Refining Dimensions for Improving Clustering-based Cross-lingual Topic Models

Chia-Hsuan Chang[1], Tien-Yuan Huang[2], Yi-Hang Tsai[2], Chia-Ming Chang[2], San-Yih Hwang[2]
[1]Department of Biomedical Informatics & Data Science, Yale University
New Haven, CT 06510, United States
[2]Department of Information Management, National Sun Yat-sen University
Kaohsiung 80424, Taiwan
Correspondence: shane.chang.tw@gmail.com, syhwang@mis.nsysu.edu.tw

Abstract

Recent works in clustering-based topic models perform well in monolingual topic identification by introducing a pipeline to cluster the contextualized representations. However, the pipeline is suboptimal in identifying topics across languages due to the presence of language-dependent dimensions (LDDs) generated by multilingual language models. To address this issue, we introduce a novel, SVD-based dimension refinement component into the pipeline of the clustering-based topic model. This component effectively neutralizes the negative impact of LDDs, enabling the model to accurately identify topics across languages. Our experiments on three datasets demonstrate that the updated pipeline with the dimension refinement component generally outperforms other state-of-the-art cross-lingual topic models [1].

1 Introduction

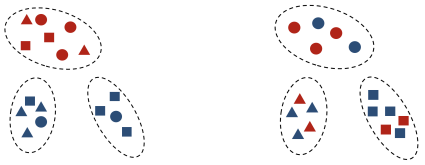
Traditional cross-lingual topic models (CLTM) rely on additional resources to identify topics across languages. Based on the types of resources, CLTMs can be categorized into document and vocabulary-linking models. The document-linking models require parallel or comparable corpora to model the co-occurring word statistics across languages and infer the cross-lingual topics (Mimno et al., 2009; Piccardi and West, 2021). The vocabulary-linking models are more resource-efficient than their document-linking counterpart because they only require a bilingual dictionary (i.e., a set of translation entries). However, vocabulary-linking models often result in monolingual topics (Hu et al., 2014; Hao and Paul, 2020; Wu et al., 2023) when the dictionary is of limited coverage to the target corpus. Several studies proposed to link word embedding spaces across languages to decrease the effort of compiling a well-covered dictionary. When the assumption of shared structures across spaces (i.e., isomorphism) holds, a small number of translation entries will be sufficient to identify topics across languages (Chang et al., 2018; Yuan et al., 2018; Chang and Hwang, 2021). However, the word spaces of different languages seldom share the same structure in practice, especially for languages that are distantly related, and iterative human involvement is still required for acquiring a quality dictionary.

The recent development of multilingual language models (MLM), e.g., mBERT, XLM-R, and GPT models, attracts attention from the natural language processing community. MLM learns the language-agnostic representations without any additional resources (Pires et al., 2019a; Dufter and Schütze, 2020), which has the potential to realize the zero-shot topic identification across languages (Bianchi et al., 2021), thereby reducing efforts on data preparation. Recent studies increasingly favor the clustering-based topic model due to its superior performance and higher efficiency (Sia et al., 2020; Grootendorst, 2022; Zhang et al., 2022). The clustering-based topic model adopts a pipeline (see Sec. 2.1) to leverage the induced representations of language models for topic identification. MLMs can be directly applied to the pipeline of clustering-based topic modeling for cross-lingual topic identification. However, the current pipeline is hindered by the existence of language-dependent dimensions (LDDs) in the representations generated by MLMs, which makes the representations sensitive to languages and hinders the pipeline from identifying topics across languages. As depicted in Fig. 1a, the current pipeline with MLM tends to cluster documents by languages rather than semantic meanings. We also report the qualitative result of misaligned topics generated using BERTopic (Grootendorst, 2022), an accessible implementation for clustering-based topic modeling, in Table 1. Ideally, topic clusters should group

[1]Our code and data are available at https://github.com/Text-Analytics-and-Retrieval/Clustering-based-Cross-Lingual-Topic-Model.

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 46–56
January 20, 2025.

documents based on their semantic meanings, as illustrated in Fig. 1b.



(a) Topic clusters grouped by languages (b) Topic clusters grouped by semantic meanings

Figure 1: Two resultant scenarios of clustering-based topic model. Different shapes indicate the documents discussing various topics, while different colors represent documents of different languages.

To mitigate such a problem, this study proposes adding a new dimension refinement component into the pipeline to neutralize the impacts of LDDs from the representations. Specifically, we utilize singular value decomposition (SVD) to identify the LDDs and offer two implementations of the dimension refining component: unscaled SVD (u-SVD) and SVD with language dimension removal (SVD-LR). The contributions of this study are threefold:

1. We observe and identify the negative impacts of LDDs on the pipeline of the clustering-based topic model in a cross-lingual topic identification task.

