InfoCTM: A Mutual Information Maximization Perspective of Cross-Lingual Topic Modeling

Xiaobao Wu[1], **Xinshuai Dong**[2], **Thong Nguyen**[3], **Chaoqun Liu**[1,4], **Liang-Ming Pan**[3], **Anh Tuan Luu**[1]
[1]Nanyang Technological University, Singapore
[2]Carnegie Mellon University, USA
[3]National University of Singapore, Singapore
[4]DAMO Academy, Alibaba Group, Singapore
{xiaobao002,chaoqun001,anhtuan.luu}@ntu.edu.sg xinshuad@andrew.cmu.edu {e0998147,liangmingpan}@u.nus.edu

Abstract

Cross-lingual topic models have been prevalent for cross-lingual text analysis by revealing aligned latent topics. However, most existing methods suffer from producing repetitive topics that hinder further analysis and performance decline caused by low-coverage dictionaries. In this paper, we propose the Cross-lingual Topic Modeling with Mutual Information (InfoCTM). Instead of the direct alignment in previous work, we propose a topic alignment with mutual information method. This works as a regularization to properly align topics and prevent degenerate topic representations of words, which mitigates the repetitive topic issue. To address the low-coverage dictionary issue, we further propose a cross-lingual vocabulary linking method that finds more linked cross-lingual words for topic alignment beyond the translations of a given dictionary. Extensive experiments on English, Chinese, and Japanese datasets demonstrate that our method outperforms state-of-the-art baselines, producing more coherent, diverse, and well-aligned topics and showing better transferability for cross-lingual classification tasks.

Introduction

Cross-lingual topic models have been popular for cross-lingual text analysis and applications (Vulić, De Smet, and Moens 2013). As shown in Figure 1, they aim to discover cross-lingual topics from bilingual corpora. Each topic is interpreted as the relevant words in the corresponding language. The same cross-lingual topics are required to be aligned (semantically consistent across languages). For example, the English Topic#3 and Chinese Topic#3 are aligned as they are both about music, and English Topic#5 and Chinese Topic#5 are aligned and both about the celebrity. These aligned topics can reveal commonalities and differences across languages and cultures, which enables cross-lingual analysis without supervision (Ni et al. 2009; Shi et al. 2016; Gutiérrez et al. 2016; Lind et al. 2019).

Since parallel corpora are often difficult to access, recent cross-lingual topic models tend to rely on vocabulary linking information from bilingual dictionaries (Shi et al. 2016; Yuan, Van Durme, and Ying 2018; Yang, Boyd-Graber, and Resnik 2019; Wu et al. 2020a). They commonly use translations of a dictionary as linked cross-lingual words and make these words belong to the same cross-lingual topics, *i.e.*, align their topic representations (what topics a word belongs to). For instance in Figure 1, the word "song" and its Chinese translation both belong to Topic#3 of English and Chinese. These methods are more practical because dictionaries are widely accessible. Recent studies (Bianchi et al. 2020; Mueller and Dredze 2021) employ multilingual BERT (Devlin et al. 2018) for multilingual corpora, but they are not traditional cross-lingual topic models since they do not discover aligned cross-lingual topics.

However, despite the practicality, these methods, *e.g.*, MCTA (Shi et al. 2016) and NMTM (Wu et al. 2020a), suffer from two issues: (**1**) They tend to generate low-quality

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

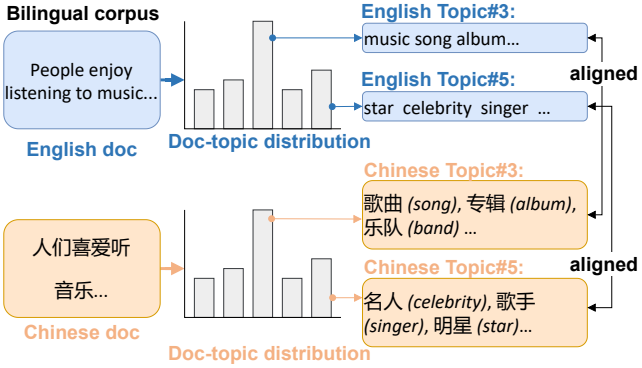


Figure 1: Illustration of cross-lingual topic models, producing aligned topics of different languages. *Words* in the brackets are the corresponding English translations.

English Topic#1:	photos	style	finds	ebay	week	vintage
Chinese Topic#1:	风格	文字	西装	模特	西服	每周
English Topic#2:	fashion	week	photos	new	style	beauty
Chinese Topic#2:	时尚	时髦	流行	时装	模特	每周
English Topic#3:	photos	fashion	new	beauty	line	week
Chinese Topic#3:	时尚	时髦	全新	金山	流行	模特

Table 1: Top related words of of repetitive cross-lingual topics produced by MCTA (Shi et al. 2016). Repetitive words are underlined.

repetitive cross-lingual topics, as exemplified in Table 1. We see they all refer to similar semantics with many repetitive words like "fashion" and "photo". Consequently, this makes the discovered topics less useful for further text analysis and also hampers the performance of downstream applications. (**2**) These methods mostly suffer from performance decline caused by *low-coverage dictionaries*. Due to cultural differences, available bilingual dictionaries can only cover a small part of the involved vocabulary, especially for low-resource languages (Chang and Hwang 2021). Low-coverage dictionaries have been shown to hinder the topic alignment of cross-lingual topic models (Jagarlamudi and Daumé 2010; Hao and Paul 2020). For example, it will be difficult to align English Topic#3 and Chinese Topic#3 in Figure 1 if we are unaware of the Chinese translations of English words like "song" or "album".

