Knowledge and Information Systems (2021) 63:1529–1555
https://doi.org/10.1007/s10115-021-01555-7

REGULAR PAPER

A word embedding-based approach to cross-lingual topic modeling

Chia-Hsuan Chang[1] · San-Yih Hwang[1]

Received: 26 February 2020 / Revised: 22 February 2021 / Accepted: 27 February 2021 /
Published online: 24 April 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract
The cross-lingual topic analysis aims at extracting latent topics from corpora of different languages. Early approaches rely on high-cost multilingual resources (e.g., a parallel corpus), which is hard to come by in many real cases. Some works only require a translation dictionary as a linkage between languages; however, when given an inappropriate dictionary (e.g., small coverage of dictionary), the cross-lingual topic model would shrink to a monolingual topic model and generate less diversified topics. Therefore, it is imperative to investigate a cross-lingual topic model requiring fewer bilingual resources. Recently, some space-mapping techniques have been proposed to help align multiple word embedding of different languages into a quality cross-lingual word embedding by referring to a small number of translation pairs. This work proposes a cross-lingual topic model, called Cb-CLTM, which incorporates with cross-lingual word embedding. To leverage the power of word semantics and the linkage between languages from the cross-lingual word embedding, the Cb-CLTM considers each word as a continuous embedding vector rather than a discrete word type. The experiments demonstrate that, when cross-lingual word space exhibits strong isomorphism, Cb-CLTM can generate more coherent topics with higher diversity and induce better representations of documents across languages for further tasks such as cross-lingual document clustering and classification. When the cross-lingual word space is less isomorphic, Cb-CLTM generates less coherent topics yet still prevails in topic diversity and document classification.

Keywords Cross-language · Cross-lingual topic model · Cross-lingual word embedding

1 Introduction

The rapid development of the Internet and the advance in information and communication technology are engaging people worldwide to form a global village. This development facilitates the dissemination of information about events and allows people to listen to opinions

✉ Chia-Hsuan Chang
sham82503@gmail.com

San-Yih Hwang
syhwang@mis.nsysu.edu.tw

1 Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan

worldwide. Members of the general public can now comment on significant events in various discussion forums and social media platforms. For some important global events, readers from different continents express opinions from different perspectives. For example, since July 6, 2018, the two largest economies (the USA and China) have been engaged in a trade war involving the mutual placement of tariffs. This event has a significant impact on the world, and there is intense interest in this issue and several related topics have been discussed globally, such as profit cuts of specific industries, moving production out of China, and competition in the future 5G market. Understanding topics discussed in different countries and markets will inevitably influence government policies and business strategies. Under the circumstances, identifying the patterns of common discussion topics across languages can provide considerable insight and is vital for both the public and private sectors. As a result, the demand for analyzing cross-lingual topics is growing in many research fields, including categorizing UGC [39], classifying multilingual texts [16], detecting cross-culture differences [44], and constructing bilingual dictionaries [29].

For yielding cross-lingual topics, cross-lingual topic models have been proposed to modify the generative process of latent Dirichlet allocation (LDA), one of the most influential topic models, by incorporating the linking between languages. There are two types of linkages: document linking and vocabulary linking. A document-linking model [23,26,36,38,50,52] depends on the availability of a parallel corpus such as the EuroParl Corpus,[1] where each document has complete translated versions, one for each language. The main drawback of this model is that the parallel corpus is difficult to acquire in practice. Although a comparable corpus (e.g., Wikipedia articles) can also be used, this might compromise the performance [36]. In contrast, a vocabulary-linking model [10,21,25,30] only requires a bilingual dictionary as input. Since translation dictionaries are widely available (e.g., MUSE project[2]), the vocabulary-linking model seems more practical. However, insufficient coverage and a low frequency of dictionary entries in the corpus have been shown to reduce the vocabulary-linking model to a union of monolingual topic models [25]. This situation has prompted increasing interest in constructing a cross-lingual topic model using fewer resources.

The inferred topic space of a cross-lingual topic model is considered to be language-agnostic [22]. In other words, even though words are language-specific, we can align those from different languages based on their themes and generate topics across languages. A similar concept applies to cross-lingual word space alignment. Several studies have proposed methods for aligning multiple monolingual word spaces into a single cross-lingual word space using only a small amount of cross-language resources [2,12,56]. Those methods assume that the semantic structure is isomorphic across languages. The comprehensive analogy is that the word spaces learned from different languages correspond to the same map from different angles, and so we can align them by learning rotations. Although the resultant aligned spaces facilitate several cross-lingual NLP tasks such as sentence translation, cross-lingual sentiment classification [58], and word translation [12,49], few studies have developed topic models using cross-lingual word embedding. The mechanism for constructing a cross-lingual topic model with cross-lingual word embedding remains underdeveloped. Also, understanding the important factors that influence the performance of such a model is vital when applying the approach to real cases. To address the above research gaps, we make the following contributions in this paper:

1. We propose a cross-lingual topic model that extends the generative process of LDA using cross-lingual word spaces.

[1] http://www.statmt.org/europarl/.

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

2. We propose a simple but effective approach to eliminate language-specific dimensions of the cross-lingual word space, which results in language-biased topics.
3. We thoroughly evaluate our model using topic coherence, topic diversity, and quality of topic representation by parallel and non-parallel corpora and find that our model outperforms other comparative models in all metrics when using the cross-lingual word space with strong isomorphism. When the cross-lingual word space is less isomorphic, our model still prevails in topic diversity and zero-shot cross-lingual document classification.

The rest of the paper is structured as follows: Sect. 2 reviews related work in cross-lingual LDA and continuous LDA, Sect. 3 illustrates our proposed model, and Sect. 4 describes the data preparation, metrics, and experimental results. We finally draw conclusions and suggest areas for future work in Sect. 5.

2 Related works

2.1 Cross-lingual LDA

With the wide adoption of LDA [8,19,31,32,53], several studies have extended LDA to cross-lingual applications. The approaches used by these models can be categorized into two types: document linking and vocabulary linking.

Document linking relies on the availability of a parallel corpus such as the Europarl Parallel Corpus, for which versions are available in 21 European languages, or a comparable corpus such as Wikipedia, which involves articles in various languages with differing degrees of detail. The most representative model is the polylingual topic model (PLTM), initially proposed by Mimno et al., and subsequently extensively extended [36,38,52]. In the settings of the PLTM, the corpus is regarded as a set of document tuples, where each tuple consists of several comparable documents written in different languages yet addressing the same topics or issues. Specifically, PLTM assumes that documents in each tuple share the same distribution over topics, and each topic has a specific distribution over words for each language. Heyman et al. [23] then introduced a Bernoulli distribution to model the probability of topic occurrence in the target and source languages, which relaxes the assumption of the PLTM and allows the extraction of language-specific topics. Nevertheless, their evaluations show that the resultant topic distribution of each document fails to achieve satisfactory performance in cross-lingual document classification. Observing that a document can often be viewed as a hierarchy of segments, Tamura and Sumita [50] incorporated the Pitman-Yor process that allows the topic distribution to be identified at the segment level. Nevertheless, document-linking models require either a parallel corpus or a comparable corpus, which might not be available in many cases.

Contrary to document linking, vocabulary-linking models rely on the use of a translation dictionary. Examples of these models include JointLDA [25] and MuTo [10]. In contrast to document-linking LDA models, which assume that each topic has a word distribution for each language, the vocabulary-linking LDA models regard each topic as a distribution over dictionary entries, where each entry is a tuple of words in different languages. Hu et al. [24] used the Dirichlet tree distribution to model the probability of the translation dictionary. Each translation entry then shares the same ancestor in the tree structure and has a similar drawing probability. To meet the nature of unaligned topics across languages, Yang et al. [54] introduced the cross-lingual topic transformation into the generative process so that a pair of topics in different languages that share more translation entries incurs higher weights, meaning that they are more similar. Yuan et al. [55] extended the anchor-based topic model for capturing multilingual contexts. The anchor-based approach derives word distribution of topics from the word co-occurrence matrix by some anchor words for each topic. The anchors, which are responsible for linking spaces of different languages, are then chosen from the translation dictionary to enlarge the topic diversity as much as possible. Hao and Paul [21] proposed extending the soft document-linking by estimating word translation overlapping between non-parallel documents. However, the lower coverage of the dictionary entries results in the less coherent topics [21,22]. Also, limited dictionary size often shrinks vocabulary-linking models into the monolingual topic model [25]. To mitigate these restrictions, our work considers dictionary entries as anchors in the continuous word spaces of different languages rather than as possible values in topic distributions. Thanks to cross-lingual word embedding techniques [15,34], we can obtain a quality cross-lingual word space with a small number of dictionary entries. Below we discuss the previous studies related to continuous LDA.

2.2 Continuous LDA

Recent developments in word vector space models (e.g., skip-gram, CBOW, and Glove [33, 35]) have succeeded in learning word representations that can capture both word semantics and their lexical relationships. Each word representation is a low-dimension vector that serves as the building block in a wide range of natural language processing (NLP) tasks. The continuous topic model is a variant of LDA that integrates with word representations, and it considers a topic as a distribution in a continuous vector space with a finite number of dimensions rather than a distribution on a large number of discrete word tokens, as assumed in LDA [5,13,37,41]. Nguyen et al. [37] proposed a topic model called latent-feature LDA (LF-LDA) that includes word embedding in the generative process. When sampling a word from a document given a particular topic, LF-LDA considers the similarity between the center of the topic and a word based on their representations. GaussianLDA [13] regards each topic as a multivariate Gaussian distribution in the word space. Given a topic, a word is chosen according to its multivariate Gaussian distribution. However, previous studies suggest that von Mises–Fisher (vMF) distribution (parameterized by cosine distances) is often a better alternative to a multivariate Gaussian distribution because the cosine distances can cope better with the large range of densities in high-dimension directional data [3,57]. For this reason, SphericalLDA [5,41] applies the vMF distribution for modeling the density of words over a unit sphere. The resultant model shows better performance than GaussianLDA in measuring the coherence. All the above continuous topic models only work in monolingual applications, and so how to apply it to cross-lingual applications still needs to be addressed.

For comparing word semantics across languages, one approach is to align pre-trained monolingual word vector spaces into a cross-lingual word space using word-alignment resources [42], such as bilingual dictionaries. A method called postmatching LDA (PMLDA) [11] relies on such cross-lingual word space to construct a cross-lingual topic model. PMLDA first constructs monolingual topic models and subsequently concatenates these models into a cross-lingual topic model. The underlying combination mechanism is to view each topic as a vector in cross-lingual word space and group topic vectors using the DBSCAN algorithm. The transformer-based language model is another approach that directly learns word representations across languages from large multilingual corpora (e.g., Wikipedia and Common Crawl). An example model is Multilingual BERT (M-BERT) that employs the transformer architecture to learn word representations across 104 languages[3]

[3] https://github.com/google-research/bert/blob/master/multilingual.md.