2. We introduce a dimension refinement component, implemented by either u-SVD or SVD-LR, into the current pipeline of the clustering-based topic model, which enables it to identify topics across languages.

3. Our updated pipeline of the clustering-based topic model is shown to outperform the other state-of-the-art CLTMs on three datasets.

2 Methodology

2.1 Background: Pipeline of Clustering-based Topic Model

The pipeline of clustering-based topic model (Grootendorst, 2022; Zhang et al., 2022) contains four steps: Document Embedding Generation → Dimension Reduction → Document Clustering → Cluster Summarization. The first step adopts a pre-trained language model to embed documents into contextualized representations. The next step, Dimension Reduction, reduces the dimension of the representations for speeding up the subsequent clustering process. The Document Clustering step applies some clustering techniques, e.g., K-Means (Zhang et al., 2022), to the reduced representations for topic cluster identification. The last step, Cluster Summarization, reconstructs topic-word distribution by using word importance ranking metric, e.g., c-TF-IDF (Grootendorst, 2022), on each topic cluster. c-TF-IDF calculate the importance of the word w in the cluster k by

$$\text{tf}_{w,k} \times \log(1 + \frac{A}{f_w}), \tag{1}$$

where $\text{tf}_{w,k}$ is the word frequency of w in the document cluster k, A is the average word frequency of all clusters, and f_w is the frequency of word w across clusters. The higher value means the word w is more representative to a cluster k.

2.2 Pipeline Adaption for Cross-lingual Topic Identification

To adapt the current pipeline for cross-lingual topic identification, MLMs, such as Distilled XLM-R (Reimers and Gurevych, 2020; Conneau et al., 2020) and Cohere multilingual model, can be used in step 1 for embedding documents into language-agnostic representations $E \in R^{m \times d}$, where m is number of documents and d is dimension of representations. However, we observe that a number of dimensions of MLMs' representations retain language information. These dimensions are denoted as language-dependent dimensions (LDDs). To illustrate, we group documents written in language $l \in \{l_1, l_2\}$ and look into their representations. Let $e_i^l \in R^{m^l \times 1}$ be the values of i'th dimension for m^l documents written in l. We compare the values of each dimension $i \in d$ across two languages l_1 and l_2 by performing a two-sample t-test on $e_i^{l_1}$ and $e_i^{l_2}$. We then sort all dimensions based on the corresponding t-statistics in descending order. As the larger t-statistic indicates the larger mean value difference across languages, we hereby identify LDDs. As shown in the upper-left subplot of Fig. 2, the original MLM embeddings show notable distinctions for documents written in two different languages, suggesting the presence of LDDs within the original embeddings. Furthermore, after applying UMAP, a dimension reduction approach used by previous cluster-based topic models (Grootendorst, 2022; Zhang et al., 2022), even more significant LDDs are present (see the upper-right subplot of Fig. 2). This is likely to occur as UMAP focuses on capturing the local structure (McInnes et al., 2020).

Table 1: Top representative words of five sampled topics generated from BERTopic (Grootendorst, 2022) with default parameters. We first use Cohere multilingual model to embed the Airiti dataset (Chang et al., 2020) and then employ BERTopic to generate topics.

Topic#1	cell, protein, expression, induce, gene, mouse, find, show, study, treatment
Topic#2	細胞(cell), 蛋白(protein), 表現(expression), 基因(gene), 抑制(inhibition) 蛋白質(protein), 我們(we), 發現(discover), 調控(control), 病毒(virus)
Topic#5	firm, market, financial, company, return, investor, investment, bank, stock, model
Topic#22	反應(reaction), 分子(molecule), 高分子(polymer), 結構(structure), 合成(synthesize) 化合物(compound), 錯合物(complex), 具有(have), 形成(form), 利用(utilize)
Topic#46	市場(market), 報酬(return), 投資(investment), 股票(stock), 指數(index) 股價(stock price), 交易(transaction), 模型(model), 公司(company), 價格(price)