To address the above problems, we in this paper propose a novel neural cross-lingual topic model, named **C**ross-lingual **T**opic **M**odeling with Mutual **Info**rmation (InfoCTM). First, to address the repetitive topic issue, we propose a Topic Alignment with Mutual Information (TAMI) method. Instead of the direct alignment in previous work (Shi et al. 2016; Yang, Boyd-Graber, and Resnik 2019; Wu et al. 2020a), TAMI maximizes the mutual information between the topic representations (what topics a word belongs to) of linked cross-lingual words. This not only aligns the topic representations of linked words but also prevents them from degenerating into similar values, which encourages words to belong to different topics. As a result, the discovered topics are distinct from each other, which alleviates the repetitive topic issue and enhances topic coherence and alignment.

Second, to find linked words for TAMI and to overcome the low-coverage dictionary issue, we propose a Cross-lingual Vocabulary Linking (CVL) method. Instead of only using the translations in a dictionary as linked words, CVL additionally links a word to the translations of its nearest neighbors in the word embedding space. This is motivated by the fact that topic models focus on what topics a word belongs to rather than accurate translations. For instance in Figure 1, the English word "album" and the Chinese translation of "song" both belong to Topic#3 of English and Chinese although they are not translations of each other. With CVL, we can obtain more linked cross-lingual words for our TAMI beyond the given dictionary, which mitigates the low-coverage dictionary issue.

The contributions of this paper can be summarized as [1]:

- We propose a novel neural cross-lingual topic model with a new topic alignment with mutual information method that can prevent degenerate topic representations and avoid generating repetitive topics.
- We further propose a novel cross-lingual vocabulary linking method, which finds more linked cross-lingual words beyond the translations and effectively alleviates the low-coverage dictionary issue.
- We conduct extensive experiments on datasets of different languages and show that our model consistently outperforms baselines, producing higher-quality topics and showing better cross-lingual transferability for downstream tasks.

[1]Our code is available at https://github.com/bobxwu/InfoCTM.

Related Work

Cross-lingual Topic Models Cross-lingual topic modeling is proposed as an extension of monolingual topic modeling (Blei, Ng, and Jordan 2003; Blei and Lafferty 2006; Wu and Li 2019). The earliest polylingual topic model (Mimno et al. 2009) uses one topic distribution to generate a tuple of comparable documents in different languages, *e.g.*, EuroParl (Koehn 2005). As it is limited by the requirement of parallel/comparable corpus to link documents, another line of work uses vocabulary linking from bilingual dictionaries (Jagarlamudi and Daumé 2010; Boyd-Graber and Blei 2012). Recent studies of this line (Shi et al. 2016; Yang, Boyd-Graber, and Resnik 2019; Wu et al. 2020a) commonly use translations in a dictionary as linked words and directly align topics by making these words belong to the same topics. Chang and Hwang (2021) induce translations by transforming cross-lingual word embeddings into the same space; however, they heavily rely on the isomorphism assumption (Conneau et al. 2017) that cannot always hold as they find. Recently, Bianchi et al. (2020); Mueller and Dredze (2021) employ multilingual BERT (Devlin et al. 2018) to infer cross-lingual topic distributions for zero-shot learning, but they cannot discover aligned cross-lingual topics as required. Different from these work, we focus on two crucial issues of cross-lingual topic modeling: the repetitive topic issue and the low-coverage dictionary issue. To address these issues, we propose the topic alignment with mutual information instead of direct alignment and propose the cross-lingual vocabulary linking method instead of only using translations of a dictionary.

Mutual Information Maximization Mutual information maximization has been prevalent to learn visual and language representations (Bachman, Hjelm, and Buchwalter 2019; Kong et al. 2020; Chi et al. 2020; Dong et al. 2021). In practice, mutual information maximization is approximated with a tractable lower bound, such as InfoNCE (Van den Oord, Li, and Vinyals 2018) and InfoMax (Hjelm et al. 2019). These are also known as contrastive learning (Arora et al. 2019; Wang and Isola 2020; Nguyen et al. 2022; Wu, Luu, and Dong 2022) that learns the representation similarity of positive and negative samples. Some recent studies (Xu et al. 2022) apply mutual information for monolingual topic modeling and focus on the representations of documents. We share the same perspective of information theory but look into a different problem, cross-lingual topic modeling. More importantly, instead of learning the representations of documents, we focus on the topic representations of words, which motivates our topic alignment with mutual information. This is also different from precedent work.

Methodology

We first introduce the problem setting of cross-lingual topic modeling. Then, we present our new method Cross-lingual Topic Modeling with Mutual Information (InfoCTM), com-

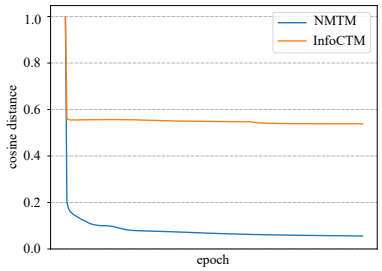


Figure 2: Cosine distance between the topic representations of words over the course of training. The results show that while the topic representations degenerate into similar values in NMTM (Wu et al. 2020a), our InfoCTM successfully avoids degenerate topic representations.

prised of Topic Alignment with Mutual Information (TAMI) and Cross-lingual Vocabulary Linking (CVL).

Problem Setting and Notations

Consider a bilingual corpus of language ℓ_1 and ℓ_2 (*e.g.*, English and Chinese). The vocabulary sets of each language are $\mathcal{V}^{(\ell_1)}$ and $\mathcal{V}^{(\ell_2)}$ with sizes as V_1 and V_2. Letting w_i denote the i-th word type in the bilingual corpus, we assume the previous V_1 words are in language ℓ_1 and the last V_2 words are in language ℓ_2: $\mathcal{V}^{(\ell_1)} = \{w_i | i = 1, \ldots, V_1\}$ and $\mathcal{V}^{(\ell_2)} = \{w_i | i = V_1 + 1, \ldots, V_1 + V_2\}$. As illustrated in Figure 1, cross-lingual topic models aim to discover K topics for each language from the bilingual corpus. Each topic of a language is defined as a distribution over words in the vocabulary (topic-word distribution). Namely, the Topic#k of language ℓ_1 and ℓ_2 are defined as $\boldsymbol{\beta}_k^{(\ell_1)} \in \mathbb{R}^{V_1}$ and $\boldsymbol{\beta}_k^{(\ell_2)} \in \mathbb{R}^{V_2}$ respectively. We require the Topic#k in language ℓ_1 and the Topic#k in language ℓ_2 to be aligned, *i.e.*, semantically consistent across languages. For example, the English Topic#3 and Chinese Topic#3 both focus on music in Figure 1. Besides, cross-lingual topic models also infer what topics a document contains, *i.e.*, the topic distributions of documents (doc-topic distributions), defined as $\boldsymbol{\theta}^{(\ell_1)}, \boldsymbol{\theta}^{(\ell_2)} \in \mathbb{R}^K$. We require the doc-topic distributions to be consistent across languages. If two documents in different languages contain similar topics, their inferred doc-topic distributions should also be similar. For instance, Figure 1 shows that the doc-topic distributions of the parallel English and Chinese document are similar.

