Learning Multilingual Topics with Neural Variational Inference

Xiaobao Wu[1,⋆], Chunping Li[1], Yan Zhu[2], and Yishu Miao[3]

[1] School of Software, Tsinghua University, Beijing, China
wxb18@mails.tsinghua.edu.cn, cli@mail.tsinghua.edu.cn
[2] School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
yzhu@swjtu.edu.cn
[3] Department of Computing, Imperial College London, London, UK
y.miao20@imperial.ac.uk

Abstract. Multilingual topic models are one of the most popular methods for revealing common latent semantics of cross-lingual documents. However, traditional approximation methods adopted by existing probabilistic models sometimes do not effectively lead to high-quality multilingual topics. Besides, as the generative processes of these models become more expressive, the difficulty of performing fast and accurate inference methods over parameters grows. In this paper, to address these issues, we propose a new multilingual topic model that permits training by backpropagation in the framework of neural variational inference. We propose to infer topic distributions via a shared inference network to capture common word semantics and an incorporating module to incorporate the topic-word distribution from another language through a novel transformation method. Thus, the networks of cross-lingual corpora are coupled together. With jointly training the coupled networks, our model can infer more interpretable multilingual topics and discriminative topic distributions. Experimental results on real-world datasets show the superiority of our model both in terms of topic quality and text classification performance.

Keywords: VAE · Multilingual Documents · Topic Modeling

1 Introduction

Nowadays, our world is swamped by affluent information composed of diverse languages, such as news, web pages, and microblogs. Originating from PLSA (Probabilistic Latent Semantic Analysis) [9] and LDA (Latent Dirichlet Allocation) [3], topic models have been adapted to these cross-lingual document collections for topic alignment, which are called multilingual topic models [17]. Topic alignment refers to exposing the corresponding words under the same topics in multiple languages. Moreover, revealing the latent relationship between

[⋆] Corresponding author.

cultures, such as wording and phrasing, is another ability of multilingual topic models.

Multilingual topic models ordinarily depend on external information to fill the gap between languages, such as word alignment [29] or dictionaries [11]. Recent multilingual topic models include MCTA (Multilingual Cultural-common Topic Analysis) [21] and MTAnchor (Multilingual Topic Anchors) [28] which adopt conventional approximation methods to infer latent variables, such as Gibbs Sampling [23] and Variational Inference [2]. However, they tend to generate repetitive or trivial topics that are worthless in the following applications. To be more specific, repetitive topics include repeated words of other topics and trivial topics are composed of irrelevant or uninformative words. It is a vital issue since these topics are worthless in the following applications. Besides, it is time-consuming and inconvenient to re-derive the corresponding inference methods for various assumptions of generative processes. Recently, the neural variational inference has been proposed to effectively and efficiently approximate complex and intractable distributions through deep neural networks without heavy derivation [13,20]. Unfortunately, directly applying neural variational inference framework for discovering multilingual topics is infeasible, since it only models the generation process of monolingual corpus and cannot learn common topics from cross-lingual corpora [15,22].

Therefore, to more effectively generate high-quality multilingual topics through neural variational inference, we propose the Neural Multilingual Topic Model (NMTM). First, to map the words in different languages into the same embedding space, we infer the topic distributions of cross-lingual texts with a shared parameter inference network, which stimulates the network to better capture common word semantics. Moreover, to bridge the gap between topic semantics, we proposed an incorporating module that focuses on incorporating topic semantics of another language with a new transformation method. The transformation method adopts a matrix to map the topic-word distribution from a language to another. Thus, the generative networks of cross-lingual texts are coupled together based on these approaches. Through jointly training the networks by backpropagation, our model can effectively discover common topics from multilingual corpora. The code is available at https://github.com/bobxwu/NMTM.