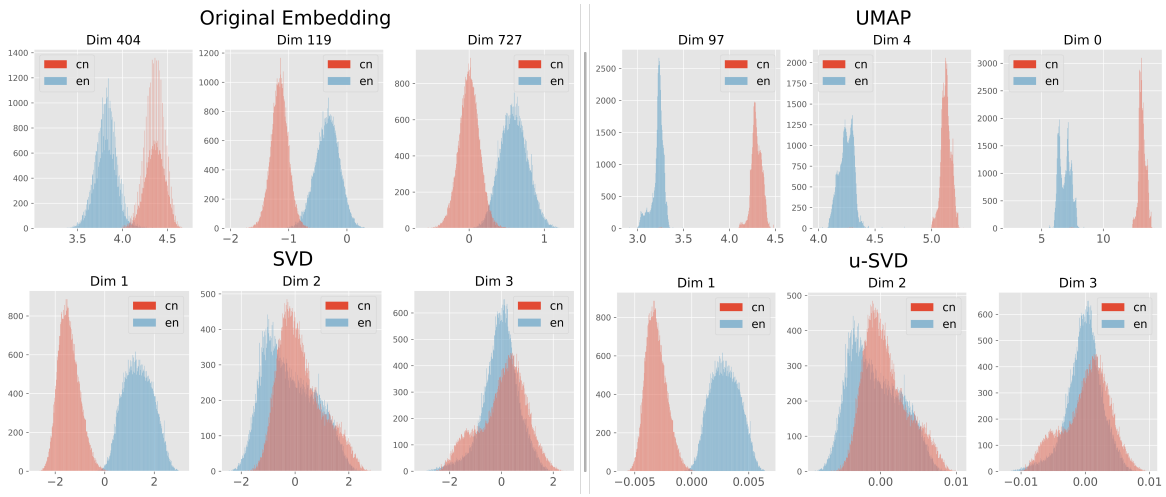


Figure 2: Top 3 language-dependent dimensions, sorted by t-statistic values, for original embeddings and embeddings reduced using UMAP, SVD and u-SVD. We utilize the Cohere multilingual model (see Section 3.2) to encode the documents in one of our experimental datasets, namely ECNews. The value distributions for Chinese (cn) and English (en) documents are indicated by red and blue, respectively. All UMAP, SVD, and u-SVD reduced the dimension size of the original representations from 768 to 100. Appendix A presents the same analysis to the other dataset, namely Rakuten Amazon.

LDDs adversely affect the subsequent document clustering process, as they disproportionately influence the distance calculations between documents during clustering. As a result, LDDs cause the algorithm to cluster documents by language rather than by their semantic meaning. In order to mitigate the negative impacts of LDDs, we repurpose the step 2 of the pipeline from a dimension reduction component to a dimension refinement component. Our dimension refinement component incorporates SVD, leveraging its notable feature that the reduced dimensions are orthogonal to one another. Note that previous researches have long applied SVD for topic modeling (Deerwester et al., 1990; Crain et al., 2012), yet its usage has been confined to monolingual topic modeling for decomposing the term-document matrix to capture the latent semantic structure. We propose a novel approach that applies SVD to neutralize LDDs from the representations generated by MLMs and further reduces the influence of languages. Owing to the orthogonal decomposition property of SVD, when one dimension retains language information, the remaining dimensions are more likely to capture other types of information. The lower-left subplot of Fig. 2 demonstrates that SVD consolidates the scattered LDDs into a concentrated set of reduced dimensions.

We explore two implementations of dimension refinement components, namely unscaled SVD (u-SVD) and SVD with Language dimension Removal (SVD-LR). Both u-SVD and SVD-LR methods follow the same decomposition manner as the standard SVD, which is represented by $E = U\Sigma V^T$. However, unlike standard SVD, u-SVD only utilizes $U \in R^{m \times r}$ to represent m documents in r reduced dimensions. Since U is an orthonormal matrix, u-SVD reduces the influence of LDDs by

ensuring that each dimension has a unit length. For instance, the lower-right subplot shows that u-SVD represents the dimensions using smaller scale (see x-axis) compared to the SVD in the lower-left subplot. By reducing the scale of dimensions, u-SVD decreases the negative contributions of LDDs in the subsequent clustering. u-SVD is a conservative approach as it reconciles the effects of LDDs without removing any dimension. In contrast, SVD-LR is more aggressive by removing the most influential LDD after performing SVD. Specifically, we represent the documents using $U\Sigma \in R^{m \times r}$ and use the two-sample t-test to identify the most influential LDD $\hat{r}$, which has the largest difference in the mean values of two languages. Then, SVD-LR removes $\hat{r}$ from $U\Sigma$.