Aligning Topics across Languages By Maximizing Mutual Information

We first analyze what causes repetitive topics with a state-of-the-art method, and then provide our solution called topic alignment with mutual information.

What Causes Repetitive Topics? In order to align topics, recent methods commonly use translations of a dictionary as linked cross-lingual words and directly align their topic representations. The topic representation of a word represents what topics this word belongs to. For example, Yang,

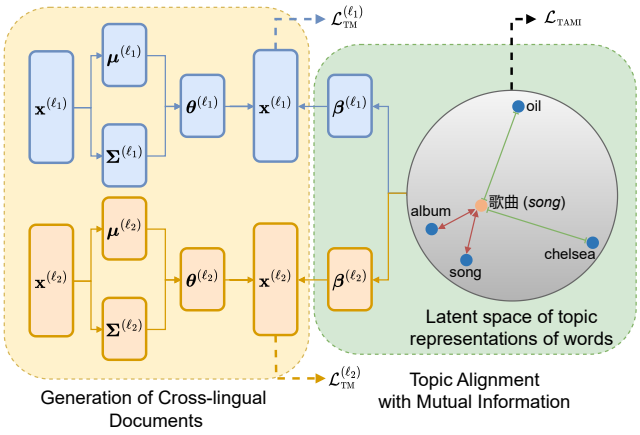


Figure 3: Illustration of InfoCTM. The generation of cross-lingual documents follows VAE. The proposed topic alignment with mutual information method aligns the topic representations of linked words ("歌曲" (*song*) and "song" or "album") and also keeps the distance between the topic representations of unlinked words ("歌曲" (*song*) and "oil" or "chelsea") to avoid degenerate topic representations.

Boyd-Graber, and Resnik (2019) computes the topic distributions of words as topic representations and aligns them through inference, and Shi et al. (2016); Wu et al. (2020a) transform topic representations of words to another vocabulary space and align them through generations. However, we find these methods using direct alignment have a severe issue: they easily produce repetitive topics (shown in Table 1 and the experiment section). To investigate the behind reason, we compute the cosine distance between the learned topic representations in a state-of-the-art method, NMTM (Wu et al. 2020a). Figure 2 shows that the cosine distance is close to 0 after training in NMTM. This means NMTM ends with a trivial solution that all topic representations become similar. This is because the direct alignment of NMTM only encourages capturing the similarity between topic representations while ignoring the dissimilarity between them. As a result, all topic representations wrongly degenerate into similar values, and the discovered topics cover similar words, which leads to repetitive topics.

Topic Alignment with Mutual Information Motivated by the above analyses, we aim to (i) capture the similarity between the topic representations of linked cross-lingual words and (ii) avoid degenerate topic representations. For these two purposes, we propose the topic alignment with mutual information (TAMI). Figure 3 illustrates the idea of our TAMI; Figure 2 shows that our InfoCTM with TAMI can effectively avoid degenerate topic representations. Specifically, we define random variables W and W' as two linked cross-lingual words with related semantics, *e.g.*, a translation pair. We can achieve the above two purposes by maximizing the mutual information between W and W' estimated by their topic representations:

$$\max I(W; W'). \tag{1}$$

Intuitively, this mutual information measures the dependency between W and W'. Maximizing this dependency can make the topic representation of linked words similar. Meanwhile, this dependency is reduced if topic representations are all similar to each other since a word will be associated with every other word. Thus, maximizing this dependency can also keep the dissimilarity between the topic representations of unlinked words and thus avoid degenerate topic representations.

Unfortunately, it is generally intractable to directly maximize mutual information when cooperating with neural networks, so we resort to a lower bound on it. One particular lower bound, InfoNCE (Logeswaran and Lee 2018; Van den Oord, Li, and Vinyals 2018) has been shown to work well in practice. Similarly, we relax the mutual information following InfoNCE as:

$$I(W;W') \geqslant \log|\mathcal{B}| + \mathbb{E}_{p(w_i,w_j)}\left[\log \frac{\exp(g(f(w_i),f(w_j)))}{\sum_{w_{j'}\in\mathcal{B}}\exp(g(f(w_i),f(w_{j'})))}\right]. \quad (2)$$

Here w_i and w_j are specific values of W and W' respectively. $f : \mathcal{V}^{(\ell_1)} \cup \mathcal{V}^{(\ell_2)} \rightarrow \mathbb{R}^K$ denotes a lookup function that maps a word type w_i into a vector $\boldsymbol{\varphi}_i$ as its topic representation. So we have $g(f(w_i), f(w_j)) = g(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j)$. Function $g(\cdot,\cdot)$ is a critic to characterize the similarity between $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varphi}_j$. We implement g as a scaled cosine function (Wu et al. 2018): $g(a,b) = \cos(a,b)/\tau$ where τ is a temperature hyper-parameter. Set $\mathcal{B}$ includes positive sample w_j and $(|\mathcal{B}|-1)$ negative samples. This is also known as contrastive learning (Chen et al. 2020; Tian, Krishnan, and Isola 2020) where we pull close the topic representations of a positive pair (w_i, w_j) and push away the topic representations of negative pairs $(w_i, w_{j'})$ $(j' \neq j)$. From the perspective of contrastive learning, the maximization of mutual information can also be justified by the alignment and uniformity following Wang and Isola (2020). Maximizing the mutual information encourages the alignment and uniformity of topic representations of words in the latent space, and thus they are prevented to degenerate into close points.

