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

Seung-Won Seo, Soon-Sun Kwon

Department of Mathematics, Ajou University

EMNLP 2025
Suzhou, China 中国苏州

아주대학교
AJOU UNIVERSITY

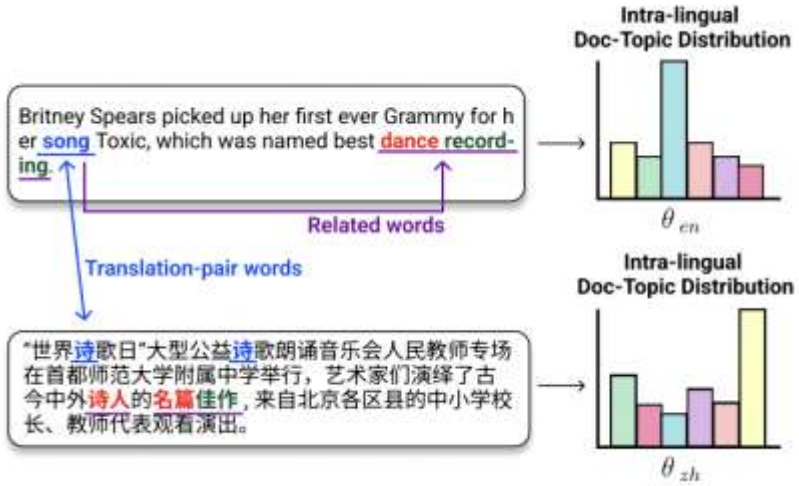
Introduction

- What is Cross-lingual Topic Model ?
  - Target data: Non-parallel bilingual corpora
  - Goal: Generate aligned topics in unsupervised manner

- What is *“aligned topic” ?*
  - *Aligned topic* defined as a pair of topics from two different languages where the *i*-th to pic in each language represents a semantically similar topic

Motivation

- Issue -1: Topic Mismatch
  - *Do translation-based word pairs always guarantee semantically similar and well-aligned topics ?*
  - Our motivating example in previous word-level alignment based cross-lingual topic modeling



Motivation

- Issue -2: Degenerating Intra-lingual topic Interpretability
  - Topic alignment objective VS. Intra-lingual topic interpretability
  - Cross-lingual topic model tends to generate poor-quality topics in each language compared to mono-lingual topic model.

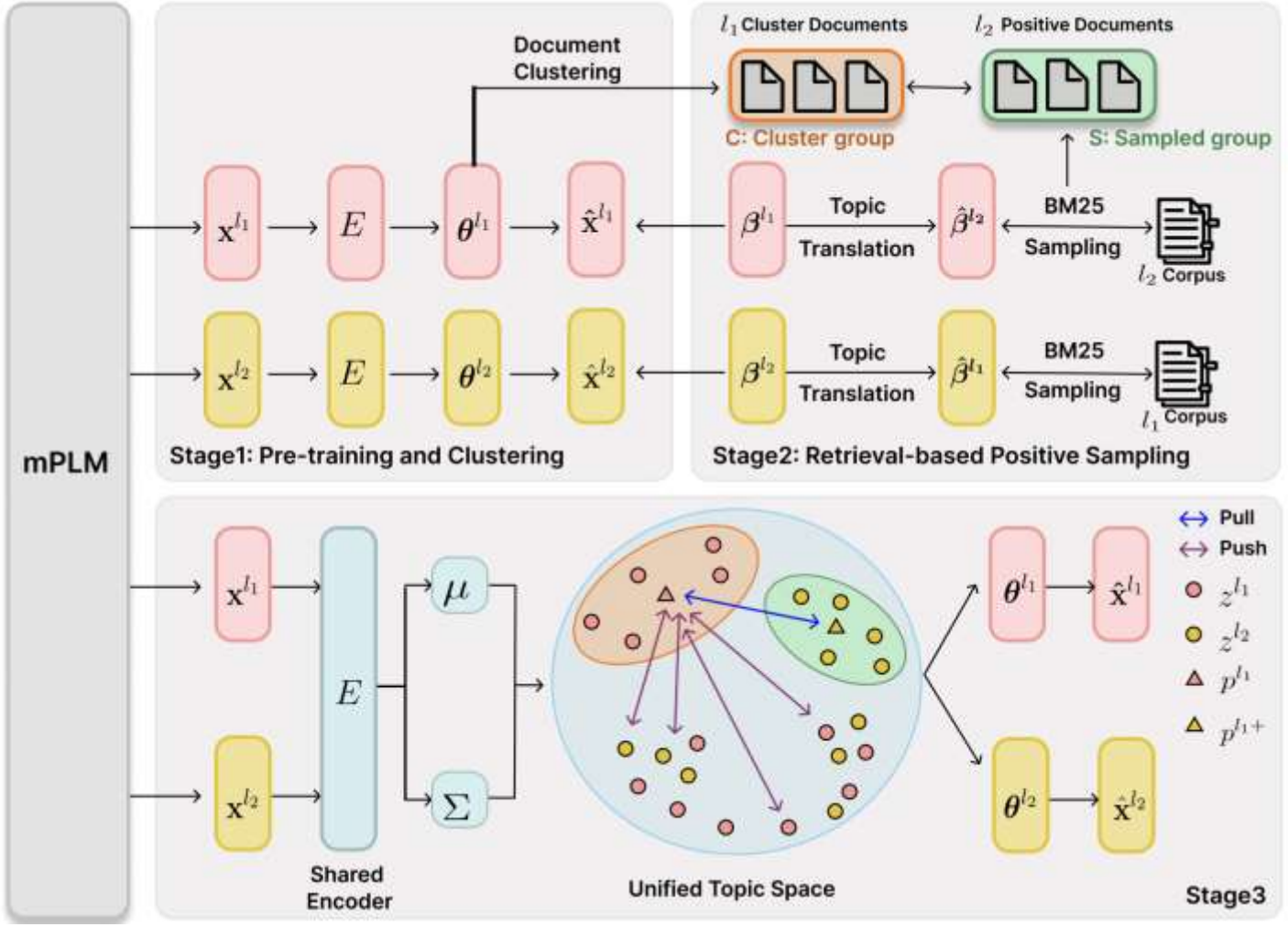
InfoCTM		BERTopic
Topic # 13		Topic # 157
EN	ZH	EN
sing	秀(show)	albums
concert	高潮(climax)	chart
exhibit	唱歌(singing)	album
artist	演出(performance)	charts
album	歌(song)	soundtrack
songs	展(exhibition)	band
rap	直播(broadcast)	musicians
broadcast	演艺(performance)	singles
song	游(tour)	dj
travel	艺术家(artist)	songs