Algorithm 1 Updated Pipeline for Cross-lingual Clustering-based Topic Model

Require: MLM, corpus, number of reduced dimensions r, number of topics K

1: Obtain E by embedding the corpus using the assigned MLM
2: $U, \Sigma, V^T = \text{SVD}(E, r)$
3: **if** u-SVD **then**
4: $\quad E^* = U$
5: **else if** SVD-LR **then**
6: $\quad$ Identify the most influential LDD $\hat{r}$ using two-sample t-test
7: $\quad$ Obtain E^* by removing $\hat{r}$ from $U\Sigma$
8: **end if**
9: $C_1, C_2, ..., C_K = \text{Kmeans}(E^*, K)$
10: $\phi_1, \phi_2, ..., \phi_K = \text{c-Tf-IDF}(C_1, C_2, ..., C_K)$
11: **return** $\phi_1, \phi_2, ..., \phi_K$

Algorithm 1 presents the updated pipeline, which is detailed as follows: (1) in line 1, documents are embedded using the MLM to obtain document representations E, (2) from line 2 to line 8, we perform the dimension refinement step [2] using either u-SVD or SVD-LR to obtain refined document representations E^*, (3) in line 9, Kmeans algorithm [3] are applied on E^* to group documents into K topic clusters, and (4) in line 10, we summarize and reconstruct the topic-word distribution for each topic cluster using c-TF-IDF (Eq. 1).

[2]We use the SVD implementation from Dask package https://www.dask.org.

[3]We use Kmeans implementation with default parameters from scikit-learn package https://scikit-learn.org/.

3 Experimental Setup

3.1 Dataset

We conduct experiments using three datasets: (1) **Airiti Thesis** which consists of 163,150 pairs of English and Chinese thesis abstracts (Chang et al., 2020). On average, each abstract contains 165 words. (2) **ECNews** comprises 50,000 Chinese news and 46,850 English news articles, with an average length of 11 words per article. (3) **Rakuten Amazon** is a compilation of 25,000 Japanese and 25,000 English product reviews, with an average of 27 words per review. ECNews and Rakuten Amazon were used in the previous research for cross-lingual topic evaluation (Wu et al., 2023). Considering that ECNews and Rakuten Amazon primarily contain short documents, we include Airiti Thesis in our experiments to evaluate the performance on identifying topics in longer documents.

3.2 Multilingual Language Model

We evaluate our proposed methods and compare them with other methods using three different MLMs: (1) **mBERT** (Devlin et al., 2019) has been investigated for its capability on cross-lingual classification tasks (Pires et al., 2019b). We use transformers[4] to load bert-base-multilingual-cased[5] and use output of special classification token ([CLS]) to get the mBERT embedding for a document. (2) **Distilled XLM-R** (Reimers and Gurevych, 2020) is designed for embedding a paragraph and is based XLM-R (Conneau et al., 2020), which is superior than mBERT in parallel sentence retrieval (Libovický et al., 2020). We use sentence-transformers[6] to access Distilled XLM-R (paraphrase-xlm-r-multilingual-v1). (3) **Cohere multilingual model** has shown its capabilities in various cross-lingual retrieval tasks (Kamalloo et al., 2023). We use the Cohere multilingual model (embed-multilingual-v2.0) by the API[7].

3.3 Baseline & Competitor

We compare three alternative baselines to show the effectiveness of using u-SVD and SVD-LR as dimension refinement step: (1) **original embedding**, referred as *OE*, which is simply generated

[4]https://github.com/huggingface/transformers

[5]https://huggingface.co/bert-base-multilingual-cased

[6]https://www.sbert.net

[7]https://txt.cohere.com/multilingual/

from the given MLM, (2) **UMAP** [8], which is the popular dimension reduction method, whose effectiveness in identifying monolingual topics has been shown (i.e., CETopic) (Zhang et al., 2022), and (3) **pure SVD**, which is used as a benchmark to compare against u-SVD and SVD-LR. Moreover, we compare two recent cross-lingual topic models: (1) **Cb-CLTM** (Chang and Hwang, 2021) incorporates a cross-lingual word space into the generative process of latent Dirichlet allocation (Blei et al., 2003). Cb-CLTM demonstrates its superior performances compared to other probabilistic cross-lingual topic models. To enable the Cb-CLTM, we use pre-aligned English-Chinese and English-Japanese word spaces from MUSE project[9]. (2) **InfoCTM** (Wu et al., 2023) is a neural topic model that identifies topics across languages based on the guidance of the given bilingual dictionary. InfoCTM is the state-of-the-art neural cross-lingual topic model. We follow the report of the InfoCTM to use a Chinese-English dictionary from MDBG[10] and Japanese-English dictionary from MUSE project to link topics across languages.

3.4 Evaluation Metric

We measure the generated topics using two metrics widely adopted in previous CLTMs: CNPMI and Diversity. For each topic $k \in K$, we select top-N represented words for l_1 and l_2 languages, denoted as $\mathcal{W}_{k,N}^{l_1}$ and $\mathcal{W}_{k,N}^{l_2}$.