Cross-lingual Vocabulary Linking Now we describe how to find linked cross-lingual word pair (w_i, w_j) (a positive pair). As previous work (Shi et al. 2016; Yuan, Van Durme, and Ying 2018; Wu et al. 2020a), (w_i, w_j) can be a translation pair sampled from a bilingual dictionary. However, dictionaries could be low-coverage in real-world applications due to cultural differences, especially for low-resources languages. Low-coverage dictionaries provide insufficient translations and incur performance decline (Jagarlamudi and Daumé 2010; Hao and Paul 2020).

To alleviate the low-coverage dictionary issue, we propose a cross-lingual vocabulary linking (CVL) method. CVL first prepares monolingual word embeddings via the commonly-used Word2Vec (Mikolov et al. 2013) for each language. As shown in Figure 4, CVL then links word w_i to the translations of its nearest neighbors in the embedding space beside its own translations. We denote $\mathrm{CVL}(w_i)$ as

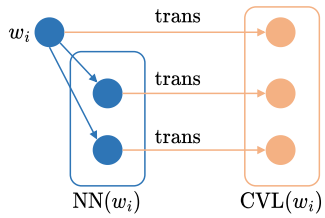


Figure 4: Illustration of Cross-lingual Vocabulary Linking.

the linked word set of w_i, which is defined as

$$\mathrm{CVL}(w_i) = \bigcup_{w} \mathrm{trans}(w), \text{where } w \in \{w_i\} \cup \mathrm{NN}(w_i). \quad (3)$$

Here, $\mathrm{NN}(w_i)$ denotes the set of nearest neighbors of word w_i. $\mathrm{trans}(w)$ denotes the translation set of word w in a given dictionary. CVL views the cross-lingual words with related semantics as linked words instead of only translations. This is justified by the fact that topic modeling focuses on what topics a word belongs to rather than accurate translations. For example, English word "album" and the Chinese translation of "song" should belong to the same topic in English and Chinese although they are not translations of each other. Accordingly, our CVL method can easily infer more linked words beyond translations in a dictionary.

Objective Function of Topic Alignment with Mutual Information Let N_{CVL} denote the number of all linked word pairs (positive pairs) found by the CVL method: $N_{\mathrm{CVL}} = \sum_{i=1}^{V_1+V_2} |\mathrm{CVL}(w_i)|$. We then sample uniformly from all linked word pairs as $p(w_i, w_j) = \frac{1}{N_{\mathrm{CVL}}}$ if $w_j \in \mathrm{CVL}(w_i)$ else 0. Given a positive pair (w_i, w_j), the negative samples of w_i in the set $\mathcal{B}$ are selected as the rest of words in the vocabulary set of w_j except $\mathrm{CVL}(w_i)$:

$$\mathcal{B} = \{w_j\} \cup (\mathcal{V}^{(\ell)} \setminus \mathrm{CVL}(w_i)) \quad (4)$$

where ℓ refers to the language of w_j, and $\mathcal{V}^{(\ell)}$ is the vocabulary set of language ℓ. Now we write the maximization of the lower bound (Eq. (2)) as minimizing $\mathcal{L}_{\mathrm{TAMI}}$:

$$\mathcal{L}_{\mathrm{TAMI}} = -\frac{1}{N_{\mathrm{CVL}}}\sum_{i=1}^{V_1+V_2}\sum_{w_j\in\mathrm{CVL}(w_i)}\log\frac{\exp(g(\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_j))}{\sum_{w_{j'}\in\mathcal{B}}\exp(g(\boldsymbol{\varphi}_i,\boldsymbol{\varphi}_{j'}))}. \quad (5)$$

Cross-lingual Topic Modeling with Mutual Information

In this section, we introduce InfoCTM by applying our proposed TAMI to the context of topic modeling through the generation of cross-lingual documents. Figure 3 illustrates the overall architecture of InfoCTM.

Generation of Cross-lingual Documents The generation process follows the framework of VAE (Kingma and Welling 2014) as previous monolingual neural topic models (Miao, Yu, and Blunsom 2016; Wu et al. 2020b; Wu, Li, and Miao 2021; Wu et al. 2022). We use document $\mathbf{x}^{(\ell_1)}$ in language ℓ_1 to describe the generation process. First,

Model	EC News		Amazon Review		Rakuten Amazon	
	CNPMI	TU	CNPMI	TU	CNPMI	TU
MCTA	0.025‡	0.489‡	0.028‡	0.319‡	0.021‡	0.272‡
MTAnchor	-0.013‡	0.192‡	0.028‡	0.323‡	-0.001‡	0.214‡
NMTM	0.031‡	0.784‡	0.042	0.732‡	0.009‡	0.679‡
InfoCTM	**0.048**	**0.913**	**0.043**	**0.923**	**0.034**	**0.870**

Table 2: Topic quality results of topic coherence (CNPMI) and diversity (TU). The best are in bold. The superscript ‡ means the improvements of InfoCTM is statistically significant at 0.05 level.