[14]. ZeroShotTM [6] composes an inference network and a decoding network for generating a cross-lingual topic model. An English corpus is required for obtaining two necessary inputs: word representations encoded using M-BERT and bag-of-words. After applying sentence transformer to obtain paragraph representations using word representations, the inference network employs the neural architecture of ProdLDA [48] to learn topic representations (i.e., document-topic distributions) from paragraph representations. The decoding network is then responsible for reconstructing bag-of-words to mimic topic-word distributions using topic representations. Because of the multilinguality of M-BERT, the learned inference network is capable of determining the topic representations of documents in other non-English documents. However, the design of decoding network prevents it from generating topic-word distributions across languages, limiting its interpretability compared to other cross-lingual topic models discussed in Sect. 2.1.

To sum up, most existing continuous topic models are proposed for the single-lingual corpus only; there is a need to investigate how to incorporate cross-lingual word embedding into a cross-lingual topic model.

3 Our approach

3.1 Background

Topic modeling is an important technique of text mining that aims to extract underlying topics from large textual data. LDA [9] is by far the most famous and influential model utilized for this task. LDA analyzes a corpus D, where each document $d \in D$ is represented as a bag of words N_d, and assumes the existence of several latent topics in corpus D that determine the generation of D. Each topic $t \in T$ is modeled by a probability distribution over vocabularies, denoted ϕ_t, and each document d is considered as a probabilistic mixture of topics, denoted θ_d. The generative process is shown as follows:

1. Initialize each topic $\phi_t \sim \text{Dir}(\beta)$
2. For each document d in D:
   (a) $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word d_i in N_d:
      i Draw a topic assignment $z_{d_i} \sim \text{Categorical}(\theta_d)$
      ii Draw a word type $w_{d_i} \sim \text{Categorical}(\phi_{z_{di}})$,

where α and β are hyper-parameters of Dirichlet distribution for controlling the level of concentration of its generated distributions.

In cross-lingual contexts, the corpus D consists of documents written in a set of different languages L. We use $l_d \in L$ to label the language of each document d. Our proposed cross-lingual topic model aims at extracting the hidden topic patterns across languages from D. Similar to monolingual LDA, the resultant cross-lingual topics are represented as two types of distributions: (1) document-topic distributions θ_d, which record the tendency of the topics conveyed in each document, and (2) topic-word distributions ϕ_t, which collect words with similar topic contexts across languages in each topic. In the following sections, we first introduce the preparation of cross-lingual word embedding and then propose a cross-lingual topic model using the cross-lingual word embedding.

3.2 Preparing the cross-lingual word embedding

To generate cross-lingual word embedding, we first construct a monolingual word space. This is achieved by applying text processing techniques to documents of the same language to extract tokens and part-of-speech tags. After processing, we remove stop words and retain only nouns and verbs to train the monolingual word vector using the skip-gram algorithm with negative sampling [33,35]. The idea of that algorithm is to learn word embedding for predicting neighbor words. We denote words embedded in language $l \in L$ as $H_l \in \mathbb{R}^{|V_l|\times|S|}$, where $|V_l|$ and $|S|$ indicate the number of vocabularies in language l and the number of dimensions of the word space, respectively. We choose the skip-gram algorithm because it has been widely studied in the field of distributed semantics [4] and has served as the building block in many NLP tasks.

To align two monolingual spaces into a cross-lingual word space, we apply an orthogonal transformation method [15,34,46] since it is a well-studied and most commonly adopted method [42,46]. More specifically, we choose a target language l' and map the word space of the other language l to that of l'. The orthogonal transformation method uses a bilingual dictionary $\{V_{l,i}, V_{l',i}\}_{i=1}^{P}$ to train a transformation matrix $\Omega \in \mathbb{R}^{|S|\times|S|}$ that allows H_l, the embedded words of V_l, to fold in $H_{l'}$, the embedded words of $V_{l'}$, with least square error, where $l \neq l' \in L$, P is the number of dictionary pairs, and $\{V_{l,i}, V_{l',i}\}$ represents the i-th translation word pair in languages l and l'. We show a training objective of Ω in Eq. 1, where $H_{l,i}$ and $H_{l',i}$ are word vectors of $V_{l,i}$ and $V_{l',i}$, respectively. We constrain Ω to an orthogonal matrix so that the transformation will be more robust to noisy dictionary entries [46]. We then solve it by applying the Procrustes solution [12,46]:

$$\arg\min_{\Omega} \sum_{i=1}^{P} \|\Omega H_{l,i} - H_{l',i}\|^2 \text{ subject to } \Omega^{\mathrm{T}}\Omega = I \tag{1}$$

We use Fig. 1 to illustrate the aligning process, where Fig. 1a, b shows the pre-trained word spaces of source language l and target language l', respectively. Given (网际网路 , internet), (科研 , research), and (竞选 , election) as dictionary pairs, the Ω can be determined so that we can rotate the source word space into target one and construct a cross-lingual word space H^{cs} by $\Omega H_l \cup H_{l'}$. With H^{cs}, we can compare the semantic distance between two words of different languages.

3.3 Center-based cross-lingual topic Model

We propose a method called the center-based cross-lingual topic model (Cb-CLTM), in which word vectors are regarded as new observational variables in the generative process. To incorporate cross-lingual word embedding, we replace the topic-word categorical distribution with topic centers in the form of word embedding. Below we first introduce the generative process and then illustrate the inference strategy of Cb-CLTM.

3.3.1 Generative process of Cb-CLTM

Figure 2 shows the plate notation of Cb-CLTM. Similar to LDA, our Cb-CLTM assumes that each document has its topic distribution represented as the Dirichlet-multinomial distribution. The key variant is that we characterize each topic t as a multivariate vector $\psi_t \in \mathbb{R}^{|S|}$ in cross-lingual word space. In other words, we consider each ψ_t as a semantic center point of topic t.

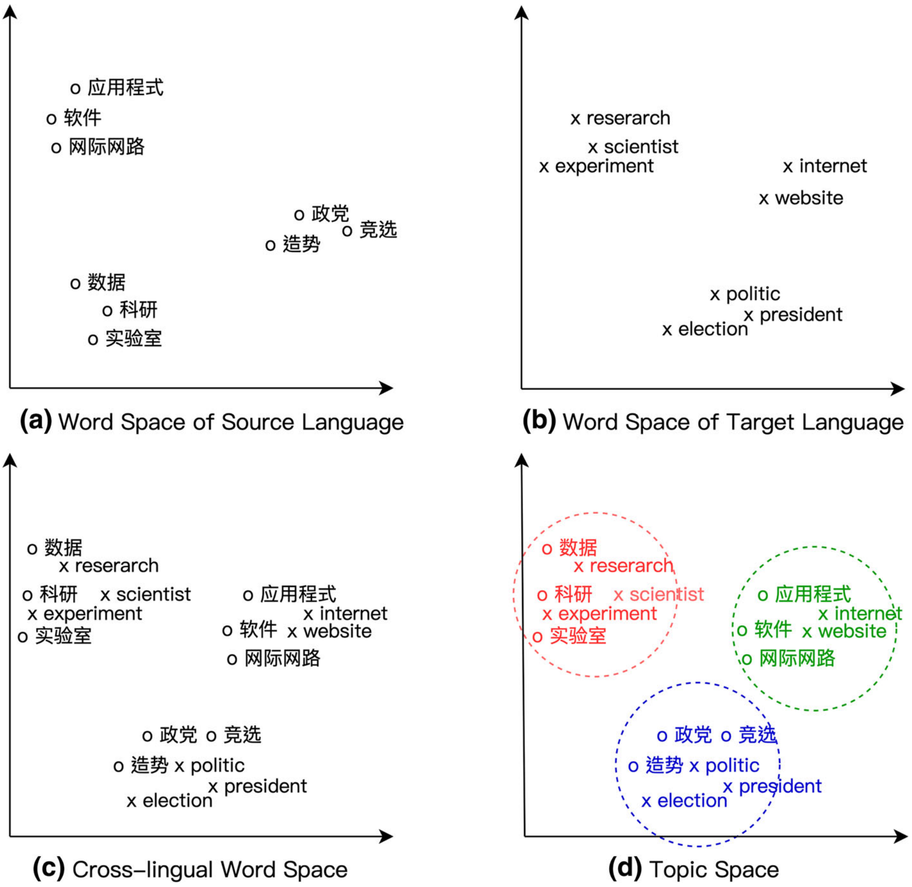

Fig. 1 Illustration of cross-lingual word space and cross-lingual topic model construction

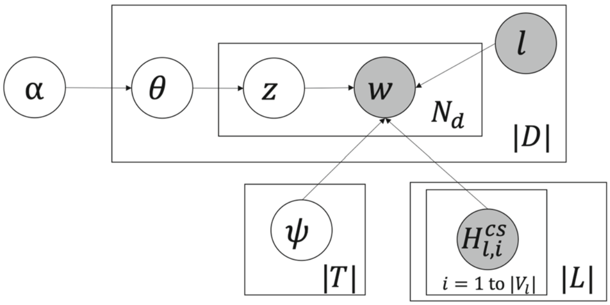

Fig. 2 The plate notation of Cb-CLTM

The absence of a topic-word distribution means that Cb-CLTM approximates the categorical word distribution ϕ_t of each topic by the softmax function:

$$\phi_t(w_{d_i}|\psi_t; H^{cs}_l) = \frac{exp(\psi_t \cdot H^{cs}_{l,w_{d_i}})}{\sum_{1\leq i\leq|V_l|} exp(\psi_t \cdot H^{cs}_{l,i})} \tag{2}$$

In contrast to most continuous LDA models, the use of the ϕ function in our model means that significantly fewer parameters and calculations are required, which helps to improve efficiency when training and inferring. For example, when compared to GaussianLDA [13], Cb-CLTM does not need the covariance matrix for each topic. Similarly, when compared to SphericalLDA [5,41], Cb-CLTM does not have a concentration parameter of the vMF distribution, and estimating this parameter for each topic has a high computational cost [47]. Although we simplify the parameters of the model, it has high efficacy for the cross-lingual topic model, as demonstrated in our experiments.

