Analyzing Bayesian Crosslingual Transfer in Topic Models

Shudong Hao
Boulder, CO
shudonghao@gmail.com

Michael J. Paul
Information Science
University of Colorado
Boulder, CO
mpaul@colorado.edu

Abstract

We introduce a theoretical analysis of crosslingual transfer in probabilistic topic models. By formulating posterior inference through Gibbs sampling as a process of language transfer, we propose a new measure that quantifies the loss of knowledge across languages during this process. This measure enables us to derive a PAC-Bayesian bound that elucidates the factors affecting model quality, both during training and in downstream applications. We provide experimental validation of the analysis on a diverse set of five languages, and discuss best practices for data collection and model design based on our analysis.

1 Introduction

Crosslingual learning is an important area of natural language processing that has driven applications including text mining in multiple languages (Ni et al., 2009; Smet and Moens, 2009), cultural difference detection (Gutiérrez et al., 2016), and various linguistic studies (Shutova et al., 2017; Barrett et al., 2016). Crosslingual learning methods generally extend monolingual algorithms by using various multilingual resources. In contrast to traditional high-dimensional vector space models, modern crosslingual models tend to rely on learning low-dimensional word representations that are more efficient and generalizable.

A popular approach to representation learning comes from the word embedding community, in which words are represented as vectors in an embedding space shared by multiple languages (Ruder et al., 2018; Faruqui and Dyer, 2014; Klementiev et al., 2012). Another direction is from the topic modeling community, where words are projected into a probabilistic topic space (Ma and Nasukawa, 2017; Jagarlamudi and III, 2010). While formulated differently, both types of models apply the same principles—low-dimensional vectors exist in a shared crosslingual space, wherein vector representations of similar concepts across languages (*e.g.,* "dog" and "hund") should be nearby in the shared space.

To enable crosslingual representation learning, knowledge is transferred from a source language to a target language, so that representations have similar values across languages. In this study, we will focus on probabilistic topic models, and "knowledge" refers to a word's probability distribution over topics. Little is known about the characteristics of crosslingual knowledge transfer in topic models, and thus this paper provides an analysis, both theoretical and empirical, of crosslingual transfer in multilingual topic models.

1.1 Background and Contributions

Multilingual Topic Models Given a multilingual corpus $D^{(1,\dots,L)}$ in languages $\ell = 1, \dots, L$ as inputs, a multilingual topic model learns K topics. Each multilingual topic $k^{(1,\dots,L)}$ ($k = 1, \dots, K$), is defined as an L-dimensional tuple $\left(\phi_k^{(1)}, \dots, \phi_k^{(L)}\right)$, where $\phi_k^{(\ell)}$ is a multinomial distribution over the vocabulary $V^{(\ell)}$ in language ℓ. From a human's perspective, a multilingual topic $k^{(1,\dots,L)}$ can be interpreted by looking at the word types that have C highest probabilities in $\phi_k^{(\ell)}$ for each language ℓ. C here is called *cardinality* of the topic. Thus, a multilingual topic can loosely be thought of as a group of word lists where each language ℓ has its own version of the topic.

Multilingual topic models are generally extended from Latent Dirichlet Allocation (Blei et al., 2003, LDA). Though many variations have been proposed, the underlying structures of multilingual topic models are similar. These models require either a parallel/comparable corpus in multiple languages, or word translations from a

Proceedings of NAACL-HLT 2019, pages 1551–1565
Minneapolis, Minnesota, June 2 - June 7, 2019.

dictionary. One of the most popular models is the polylingual topic model (Mimno et al., 2009, PLTM), where comparable document pairs share distributions over topics θ, while each language ℓ has its own distributions $\{\phi_k^{(\ell)}\}_{k=1}^K$ over the vocabulary $V^{(\ell)}$. By re-marginalizing the estimations $\{\widehat{\phi}_k^{(\ell)}\}_{k=1}^K$, we obtain word representations $\widehat{\varphi}^{(w)} \in \mathbb{R}^K$ for each word w, where $\widehat{\varphi}_k^{(w)} = \Pr(z_w = k|w)$, *i.e.,* the probability of topic k given a word type w.

Crosslingual Transfer Knowledge transfer through crosslingual representations has been studied in prior work. Smet and Moens (2009) and Heyman et al. (2016) show empirically how document classification using topic models implements the ideas of crosslingual transfer, but to date there has been no theoretical framework to analyze this transfer process in detail.