CNPMI (Hao and Paul, 2020; Chang and Hwang, 2021; Wu et al., 2023) measures the coherence of generated topic words across languages:

$$-\frac{1}{N^2} \sum_{w_i \in \mathcal{W}_{k,N}^{l_1}, w_j \in \mathcal{W}_{k,N}^{l_2}} \frac{\log \frac{Pr(w_i,w_j)}{Pr(w_i)Pr(w_j)}}{\log Pr(w_i,w_j)}, \quad (2)$$

where $Pr(w_i, w_j)$ is the co-occurring probability of words w_i and w_j and $Pr(w_i)$ is the marginal probability of w_i. For Airiti Thesis, we estimate the probability using the comparable abstracts in the Airiti Thesis. For ECNews and Rakuten, we measure the probability using comparable Wikipedia corpus [11]. The CNPMI ranges from -1 (least coherent) to 1 (most coherent), and we report the average CNPMI scores across K topics.

Diversity (Dieng et al., 2020) measures the uniqueness of generated topic words across K topics:

$$\frac{|\bigcup_{1 \le k \le K} \mathcal{W}_{k,N}^{l_1}| + |\bigcup_{1 \le k \le K} \mathcal{W}_{k,N}^{l_2}|}{K \times 2 \times N}, \quad (3)$$

which ranges between 0 (the least diversity) and 1 (the highest diversity). To combine the two aspects, we further compute **Topic Quality (TQ)** (Dieng et al., 2020) as the product of max(0, CNPMI) and Diversity, providing a cohesive measure for our analysis. Note that positive CNPMI contributes to TQ because NPMI measurement positively correlates with human interpretability (Lau et al., 2014). The topic with negative CNPMI are considered to be uninterpretable.

We evaluate top 15 words ($N = 15$) of each topic for CNPMI and Diversity. For more robust comparison, we re-run every method five times using different seeds and report the average performance.

4 Results & Analysis

4.1 Performance of Cross-lingual Topic Model

Table 2 shows the performance of different methods on three datasets. We adopt the following settings. Cohere multilingual model is chosen as the MLM, which embeds every document into 768 dimensional representations. All dimension reduction methods reduce the original embedding from 768 to 100 dimensions. The number of topics (clusters) is set to 50 because InfoCTM (Wu et al., 2023) reports performances on this number for both ECNews and Rakuten Amazon.

The results clearly indicate that incorporating a clustering-based topic model pipeline with three baseline embeddings, including original embedding, UMAP, and SVD, does not perform well in terms of CNPMI and Diversity. We also use feature-wise min-max normalization on UMAP, resulting in UMAP-norm. However, UMAP-norm does not enhance performance. Both Cb-CLTM and InfoCTM exhibit high diversity scores. However, when applied to the Airiti dataset, they generate topics with negative CNPMI scores, suggesting that their generated topics are difficult to be interpreted by human (Lau et al., 2014). The pipelines

[8] We use the implementation from umap-learn package https://github.com/lmcinnes/umap.

[9] https://github.com/facebookresearch/MUSE

[10] https://www.mdbg.net/chinese/dictionary?page=cc-cedict

[11] We use the implementation https://github.com/BobXWu/CNPMI from the authors of InfoCTM (Wu et al., 2023).

Table 2: Comparison of topic quality for baselines, competitors, and our proposed methods.

Dataset	Airiti			ECNews			Rakuten Amazon		
Metric	**CNPMI**	**Diversity**	**TQ**	**CNPMI**	**Diversity**	**TQ**	**CNPMI**	**Diversity**	**TQ**
OE	-0.244	0.570	0.000	0.022	0.554	0.012	0.009	0.290	0.003
UMAP	-0.202	0.572	0.000	0.019	0.598	0.011	0.003	0.265	0.001
UMAP-norm	-0.207	0.585	0.000	0.019	0.613	0.012	0.003	0.264	0.001
SVD	-0.251	0.564	0.000	0.026	0.567	0.015	0.009	0.282	0.003
Cb-CLTM	-0.145	**0.941**	0.000	0.021	0.774	0.016	0.008	0.699	0.006
InfoCTM	-0.087	0.917	0.000	0.044	**0.905**	0.040	0.033	**0.856**	**0.028**
SVD-LR	**0.179**	0.571	**0.103**	**0.087**	0.741	0.065	0.032	0.607	0.019
u-SVD	0.171	0.603	**0.103**	0.086	0.823	**0.071**	**0.037**	0.665	0.025

with u-SVD and SVD-LR result in less diverse topics than Cb-CLTM and InfoCTM but have better CNPMI and TQ on the Airiti and ECNews datasets. Moreover, InfoCTM, SVD-LR, and u-SVD reach comparable CNPMI and TQ on the Rakuten Amazon dataset. These results suggest that u-SVD and SVD-LR can generalize to datasets of different lengths.