The softmax function is used to convert a set of numbers into a probability distribution. Thus, given topic t and language l, the probability of a word w_{d_i} is determined using Eq. 2. In other words, if the cross-lingual word vector of a word is close to the center of a topic ψ_t, it will have higher probability of being selected. Take Fig. 1d for example, three dashed circles with different colors represent distinct topics. Assuming that we determine the ψ_t at the center of the red scientific circle, it should assign more probabilities for words in the red circle compared to those outside the red circle. From the above specifications, the generative process of Cb-CLTM can be described as below:

1. For each document d in D:
   (a) $\theta_d \sim \mathrm{Dir}(\alpha)$
   (b) l_d is observable
   (c) For each word d_i in N_d:
      i Draw a topic assignment $z_{d_i} \sim \mathrm{Categorical}(\theta_d)$
      ii Draw a word type $w_{d_i} \sim \phi_{z_{d_i}}(\,.\,|\psi_{z_{d_i}};\, H^{cs}_{l_d})$,

3.3.2 Inferencing parameters of Cb-CLTM

To infer θ and ψ of Cb-CLTM, we apply the Bayesian and frequentist methods simultaneously. Since the Gibbs sampling, a type of Markov chain Monte Carlo, is asymptotically close to real posterior distribution in theory, we adopt collapsed Gibbs sampling to approximate θ by drawing the samples of topic assignments for each document. After integrating out θ_d from the conditional distribution of z_{d_i} in Cb-CLTM, the sampler of the topic index z_{d_i} for the ith word of document d can be written as below:

$$p(z_{d_i} = t|\mathbf{z}_{\neg d_i}, \mathbf{w}) \propto \frac{(N^t_{\neg d_i} + \alpha)}{\sum_{t=1}^{T} N^t_d + \alpha_t} \cdot \phi_t(w_{d_i}|\psi_t;\, H^{cs}_{l_d}), \tag{3}$$

where $N^t_{\neg d_i}$ is the number of words pertaining to topic t in document d except for the current observed word w_{d_i}. In addition, the word probability in topic t, $p(w_{d_i}|\psi_t)$, is approximated by ϕ_t in Eq. 2. Equation 3 can be simplified to $(N^t_{\neg d_i} + \alpha) \cdot \phi_t(w_{d_i}|\psi_t;\, H_{l_d})$ because given document d, $\sum_{t=1}^{T} N^t_d + \alpha_t$ is a constant for each word. Intuitively, the topic assignment of word w_{d_i} is controlled by two factors: (1) the proportion of topics in document d, and (2) the closeness between w_{d_i} and ψ_t in the word space. The complete derivation of sampler in Eq. 3 begins with the equation $p(z_{d_i} = t|\mathbf{z}_{\neg d_i}, \mathbf{w}) \propto p(z_{d_i} = t, w_{d_i}|\mathbf{z}_{\neg d_i}, \mathbf{w}_{\neg d_i})$ [4], and the details are shown in Appendix A. As a result, after sampling topic assignments of each document d, we use expectation of the categorical distribution to infer its topic distribution θ_d.

To derive ψ_t for each topic, we utilize maximum likelihood estimation and strip the language index of H^{cs} because different languages now share the same cross-lingual word

[4] $p(z_{d_i} = t|\mathbf{z}_{\neg d_i}, \mathbf{w}) = \frac{p(\mathbf{z},\mathbf{w})}{p(\mathbf{z}_{\neg d_i},\mathbf{w})} = \frac{p(\mathbf{z},\mathbf{w})}{p(\mathbf{z}_{\neg d_i},\mathbf{w}_{\neg d_i})p(w_{d_i})}$, $p(z_{d_i} = t, w_{d_i}|\mathbf{z}_{\neg d_i}, \mathbf{w}_{\neg d_i}) = \frac{p(\mathbf{z},\mathbf{w})}{p(\mathbf{z}_{\neg d_i},\mathbf{w}_{\neg d_i})}$

space after orthogonal projection as described in Sect. 3.2. The size of H^{cs} corresponds to the total number of vocabularies across different languages. Thus, the likelihood function of ψ_t is

$$L(\psi_t) = \prod_{v \in V} (\phi_t(v|\psi_t; H^{cs}))^{N^{t,v}}, \tag{4}$$

where $N^{t,v}$ is the number of times word v is assigned to topic t. The likelihood function can then be transformed into a negative log-likelihood function for optimization purposes. Referring to the form of ϕ_t, we represent the negative log-likelihood(NLL) of ψ_t as

$$NLL(\psi_t) = -\sum_{v \in V} N^{t,v} \left(\psi_t \cdot H_v^{cs} - log \left(\sum_{v \in V} exp(\psi_t \cdot H_v^{cs}) \right) \right) + \lambda \|\psi_t\|_2^2 \tag{5}$$

Note that in Eq. 5, L2 regularization is added to avoid overfitting. The gradient of each topic vector ψ_t is

$$\frac{\partial NLL(\psi_t)}{\partial \psi_t} = -\sum_{v \in V} N^{t,v} \left(H_v^{cs} - \sum_{v \in V} H_v^{cs} \underbrace{\frac{exp(\psi_t \cdot H_v^{cs})}{\sum_{v \in V} exp(\psi_t \cdot H_v^{cs})}}_{\phi_t(v|\psi_t; H^{cs})} \right) + 2\lambda \psi_t \tag{6}$$

By providing gradients, we apply the quasi-Newton L-BFGS[5] optimization method to minimize $NLL(\psi_t)$ and search ψ_t. Our use of L-BFGS optimization avoids the need to tune the appropriate learning rate, in contrast to using deepest gradient descent, and it generally works in both nonlinear and nonsmooth optimization cases.

Connection to expectation–maximization algorithm Our strategy for parameter inference also shares the same spirit with expectation-maximization (EM) algorithm. The goal of EM is to optimize the likelihood function $p(D, Z|\Theta)$, where D is corpus, Z is all topic assignments for all words in the corpus, and Θ is a set of parameters. EM iteratively employs two following steps: (1) E step: fixing Θ to optimize Z using the fact that $Z = p(Z|D, \Theta)$, and (2) M step: fixing Z to optimize Θ. Apparently, we can use Eq. 3 to fulfill the objective of E step, which assigns topic assignments for all words. Since the topic-document distribution θ has been integrated out, the topic vector of each topic ψ_t is the only remaining parameter in Θ. Therefore, we can carry out the M step by optimizing Eq. 5, which updates all topic vectors. We can guarantee that the inference strategy of Cb-CLTM has the same convergence property as EM because of this connection. Both determined topic assignments and topic vectors of each iteration will better fit the corpus (i.e., observed data likelihood) than those from the previous iteration.

3.3.3 Language dimension reduction of embedding

We notice that the induced topics of Cb-CLTM would potentially bias towards a specific language when purely using the pretrained cross-lingual word vectors. The cause of such language bias is because some dimensions in cross-lingual word space could be language-specific. Thus, words that are close to a semantic center tend to have similar values in these language-specific dimensions, resulting in the phenomenon of "clustering by language" [17]. Table 1 presents the sample topics inferred from UM-Corpus [51] using the original pretrained cross-lingual word vectors H^{cs} by Cb-CLTM. The results show that each topic

[5] L-BFGS is the abbreviation of Limited–memory Broyden–Fletcher–Goldfarb–Shanno algorithm.

Table 1 Sample topics from full-dimension Cb-CLTM

Topic number	Representative words
1	公司, 计划, 项目, 企业, 亿美元, 基金, 投资, 成功, 投入, 研发
2	year money company business day time job end pay hold
3	政府, 组织, 国家, 叙利亚, 地区, 控制, 活动, 非洲, 城市, 袭击
4	president government obama security party military iran
5	university school professor institute york book science
6	美国, 地方, 来自, 大学, 纽约, 现在, 人们, 学校, 教授, 东西

Algorithm 1 The pseudocode of Cb-CLTM

Initialize all z randomly and record in N_d^t, $N^{t,v}$
Apply language dimension removal to generate H_l^{cs}
for each iteration in I **do**
  for each topic t in T **do**
    Update ψ_t using L-BFGS
  end for
  for each document d in D **do**
    for each word d_i in N_d **do**
      $z_{d_i}^{new} \sim P(z_{d_i} = t | \mathbf{z}_{\neg d_i}; \alpha, \theta, \psi, H^{cs}, l_d)$
      Update the counts in N_d^t, $N^{t,v}$
    end for
  end for
end for

clusters words with similar concepts in the same language rather than across languages. Under this circumstance, Cb-CLTM hardly aligns similar topics across languages such as topic 1 versus topic 2, topic 3 versus topic 4, and topic 5 versus topic 6. To remedy this problem, we propose eliminating the dimensions that are language-specific. Specifically, we assign a label y to each embedding $H_{l,i}^{cs}$ according to its language l, and estimate the predictive power of each dimension in S. A dimension that has high predictive power is considered language-specific and will be removed. Thus, we obtain a subset S^* by removing dimensions with the maximum predictive power. Several algorithms, including logistic regression and tree-based methods, can be used to identify these language-specific dimensions. After removing these dimensions, we apply L2 normalization to normalize each row in $H_l^{cs} \in \mathbb{R}^{|V_l| \times |S^*|}$. This normalized H_l^{cs} would be used as the input of Cb-CLTM. The number of removed dimensions is the hyperparameter that controls the trade-off between the semantic completeness of cross-lingual space and the performance of the cross-lingual topic model. We examined this effect in our experiments.

The pseudocode of Cb-CLTM is presented in Algorithm 1. We randomly assign a topic index for each word and record the number of words in document d that are assigned to topic t in N_d^t, as well as the number of times word v is assigned to topic t in $N^{t,v}$. Language dimensions are removed to generate H_l^{cs}. In each iteration, we optimize each ψ_t only once in order to improve efficiency. Then, every word follows the generative process to be reassigned its topic index according to the conditional distribution of z. After drawing a new topic index, we update N_d^t and $N^{t,v}$ and subsequently derive new θ_d and ψ_t.

4 Experimental results

4.1 Description of datasets

We used two corpora in our experimental evaluation: UM-Corpus [51] and Reuters Corpus Volume 2 (RCV2) [27]. UM-Corpus is a parallel corpus that contains a large number of pairs of English and Chinese sentences. We selected the news domain of the corpus as our dataset, which comprises 450,000 pairs of bilingual sentences involving categories such as politics, economics, technology, education, agriculture, and society. RCV2 is a nonparallel and noncomparable corpus that includes numerous news articles in 13 languages. In our work, we chose articles in three languages for our cross-lingual topic modeling experiments, namely English, Chinese, and Japanese. Each news article is categorized into one of the following categories: CCAT (corporate/industrial), ECAT (economics), GCAT (government/social), and MCAT (markets). This dataset has been widely used to evaluate algorithms related to cross-lingual document classification [23,43]. English text was processed using spaCy, while Stanford CoreNLP and Mecab were used for the Chinese texts and Japanese texts, respectively. After applying tokenization and tagging the parts of speech, we only retained nouns and verbs for further analysis.