we specify the prior and variational distribution. Following Srivastava and Sutton (2017), we use a latent variable $\mathbf{r}^{(\ell_1)}$ with a logistic normal distribution as prior: $p(\mathbf{r}^{(\ell_1)}) = \mathcal{LN}(\boldsymbol{\mu}_0^{(\ell_1)}, \boldsymbol{\Sigma}_0^{(\ell_1)})$ where $\boldsymbol{\mu}_0^{(\ell_1)}$ and $\boldsymbol{\Sigma}_0^{(\ell_1)}$ are the mean and the diagonal covariance matrix. The variational distribution is modeled as $q_{\Theta_1}(\mathbf{r}^{(\ell_1)}|\mathbf{x}^{(\ell_1)}) = \mathcal{N}(\boldsymbol{\mu}^{(\ell_1)}, \boldsymbol{\Sigma}^{(\ell_1)})$ where $\boldsymbol{\mu}^{(\ell_1)}$ and $\boldsymbol{\Sigma}^{(\ell_1)}$ are the outputs of an encoder neural network with Θ_1 as parameters. By applying the reparameterization trick (Kingma and Welling 2014), we sample as $\mathbf{r}^{(\ell_1)} = \boldsymbol{\mu}^{(\ell_1)} + (\boldsymbol{\Sigma}^{(\ell_1)})^{1/2}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The doc-topic distribution $\boldsymbol{\theta}^{(\ell_1)}$ is modeled as $\boldsymbol{\theta}^{(\ell_1)} = \mathrm{softmax}(\mathbf{r}^{(\ell_1)})$.

To generate the document with $\boldsymbol{\theta}^{(\ell_1)}$, we model the topic-word distribution matrices $\boldsymbol{\beta}^{(\ell_1)} \in \mathbb{R}^{V_1 \times K}$ of language ℓ_1 and $\boldsymbol{\beta}^{(\ell_2)} \in \mathbb{R}^{V_2 \times K}$ of language ℓ_2 by the topic representations of words as:

$$\boldsymbol{\beta}^{(\ell_1)} = (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{V_1})^\top \tag{6}$$

$$\boldsymbol{\beta}^{(\ell_2)} = (\boldsymbol{\varphi}_{V_1+1}, \ldots, \boldsymbol{\varphi}_{V_1+V_2})^\top. \tag{7}$$

Then, we typically generate words in $\mathbf{x}^{(\ell_1)}$ by sampling from a multinomial distribution: $x \sim \mathrm{Mult}(\mathrm{softmax}(\boldsymbol{\beta}^{(\ell_1)}\boldsymbol{\theta}^{(\ell_1)}))$ (Miao, Yu, and Blunsom 2016). Similarly, the generation of document $\mathbf{x}^{(\ell_2)}$ in language ℓ_2 is formulated as $\mathrm{softmax}(\boldsymbol{\beta}^{(\ell_2)}\boldsymbol{\theta}^{(\ell_2)})$ with parameter Θ_2.

Objective Function for Generation of Topic Modeling Following the ELBO of VAE (Kingma and Welling 2014), we formulate the generation objective of topic modeling as:

$$\begin{aligned}\mathcal{L}_{\text{TM}}^{(\ell_1)}(\mathbf{x}^{(\ell_1)}) = &- (\mathbf{x}^{(\ell_1)})^\top \log(\mathrm{softmax}(\boldsymbol{\beta}^{(\ell_1)}\boldsymbol{\theta}^{(\ell_1)})) \\ &+ \mathrm{KL}\left[q_{\Theta_1}(\mathbf{r}^{(\ell_1)}|\mathbf{x}^{(\ell_1)}) \| p(\mathbf{r}^{(\ell_1)})\right].\end{aligned} \tag{8}$$

The first term measures the reconstruction error with $\mathbf{x}^{(\ell_1)}$ in the form of Bag-of-Words as previous work (Miao, Yu, and Blunsom 2016). The second term is the KL divergence between the prior and variational distribution. Similar to Eq. (8), we can easily write the objective function $\mathcal{L}_{\text{TM}}^{(\ell_2)}(\mathbf{x}^{(\ell_2)})$ for document $\mathbf{x}^{(\ell_2)}$.

Overall Objective Function for InfoCTM Letting $\mathcal{S}$ denote a set of cross-lingual documents, we write the overall objective function of InfoCTM with Eq. (5) and Eq. (8) as

$$\min_{\Theta_1, \Theta_2, \boldsymbol{\varphi}} \lambda_{\text{TAMI}}\mathcal{L}_{\text{TAMI}} + \sum_{(\mathbf{x}^{(\ell_1)}, \mathbf{x}^{(\ell_2)}) \in \mathcal{S}} \frac{1}{|\mathcal{S}|}(\mathcal{L}_{\text{TM}}^{(\ell_1)}(\mathbf{x}^{(\ell_1)}) + \mathcal{L}_{\text{TM}}^{(\ell_2)}(\mathbf{x}^{(\ell_2)}))$$

where λ_{TAMI} is a weight hyper-parameter. The $\mathcal{L}_{\text{TAMI}}$ objective works as a regularization of the generation objective of topic modeling, which aligns the topics across languages and meanwhile prevents degenerate topic representations.

Experiment

In this section, we conduct extensive experiments to show the effectiveness of our method.

Experiment Setup

Datasets and Dictionaries We use the following benchmark datasets in our experiments:

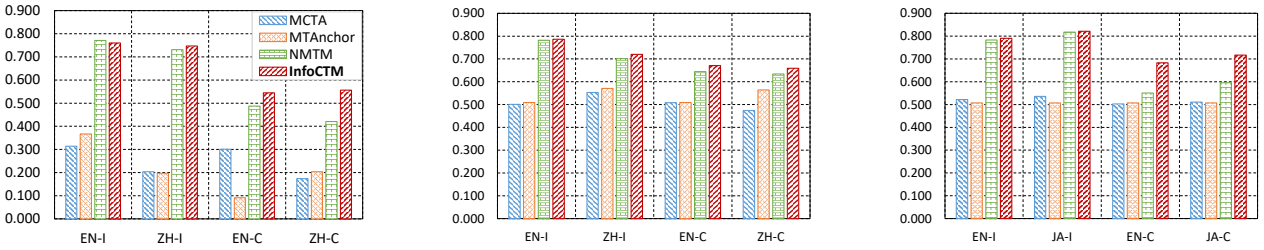
- **EC News** is a collection of English and Chinese news (Wu et al. 2020a) with 6 categories: business, education, entertainment, sports, tech, and fashion.
- **Amazon Review** includes English and Chinese reviews from the Amazon website where each review has a rating from one to five. We simplify it as a binary classification task by labeling reviews with ratings of five as "1" and the rest as "0" following Yuan, Van Durme, and Ying (2018).
- **Rakuten Amazon** contains Japanese reviews from Rakuten (a Japanese online shopping website, Zhang and LeCun 2017), and English reviews from Amazon (Yuan, Van Durme, and Ying 2018). Similarly, it is also simplified as a binary classification task according to the rating.