4.2 Performance on Different MLMs

To test the generalizability of u-SVD and SVD-LR, we evaluate and compare performances on three MLMs, namely mBERT, Distilled XLM-R, and Cohere Multilingual Model, on the Airiti Thesis. All three MLMs generate document embedding with 768 dimensions. To benchmark with the results shown in Table 2, each document embedding is also reduced or refined to 100 dimensions, and the number of topic clusters is set to 50.

Table 3 reveals that when using mBERT, both SVD-LR and u-SVD achieve only marginal improvement, if any, on topic quality compared to other three baselines. This may be attributed to limited cross-lingual capability of mBERT because it is the first generation MLM. On the other hand, with the document representations generated by more capable MLMs, namely Distilled XLM-R and Cohere Multilingual Model, SVD-LR and u-SVD consistently demonstrates their robust performances and generate topic clusters with better topic quality.

4.3 Sensitivity Analysis on the Size of Reduced Embeddings

To better understand u-SVD and SVD-LR, we conduct sensitivity analysis on the size of embeddings. In this analysis, we use all three datasets and fix the number of cluster topics at 50. We reduce the document representations generated by Cohere Multilingual Model from 768 to 100, 200, and 500 to see their influence on the CNPMI.

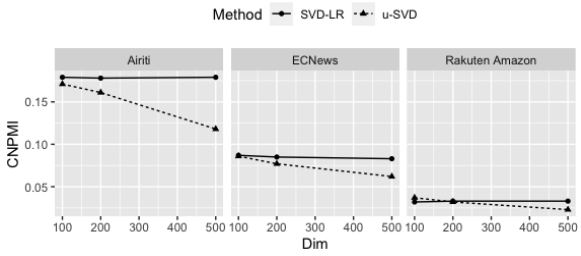


Figure 3: Sensitivity analysis of u-SVD and SVD-LR on different dimensions.

Fig. 3 shows that SVD-LR has a more robust result across different embedding dimensions. SVD-LR preserves the importance weight (i.e., Σ) of each dimension except for the most influential LDD, resulting in robust performance across various dimensions. On the contrary, u-SVD abandons the importance weight of dimensions from SVD to lessen the effect of LDDs. Thus, u-SVD is affected by those dimensions that originally had small singular values, leading to poorer outcomes when more dimensions are utilized. In summary, while both u-SVD and SVD-LR lose some information due to the elimination of LDDs, SVD-LR seems to lose fewer information when more dimensions are introduced.

4.4 Qualitative Result

We apply the Cohere multilingual model to embed the Airiti dataset and use BERTopic (Grootendorst, 2022), which implements the previous pipeline of clustering-based topic model. Table 1 shows the representative words for ten manually sampled topics generated by BERTopic. Each topic consists of top words purely from a single language and is misaligned by the semantic meaning. For in-

Table 3: Topic quality of using three different MLMs.

Method	mBERT			Distilled XLM-R			Cohere Multilingual Model		
	CNPMI	Diversity	TQ	CNPMI	Diversity	TQ	CNPMI	Diversity	TQ
OE	-0.122	0.478	0.000	-0.211	0.600	0.000	-0.244	0.570	0.000
UMAP	-0.190	0.421	0.000	-0.198	0.536	0.000	-0.202	0.572	0.000
SVD	-0.117	0.476	0.000	-0.208	0.580	0.000	-0.251	0.564	0.000
SVD-LR	-0.149	0.492	0.000	0.172	0.527	0.091	**0.179**	0.571	**0.103**
u-SVD	**0.001**	**0.591**	**0.000**	**0.182**	**0.629**	**0.115**	0.171	**0.603**	**0.103**

Table 4: Top representative words of 10 sampled topics from updated pipeline with u-SVD and SVD-LR

u-SVD	
Topic#2	optical, 光學(optics), 雷射(laser), 發光(glow), laser, light, 元件(component), led, 我們(we), 結構(structure)
Topic#7	影像(image), image, 我們(we), 演算法(algorithm), 方法(method), propose, algorithm, 提出(propose), method, video
Topic#9	網路(network), 無線(wireless), 傳輸(transmission), 通訊(communication), network, 我們(we), 使用(use), 系統(system), 提出(propose), propose
Topic#20	polymer, 高分子(polymer), 材料(material), surface, film, increase, high, property, 結構(structure)
Topic#21	投資(investment), 市場(market), 報酬(return), market, return, 股票(stock), 交易(transaction), 指數(index), stock, 投資人(investor)
SVD-LR	
Topic#1	optical, 發光(glow), 光學(optics), 雷射(laser), 元件(component), led, laser, light, 結構(structure), 我們(we)
Topic#6	影像(image), image, 我們(we), 演算法(algorithm), 方法, propose, algorithm, 提出(propose), method, video
Topic#12	polymer, 高分子(polymer), 材料(material), surface, 表面(surface), film, 結構(structure), increase, high, material
Topic#17	網路(network), 無線(wireless), network, 傳輸(transmission), 我們(we), 使用(use), 節點(node), 通訊(communication), 提出(propose), 服務(service)
Topic#21	投資(investment), 市場(market), market, 報酬(return), return, 指數(index), 交易(transaction), 股票(stock), stock, investor