Preparing datasets for the topic model In our experiments, we generated cross-lingual topic models on four different datasets. The first two were the whole UM-Corpus and 25,000 sampled document pairs from UM-Corpus (called UM-Corpus 25K). We created UM-Corpus 25K for evaluating the performance on a small-scale parallel dataset. The last two were datasets created from RCV2. Since the class distributions differed significantly between the English, Chinese, and Japanese corpus, we utilized the MLDoc scripts [43] to sample documents uniformly across classes in three languages, resulting in two subsets of RCV2, called MLDoc En-Zh and MLDoc En-Ja. Each subset consists of 10,000, 1,000, and 4,000 news articles for the training, validating, and testing for text classification task in each language, respectively. We present the descriptive statistics of all datasets in Table 2.

Preparing for cross-lingual word embedding To obtain the cross-lingual word space H^{cs} required for Cb-CLTM, we applied the approach described in Sect. 3.2 to UM-Corpus and RCV2 and initially set the number of word dimensions S to 100. Since word vector space tends to be more robust when training it from the large-scale corpus, we determined the word vector spaces from UM-Corpus and RCV2 rather than UM-Corpus 25K and two MLDoc subsets. To prepare the anchors across languages, we used the Chinese–English and Japanese–English bilingual dictionary from the Facebook MUSE project [12] that is available at https://github.com/facebookresearch/MUSE. The coverage ratios of the Chinese-English dictionary in UM-Corpus and RCV2 were 8.7% and 4.6%, respectively. The Japanese–English dictionary covered 7.2% vocabulary in RCV2. We did not use additional dictionaries to increase the coverage because this is a common situation in real-world applications, and we wanted to determine the impact of a low dictionary coverage on our model and other vocabulary-linking models. After aligning the cross-lingual word spaces, we then used logistic regression to estimate the language effect of each dimension and find a subset of dimensions S^* based on our discussion in Sect. 3.3.3.

4.2 Performance metrics

Coherence metric The normalized pointwise mutual information (NPMI) score is a well-recognized metric for evaluating the coherence of topic-word distribution ϕ in a topic model

Table 2 Descriptive statistics of datasets, where those started with † are used for determining word vector spaces

Dataset	#Chinese documents (#word types)	#English documents (#word types)	#Japanese document (#word types)	#Average tokens per document
†UM-Corpus	450,000 (31,287)	450,000 (21,199)		8.62
UM-Corpus 25K	25,000 (18,449)	25,000 (9,695)		8.62
†RCV2	24,533 (41,344)	673,765 (19,982)	58,599 (32,876)	71.67
MLDoc En-Zh	15,000 (6,760)	15,000 (14,254)		74.21
MLDoc EN-Ja		15,000 (14,254)	15,000 (12,800)	81.86

because it strongly correlated with human judgment [28]. As shown in Eq. 7, the NPMI score quantifies the correlation between two words w_i and w_j as well as represents an estimate of word probability $Pr(.)$ and word joint probability $Pr(.,.)$ at the document level. The numerator determines the dependency between the two words, with 0 indicating their independence, and the denominator, $-log(Pr(w_i, w_j))$, is used to normalize the score into the range [–1, 1], with a higher NPMI score indicating a higher dependence between the two words:

$$NPMI(w_i, w_j) = \frac{log(\frac{Pr(w_i,w_j)}{Pr(w_i)Pr(w_j)})}{-log(Pr(w_i, w_j))} \tag{7}$$

To adjust the NPMI score for measuring the closeness of words in different languages, we followed the strategy of Hao et al. [20] based on a large number of comparable Wikipedia documents as a reference corpus. For this purpose, we used a 405K English–Chinese Wikipedia comparable corpus and a 393K English–Japanese corpus downloaded from https://linguatools.org/ as the reference documents. We merged each pair of bilingual documents into a single cross-lingual document for estimating $Pr(.)$ and $Pr(.,.)$.

Equation 8 shows how we determine the coherence score when given a cross-lingual topic t and the C top contributed words from topic-word distribution ϕ_t^l. The cross-lingual NPMI (CNPMI) score of a topic is the average of the NPMI scores for all word pairs of different languages. For instance, given a topic t, let the top-two contributed topic words of $\phi_t^{l=\text{English}}$ and $\phi_t^{l'=\text{Chinese}}$ be {disease, treatment} and {疾病, 治疗}, respectively. We then calculate the NPMI score of (disease, 疾病), (disease, 治疗), (treatment, 疾病), and (treatment, 治疗). The CNPMI score of a topic model is then the average of CNPMI scores for all topics:

$$CNPMI(t, C, l, l') = \frac{\sum_{w_i^l \in \text{Top}(t,l), w_j^{l'} \in \text{Top}(t,l')} NPMI(w_i^l, w_j^{l'})}{C^2}, \tag{8}$$

where $\text{Top}(t,l)$ is the set of top-C words in language l according to ϕ_t^l.

Diversity metric A good topic model should contain distinguished (i.e., diversified) topics. Moreover, the inferred topics are informative when the top contributed words of topics are exclusive to others [7]. Therefore, we leverage the inverse Jaccard index to measure the divergence between topic-word distributions. The inverse average Jaccard similarity (inverse-AJS) of a topic model is defined as

$$\text{inverse-AJS}(T) = 1 - \frac{\sum_{l \in L} \sum_{t,t' \in T} \frac{\text{Top}(t,l) \cap \text{Top}(t',l)}{\text{Top}(t,l) \cup \text{Top}(t',l)}}{|L| \times |T| \times (|T| - 1)/2}, \tag{9}$$

where $t \neq t'$. A higher inverse-AJS indicates that more diverse topics have been generated by a topic model.

Metric for the quality of cross-lingual document representation We evaluated the quality of a document-topic distribution inferred by topic models by adopting different metrics for the two datasets. For the parallel datasets UM-Corpus and UM-Corpus 25K, we calculated the divergence between the topic distributions $(\theta_d^l, \theta_d^{l'})$ for each parallel sentence d using the Jensen-Shannon divergence (JSD). For consistency, we reported the inverse-JSD defined as 1−JSD, where a higher score indicates greater similarity of the topic distributions. For measuring the nonparallel documents in two MLDoc datasets, we constructed a news category classifier that uses document-topic distributions as features to assess the prediction accuracy in cross-lingual document classification. More specifically, we adopted the zero-shot strategy that trains a classifier based on document features θ_d^l from the source language (i.e., English) and used it to classify document features $\theta_d^{l'}$ in other language (i.e., Chinese and Japanese). Applying the zero-shot strategy allows evaluation of how well the topic model recognizes the topics across languages.

4.3 Parameter settings

Comparative models and Bayes prior settings We benchmarked the performance of Cb-CLTM with the other four existing cross-lingual topic models, including one document-linking model, one vocabulary-linking model, one model using the cross-lingual word embedding, and one anchor-based model. These models are chosen for two reasons: (1) their characteristics were well studied such as PLTM and JointLDA, and (2) they were proposed recently and had accessible implementations such as MTAnchor and PMLDA. These models are detailed as below:

1. PLTM by Mimno et al. [36,38,52], which is a representative document-linking model. It requires a parallel or comparable corpus as inputs and assumes that the documents in the same pair share the same topic distribution. We used the implementation from https://github.com/mimno/Mallet.
2. JointLDA by Jagarlamudi and Daumé [25], which is a well-studied vocabulary-linking model. This model represents each entry of a bilingual dictionary as a word concept in the topic-word distribution for catching the cross-lingual topics. We reconstructed JointLDA as described at https://github.com/ponshane/python-topic-model.
3. PMLDA by Chang et al. [11], which also uses cross-lingual word spaces for connecting topics across languages. It first determines monolingual topics and then constructs cross-lingual topics using the clustering to link topics with semantic meaning. We used the program implemented by the authors.
4. MTAnchor by Yuan et al. [55], which is a multilingual extension of the anchor-based topic model. When given a bilingual dictionary, it first finds the bilingual topic anchors from dictionary by searching the convex hull on low-dimensional word spaces. Then, the topic-word distributions are recovered by RecoverL2 algorithm [1]. The implementation can be found at https://github.com/forest-snow/anchor-topic.

All statistical topic models, namely Cb-CLTM, PLTM, JointLDA, and PMLDA, share common parameters including Dirichlet prior α of the document-topic distribution, Dirichlet prior β of the topic-word distribution,[6] and the number of Gibbs iterations (I). To ensure fair comparisons, we fixed the same settings across models, with α and β set at $50/T$ and 0.1, respectively [18], and I set to 1,000 for the convergence of the sampling process.

Effects of language dimension reduction for Cb-CLTM Before comparing the performance between models, we first investigated the effects of removing the language dimensions for Cb-CLTM based on coherence measurement, CNPMI. Specifically, for each dataset, we fixed the number of topics, $|T|$, to 20 and experimented with the effect of Cb-CLTM. The size of the embedding dimension is set at 100. Table 3 reports that, in all datasets, when removing more dimensions, more semantic relationships will disappear, resulting in a lower and more unsteady CNPMI score. Nevertheless, without removing any dimensions (i.e., $S^* = 100$), Cb-CLTM only generated language-biased topics as presented in Table 4. That is, each inferred topic center of Cb-CLTM is biased towards a particular language, which in turn harms the coherence performance when $S^* = 100$. Notice that we selected the top-100 contributed words for each topic t and determined a language-biased topic if more than 70

[6] Note that the Cb-CLTM does not have this parameter.