The main contributions of this paper can be concluded as follows:

1. We propose a neural variational inference model for multilingual topic modeling with incorporating topic-word distributions of other languages and jointly training instead of conventional probabilistic models;
2. We conduct extensive experiments on real-world datasets and demonstrate that our model outperforms the baselines concerning topic quality and classification performance.

2 Related Work

Conventional Topic Models Conventional monolingual topic models [9,3] are extended for various scenes, such as online topic modeling [8] and short text topic modeling [27,25]. One important extension is multilingual topic modeling to discover common latent topics from cross-lingual documents. The Polylingual Topic Model [17] takes the assumption that cross-lingual documents are topically aligned and infers the same topic distributions of them. Joint LDA [11] proposes "concepts" to connect words in different languages with a bilingual dictionary. Polylingual Tree-based Topic Model [10] adopts tree priors to incorporate word relationships and document alignment information. MCTA [21] is another generative multilingual topic model using dictionary entries to capture common topics across culture from the news. MTAnchor [28] firstly obtains anchor words [14] through a multilingual anchoring algorithm and then infers the topic distributions with fixed topic-word distributions. These conventional models usually adopt Gibbs Sampling or Variational Inference methods to approximate the posterior.

Neural Topic Models In another vein, several neural topic models are proposed due to the success of neural variational inference [13,20]. In the framework of Variational Auto-Encoder (VAE), Neural Variational Document Model (NVDM) [16] infers topic distributions and generates texts through MLPs. The objective function of NVDM includes the reconstruction error between generated texts and input texts, and the KL divergence between the prior distribution and the inferred variational distribution. Product-Expert LDA (ProdLDA) [22] is a black-box neural variational inference framework for monolingual topic modeling using the logistic normal distribution. More other neural topic models are also proposed, like for short texts [26] and supervised versions [24]. Borrowing the idea of neural monolingual topic modeling, we propose our new neural multilingual model with a novel incorporating module to bridge the gap of topic semantics.

3 Preliminary

We follow the basic assumptions of LDA [3], one of the most widely used topic models, to design our model. In its formulation, a topic is defined as the distribution of words, and words in a text are drawn from the mixtures of Multinomial distributions with a Dirichlet distribution as the prior. The latent variable z denotes the topic assignment of word x_i and $\boldsymbol{\theta}$ in LDA is the topic distribution of a text. According to the generation procedure of LDA, the marginal likelihood of text $\boldsymbol{x}$ is

$$p(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \int_{\boldsymbol{\theta}} \left(\prod_{i=1}^{N} \sum_{z=1}^{K} p\left(x_i|z,\boldsymbol{\beta}\right) p\left(z|\boldsymbol{\theta}\right) \right) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} \tag{1}$$

where N refers to the number of words in text $\boldsymbol{x}$, K is the topic number and $\boldsymbol{\alpha}$ is the hyperparameter of Dirichlet distribution. Moreover, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$ is the topic-word distribution matrix composed of word probability vectors of all topics.

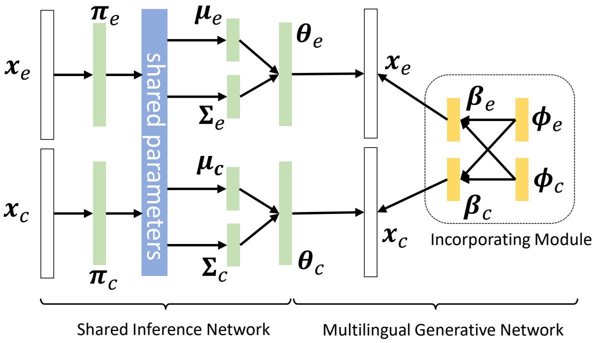

Fig. 1: Network structure of NMTM.

Differently, with the help of neural variational inference, several neural topic models [15,22] have been proposed. These models adopt a simplification that the discrete latent variable z is integrated out in the marginal likelihood as

$$p(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\beta})=\int_{\boldsymbol{\theta}}\left(\prod_{i=1}^{N}p\left(x_i|\boldsymbol{\theta},\boldsymbol{\beta}\right)\right)p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta} \tag{2}$$

Thus, this adaption allows to infer topics and generate texts through neural networks which can be directly updated by gradient backpropagation. Based on these preceding monolingual neural topic models, we present our new model for multinomial topic modeling.