stance, topics #1 & #2 discuss the same topic but are separated into two topics. Table 4 uses the same setting as Table 1 but apply u-SVD and SVD-LR for dimension refinement. Most topics contain representative words across languages and are grouped by the semantic meanings of topics. For example, the concept of "Financial Market" is separated into two topics in Table 1, namely topics #5 & #46, based on languages. On the contrary, as shown in Table 4, topic #21 from u-SVD and topic #21 from SVD-LR include the words of different languages yet with similar concept.

5 Related Work

5.1 Clustering-based Topic Model

Recent works (Sia et al., 2020; Zhang et al., 2022; Grootendorst, 2022) have explored methods that cluster contextualized representations to identify topics from a corpus. Sia et al. (2020) used the BERT model to encode each token into a representation, averaging these representations to obtain a document-level representation. They then applied K-means clustering to these document representations and reconstructed the topic-word distributions using a tf-idf weighting scheme. The coherence performance of their resultant topics was comparable to that of the traditional topic model, LDA (Blei et al., 2003). Similarly, Zhang et al. (2022) and Grootendorst (2022) proposed a pipeline consisting of four steps. First, they used language models, such as sentence BERT (SBERT), to encode documents into representations. Next, they applied the dimension reduction technique UMAP to these representations. In the third step, they used K-means clustering on the reduced representations to gen-

erate document clusters, each considered a topic cluster. Finally, they employed a word importance ranking method, c-Tf-IDF, to identify representative topic words. Their pipelines outperformed neural topic models in terms of both efficiency and topic quality. However, the proposed pipeline hasn't been evaluated in cross-lingual settings. Our study aims to fill this gap.

5.2 Language-dependent Component

Several studies (Libovický et al., 2020; Zhao et al., 2021; Chang and Hwang, 2021) have shown that MLM-generated representations contain language-dependent components (LDDs), which signal language identity and hinder cross-lingual transfer. To mitigate such LDDs, Libovický et al. (2020) noted that representations of the same language are closely located in the space. They recommend removing the language-specific mean from the mBERT representations as a solution. However, even after this adjustment, the resulting representations can still be utilized as features to predict the language accurately, suggesting that simply removing the language-specific means from the representations is insufficient. Zhao et al. (2021) propose a method that requires parallel corpus to fine-tune mBERT and XLM-R for generating language-agnostic representations. The method fine-tunes the language model to align the sentence pairs from the parallel corpus. To further close the gap between languages, the method also constrains the representations of different languages to be distributed with zero mean and unit variance. Such an idea is close to our proposed u-SVD; however, u-SVD is a more efficient and appropriate method for models with ample parameters because it does not require parallel corpus and fine-tuning. Chang and Hwang (2021) observed that LDDs prevent their topic model from identifying topics across languages. They proposed training a logistic regression to identify the contributed dimensions (i.e., LDDs) for language identity and removed them from the representations. They found that removing the LDDs helped identify more cross-lingual topics. However, removing the LDDs directly from the original representations comes with the cost of losing semantic completeness. Our SVD-LR eases this issue because utilizing SVD helps us to consolidate the scattered language-dependent dimensions into one specific dimension. Therefore, SVD-LR only removes the most contributed LDD, potentially minimizing the risk of losing other semantic meanings.

6 Conclusion

We investigate the problem with the current pipeline of clustering-based topic model when applied on multilingual corpus, which is caused by language-dependent dimensions in the multilingual contextualized embedding. To solve this problem, we propose two methods for dimension refinement, namely u-SVD and SVD-LR. Our experiments suggest that the updated pipeline with our proposed refinement component is effective in cross-lingual topic identification and results in more coherent topics than existing cross-lingual topic models.

Limitations

This study only evaluates proposed dimension refinement components, u-SVD and SVD-LR, on three MLMs, namely mBERT, XLM-R, and Cohere Multilingual Model. We chose these three MLMs because of their extensive investigations in cross-lingual retrieval tasks. The future work may investigate more other MLMs such as LASER[12], Universal Sentence Encoder[13], and OpenAI embedding API [14]. Extensive experiments on more language pairs are another future work since we only evaluate two English-Chinese datasets and one English-Japanese dataset. It is worth noting that our proposed methods are effective in language pairs from distant and different language families. Furthermore, it's also crucial to investigate our methods for datasets with more than two languages, such as EuroParl.