In this paper, we describe two types of transfer—on-site and off-site—based on the nature of where and how the transfer takes place. We refer to transfer that happens while training topic models (*i.e.,* during representation learning) as *on-site*. Once we obtain the low-dimensional representations, they can be used for downstream tasks. We refer to transfer in this phase as *off-site*, since the crosslingual tasks are usually detached from the process of representation learning.

Contributions Our study provides a theoretical analysis of crosslingual transfer learning in topic models. Specifically, we first formulate on-site transfer as circular validation, and derive an upper bound based on PAC-Bayesian theories (Section 2). The upper bound explicitly shows the factors that can affect knowledge transfer. We then move on to off-site transfer, and focus on crosslingual document classification as a downstream task (Section 3). Finally, we show experimentally that the on-site transfer error can have impact on the performance of downstream tasks (Section 4).

2 On-Site Transfer

On-site transfer refers to the training procedure of multilingual topic models, which usually involves Bayesian inference techniques such as variational inference and Gibbs sampling. Our work focuses on the analysis of collapsed Gibbs sampling (Griffiths and Steyvers, 2004), showing how knowledge is transferred across languages and how a topic space is formed through the sampling process. To this end, we first describe a specific formulation of knowledge transfer in multilingual topic models as a starting point of our analysis (Section 2.1). We then formulate Gibbs sampling as circular validation and quantify a loss during this phase (Section 2.2). This formulation leads us to a PAC-Bayesian bound that explicitly shows the factors that affect the crosslingual training (Section 2.3). Lastly, we look further into different transfer mechanisms in more depth (Section 2.4).

2.1 Transfer through Priors

Priors are an important component in Bayesian models like PLTM. In the original generative process of PLTM, each comparable document pair (d_S, d_T) in the source and target languages (S, T) is generated by the same multinomial $\theta \sim \mathrm{Dir}(\alpha)$.

Hao and Paul (2018) showed that knowledge transfer across languages happens through priors. Specifically, assume the source document is generated from $\theta^{(d_S)} \sim \mathrm{Dir}(\alpha)$, and has a sufficient statistics $\mathbf{n}_{d_S} \in \mathbb{N}^K$ where each cell $n_{k|d_S}$ is the count of topic k in document d_S. When generating the corresponding comparable document d_T, the Dirichlet prior of the distribution over topics $\theta^{(d_T)}$, instead of a symmetric α, is parameterized by $\alpha + \mathbf{n}_{d_S}$. This formulation yields the same posterior estimation as the original joint model and is the foundation of our analysis in this section.

To see this transfer process more clearly, we look closer to the conditional distributions during sampling, and take PLTM as an example. When sampling a token in target language x_T, the Gibbs sampler calculates a conditional distribution $\mathcal{P}_{x_T}$ over K topics, where a topic k is randomly drawn and assigned to x_T (denoted as z_{x_T}). Assume the token x_T is in document d_T whose comparable document in the source language is d_S. The conditional distribution for x_T is

$$\begin{aligned} \mathcal{P}_{x,k} &= \Pr(z_x = k; \mathbf{w}_-, \mathbf{z}_-) \qquad (1) \\ &\propto \left(n_{k|d_T} + n_{k|d_S} + \alpha\right) \cdot \frac{n_{w_T|k} + \beta}{n_{\cdot|k} + V^{(T)}\beta}, \end{aligned}$$

where the quantity $n_{k|d_S}$ is added and thus transferred from the source document. Thus, the calculation of $\mathcal{P}_x$ incorporates the knowledge transferred from the other language.

Now that we have identified the transfer process, we provide an alternative view of Gibbs sampling, *i.e.,* circular validation, in the next section.

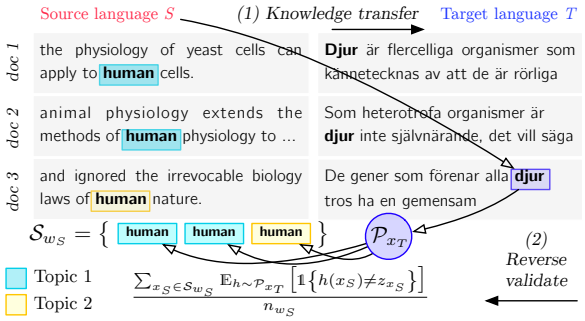

Figure 1: The Gibbs sampler is sampling the token "djur" (animal). Using the classifier h_k sampled from its conditional distribution $\mathcal{P}_{x_T}$, circular validation evaluates h_k on all the tokens of type "human".