4 Proposed Method

In this section, as depicted in Figure 1, we detail our proposed Neural Multilingual Topic Model (NMTM) aligning topics of English and Chinese corpora.

4.1 Shared Inference Network

We first infer the topic distributions through the shared inference network. The representations of an English text $\boldsymbol{x}_e$ and a Chinese text $\boldsymbol{x}_c$ are in the form of bag-of-words that ignores word sequences and is commonly adopted by previous topic models [3]. Then, we obtain the intermediate representation as

$$\boldsymbol{\pi}_e = f_{W_e}(\boldsymbol{x}_e) \tag{3}$$

where f_{W_e} is a fully connected network with weight W_e as the word embedding matrix of English. Simialry, we have $\boldsymbol{\pi}_c = f_{W_c}(\boldsymbol{x}_c)$ for Chinese texts. Next, to effectively capture the common word semantics of different languages, we make the inference network share the same parameters. In detail, we define two

MLPs to infer the mean and variance of the variational distribution, f_μ and f_Σ. Accordingly, the mean and variance of $\boldsymbol{x}_e$ are

$$\boldsymbol{\mu}_e = f_\mu(\boldsymbol{\pi}_e) \tag{4}$$

$$\boldsymbol{\Sigma}_e = diag(f_\Sigma(\boldsymbol{\pi}_e)) \tag{5}$$

In the same way, the mean and variance of $\boldsymbol{x}_c$ are caculated through the same inference network, $\boldsymbol{\mu}_c = f_\mu(\boldsymbol{\pi}_c)$ and $\boldsymbol{\Sigma}_c = diag(f_\Sigma(\boldsymbol{\pi}_c))$. Then, according to the reparameterization trick for variance reduction [13], $\boldsymbol{\epsilon}$ is sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ to generate the latent representation as

$$\boldsymbol{\theta}_e = \sigma(\boldsymbol{\mu}_e + \boldsymbol{\Sigma}_e^{1/2}\boldsymbol{\epsilon}) \tag{6}$$

$$\boldsymbol{\theta}_c = \sigma(\boldsymbol{\mu}_c + \boldsymbol{\Sigma}_c^{1/2}\boldsymbol{\epsilon}) \tag{7}$$

where $\sigma(\cdot)$ means softmax function for normalization. Thus, with shared parameters, inference work can capture the common word semantics in different languages. Besides, to better approximate the Dirichlet priori distribution in topic modeling, the priori distribution is modeled as a logistic normal distribution $\mathcal{LN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ [7,22].

4.2 Multilingual Generation Network

After the inference network, topic distributions of texts are fed to the multilingual generation network for reconstruction.

Incroporating Module for Topic-word Distributions For the reconstruction of monolingual neural topic models[16,22], topic distributions typically are multiplied with the topic-word distribution matrix which is randomly initialized then optimized during training processes [16,22]. Differently, in our multilingual generation network, we propose an incorporating module to incorporate the topic-word distributions from another language for topic alignment; thus, a transformation mechanism is necessary to connect the topic-word distribution matrices of two languages. Shi et al. [21] has proposed a transformation method for probabilistic models according to the word frequencies in the target language. Nevertheless, this approach merely considers the word pairs in the existing dictionary; therefore, it encounters the sparsity problem that hinders the optimization of the topic-word distribution matrix in neural networks. To this end, we propose a new transformation approach for the incorporating module. We define $\boldsymbol{M}$ as the cross-lingual transformation matrix where M_{ij} denotes the mapping probability from x_i in the source language to x_j in the target, computed as