Table 3 Coherence performances of Cb-CLTM with different S^* values for four datasets

	$S^* = 40$	$S^* = 60$	$S^* = 80$	$S^* = 90$	$S^* = 100$
UM-Corpus	0.175 (0.093)	0.174 (0.092)	0.171 (0.090)	**0.174 (0.087)**	0.144 (0.094)
UM-Corpus 25K	0.162 (0.103)	0.160 (0.103)	0.165 (0.097)	**0.168 (0.091)**	0.153 (0.088)
MLDoc En-Zh	0.087 (0.177)	0.087 (0.172)	0.095 (0.161)	**0.100 (0.162)**	0.099 (0.154)
MLDoc En-Ja	0.076 (0.155)	0.087 (0.141)	0.085 (0.137)	0.091 (0.129)	**0.093 (0.128)**

The highest coherence of each dataset is bold. The standard deviation is in the parenthesis

Table 4 Proportion of non-language-biased topics generated by Cb-CLTM when $S^* = 90$ and $S^* = 100$

	$S^* = 90$ (%)	$S^* = 100$ (%)
UM-Corpus	35	0
UM-Corpus 25K	45	0
MLDoc En-Zh	20	5
MLDoc En-Ja	10	0

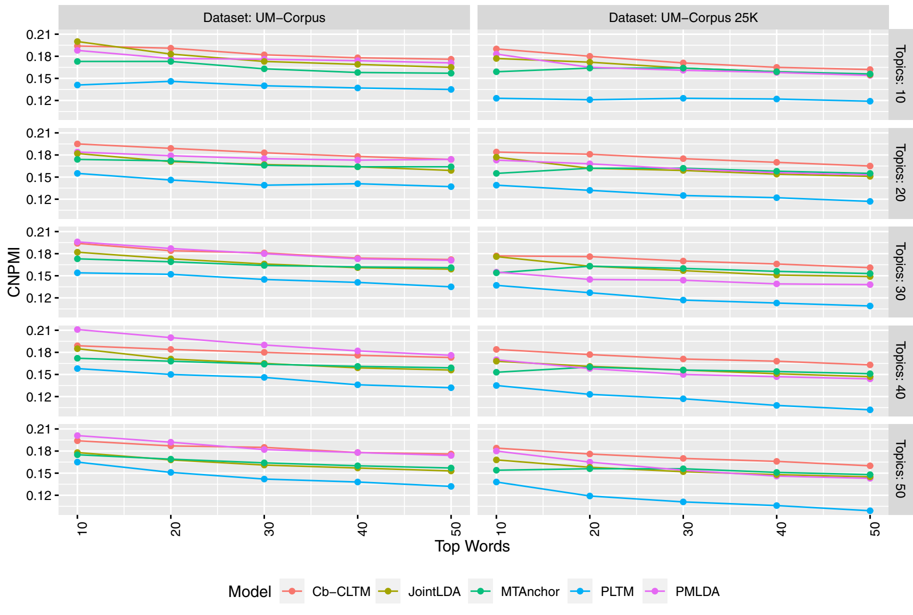

Fig. 3 Coherence performances of Cb-CLTM for different sizes of UM-Corpus

words are from the same language. Table 4 also reveals that Cb-CLTM determined more non-language-biased topics from UM-Corpus than those from two MLDoc datasets because both UM-Corpus datasets are parallel corpora.

A cross-lingual topic model shall generate coherent topics and avoid from clustering topics by languages. As a result, we adopted word spaces with $S^* = 90$ to Cb-CLTM for further model comparisons because this setting achieves the almost highest coherence score in four datasets and avoids simply generating only language-biased topics.

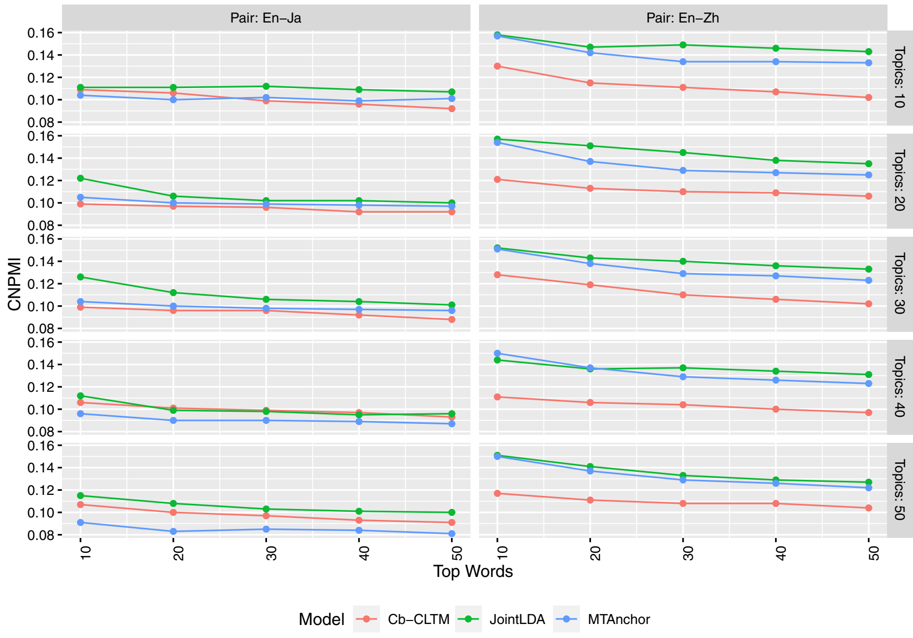

Fig. 4 Coherence performances in two MLDoc datasets

4.4 Coherence performance

UM-Corpus Figure 3 reports the CNPMI scores of each model for UM-Corpus and UM-Corpus 25K. Cb-CLTM outperforms the other models in UM-Corpus 25K and performs competitively with PMLDA in UM-Corpus. Because Cb-CLTM and PMLDA use the same cross-lingual word space, their performances are close in UM-Corpus. Nevertheless, Cb-CLTM tended to generate more coherent topics in the smaller dataset, namely UM-Corpus 25K. The good performance of Cb-CLTM demonstrates that the continuous topic models are useful in generating more coherent topics across languages. Moreover, it is an encouraging result because Cb-CLTM needs neither a parallel/comparable corpus nor many dictionary entries for inferring the topic and it benefits from lexical semantics of embedding to generate a more coherent ϕ than JointLDA and MTAnchor. The cross-lingual word space could provide more information, even using a low coverage bilingual dictionary (8.6%) as anchors. A particularly interesting observation is that the PLTM performs the worst, and we attribute this to the short text characteristic of UM-Corpus, in which the average sentence comprises only 8.62 tokens. Besides, the PLTM is the only model that assumes documents in each pair share the same topic distribution. Hence, short texts could destabilize the allocation of θ, which in turn decreases the performance of ϕ due to the Gibbs sampling mechanism. When increasing the size of the dataset, the CNPMI scores of all models increased, which is reasonable because a larger number of observed documents helps the sampling process.

MLDoc Figure 4 reports the CNPMI scores of each model for two MLDoc datasets. Note that the PLTM is not included in the comparison because it is not applicable to a nonparallel corpus. Also, we excluded PMLDA because it fails to links topics across languages, resulting into very few cross-lingual topics. For reference, in MLDoc En-Zh (MLDoc En-Ja), PMLDA generated only 2(1), 0(0), 2(0), 1(3), 1(1) cross-lingual topics at $|T| = 10, 20, 30, 40$, and

50, respectively. However, Fig. 4 indicates that Cb-CLTM did not generate the most coherent topics and even cause the worst performance in the MLDoc En-Zh dataset. The degradation of Cb-CLTM is caused by the poor quality of cross-lingual word embedding induced from RCV2. In the RCV2 dataset, we observed the large differences of class distributions between languages. For example, the class distributions on economics, corporate/industry, government/social and markets of Chinese corpus are 19.7%, 18.2%, 2.8%, and 59.4%, but those of English corpus are 6.2%, 39.8%, 29.5%, and 24.5% [43]. The dramatically different class distributions between languages make it difficult to fit a language transformation mapping. This issue has been reported previously as the ”isomorphism” problem between the word vector spaces of different languages, which has been regarded as a prerequisite for learning language transformation [12]. Figure 5 shows the 2D projection of three resultant cross-lingual word spaces trained using UM-Corpus and RCV2, in which different colors represent different languages. It can be seen from the figure that the English–Chinese word space of RCV2 contains fewer overlaps between languages; so it cannot effectively provide language links. This poor alignment explains the significant CNPMI drops of Cb-CLTM in MLDoc En-Zh. This non-aligned cross-lingual word space results in "clustering by languages" phenomenon [17], which will harm the generative process and topic assignments of Cb-CLTM. That is to say, the center of a topic ψ_t could vary significantly based on the given language, which impedes generating coherent topics across languages.

To complement the qualitative 2D visualization of the cross-lingual word spaces, we adopted the modularity metric to measure the quality of our induced cross-lingual word spaces. Modularity was proposed by Fujinuma et al. [17] for measuring the quality of a cross-lingual word space. Their empirical experiments found that a bad cross-lingual word space tends to have high modularity and clusters words by languages, while a good one has lower modularity and clusters words in a more language-agnostic fashion. In other words, a good cross-lingual word space shall position words with similar meanings closely regardless of their languages. It is also found that the modularity of a cross-lingual word space is negatively related to the performance in downstream tasks (i.e., the cross-lingual word space with lower modularity tends to have a better performance in downstream tasks such as document classification, bilingual lexicon induction, and document retrieval). With this metric implemented in https://github.com/akkikiki/modularity_metric, the modularities for the H^{cs} determined from UM-Corpus, En-Zh documents of RCV2, and En-Ja documents of RCV2 are 0.116, 0.279, and 0.278, respectively. It implies that smaller modularity of H^{cs} in UM-Corpus leads to better CNPMI, while larger modularities in MLDoc En-Zh and MLDoc En-Ja incur inferior CNPMI. To sum up, both Cb-CLTM and PMLDA rely on the whole cross-lingual word space to infer the topic patterns across languages, and its performance is strongly correlated with the quality of the cross-lingual word space. Given a bad cross-lingual word space, PMLDA would merely produce monolingual topics, and Cb-CLTM would generate less coherent topics.

4.5 Topic diversity

UM-Corpus Figure 6 compares the topic diversity across the models. While the mean diversities (i.e., inverse-AJS values) are high for most models, Cb-CLTM has the smallest standard deviation in two UM-Corpus datasets. Previous studies have shown that high-frequency words often dominate inferred topics of discrete topic models due to ignoring low-frequency words in the generative process [7,45], which in turn results in the wider standard deviation of the PLTM and JointLDA. Also, PMLDA suffers from the same problem because it first

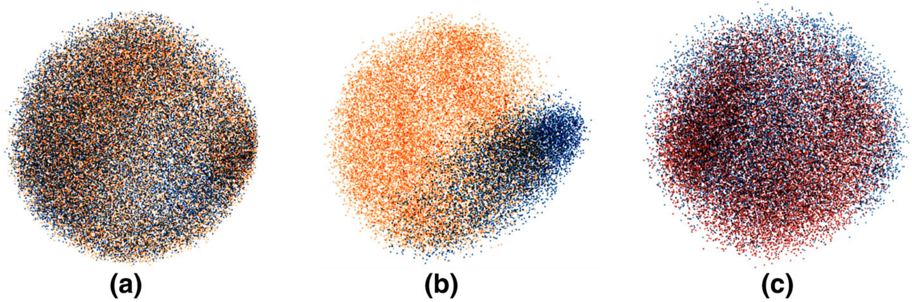

Fig. 5 2D projections of H^{cs} trained using UM-Corpus and RCV2. Different colors indicate different languages. We used principle-components analysis to reduce dimensions. **a** English–Chinese space from UM-Corpus, **b** English–Chinese space trained from RCV2, **c** English–Japanese space trained from RCV2