2.2 Circular Validation

Circular validation (or reverse validation) was proposed by Zhong et al. (2010) and Bruzzone and Marconcini (2010) in transfer learning. Briefly, a learning algorithm $\mathcal{A}$ is trained on both source and target datasets (D_S and D_T), where the source is labeled and target is unlabeled. After predicting the labels for the target dataset using $\mathcal{A}$ (predictions denoted as $\mathcal{A}(D_T)$), circular validation trains another algorithm $\mathcal{A}'$ in the reverse direction, *i.e.,* uses $\mathcal{A}(D_T)$ and D_T as the labeled dataset and D_S as the unlabeled dataset. The error is then evaluated on $\mathcal{A}'(D_S)$. This "train-predict-reverse-repeat" cycle has a similar flavor to the iterative manner of Gibbs sampling, which inspires us to look at the sampling process as circular validation.

Figure 1 illustrates this process. Suppose the Gibbs sampler is currently sampling x_T of word type w_T in target language T. As discussed for Equation (1), the calculation of the conditional distribution $\mathcal{P}_{x_T}$ incorporates the knowledge transferred from the source language. We then treat the process of drawing a topic from $\mathcal{P}_{x_T}$ as a *classification* of the token x_T. Let $\mathcal{P}_{x_T}$ be a distribution over K unary *classifiers*, $\{h_k\}_{k=1}^K$, and the k-th classifier labels the token as topic k with a probability of one:

$$h_k \sim \mathcal{P}_{x_T}, \text{ and } \Pr(z_{x_T} = k; h_k) = 1. \quad (2)$$

This process is repeated between the two languages until the Markov chain converges.

The training of topic models is unsupervised, *i.e.,* there is no ground truth for labeling a topic, which makes it difficult to analyze the effect of transfer learning. Thus, after calculating $\mathcal{P}_{x_T}$, we take an additional step called *reverse validation*, where we design and calculate a measure—circular validation loss—to quantify the transfer.

Definition 1 (Circular validation loss, CVL)**.** *Let $\mathcal{S}_w$ be the set containing all the tokens of type w throughout the whole training corpus, and call it the sample of w. Given a bilingual word pair (w_T, w_S) where w_T is in target language T while w_S in source S, let $\mathcal{S}_{w_T}$ and $\mathcal{S}_{w_S}$ be the samples for the two types respectively, and n_{w_T} and n_{w_S} the sizes of them. The empirical circular validation score ($\widehat{\text{CVL}}$) is defined as*

$$\begin{aligned}
\widehat{\text{CVL}}(w_T, w_S) &= \frac{1}{2} \mathop{\mathbb{E}}_{x_S, x_T} \left[\widehat{\mathcal{L}}(x_T, w_S) + \widehat{\mathcal{L}}(x_S, w_T) \right], \\
\widehat{\mathcal{L}}(x_T, w_S) &= \frac{1}{n_{w_S}} \sum_{x_S \in \mathcal{S}_{w_S}} \mathbb{E}_{h \sim \mathcal{P}_{x_T}} \left[\mathbb{1}\{h(x_S) \neq z_{x_S}\} \right] \\
&= \frac{1}{n_{w_S}} \sum_{x_S \in \mathcal{S}_{w_S}} \left(1 - \mathcal{P}_{x_T, z_{x_S}} \right),
\end{aligned}$$

where $\mathcal{P}_{x_T,k}$ is the conditional probability of token x_T assigned with topic k. Taking expectations over all tokens x_S and x_T, we have general CVL*:*

$$\begin{aligned}
\text{CVL}(w_T, w_S) &= \frac{1}{2} \mathop{\mathbb{E}}_{x_S, x_T} \left[\mathcal{L}(x_T, w_S) + \mathcal{L}(x_S, w_T) \right], \\
\mathcal{L}(x_T, w_S) &= \mathbb{E}_{x_S} \mathbb{E}_{h \sim \mathcal{P}_{x_T}} \left[\mathbb{1}\{h(x_S) \neq z_{x_S}\} \right].
\end{aligned}$$

When sampling a token x_T, we still follow the two-step process as in Equation (2), but instead of labeling x_T itself, we use its conditional $\mathcal{P}_{x_T}$ to label the entire sample of a word type w_S in the source language. Since all the topic labels for the source language are fixed, we take them as the assumed "correct" labelings, and compare x_S's labels and the predictions from $\mathcal{P}_{x_T}$. This is the intuition behind CVL.

