ProtoXTM: Cross-Lingual Topic Modeling with Document-Level Prototype-based Contrastive Learning

EMNLP 2025 Suzhou, China | 中国苏州

아주대학교 AJOU UNIVERSITY

Seung-Won Seo Soon-Sun Kwon

Department of Mathematics, Ajou University

Motivation

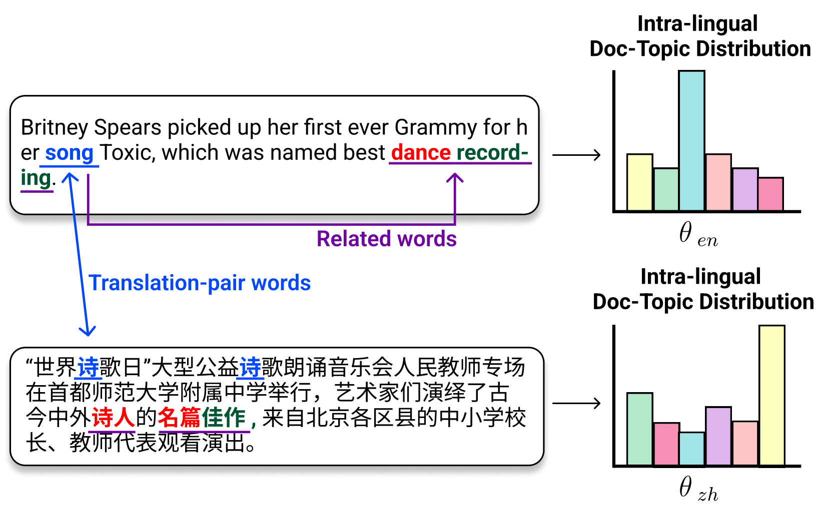

Figure 1. A motivating example of topic mismatch issue in cross-lingual topic modeling.

- *Do translation-based word pairs always guarantee semantically similar and well-aligned topics?*
- We observe a case where translation word pairs appear in two semantically distinct negative bilingual documents.
- The two documents represent divergent topic distributions within their respective intra-lingual corpora.
- Topic mismatch issue arises due to linguistic diversity and cultural differences.

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 and BERTopic on ECNews dataset. The degenerating intra-lingual topic interpretability issue in cross-lingual topic modeling

- We investigate the topics generated by a stateof-the-art cross-lingual neural topic model, InfoCTM and a mono-lingual neural topic model, BERTopic.
- This observation suggests that the objective of alignment in cross-lingual topic models such as InfoCTM can compromise intra-lingual topic interpretability.

Contributions

- We identify two critical issues in cross-lingual topic modeling, the topic mismatch issue and the degeneration of intra-lingual topic interpretability.
- We propose **DPCL** method, a new Document-level Prototype-based Contrastive Learning paradigm tailored for effective cross-lingual topic modeling. Furthermore, we design Retrieval-based Positive Sampling **(RPS)** 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 intralingual topic interpretability.
- We conduct extensive experiments on nonparallel 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.

Main Results

	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	**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**.