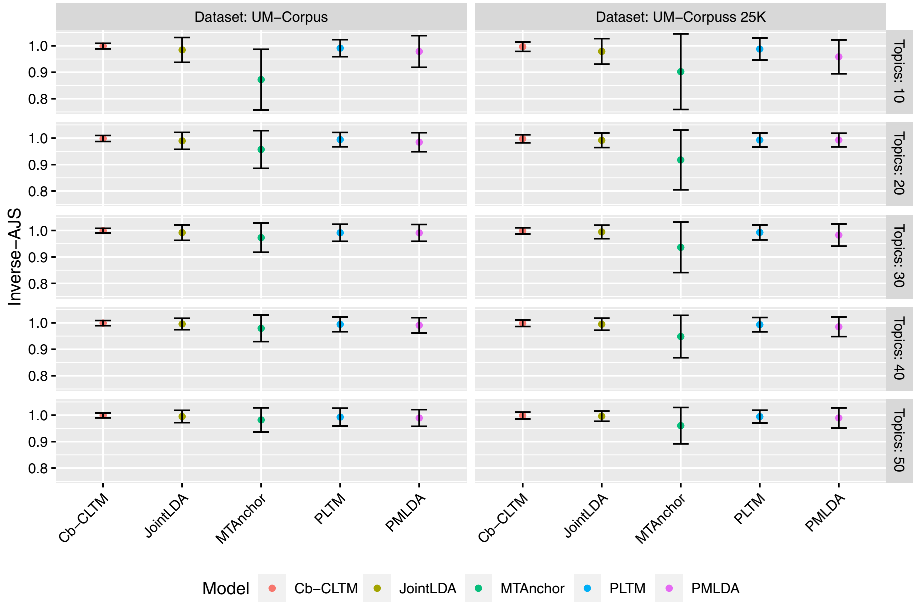

Fig. 6 Diversity performances of comparative models for different sizes of UM-Corpus

constructs a monolingual LDA and link monolingual topics across languages. MTAnchor performs the worst in diversity measurement. Even though MTAnchor applies the orthogonal projection to search the topic anchors iteratively, those topics seem duplicated after recovering the topic-word distributions. Rather than observing discrete word types, Cb-CLTM observes continuous word embedding that prevents focusing on frequent words.

MLDoc Figure 7 shows that Cb-CLTM still prevails in diversity measurement on the two MLDoc datasets, and MTAnchor remains the worst. We attribute the failure of generating diversified topics to the design of MTAnchor. Since MTAnchor is initially designed to involve the manual selection process, it is suboptimal in selecting topic anchors from a set of bilingual dictionary entries and generating distinct topics. Both Figs. 6 and 7 show that Cb-CLTM generates the most diversified topics in both the comparable and noncomparable corpus. Comparing to JointLDA and MTAnchor, Cb-CLTM is not constrained by the given bilingual

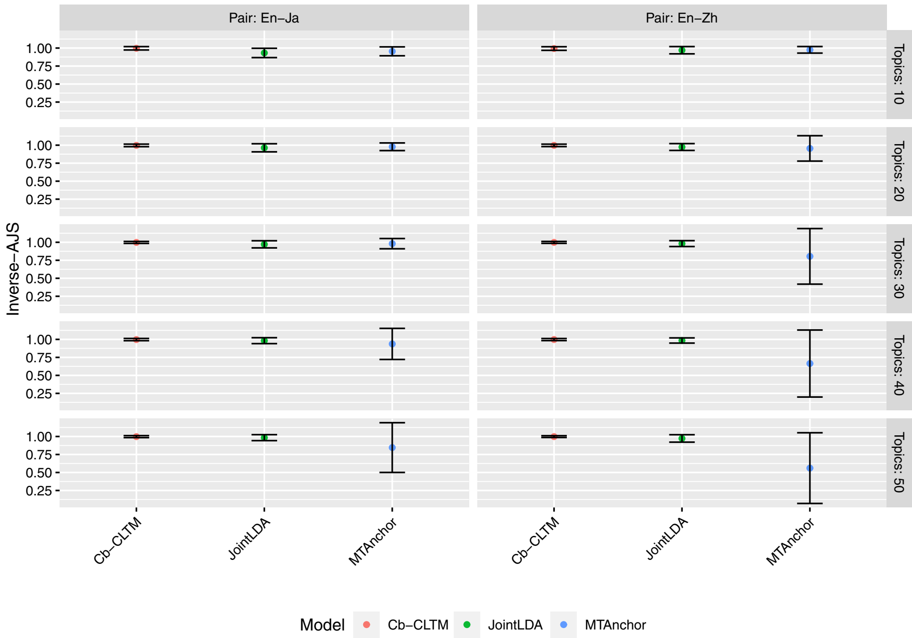

Fig. 7 Diversity performances of comparative models for two MLDoc datasets

dictionary when learning cross-lingual topics. Instead, the bilingual dictionary is used only to construct cross-lingual word spaces, which prevents Cb-CLTM from duplicating similar word allocations across the resultant topics.

4.6 Performance in cross-lingual document representation

UM-Corpus Figure 8 shows the inverse-JSD of each model for two UM-Corpus datasets. Cb-CLTM stands out at all settings except for $|T| = 10$, in which Cb-CLTM still has comparable performance as the PLTM. We observe that when Cb-CLTM and JointLDA categorize the dataset into more topics, their inverse-JSD increase. This behavior is attributed to their highly coherent topic-word distribution ϕ as listed in Fig. 3. It helps the model to result in a better document-topic distribution θ [37]. Likewise, despite the less diversity shared between topics induced from MTAnchor, it still has an increased performance when modeling more topics across languages. Conversely, the inverse-JSD of the PLTM decreases as the number of topics increases. This behavior is caused by low coherence ϕ of the PLTM, and it conforms to the original report of the PLTM that more topics would decrease the closeness between parallel documents [36]. Furthermore, when increasing the number of topics, PMLDA tends to produce only monolingual topics, resulting in some dimensions of θ being language-specific, which dramatically decreases the inverse-JSD of each parallel pair.

MLDoc We follow the zero-shot learning strategy discussed in Sect. 4.2 and provide the results for MLDoc in Fig. 9. We use the English dataset for training a multiclass regularized logistic regression and tune the hyperparameters using the English validation set. The intralingual prediction accuracy is obtained by applying the classifier to the English test set, and the interlingual prediction accuracy is computed by applying the classifier to the Chi-

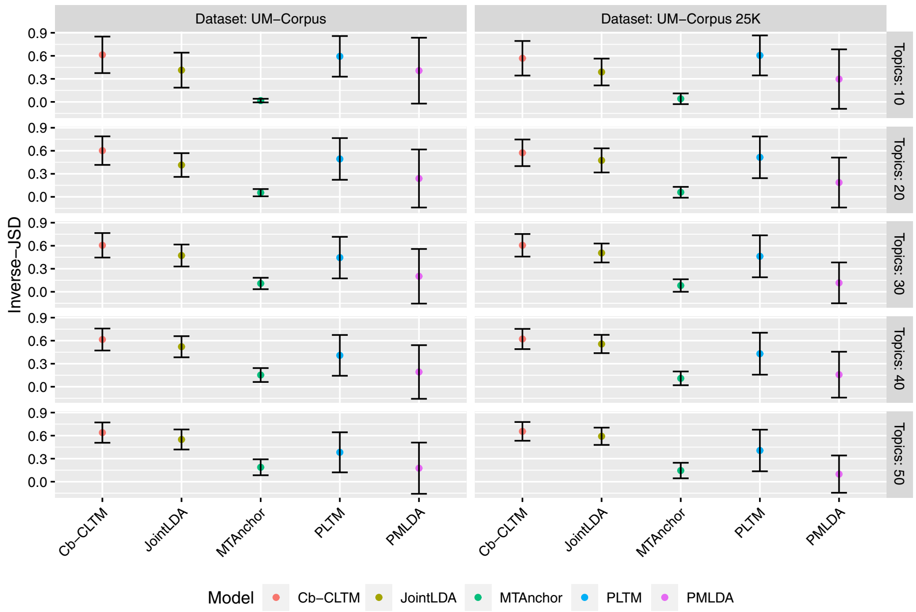

Fig. 8 Inverse-JSD of each model for UM-Corpus datasets

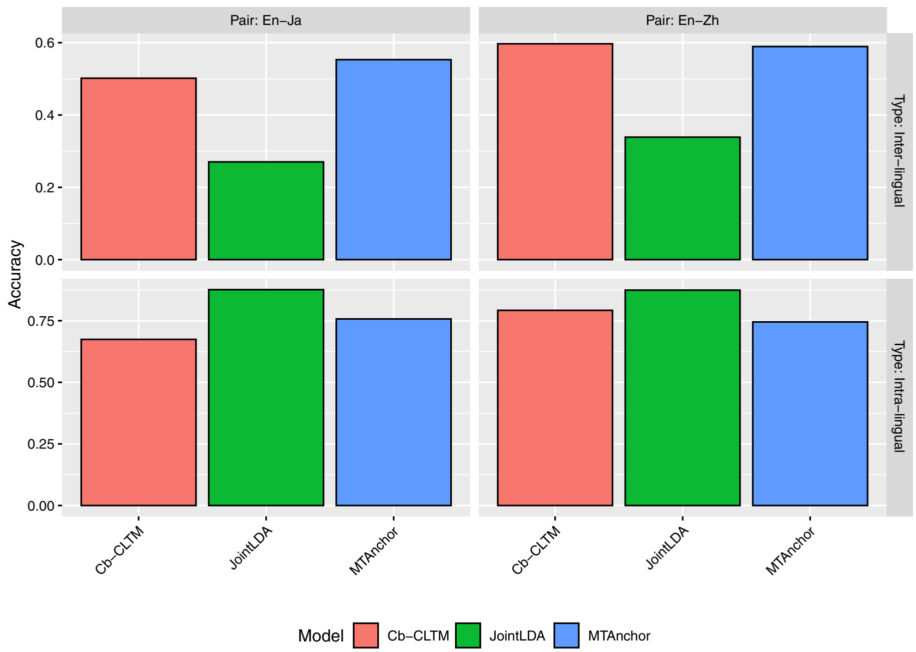

Fig. 9 Prediction accuracy of models for two MLDoc datasets

Table 5 Sample topic results for UM-Corpus 25K from Cb-CLTM, PMLDA, PLTM, and JointLDA