Note that the choice of word types w_T and w_S to calculate $\widehat{\text{CVL}}$ is arbitrary. However, $\widehat{\text{CVL}}$ is only meaningful when the two word types are semantically related, such as word translations, because those word pairs are where the knowledge transfer takes place. On the other hand, the Gibbs sampler does not calculate this $\widehat{\text{CVL}}$ explicitly, and thus adding reverse validation step does not affect the training of the model. It does, however, help us to expose and analyze the knowledge transfer mechanism. In fact, as we show in the next theorem, sampling is also a procedure of optimizing $\widehat{\text{CVL}}$.

Theorem 1. *Let $\widehat{\text{CVL}}^{(t)}(w_T, w_S)$ be the empirical circular validation loss of any bilingual word pair at iteration t of Gibbs sampling. Then $\widehat{\text{CVL}}^{(t)}(w_T, w_S)$ converges as $t \to \infty$.*

Proof. See Appendix. □

2.3 PAC-Bayes View

A question following the formulation of $\widehat{\text{CVL}}$ is, what factors could lead to better transfer during this process, particularly for semantically related words? To answer this, we turn to theory that bounds the performance of classifiers and apply this theory to this formulation of topic sampling as classification.

The PAC-Bayes theorem was introduced by McAllester (1999) to bound the performance of Bayes classifiers. Given a hypothesis set $\mathcal{H}$, the majority vote classifier (or Bayes classifier) uses every hypothesis $h \in \mathcal{H}$ to perform binary classification on an example $\mathbf{x}$, and uses the majority output as the final prediction. Since minimizing the error by Bayes classifier is NP-hard, an alternative way is to use a *Gibbs classifier* as approximation. The Gibbs classifier first draws a hypothesis $h \in \mathcal{H}$ according to a posterior distribution over $\mathcal{H}$, and then uses this hypothesis to predict the label of an example $\mathbf{x}$ (Germain et al., 2012). The generalization loss of this Gibbs classifier can be bounded as follows.

Theorem 2 (PAC-Bayes theorem, McAllester (1999))**.** *Let $\mathcal{P}$ be a posterior distribution over all classifiers $h \in \mathcal{H}$, and Q a prior distribution. With a probability at least $1-\delta$, we have*

$$\mathcal{L} \;\le\; \widehat{\mathcal{L}} + \sqrt{\frac{1}{2n}\left(\mathrm{KL}\left(\mathcal{P}||Q\right) + \ln\frac{2\sqrt{n}}{\delta}\right)},$$

where $\mathcal{L}$ and $\widehat{\mathcal{L}}$ are the general loss and the empirical loss on a sample of size n.

In our framework, a token x_T provides a posterior $\mathcal{P}_{x_T}$ over K classifiers. The loss $\widehat{\mathcal{L}}(x_T, w_S)$ is then calculated on a sample of $\mathcal{S}_{w_S}$ in language S. The following theorem shows that for a bilingual word pair (w_T, w_S), the general CVL can be bounded with several quantities.

Theorem 3. *Given a bilingual word pair (w_T, w_S), with probability at least $1-\delta$, the following bound holds:*

$$\text{CVL}(w_T, w_S) \;\le\; \widehat{\text{CVL}}(w_T, w_S) \;+ \tag{3}$$
$$\frac{1}{2}\sqrt{\frac{1}{n}\left(\mathrm{KL}_{w_T} + \mathrm{KL}_{w_S} + 2\ln\frac{2}{\delta}\right) + \frac{\ln n^\star}{n}},$$
$$n = \min\left\{n_{w_T}, n_{w_S}\right\}, \quad n^\star = \max\left\{n_{w_T}, n_{w_S}\right\}.$$

For brevity we use KL_w to denote $\mathrm{KL}(\mathcal{P}_x||Q_x)$, where $\mathcal{P}_x$ is the conditional distribution from Gibbs sampling of token x with word type w that gives highest loss $\widehat{\mathcal{L}}(x, w)$, and Q_x a prior.