References

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. 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(Jan):993–1022.

Chia-Hsuan Chang and San-Yih Hwang. 2021. A word embedding-based approach to cross-lingual topic

[12] https://github.com/facebookresearch/LASER

[13] https://www.kaggle.com/models/google/universal-sentence-encoder/

[14] https://platform.openai.com/docs/api-reference/embeddings

modeling. *Knowledge and Information Systems*, 63(6):1529–1555.

Chia-Hsuan Chang, San-Yih Hwang, and Tou-Hsiang Xui. 2018. Incorporating Word Embedding into Cross-Lingual Topic Modeling. In *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 17–24, San Francisco, CA, USA. IEEE.

Chia-Ming Chang, Chia-Hsuan Chang, and San-Yih Hwang. 2020. Employing word mover's distance for cross-lingual plagiarized text detection. In *Proceedings of the Association for Information Science and Technology*, volume 57, page e229.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Steven P. Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. 2012. Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 129–161. Springer US, Boston, MA.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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.

Philipp Dufter and Hinrich Schütze. 2020. Identifying Elements Essential for BERT's Multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Shudong Hao and Michael J Paul. 2020. An Empirical Study on Crosslingual Transfer in Probabilistic Topic Models. *Comput. Linguist.*, 46(1):95–134.

Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan Boyd-Graber. 2014. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1166–1176, Baltimore, Maryland. Association for Computational Linguistics.

Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, Nandan Thakur, David Alfonso-Hermelo, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Evaluating Embedding APIs for Information Retrieval.

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.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

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 Volume 2 - EMNLP '09*, volume 2, page 880, Singapore. Association for Computational Linguistics.

Tiziano Piccardi and Robert West. 2021. Crosslingual Topic Modeling with WikiPDA. In *Proceedings of the Web Conference 2021*, pages 3032–3041, Ljubljana Slovenia. ACM.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019a. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019b. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. 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.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liang-Ming Pan, and Anh Tuan Luu. 2023. 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*, volume 37, pages 13763–13771. AAAI Press.

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.

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.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing Language-Agnostic Multilingual Representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

A Two-sample t-test on MLM embeddings for Rakuten Amazon dataset

We use the same setting as in Fig 2 to display the top three language-dependent dimensions of the Rakuten Amazon dataset in Fig 4.

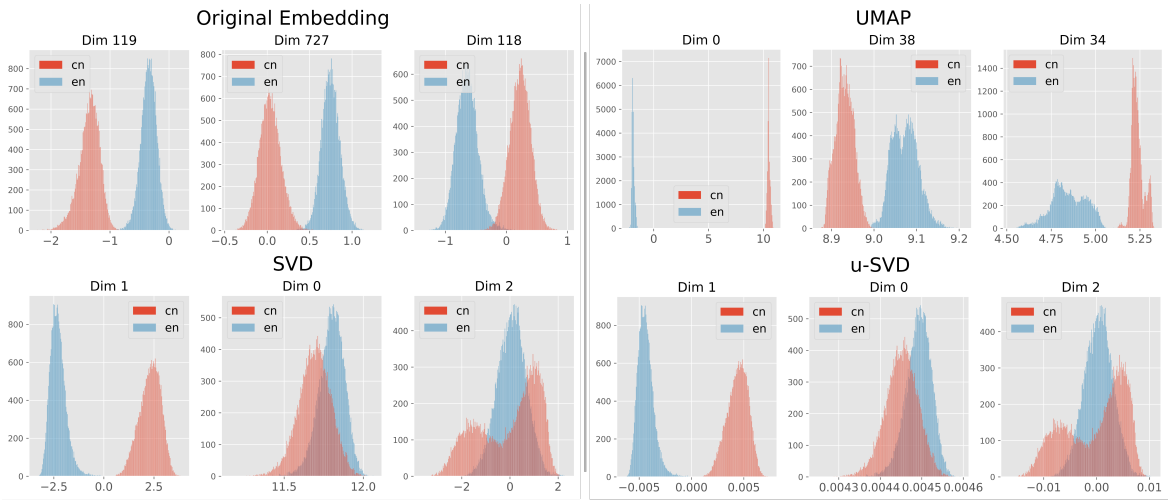


Figure 4: Top 3 language-dependent dimensions, sorted by t-statistic values, for original embeddings and embeddings reduced using UMAP, SVD and u-SVD on Rakuten Amazon dataset.