We employ the entries from MDBG [2] as the Chinese-English dictionary for EC News and Amazon Review, and we use the Japanese-English dictionary from MUSE [3] (Conneau et al. 2017) for Rakuten Amazon.

Baseline Models We compare our method with the following state-of-the-art baseline models: (i) **MCTA** (Shi et al. 2016), a probabilistic cross-lingual topic model that detects cultural differences. (ii) **MTAnchor** (Yuan, Van Durme, and Ying 2018), a multilingual topic model based on multilingual anchor words. (iii) **NMTM** (Wu et al. 2020a), a neural multilingual topic model which aligns topic representations by transforming them into the same vocabulary space. We do not consider recent studies (Bianchi et al. 2020; Mueller and Dredze 2021) because they do not discover aligned cross-lingual topics as required.

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

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



(a) EC News (b) Amazon Review (c) Rakuten Amazon

Figure 5: Document classification accuracy where "-I" means intra-lingual classification, and "-C" is cross-lingual classification. Involved languages are English (EN), Chinese (ZH) and Japanese (JA). The improvements of InfoCTM on cross-lingual classification (EN-C,ZH-C,JA-C) are statistically significant at 0.05 level.

Cross-lingual Topic Quality

Evaluation Metrics Following Wu et al. (2020a); Chang and Hwang (2021), we evaluate topic quality from two perspectives: (i) **Topic coherence** evaluates the coherence and alignment of cross-lingual topics. We use **CNPMI** (Cross-lingual NPMI, Hao, Boyd-Graber, and Paul 2018), a popular metric for cross-lingual topics based on NPMI (Chang et al. 2009; Newman et al. 2010). CNPMI measures the coherence between the words in each topic of different languages, *e.g.*, between words in English Topic#k and words in Chinese Topic#k. Higher CNPMI indicates topics are more coherent and well-aligned across languages. (ii) **Topic diversity** evaluates the difference between discovered topics to verify if they are repetitive. We employ Topic Uniqueness (TU) (Nan et al. 2019), which calculates the proportion of different words in the discovered topics. We report the average TU of different languages for each dataset. We select the top 15 related words of each topic for coherence and diversity evaluation.

Result Analysis Table 2 summaries the topic coherence (CNPMI) and diversity (TU) results under 50 topics. We observe that baseline models generally suffer from repetitive topics: their TU scores are quite low. As aforementioned, these repetitive topics are of low quality and can hinder further text analysis. In contrast, our InfoCTM consistently has much higher TU under all the settings. *E.g.*, InfoCTM achieves a TU score of 0.913 on EC News while the runner-up is only 0.784. This is because InfoCTM adopts our topic alignment with mutual information, which prevents degenerate topic representations, alleviates repetitive topics, and thus improves topic diversity. In addition, InfoCTM achieves the best CNPMI scores on all datasets as in Table 2. For example, InfoCTM has a CNPMI score of 0.034 on Rakuten Amazon, while the runner-up only has 0.021. Although on Amazon Review the CNPMI score of InfoCTM is only marginally larger than the runner-up, we note the TU of InfoCTM is much better (0.923 v.s. 0.732), and thus the overall topic quality of InfoCTM is higher. In summary, these results validate that InfoCTM can mitigate the repetitive topic issue and produce higher-quality cross-lingual topics than all the baselines. This advantage is crucial for further cross-lingual text analysis and applications.

Dict Size	Model	Topic Quality		Classification			
		CNPMI	TU	EN-I	ZH-I	EN-C	ZH-C
25%	NMTM	0.019‡	0.763‡	0.775	0.733	0.351‡	0.348‡
	w/o CVL	0.035	0.795‡	**0.778**	**0.763**	0.403‡	0.356‡
	InfoCTM	**0.036**	**0.895**	0.769	0.755	**0.472**	**0.448**
50%	NMTM	0.025‡	0.789‡	**0.775**	0.730	0.403‡	0.401‡
	w/o CVL	**0.041**	0.862‡	0.772	**0.753**	0.433‡	0.449‡
	InfoCTM	0.040	**0.905**	0.765	0.746	**0.490**	**0.520**
75%	NMTM	0.029‡	0.803‡	**0.776**	0.731	0.479‡	0.441‡
	w/o CVL	**0.045**	0.884‡	0.767	0.743	0.476‡	0.462‡
	InfoCTM	**0.045**	**0.909**	0.761	**0.748**	**0.519**	**0.537**
100%	NMTM	0.031‡	0.784‡	**0.771**	0.731	0.487‡	0.420‡
	w/o CVL	**0.050**	0.899	0.768	0.739	0.511	0.544
	InfoCTM	0.048	**0.913**	0.760	**0.747**	**0.545**	**0.556**

Table 3: Experiment with low-coverage dictionaries and ablation study. Here w/o CVL means InfoCTM without the cross-lingual vocabulary linking method and only uses the translation pairs from a dictionary as linked words. The superscript ‡ denotes that the improvements of InfoCTM are statistically significant at 0.05 level.

Intra-lingual and Cross-lingual Classification

As mentioned previously, doc-topic distributions of a cross-lingual topic model should be cross-lingually consistent and provide transferable features for cross-lingual tasks. To evaluate this performance, we train SVM classifiers with doc-topic distributions as features and compare their accuracy with F1 scores following Yuan, Van Durme, and Ying (2018). Specifically, we evaluate the classification performance from two perspectives. (i) **Intra-lingual** classification (**-I**): we train and test the classifier on the *same* language. (ii) **Cross-lingual** classification (**-C**): we train the classifier on one language and test it on another. For example, Amazon Review dataset includes English (EN) and Chinese (ZH) documents; "ZH-I" denotes the classifier is trained and tested both on Chinese, while "ZH-C" denotes the classifier is trained on English and tested on Chinese.