$$M_{ij} = \frac{\mathbb{I}(x_j)n(x_i) + 1}{|T(x_i)|\, n(x_i) + V} \tag{8}$$

where $n(x_i)$ is the number of occurrences of x_i in the corresponding corpus and $T(x_i)$ is the set of translations of x_i found in the bilingual dictionary. $|T(x_i)|$ is the size of $T(x_i)$ and V is the vocabulary size of the target language. $\mathbb{I}(x_j)$ is an indicator function which equals to 1 if x_j is in $T(x_i)$ and 0 otherwise. We can see that our mapping method transfers x_i to its translation x_j with the same probability according to its frequency. In addition, inspired by the design of language modeling, we also invoke the add-one smoothing (Laplace smoothing) [6] for each word in the target language to avoid sparsity, which makes the neural network more flexible to refine the probable inappropriate or missing entries in the provided dictionary. We demonstrate its necessity in Section 5.4.

Then, assuming $\boldsymbol{\phi}_e$, $\boldsymbol{\phi}_c$ are the original topic-word distribution matrices of English and Chinese respectively, with the mapping matrix from Chinese to English $\boldsymbol{M}_{c\rightarrow e}$, the incorporated topic-word distribution matrix of English $\boldsymbol{\beta}_e$ can be written as

$$\boldsymbol{\beta}_e = \lambda \boldsymbol{\phi}_c^T \boldsymbol{M}_{c\rightarrow e} + (1-\lambda)\boldsymbol{\phi}_e \tag{9}$$

where λ is the hyper-parameter balancing the weights. Similarly, the topic-word distribution matrix of Chinese texts is

$$\boldsymbol{\beta}_c = \lambda \boldsymbol{\phi}_e^T \boldsymbol{M}_{e\rightarrow c} + (1-\lambda)\boldsymbol{\phi}_c \tag{10}$$

where $\boldsymbol{M}_{e\rightarrow c}$ means the transformation matrix from English to Chinese.

As explained in Srivastava and Sutton [22], it is inappropriate to use the normalization method on the topic-word distribution matrix since it leads to repetitive topics. Therefore, according to Equation (2), with the incorporated topic-word distribution matrices $\boldsymbol{\beta}_e$ and $\boldsymbol{\beta}_c$, we model the generation of English words as $x_e \sim \text{Mult}\left(\sigma(\boldsymbol{\beta}_e\boldsymbol{\theta}_e)\right)$ and Chinese words as $x_c \sim \text{Mult}\left(\sigma(\boldsymbol{\beta}_c\boldsymbol{\theta}_c)\right)$.

Multilingual Objective Function According to the above multilingual generation network, the objective function of $\boldsymbol{x}_e$ can be written as

$$\begin{aligned}\mathcal{L}_e(\boldsymbol{x}_e) = {} & KL\left(q(\boldsymbol{\theta}_e|\boldsymbol{x}_e)\|p(\boldsymbol{\theta}_e)\right) \\ & - \mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\boldsymbol{I})}\left[\boldsymbol{x}_e^T \log\sigma\left(\lambda\phi_c^T \boldsymbol{M}_{c\rightarrow e}\boldsymbol{\theta}_e + (1-\lambda)\boldsymbol{\phi}_e\boldsymbol{\theta}_e\right)\right]\end{aligned} \tag{11}$$

where the first term is the KL divergence of variational and prior distribution acting as a regularizer, and the second term is the reconstruction error incorporating the topic-word distributions from the Chinese corpus. Similarly, we can have the objective function of a Chinese text $\boldsymbol{x}_c$ as

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{x}_c) = {} & KL\left(q(\boldsymbol{\theta}_c|\boldsymbol{x}_c)\|p(\boldsymbol{\theta}_c)\right) \\ & - \mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\boldsymbol{I})}\left[\boldsymbol{x}_c^T \log\sigma\left(\lambda\phi_e^T \boldsymbol{M}_{e\rightarrow c}\boldsymbol{\theta}_c + (1-\lambda)\boldsymbol{\phi}_c\boldsymbol{\theta}_c\right)\right].\end{aligned} \tag{12}$$