Proposed Methodology

- Our Approach: ProtoXTM
  - Document-Level Prototype-based Contrastive Learning (DPCL) without data augmentation
  - Apply pre-trained mono-lingual topic model and Retrieval-based Positive Sampling (RPS) for DPCL

Proposed Methodology

- Our Approach: `ProtoXTM`



Experiments

- Experimental Setup
  - Datasets: ECNews, Amazon Review
  - Baselines
    - Cross-lingual topic models: InfoCTM, NMTM
    - Mono-lingual topic models: ProdLDA, ETM, ZeroshotTM, BERTopic, ECRTM
  - Evaluation metrics for topic quality: CNPMI, NPMI, Cv
  - Evaluation metrics for doc-topic distribution quality: Purity, NMI

Experiments

- Quantitative Analysis: Cross-lingual and Intra-lingual Topic Quality

	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

- `ProtoXTM` improves CNPMI performance by up to 93.8% and outperforms other cross-lingual topic mode l baselines in every settings by solving the problem of topic mismatch between translated words across la nguages through document-level topic alignment.

Experiments

- Quantitative Analysis: Doc-Topic Distribution Quality

	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**

- `ProtoXTM` outperforms clustering performances with the previous cross-lingual topic mod el baselines.
- Our document-level topic alignment approach is more effective in inferring common topic distributions within documents compared to previous word-level approaches.

Experiments

- Qualitative Analysis: Topic Quality

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)

- We observe that the topics generated by `ProtoXTM` contain semantically coherent words and consistentl y express similar topic across languages.

Experiments

- Contrastive Learning Strategy Analysis: Topic quality perspective
- Denoted by `ProtoXTM(I)` is the standard instance-wise contrasitve learning method and `ProtoXTM(P)` is our `DPCL` method.

	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**

- We compare standard instance-wise contrastive learning with our `DPCL` method in terms of topic coherence quality and runtime performance on ECNews dataset.

Experiments

- Contrastive Learning Strategy Analysis: Efficiency perspective

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

- Our `DPCL` method remains robust even with large batch sizes, indicating its potential for ef fective topic alignment and inference on large-scale datasets.

Conclusion and Future Work

- Conclusion
  - We identify two critical issues in cross-lingual topic modeling
  - `ProtoXTM` effectively mitigates topic mismatch issue and intra-lingual topic degradati on issue 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 docume nt-topic distributions with high transferability

Conclusion and Future Work

- Future Work
  - Cross-lingual topic modeling for truly low-resource languages (*e.g.*, *Quechua*, *Bengali* , etc.)
  - Selection method for optimal number of aligned cross-lingual topics