As shown in Figure 5, the intra-lingual classification accuracy (EN-I,ZH-I,JA-I) of InfoCTM is much higher than MCTA and MTAnchor, and is close to NMTM. This is reasonable since InfoCTM and NMTM both infer doc-topic distributions in the framework of VAE. Nevertheless, InfoCTM achieves clearly higher cross-lingual classification accuracy (EN-C,ZH-C,JA-C), and the improvements are statistically significant at 0.05 level. The reason lies in that InfoCTM uses our proposed topic alignment with mutual information method instead of the direct alignment of NMTM. This new method enhances topic alignment across languages and thus produces more consistent and transferable doc-topic distributions than NMTM. In a word, these results show that InfoCTM has better transferability for cross-lingual classification tasks.

Low-coverage Dictionary and Ablation Study

To evaluate the performance with low-coverage dictionaries, we experiment with different dictionary sizes following Hao and Paul (2018). Meanwhile, we conduct an ablation study on the proposed cross-lingual vocabulary linking (CVL) method. Let w/o CVL denote InfoCTM but without CVL and using the translation pairs from dictionaries as linked words only. Table 3 reports the topic quality and classification results under different dictionary sizes (25%, 50%, 75%, and 100%) on EC News. We only include NMTM in this study as NMTM outperforms all other baselines.

We have the following observations from Table 3: (**i**) InfoCTM can perform well with low-coverage dictionaries. Compared to NMTM, InfoCTM achieves better topic quality concerning CNPMI and TU. Similar to previous experiments, the intra-lingual accuracy (EN-I, ZH-I) of InfoCTM is close to NMTM, but its cross-lingual accuracy (EN-C, ZH-C) is obviously higher. We also see InfoCTM with 25% of the dictionary achieves close performance to NMTM with 100% of the dictionary. (**ii**) Our proposed CVL method effectively mitigates the low-coverage dictionary issue. InfoCTM and w/o CVL have similar CNPMI scores, but InfoCTM has increasingly higher TU and cross-lingual accuracy along with smaller dictionary sizes. These show our CVL method can improve the performance when only low-coverage dictionaries are available.

Case Study of Discovered Topics

To qualitatively evaluate the topic quality, we conduct a case study of discovered topics selected by querying keywords "soccer" and "exercise". They are shown in Table 4 (translations are in the brackets for easier understanding, and they are *not* the words of discovered topics). Recall that well-aligned topics should be semantically consistent across languages. For the topic "soccer" from EC News, NMTM produces repetitive topics with repeated words like "sports" and "episode". In contrast, InfoCTM only generates one relevant topic about soccer and the words are clearly coherent with the words "club", "milan", and "chelsea". For the topic "exercise" from Rakuten Amazon, InfoCTM aligns the topics well with relevant words in English and Japanese, *e.g.*, "yoga", "exercise" and "drinking". But NMTM wrongly aligns the topics with irrelevant and incoherent words.

	Top related words of topics				
			NMTM		
EN Topic#1:	sport	thrones	soccer	episode	bachelor
ZH Topic#1:	球队	球员	球迷	巴萨	穆里尼奥
translations:	*team*	*player*	*fans*	*abrcelona*	*mourinho*
EN Topic#2:	sport	thrones	episode	hes	wars
ZH Topic#2:	球队	球迷	球员	穆里尼奥	皇马
translations:	*team*	*fans*	*player*	*mourinho*	*real adrid*)
			InfoCTM		
EN Topic#1:	club	rent	abrcelona	milan	chelsea
ZH Topic#1:	转会	租借	米兰	俱乐部	切尔西
translations:	*transfer*	*rent*	*milan*	*club*	*chelsea*
			NMTM		
EN Topic#1:	learn	book	sweat	teach	exercise
Topic#1:	愛用	年	使い	助かり	シャンプ
translations:	*favorite*	*year*	*use*	*help*	*shampoo*
			InfoCTM		
EN Topic#1:	yoga	workout	exercise	drinking	drink
JP Topic#1:	ヨガ	飲ん	運動	飲み物	肌
translations:	*yoga*	*drinking*	*exercise*	*drinking*	*body*

Table 4: Top related words of discovered topics in each row. Repetitive words are underlined. *Words* in italics are the translations of the above Chinese or Japanese words.

Visualization of Latent Space

We use t-SNE (van der Maaten and Hinton 2008) to visualize the learned topic representations of the top related words of English and Chinese topics discovered by our InfoCTM from EC News. The Appendix shows the topics are well-aligned across languages, and the topic representations of words are well-grouped and separately scattered in the latent space. For example, English Topic#2 and Chinese Topic#2 are both about music, including the words "song", "album", and "sing". These words are close to each other while away from words of other topics. We notice Topic#2 about music and Topic#3 about the movie are closer to each other on the canvas as they are relatively more related. This qualitatively verifies our InfoCTM indeed properly aligns the topic representations and prevents degenerate topic representations.

Conclusion

In this paper, we propose InfoCTM to discover aligned latent topics of cross-lingual corpora. InfoCTM uses the novel topic alignment with mutual information method that avoids the repetitive topic issue and uses a new cross-lingual vocabulary linking method that alleviates the low-coverage issue. Experiments show that InfoCTM can consistently outperform baselines, producing higher-quality topics and showing better transferability for cross-lingual downstream tasks. Especially, InfoCTM can perform well under low-coverage dictionaries, making it applicable for more scenarios like low-resource languages.

Acknowledgements

We want to thank all anonymous reviewers for their helpful comments. This research is supported by AI Singapore technology grant AISG2-TC-2022-005.