Then, the overall multilingual objective function of both corpora is

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathcal{L}_e(\boldsymbol{x}_e^{(i)}) + \frac{1}{N_c}\sum_{i=1}^{N_c}\mathcal{L}_c(\boldsymbol{x}_c^{(i)}) \tag{13}$$

where $\boldsymbol{\Theta}$ denotes the set of all model parameters; N_e and N_c are the sizes of English and Chinese corpus. By jointly optimizing the networks of English and Chinese on the multilingual objective function, our model can infer the multilingual topics and topic distributions.

Table 1: Statistics of datasets.

Dataset	Language	Number of docs	Average length	Vocabulary size	Number of labels
Amazon Review	English	25,000	30.75	5,000	2
	Chinese	25,000	42.96	5,000	
EC-News	English	46,870	11.99	5,000	6
	Chinese	50,000	10.72	5,000	

5 Experiment

5.1 Experiment Setup

Two real-world datasets are adopted in our experiments: (1) **Amazon Review** includes English and Chinese reviews from the Amazon website. Each review has a rating ranging from one to five and we simplify it as a binary classification task by labeling reviews with ratings of five as "1" and the rest as "0" following Yuan et al. [28]. (2) **EC-News** (English and Chinese News) is collected from English news category dataset [1] and Chinese News [1]. Six labels are included in this dataset: business, education, entertainment, sports, tech, and fashion. For preprocessing these datasets, we conduct the following steps: (1) convert English characters to lower and tokenize Chinese texts with jieba [2] ; (2) remove illegal characters and stopwords; (3) remove English words with length less than 2; (4) retain the top 5000 frequent words. After preprocessing, the statistics of datasets are reported in Table 1. For Chinese-English dictionary, we use entries from MDBG [3] .

As to baseline models, we consider recently proposed MCTA [4] [21] and MTAnchor [5] [28]. For our model NMTM, we adopt Adam [12] for parameter optimization. We mainly evaluate these models from two aspects, multilingual topic quality and classification performance for each language: English (**EN**) and Chinese (**CN**).

5.2 Multilingual Topic Quality

Topic Quality Metric As mentioned in Section 1, to evaluate the performance of alleviating repetitive and trivial topics, topic coherence and diversity are two central aspects of multilingual topic quality. Topic coherence indicates the words in a topic should be as coherent as possible. We use the widely-used metric called NPMI (Normalized Pointwise Mutual Information) [4,5] which assumes coherent words should co-occur within a certain distance for evaluation. Given the top

[1] https://www.kaggle.com/rmisra/news-category-dataset/data

[2] https://github.com/fxsjy/jieba

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

[4] https://github.com/shibei00/Cross-Lingual-Topic-Model

[5] https://github.com/forest-snow/mtanchor_demo

Table 2: Topic quality performance including topic coherence (NPMI) and unique scores (TU). "-I" means using corresponding training set as the reference corpus and "-E" means using external reference corpus.

Dataset	Model	K=20						K=50					
		NPMI				TU		NPMI				TU	
		EN-I	CN-I	EN-E	CN-E	EN	CN	EN-I	CN-I	EN-E	CN-E	EN	CN
Amazon Review	MCTA	0.027	0.025	0.023	0.042	0.410	0.460	0.030	0.028	0.021	0.045	0.293	0.345
	MTAnchor	0.053	0.014	0.036	0.036	0.247	0.380	0.041	0.035	0.035	0.037	0.283	0.363
	NMTM	**0.149**	**0.202**	**0.090**	**0.099**	**0.913**	**0.997**	**0.122**	**0.152**	**0.074**	**0.079**	**0.664**	**0.800**
EC-News	MCTA	0.081	0.050	0.027	0.030	0.570	0.693	0.087	0.055	0.030	0.036	0.448	0.529
	MTAnchor	0.057	0.063	0.021	0.025	0.363	0.450	0.046	0.069	0.016	0.019	0.172	0.211
	NMTM	**0.177**	**0.253**	**0.079**	**0.156**	**0.973**	**0.993**	**0.166**	**0.206**	**0.073**	**0.113**	**0.781**	**0.787**

Table 3: Multilingual topics examples. Bold words are the translations.

