    Compute total loss by Eq.9
    Update all parameters with gradient $\nabla\mathcal{L}$
  end for
end for