References

Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; and Saunshi, N. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *36th International Conference on Machine Learning, ICML 2019*, 9904–9923. International Machine Learning Society (IMLS).

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Bianchi, F.; Terragni, S.; Hovy, D.; Nozza, D.; and Fersini, E. 2020. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.

Blei, D. M.; and Lafferty, J. D. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, 113–120.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan): 993–1022.

Boyd-Graber, J.; and Blei, D. 2012. Multilingual topic models for unaligned text. *arXiv preprint arXiv:1205.2657*.

Chang, C.-H.; and Hwang, S.-Y. 2021. A word embedding-based approach to cross-lingual topic modeling. *Knowledge and Information Systems*, 63(6): 1529–1555.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J. L.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, 288–296.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chi, Z.; Dong, L.; Wei, F.; Yang, N.; Singhal, S.; Wang, W.; Song, X.; Mao, X.-L.; Huang, H.; and Zhou, M. 2020. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2017. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, X.; Luu, A. T.; Lin, M.; Yan, S.; and Zhang, H. 2021. How Should Pre-Trained Language Models Be Fine-Tuned Towards Adversarial Robustness? *Advances in Neural Information Processing Systems*, 34: 4356–4369.

Gutiérrez, E. D.; Shutova, E.; Lichtenstein, P.; de Melo, G.; and Gilardi, L. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4: 47–60.

Hao, S.; Boyd-Graber, J. L.; and Paul, M. J. 2018. Lessons from the bible on modern topics: adapting topic model evaluation to multilingual and low-resource settings. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT*, 1–6.

Hao, S.; and Paul, M. 2018. Learning multilingual topics from incomparable corpora. In *Proceedings of the 27th international conference on computational linguistics*, 2595–2609.

Hao, S.; and Paul, M. J. 2020. An empirical study on crosslingual transfer in probabilistic topic models. *Computational Linguistics*, 46(1): 95–134.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Jagarlamudi, J.; and Daumé, H. 2010. Extracting multilingual topics from unaligned comparable corpora. In *European Conference on Information Retrieval*, 444–456. Springer.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *The International Conference on Learning Representations (ICLR)*.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, 79–86.

Kong, L.; de Masson d'Autume, C.; Yu, L.; Ling, W.; Dai, Z.; and Yogatama, D. 2020. A Mutual Information Maximization Perspective of Language Representation Learning. In *International Conference on Learning Representations*.

Lind, F.; Eberl, J.-M.; Galyga, S.; Heidenreich, T.; Boomgaarden, H. G.; Jiménez, B. H.; and Berganza, R. 2019. A bridge over the language gap: Topic modelling for text analyses across languages for country comparative research. *University of Vienna: Working Paper of the REMINDER-Project*.

Logeswaran, L.; and Lee, H. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *International conference on machine learning*, 1727–1736.

Mikolov, T.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Mimno, D.; Wallach, H.; Naradowsky, J.; Smith, D. A.; and McCallum, A. 2009. Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, 880–889.

Mueller, A.; and Dredze, M. 2021. Fine-tuning Encoders for Improved Monolingual and Zero-shot Polylingual Neural Topic Modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3054–3068.

Nan, F.; Ding, R.; Nallapati, R.; and Xiang, B. 2019. Topic Modeling with Wasserstein Autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6345–6381. Florence, Italy: Association for Computational Linguistics.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. Association for Computational Linguistics. ISBN 1932432655.

Nguyen, T.; Wu, X.; Luu, A.-T.; Nguyen, C.-D.; Hai, Z.; and Bing, L. 2022. Adaptive Contrastive Learning on Multimodal Transformer for Review Helpfulness Predictions. *arXiv preprint arXiv:2211.03524*.

Ni, X.; Sun, J.-T.; Hu, J.; and Chen, Z. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th international conference on World wide web*, 1155–1156.

Shi, B.; Lam, W.; Bing, L.; and Xu, Y. 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)*, 676–685.

Srivastava, A.; and Sutton, C. 2017. Autoencoding variational inference for topic models. In *ICLR*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *European conference on computer vision*, 776–794. Springer.

Van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, arXiv–1807.

van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Vulić, I.; De Smet, W.; and Moens, M.-F. 2013. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 16(3): 331–368.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Wu, X.; Dong, X.; Nguyen, T. T.; and Luu, A. T. 2022. Neural Topic Modeling with Embedding Clustering Regularization. Forthcoming.

Wu, X.; and Li, C. 2019. Short Text Topic Modeling with Flexible Word Patterns. In *International Joint Conference on Neural Networks*.

Wu, X.; Li, C.; and Miao, Y. 2021. Discovering Topics in Long-tailed Corpora with Causal Intervention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 175–185. Online: Association for Computational Linguistics.

Wu, X.; Li, C.; Zhu, Y.; and Miao, Y. 2020a. Learning Multilingual Topics with Neural Variational Inference. In *International Conference on Natural Language Processing and Chinese Computing*.

Wu, X.; Li, C.; Zhu, Y.; and Miao, Y. 2020b. Short Text Topic Modeling with Topic Distribution Quantization and Negative Sampling Decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1772–1782. Online.

Wu, X.; Luu, A. T.; and Dong, X. 2022. Mitigating Data Sparsity for Short Text Topic Modeling by Topic-Semantic Contrastive Learning. arXiv:2211.12878.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xu, K.; Lu, X.; Li, Y.-f.; Wu, T.; Qi, G.; Ye, N.; Wang, D.; and Zhou, Z. 2022. Neural Topic Modeling with Deep Mutual Information Estimation. *arXiv preprint arXiv:2203.06298*.

Yang, W.; Boyd-Graber, J.; and Resnik, P. 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)*, 1243–1248.

Yuan, M.; Van Durme, B.; and Ying, J. L. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. *Advances in neural information processing systems*, 31.

Zhang, X.; and LeCun, Y. 2017. Which encoding is the best for text classification in Chinese, English, Japanese and Korean? *arXiv preprint arXiv:1708.02657*.