Dataset	Model	Topic
Amazon Review	MCTA	cd album music songs like song great **画册 (picture album) 光碟 (cd) 音乐 (music)** 伴奏 编曲 乐器
	MTAnchor	movie like love album music story cd 听 歌 张 **爱 (love) 小说 (novel) 故事 (story) 音乐 (music)**
	NMTM	song sing album lyrics singer tracks band **专辑 (album) 歌曲 (song)** 唱片 **乐队 (band) 唱 (sing) 歌手 (singer) 歌词 (lyrics)**
EC-News	MCTA	game video football player players national sports 奥运 奥运会 **网游 (online game)** 热身赛 亚运会 盛大 **游戏 (game) 全世界 (national)**
	MTAnchor	lebron new franco corden first one james time says like video film nba **詹姆斯 (james)** 比赛 姚明 **爱 (like)** 球 事情 分钟 表现
	NMTM	sport match tournament ncaa championship cup olympic soccer 主场 击败 **比赛 (match) 联赛 (tournament) 冠军 (champoinship)** 客场 决赛 局

T topic words $(x_1, x_2, \ldots, x_T)$ ordered by the probability, the NPMI score is caculated as

$$\text{NPMI}(x_i, x_j) = \frac{\log \frac{p(x_i, x_j)+\epsilon}{p(x_i)p(x_j)}}{-\log(p(x_i, x_j)+\epsilon)} \tag{14}$$

where $p(x_i)$ is the probability of x_i, $p(x_i, x_j)$ the coocurrance probability of x_i, x_j within a window in the reference corpus and ϵ is used to avoid zero. We adopt the public tool [6] provided by [5] for evaluation. To adequately evaluate topic coherence, we consider both intrinsic and extrinsic scores. For intrinsic score (**I**), the corresponding training set is employed as the reference corpus in NPMI and for extrinsic score (**E**), we adopt external Wikipedia articles. Apart from topic coherence, topic diversity refers to that topics are supposed to be distinguished from each other. We utilize TU (Topic Unique) [18] for assessment. Given the top T topic words, TU is defined as $TU = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{cnt(x_i)}$ where $cnt(x_i)$ is the occurrence count of x_i in the top T words of all topics. Therefore, higher TU scores indicate topics are more diverse since fewer words are repeated across other topics. For both NPMI and TU, T is set to 15 in our experiments.

[6] https://github.com/jhlau/topic_interpretability

Table 4: Intra-lingual and cross-lingual classification performance. "-I" means intra-lingual classification and "-C" means cross-lingual classification.

Dataset	Model	K=20				K=50			
		EN-I	CN-I	EN-C	CN-C	EN-I	CN-I	EN-C	CN-C
Amazon Review	MCTA	0.502	0.547	0.501	0.471	0.501	0.553	0.508	0.474
	MTAnchor	0.589	0.636	0.517	0.582	0.509	0.571	0.509	0.564
	NMTM	**0.683**	**0.657**	**0.610**	**0.591**	**0.634**	**0.593**	**0.513**	**0.588**
EC-News	MCTA	0.295	0.197	0.290	0.176	0.314	0.204	0.301	0.174
	MTAnchor	0.428	0.432	0.204	0.198	0.367	0.197	0.091	0.204
	NMTM	**0.575**	**0.556**	**0.344**	**0.280**	**0.437**	**0.519**	**0.349**	**0.211**

Table 5: Ablation study.