Proposed Methodology

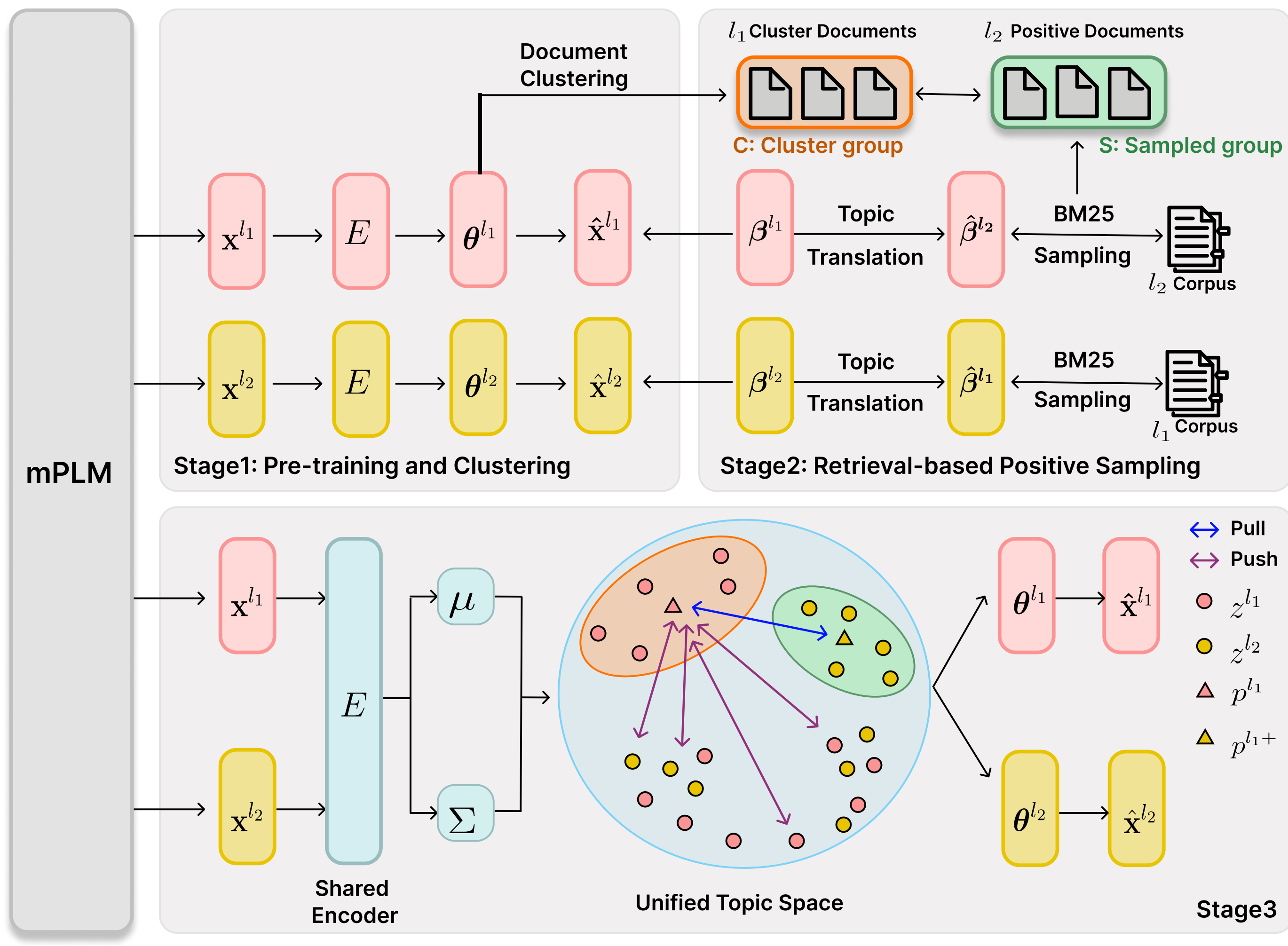

Figure 2. The overall process of ProtoXTM. We utilize the labels of positive sample documents pre-computed in Stage 1 and Stage 2 to perform cross-lingual topic alignment in Stage 3 through our DPCL method.

- **Stage1:** We assign each document to the topic with the highest probability from intra-lingual doc-topic distribution.
- **Stage2:** We utilize BM25 algorithm to retrieve semantically similar documents from the target corpus as positive samples for cross-lingual topic alignment.
- **Stage3:** We propose document-level prototype-based contrastive learning approach, effectively align cross-lingual topics by applying the prototype representations of document clusters with their corresponding positive samples across languages.

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

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

Learning Strategy Analysis

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.

	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 3. Comparison of contrastive learning strategy in ProtoXTM using topic coherence metrics.

- Our **DPCL** method outperforms the standard instance-wise contrastive learning approach in both cross-lingual and intra-lingual topic coherence.
- **DPCL** is tailored toward effective topic alignment and inference for cross-lingual topic modeling, rather than learning representations of each documents.

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 4. Comparison of runtime performance on contrastive learning perspective.

- **DPCL** maintains a fixed number of prototypes representing topics, regardless of batch size, with only the number of negative samples increasing within the mini-batch.
- **DPCL** remains robust even with large batch sizes, indicating its potential for effective topic alignment and inference on large-scale datasets.