Topic 1: science				**Topic 2: elections**				**Topic 3: technology**			
Cb-CLTM		PMLDA		Cb-CLTM		PMLDA		Cb-CLTM		PMLDA	
研究	brain	研究	people	总统	president	总统	president	公司	information	公司	online
发现	change	发现	disease	奥巴马	country	美国	party	信息	internet	网站	world
进行	human	一个	found	国家	government	支持	mr	产品	computer	信息	social
人员	scientist	可能	time	选举	year	选举	obama	网络	news	新闻	information
人类	study	这种	risk	国会	party	奥巴马	republican	行业	agency	可以	company
科学家	researcher	影响	dr	大选	obama	成为	state	数据	apple	媒体	software
通过	blood	问题	human	候选人	east	候选人	romney	发布	google	互联网	mr
影响	collect	国家	study	竞选	state	共和党	time	网站	card	网络	business
一项	result	疾病	blood	领导人	week	通过	house	新闻	smartphone	一个	people
调查	research	治疗	treatment	共和党	america	政治	washington	设备	medium	用户	google
导致	gene	美国	life	举行	africa	认为	support	环境	making	社交	market
结果	finding	认为	health	贸易	campaign	选民	candidate	互联网	app	报纸	service
实验	dr	博士	age	反对	month	法案	law	媒体	newspaper	网上	news
基因	journal	健康	heart	宣布	west	一位	congress	网上	brand	提供	internet
原因	skin	进行	young	议会	korea	竞选	american	技术	software	组织	store
PLTM		JointLDA		PLTM		JointLDA		PLTM		JointLDA	
科学家	people	研究	find	奥巴马	president	美国	people	肯尼亚	english	公司	technology
帕金森病	disease	发现	government	共和党	obama	人们	president	索马里	podcast	技术	enterprise
干细胞	found	这种	result	候选人	state	总统	national	一部分	second	企业	computer
心脏病	high	进行	gene	罗姆尼	american	全国	female	制造商	language	电脑	apple
实验室	study	结果	personnel	肯尼迪	party	支持	marriage	智能手机	episode	苹果	world
研究员	human	基因	security	参议员	time	女性	think	埃塞俄比亚	number	世界	software
弗洛伊德	risk	一种	treatment	民主党	support	婚姻	united	特斯拉	look	软件	mr
可能性	brain	影响	method	参议院	public	认为	american	消费者	learning	一家	product
研究者	long	博士	disease	约翰逊	house	团结	country	苹果公司	guide	先生	system
感兴趣	heart	人员	humanity	基督教	republican	国家	year	伊斯兰	business	产品	internet
结核病	health	保安	experiment	共和党人	law	年度	obama	竞争对手	help	系统	google
化学物质	blood	治疗	cell	支持者	year	奥巴马	man	乔布斯	welcome	上网	mobile
其他人	system	方法	international	众议院	election	成为	life	诺基亚	like	谷歌	business
意味着	body	疾病	report	巴拉克	candidate	重要	faith	维基百科	website	移动	market
潜意识	cause	人类	patient	同性恋	romney	政治	family	操作系统	going	业务	social

nese and Japanese test set. The figures show that JointLDA performs well for intralingual document classification, whereas both Cb-CLTM and MTAnchor are the best two models for interlingual document classification. The reason for the poor interlingual classification of JointLDA is the low coverage of the dictionary. For the dictionary entries encompass only a small proportion of the words in the corpus, those entries cannot effectively bridge across languages, which could reduce JointLDA into a monolingual LDA [21,25]. This explains why JointLDA performs well in intralingual classification yet fails in interlingual classification. Although Cb-CLTM is not the best model for generating coherent ϕ for two MLDoc datasets, it can learn the document-topic distribution that works well in both intralingual and interlingual document classification using a dictionary with a small coverage.

4.7 Qualitative analysis

Table 5 provides qualitative results of topic-word distribution ϕ learned from UM-Corpus 25K for the four models for three sample topics: science, elections, and technology. We exclude MTAnchor from the comparison because it does not generate topics in science, elections, and technology. Also, we found that MTAnchor duplicated several common words like "people, going, think, means, know, like, good, want, way, come, lot, world, person, talk" across topics, resulting in less diversity (see Fig. 6). The results indicate that all of the models are capable of grouping similar semantic words into a topic, yet Cb-CLTM, PMLDA, and JointLDA generate good results for topic-word distributions. Although the PLTM also generates good results for topics 1 and 2, it fails to produce coherent cross-lingual topic words for the technology topic. Some words are irrelevant to technology such as "肯尼亚 ", "索马里 ", "伊斯兰 ', "English", "language" and "episode". These results also support the performance result presented in Fig. 3, that PLTM is the least coherent model.

Table 6 presents the qualitative results obtained for MLDoc En-Zh. We select the government topic and markets topic of Cb-CLTM, PMLDA, MTAnchor, and JointLDA for the

Table 6 Sample topic results for MLDoc En-Zh from Cb-CLTM, PMLDA, MTAnchor and JointLDA, where NA in PMLDA means that there are missing connections to either the Chinese topics or the English topics

Topic 1: government				Topic 2: markets			
Cb-CLTM		PMLDA		Cb-CLTM		PMLDA	
准备	Government	NA	Government	今年	Rand	利率	NA
预订	Country	NA	Election	国内	Lift	市场	NA
罗慕斯	Tell	NA	Party	上升	Stock	今日	NA
标售	Election	NA	Talk	维持	Drop	成交	NA
议长	Talk	NA	Leader	下降	Cash	天期	NA
政府	Plan	NA	Country	上扬	Restriction	表示	NA
委员会	Minister	NA	Vote	下跌	Shortfall	交易员	NA
国际化	President	NA	State	成为	Result	合约	NA
时分	Opposition	NA	Minister	主要	Begining	台币	NA
军事	News	NA	Meeting	券商	Shrink	央行	NA
桥本龙太郎	Rule	NA	Year	出现	Decline	资金	NA
可兑换	Reporter	NA	Opposition	出口	Formation	人民币	NA
选举	City	NA	President	有限	Msci	上海	NA
投票	Right	NA	Official	进行	Month	下跌	NA
两岸	Party	NA	Rule	受到	Curtail	日电	NA
政情	Nation	NA	Parliament	利多	Contraction	路透社	NA
宣告	Leader	NA	Member	增加	Surplus	可能	NA
英国政府	Reform	NA	Week	现货	Scarcity	认为	NA
国会	Development	NA	Peace	持平	Petroleum	收盘	NA
深发展	Car	NA	Hold	获利	Swap	人士	NA
MTAnchor		JointLDA		MTAnchor		JointLDA	
表示	Say	合约	Election	指数	Percent	年度	Percent
路透社	Government	政府	Party	上升	Rise	利率	Year
日电	State	市场	Government	调整	Price	上升	Rate
经济	Tell	交易员	Say	去年同期	Index	央行	Rise
央行	Country	天期	Vote	销售	Fall	月份	Say
日本	Official	投票	Opposition	路透社	Inflation	台币	Month
德国	Year	成交	Leader	日电	Point	秋天	Price
英国	Budget	人民币	Rate	季节	Newsroom	增加	Growth
美国	Lead	利率	Power	美国	Interest	资金	Quarter
政府	Election	主席	President	数据	Compare	今日	Fall
指出	Plan	今日	Parliament	公布	Consumer	表示	Increase
成长	Minister	国债	Rule	修正	Forecast	市场	Figure
目前	People	下跌	Poll	表示	Week	票券	Expect
官员	Time	大臣	Minister	初值	Yield	准备	Forecast
预期	Party	年度	Year	物价	Growth	拆款	Inflation
东京	Meeting	上海	Lead	第季	Output	天期	Sale
可能	Leader	活跃	Win	工业生产	Figure	买票	Report
周三	Include	支援	Shanghai	零售	Stock	行库	Compare
周二	Group	日电	Campaign	消费者	Measure	银行	Period
香港	News	路透社	State	日本	Turnover	成交	Show

comparisons. Because the articles in MLDoc are from Reuters news, these topics are related to economics issues. Except for PMLDA, which only generates Chinese government topics and only English markets topics, other models generate topics with explainable cross-lingual connections. This phenomenon indicates that PMLDA fails to generate fully cross-lingual topics in MLDoc datasets. Besides, we also observed that there are duplicated topics induced by MTAnchor. Similar to its inferred topics in UM-Corpus, those duplicated topics result in poor diversity (see Fig. 7).

5 Conclusion and future work

This paper has proposed the Cb-CLTM, a cross-lingual topic model, that extends the monolingual LDA by utilizing cross-lingual word embedding for inferencing topics across languages.

We benchmarked Cb-CLTM against four existing cross-lingual topic models, namely PLTM, JointLDA, PMLDA, and MTAnchor, and measured their performance using topic coherence, topic diversity, and document classification as metrics. For the parallel corpora—UM-Corpus and UM-Corpus 25K, we found that Cb-CLTM outperforms the other models in all metrics in most settings, indicating that the semantic relations of words represented by cross-lingual word embedding indeed help construct a better cross-lingual topic model. Cb-CLTM does not require a parallel/comparable corpus and is only dependent on a few bilingual dictionary entries. For a small number of dictionary entries, Cb-CLTM outperforms JointLDA and MTAnchor in inducing coherent topics, generating divergent topics, and learning document representations across languages. Cb-CLTM also generated more coherent topics than PMLDA on the UM-Corpus 25K, which shows its robustness on the small-scale dataset.

However, for the nonparallel corpora—MLDoc En-Zh and MLDoc En-Ja, the themes of articles have very different distributions across languages in original RCV2 corpora, which causes the induced cross-lingual word spaces less isomorphic in structure between the language spaces. With non-aligned cross-lingual word spaces as inputs, the coherent performances of Cb-CLTM are lower on two MLDoc datasets, yet Cb-CLTM still prevails in topic diversity and zero-shot cross-lingual document classification. Hence, the preprocessing steps need further investigation to mitigate this problem. For reference, we attempted to improve the coherence performance by increasing the quality and number of bilingual dictionary entries. While Cb-CLTM still cannot stand out from other comparative models, our strategy did increase the coherence score.

Since it is more challenging to extract common topics across languages from different language families, we first evaluated the English–Chinese corpora and English–Japanese corpora in our experiments. It is part of our future work to see whether Cb-CLTM works well in languages from the same family, such as Indo-European languages.

Last but not least, Cb-CLTM requires a cross-lingual word vector space as a language bridge for linking topics across languages. In this study, we adopted the orthogonal transformation to align two pre-trained monolingual word spaces since it is a well-studied approach and has a solid theoretical foundation [42,46]. Nonetheless, transformer-based language model becomes a rising trend and has shown its capability of learning cross-lingual word representations in recent studies [14,40]. To the best of our knowledge, only few works (e.g., ZeroShotTM [6]) develop cross-lingual topic model based on such a language model. Therefore, incorporating the cross-lingual transformer-based language model is a possible future extension of Cb-CLTM since it could potentially bring more deep relations between languages that may help generating better cross-lingual topics.