Model	Amazon Review				EC-News			
	EN-I	CN-I	EN-C	CN-C	EN-I	CN-I	EN-C	CN-C
NMTM	0.683	0.657	0.610	0.591	0.575	0.556	0.344	0.280
w/o add-one smoothing	0.694	0.659	0.570	0.530	0.570	0.561	0.299	0.251
w/o incorporating module	0.712	0.655	0.511	0.488	0.577	0.549	0.231	0.234

Result Analysis Table 2 reports the topic coherence and diversity performance under the topic number $K = 20$ and $K = 50$. We can observe that in terms of topic coherence, NMTM achieves higher NPMI scores than baseline models regardless of intrinsic or extrinsic scores. Significantly, NMTM surpasses MCTA and MTAnchor in TU by a large margin. It indicates the topics generated by NMTM are more diverse, while MCTA and MTAnchor yield several repetitive ones. To further illustrate the topic quality performance, Table 3 shows the learned multilingual topics of different models. We can observe MCTA and MTAnchor output some incoherent topic words. More precisely, the topic "sport" of MCTA is composed of the "online game" and in the words of topic "songs", MTAnchor generates some irrelevant words like "novel" and "story". On the contrary, the topics yielded by NMTM include more corresponding translation words and are more coherent, such as "band", "lyrics" and "tournament". With the above discussion, we can observe that NMTM can generate higher quality topics from multilingual corpora.

5.3 Intra-lingual and Cross-lingual Classification with Topic Distributions

Besides, we utilize the topic distributions outputted by models as low-dimensional features to train SVM classifiers [19] for text classification. Intra-lingual (**I**) and Cross-lingual (**C**) classifications are both considered as described in Yuan et al. [28]. Intra-lingual evaluation trains and tests the classifiers on the same language while the cross-lingual evaluation trains the classifiers on one language and tests on another to test the ability to generalization ability of the model. Classification $F1$ results under different topic numbers are reported in Table 4. We can

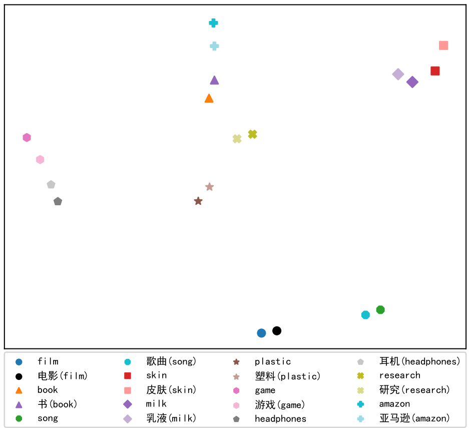

Fig. 2: t-SNE visualization of learned word embeddings.

see that NMTM performs better than MCTA and MTAnchor in terms of both intra-lingual and cross-lingual accuracy. It reveals NMTM can infer more discriminative topic distributions and the correlated cross-lingual texts are better modeled. Thus, NMTM has a better generalization ability.

5.4 Ablation Study

We also conduct an ablation study to evaluate the effectiveness of the proposed incorporating module in NMTM. Table 5 reports the results of NMTM without add-one smoothing and without incorporating module. We can first observe that the intra-lingual classification performance of these three variants is generally close to each other. However, if without incorporating module, the method degrades into two monolingual topic models and we can see cross-lingual classification results are lower, which indicates that the generalization ability is limited since no topic semantics are bridged. Besides, the cross-lingual classification performance is also impaired if add-one smoothing is removed. Therefore, we can conclude our incorporating module is effective for multilingual topic modeling.

5.5 Visualization of Word Embeddings

Figure 2 shows the t-SNE visualization of the embeddings of words in English and Chinese (W_e and W_c in Section 4.1) learned by NMTM. We can observe that embeddings of English-Chinese word pairs and relevant words are clearly close to each other. For example, the point of "film" is close to its translation and the point of "skin" is near to "milk"(a cosmetic here). Besides, the points of "game" and "headphones" are clustered together and the points of "book" are close to "amazon". Thus, with coupled networks, NMTM can learn to capture the common word semantics in various languages to improve the multilingual topic modeling performance.