Acknowledgements This work has been supported in part by Ministry of Science and Technology in Taiwan, under grant no. MOST 106-2410-H-110-017-MY3.

Declaration

Conflict of interest None.

A Collapsed Gibbs Sampler for Topic Assignment

Notice that we omit $\alpha, \theta, \psi, H^{cs}, l_d$ from distribution $p(z_{d_i} = t|\mathbf{z}_{\neg d_i}, \mathbf{w}; \alpha, \theta, \psi, H^{cs}, l_d)$ and instead use $p(z_{d_i} = t|\mathbf{z}_{\neg d_i}, \mathbf{w})$ for brevity, where $\mathbf{w}$ contains the words of a document.

$$
\begin{aligned}
p(z_{d_i} = t|\mathbf{z}_{\neg d_i}, \mathbf{w}) &\propto p(z_{d_i} = t, w_{d_i}|\mathbf{z}_{\neg d_i}, \mathbf{w}_{\neg d_i}) \\
&= \int p(z_{d_i} = t, w_{d_i}, \theta_d|\mathbf{z}_{\neg d_i}, \mathbf{w}_{\neg d_i})\mathrm{d}\theta_d \\
&= \int p(z_{d_i} = t, \theta_d|\mathbf{z}_{\neg d_i}, \mathbf{w}_{\neg d_i})\mathrm{d}\theta_d \cdot p(w_{d_i}|\mathbf{z}_{\neg d_i}, \mathbf{w}_{\neg d_i}) \\
&\propto \underbrace{\int p(z_{d_i} = t|\theta_d)p(\theta_d|\mathbf{z}_{\neg d_i}, \mathbf{w}_{\neg d_i})\mathrm{d}\theta_d}_{E(\theta_{d,t})\ of\ Dirichlet} \cdot \phi_t(w_{d_i}|\psi_{z_{d_i}=t}; H_{l_d}^{cs}) \\
&= \frac{(N_{\neg d_i}^t + \alpha)}{\sum_{t=1}^T N_d^t + \alpha_t} \cdot \phi_t(w_{d_i}|\psi_t; H_{l_d}^{cs})
\end{aligned}
$$

References

1. Arora S, Ge R, Halpern Y, Mimno D, Moitra A, Sontag D, Wu Y, Zhu M (2013) A practical algorithm for topic modeling with provable guarantees. In: International conference on machine learning, pp 280–288. www.jmlr.org
2. Artetxe M, Labaka G, Agirre E (2018) A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. arXiv:1805.06297
3. Banerjee A, Dhillon IS, Ghosh J, Sra S (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions. J Mach Learn Res 6(Sep):1345–1382
4. Baroni M, Dinu G, Kruszewski G (2014) Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Vol 1: long papers), pp 238–247
5. Batmanghelich K, Saeedi A, Narasimhan K, Gershman S (2016) Nonparametric spherical topic modeling with word embeddings. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics, ACL 2016 (Vol 2: short papers), p 537
6. Bianchi F, Terragni S, Hovy D, Nozza D, Fersini E (2020) Cross-lingual contextualized topic models with zero-shot learning. arXiv:2004.07737
7. Bischof J, Airoldi EM (2012) Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th international conference on machine learning (ICML-12), pp 201–208
8. Blei DM, Jordan MI (2003) Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 127–134
9. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
10. Boyd-Graber J, Blei DM (2009) Multilingual topic models for unaligned text. In: Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, pp 75–82
11. Chang CH, Hwang SY, Xui TH (2018) Incorporating word embedding into cross-lingual topic modeling. In: 2018 IEEE international Congress on big data (BigData Congress). IEEE, pp 17–24
12. Conneau A, Lample G, Ranzato M, Denoyer L, Jégou H (2017) Word translation without parallel data. arXiv:1710.04087
13. Das R, Zaheer M, Dyer C (2015) Gaussian lda for topic models with word embeddings. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Vol 1: long papers), pp 795–804
14. Devlin J, Chang MW, Lee K, Toutanova K (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 (Vol 1: long and short papers). Association for Computational Linguistics, pp 4171–4186
15. Dinu G, Lazaridou A, Baroni M (2014) Improving zero-shot learning by mitigating the hubness problem. arXiv:1412.6568
16. Esuli A, Moreo A, Sebastiani F (2019) Funnelling: a new ensemble method for heterogeneous transfer learning and its application to cross-lingual text classification. ACM Trans Inf Syst: TOIS 37(3):1–30
17. Fujinuma Y, Boyd-Graber J, Paul MJ (2019) A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics, pp 4952–4962
18. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235

19. Hall D, Jurafsky D, Manning CD (2008) Studying the history of ideas using topic models. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 363–371
20. Hao S, Boyd-Graber JL, Paul MJ (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, pp 1–6
21. Hao S, Paul MJ (2018) Learning multilingual topics from incomparable corpora. In: Proceedings of the 27th international conference on computational linguistics, pp 2595–2609
22. Hao S, Paul MJ (2020) An empirical study on crosslingual transfer in probabilistic topic models. Comput Linguist 46(1):1–40
23. Heyman G, Vulić I, Moens MF (2016) C-bilda extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content. Data Min Knowl Discov 30(5):1299–1323
24. Hu Y, Zhai K, Eidelman V, Boyd-Graber J (2014) Polylingual tree-based topic models for translation domain adaptation. In: 52nd annual meeting of the Association for Computational Linguistics, vol 1, pp 1166–1176
25. Jagarlamudi J, Daumé H (2010) Extracting multilingual topics from unaligned comparable corpora. In: European conference on information retrieval. Springer, pp 444–456
26. Jiang D, Tong Y, Song Y (2016) Cross-lingual topic discovery from multilingual search engine query log. ACM Trans Inf Syst: TOIS 35(2):9
27. Klementiev A, Titov I, Bhattarai B (2012) Inducing crosslingual distributed representations of words. In: Proceedings of COLING, vol 2012, pp 1459–1474
28. Lau JH, Newman D, Baldwin T (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, pp 530–539
29. Liu X, Duh K, Matsumoto Y (2015) Multilingual topic models for bilingual dictionary extraction. ACM Trans Asian Low Resour Lang Inf Process 14(3):11
30. Ma T, Nasukawa T (2016) Inverted bilingual topic models for lexicon extraction from non-parallel data. arXiv:1612.07215
31. Mann GS, Mimno D, McCallum A (2006) Bibliometric impact measures leveraging topic analysis. In: Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries. ACM, pp 65–74
32. McCallum A, Wang X, Corrada-Emmanuel A (2007) Topic and role discovery in social networks with experiments on enron and academic email. J Artif Intell Res 30:249–272
33. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
34. Mikolov T, Le QV, Sutskever I (2013) Exploiting similarities among languages for machine translation. arXiv:1309.4168
35. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
36. Mimno D, Wallach HM, Naradowsky J, Smith DA, McCallum A (2009) Polylingual topic models. In: Proceedings of the 2009 conference on empirical methods in natural language processing, EMNLP 2009. ACL, pp 880–889
37. Nguyen DQ, Billingsley R, Du L, Johnson M (2015) Improving topic models with latent feature word representations. Trans Assoc Comput Linguist 3:299–313
38. Ni X, Sun JT, Hu J, Chen Z (2009) Mining multilingual topics from Wikipedia. In: Proceedings of the 18th international conference on World wide web. ACM, pp 1155–1156
39. Peng N, Wang Y, Dredze M (2014) Learning polylingual topic models from code-switched social media documents. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Vol 2: short papers), pp 674–679
40. Pires T, Schlinger E, Garrette D (2019) How multilingual is multilingual BERT? arXiv:1906.01502
41. Reisinger J, Waters A, Silverthorn B, Mooney RJ (2010) Spherical topic models. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 903–910
42. Ruder S, Vulić I, Søgaard A (2019) A survey of cross-lingual word embedding models. J Artif Intell Res 65:569–631
43. Schwenk H, Li X (2018) A corpus for multilingual document classification in eight languages. arXiv:1805.09821
44. Shi B, Lam W, Bing L, 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 (Vol 1: long papers), pp 676–685

45. Sievert C, Shirley K (2014) Ldavis: a method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces, pp 63–70
46. Smith SL, Turban DH, Hamblin S, Hammerla NY (2017) Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv:1702.03859
47. Sra S (2012) A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of i s (x). Comput Stat 27(1):177–190
48. Srivastava A, Sutton C (2017) Autoencoding variational inference for topic models. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings. www.OpenReview.net
49. Stajner T, Mladenic D (2019) Cross-lingual document similarity estimation and dictionary generation with comparable corpora. Knowl Inf Syst 58(3):729–743
50. Tamura A, Sumita E (2016) Bilingual segmented topic model. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (Vol 1: long papers), pp 1266–1276
51. Tian L, Wong DF, Chao LS, Quaresma P, Oliveira F, Yi L (2014) Um-corpus: a large English–Chinese parallel corpus for statistical machine translation. In: LREC, pp 1837–1842
52. Vulić I, De Smet W, Moens MF (2013) Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. Inf Retr 16(3):331–368
53. Wei X, Croft WB (2006) Lda-based document models for ad-hoc retrieval. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp 178–185
54. Yang W, Boyd-Graber J, 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). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1243–1248
55. Yuan M, Van Durme B, Boyd-Graber J (2018) Multilingual anchoring: interactive topic modeling and alignment across languages. In: Advances in neural information processing systems, vol 2018, pp 8653–8663
56. Zhang M, Liu Y, Luan H, Sun M (2017) Adversarial training for unsupervised bilingual lexicon induction. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Vol 1: long papers), pp 1959–1970
57. Zhong S, Ghosh J (2005) Generative model-based document clustering: a comparative study. Knowl Inf Syst 8(3):374–384
58. Zhou G, Zhu Z, He T, Hu XT (2016) Cross-lingual sentiment classification with stacked autoencoders. Knowl Inf Syst 47(1):27–44

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chia-Hsuan Chang is a Ph.D. candidate of the Department of Information Management at National Sun Yat-sen University. His research interests include information retrieval, health informatics, and text mining, especially in cross-lingual text analysis.

San-Yih Hwang received a Ph.D. degree in computer science from the University of Minnesota, Minneapolis, in 1994. He joined the Department of Information Management at National Sun Yat-sen University (NSYSU), Taiwan, in 1995 and is presently a professor and dean of the College of Management, NSYSU. His research interests lie in technical research in information systems, including text mining, recommendation, workflow management, and service computing. He has published papers in prestigious journals such as INFORMS Journal on Computing, IEEE Transactions on Service Computing, and VLDB Journal.