6 Conclusion

In this paper, to have a better generation ability to produce high-quality multilingual topics, we propose the Neural Multilingual Topic Model (NMTM). Based on the framework of neural variational inference, NMTM maps words in different languages into the same embedding space and incorporates cross-lingual topic semantics via the incorporating module to discover multilingual topics by jointly training the coupled networks. Experimental results show that our model outperforms baselines by a large margin and especially, our approach can retain topic diversity with larger topic numbers.

Acknowledgement

This work is supported by China NSFC under Grant 61672309 and Sichuan Science and Technology Program under Grant 2019YFSY0032.

References

1. Bai, Y., Tao, J., Yi, J., Wen, Z., Fan, C.: Clmad: A chinese language model adaptation dataset. In: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 275–279. IEEE (2018)
2. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American statistical Association **112**(518), 859–877 (2017)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
4. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL pp. 31–40 (2009)
5. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. pp. 288–296 (2009)
6. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech & Language **13**(4), 359–394 (1999)
7. Hennig, P., Stern, D., Herbrich, R., Graepel, T.: Kernel topic models. In: Artificial Intelligence and Statistics. pp. 511–519 (2012)
8. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent dirichlet allocation. In: advances in neural information processing systems. pp. 856–864 (2010)
9. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. pp. 289–296. Morgan Kaufmann Publishers Inc. (1999)
10. Hu, Y., Zhai, K., Eidelman, V., Boyd-Graber, J.L.: Polylingual tree-based topic models for translation domain adaptation. In: ACL (2014)
11. Jagarlamudi, J., Daumé, H.: Extracting multilingual topics from unaligned comparable corpora. In: European Conference on Information Retrieval. pp. 444–456. Springer (2010)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: The International Conference on Learning Representations (ICLR) (2014)
14. Lund, J., Cook, C., Seppi, K., Boyd-Graber, J.: Tandem anchoring: A multiword anchor approach for interactive topic modeling. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 896–905 (2017)
15. Miao, Y., Grefenstette, E., Blunsom, P.: Discovering discrete latent topics with neural variational inference. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 2410–2419. JMLR. org (2017)
16. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: International conference on machine learning. pp. 1727–1736 (2016)
17. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. pp. 880–889. Association for Computational Linguistics (2009)
18. Nan, F., Ding, R., Nallapati, R., Xiang, B.: Topic Modeling with Wasserstein Autoencoders. arXiv preprint arXiv:1907.12374 (2019)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
20. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In Proceedings ofthe 31th International Conference on Machine Learning (2014)
21. Shi, B., Lam, W., Bing, L., Xu, Y.: 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). pp. 676–685 (2016)
22. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. In: ICLR (2017)
23. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of latent semantic analysis **427**(7), 424–440 (2007)
24. Wang, X., Yang, Y.: Neural topic model with attention for supervised learning. In: International Conference on Artificial Intelligence and Statistics. pp. 1147–1156 (2020)
25. Wu, X., Li, C.: Short Text Topic Modeling with Flexible Word Patterns. In: International Joint Conference on Neural Networks (2019)
26. Wu, X., Li, C., Zhu, Y., Miao, Y.: Auto-encoding quantization model for short texts. In: NeurIPS Workshop on Information Theory and Machine Learning (2019)
27. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web. pp. 1445–1456. ACM (2013)
28. Yuan, M., Van Durme, B., Ying, J.L.: Multilingual anchoring: Interactive topic modeling and alignment across languages. In: Advances in Neural Information Processing Systems. pp. 8653–8663 (2018)
29. Zhao, B., Xing, E.P.: Bitam: Bilingual topic admixture models for word alignment. In: Proceedings of the COLING/ACL on Main conference poster sessions. pp. 969–976. Association for Computational Linguistics (2006)