Proof. See Appendix. □

2.4 Multilevel Transfer

Recall that knowledge transfer happens through priors in topic models (Section 2.1). Because the KL-divergence terms in Theorem 3 include this prior Q, we can use this theorem to analyze the transfer mechanisms more concretely.

The conditional distribution for sampling a topic z_x for a token x during sampling can be factorized into document-topic and topic-word levels:

$$\begin{aligned}
\mathcal{P}_{x,k} &= \Pr\left(z_x = k | w_x = w, \mathbf{w}_-, \mathbf{z}_-\right) \\
&= \Pr\left(z_x = k|\mathbf{z}_-\right) \cdot \Pr\left(w_x = w | z_x = k, \mathbf{w}_-, \mathbf{z}_-\right) \\
&\propto \underbrace{\Pr\left(z_x = k|\mathbf{z}_-\right)}_{\text{document level}} \cdot \underbrace{\Pr\left(z_x = k|w_x = w, \mathbf{w}_-\right)}_{\text{word level}} \\
&\triangleq \mathcal{P}_{\theta,x,k} \cdot \mathcal{P}_{\varphi,x,k}, \\
\mathcal{P}_x &\triangleq \mathcal{P}_{\theta,x} \otimes \mathcal{P}_{\varphi,x},
\end{aligned}$$

where $\otimes$ is element-wise multiplication. Thus, we have the following inequality:

$$\begin{aligned}
\mathrm{KL}\left(\mathcal{P}_x||Q_x\right) &= \mathrm{KL}\left(\mathcal{P}_{\theta,x} \otimes \mathcal{P}_{\varphi,x}||Q_{\theta,x} \otimes Q_{\varphi,x}\right) \\
&\le \mathrm{KL}\left(\mathcal{P}_{\theta,x}||Q_{\theta,x}\right) + \mathrm{KL}\left(\mathcal{P}_{\varphi,x}||Q_{\varphi,x}\right),
\end{aligned}$$

and the KL-divergence term in Theorem 3 is simply the sum of the KL-divergences between the conditional and prior distributions on all levels.

Recall that PLTM transfers knowledge at the document level, through $Q_{\theta,x}$, by linking document translations together (Equation (1)). Assume the current token x is from a target document linked to a document d_S in the source language. Then the prior for $\mathcal{P}_{\theta,x}$ is $\widehat{\theta}^{(d_S)}$, *i.e.,* the normalized empirical distribution over topics of d_S.

Since the words are generated *within* each language under PLTM, *i.e.,* $\phi_k^{(S)}$ is irrelevant to $\phi_k^{(T)}$, no transfer happens at the word level. In this case, $Q_{\varphi,x}$, the prior for $\mathcal{P}_{\varphi,x}$, is simply a K-dimensional uniform distribution $\mathcal{U}$. Then:

$$\begin{aligned}
\mathrm{KL}_w &\le \mathrm{KL}\left(\mathcal{P}_{\theta,x}||\widehat{\theta}^{(d_S)}\right) + \mathrm{KL}\left(\mathcal{P}_{\varphi,x}||\mathcal{U}\right) \\
&= \underbrace{\mathrm{KL}\left(\mathcal{P}_{\theta,x}||\widehat{\theta}^{(d_S)}\right)}_{\text{crosslingual entropy}} + \log K - \underbrace{H(\mathcal{P}_{\varphi,x})}_{\text{monolingual entropy}}.
\end{aligned}$$

Thus, at levels where transfer happens (document- or word-level), a low crosslingual entropy is preferred, to offset the impact of monolingual entropy where no transfer happens.

Most multilingual topic models are generative admixture models in which the conditional probabilities can be factorized into different levels, thus KL-divergence term in Theorem 3 can be decomposed and analyzed in the same way as in this section for models that have transfer at other levels, such as Hao and Paul (2018), Heyman et al. (2016), and Hu et al. (2014). For example, if a model has word-level transfer, *i.e.,* the model assumes that word translations share the same distributions, we have a KL-divergence term as,

$$\begin{aligned}\mathrm{KL}_w &\leq \mathrm{KL}\left(\mathcal{P}_{\varphi,x}||\widehat{\varphi}^{(w_S)}\right) + \mathrm{KL}(\mathcal{P}_{\theta,x}||\mathcal{U}) \\ &= \mathrm{KL}\left(\mathcal{P}_{\varphi,x}||\widehat{\varphi}^{(w_S)}\right) + \log K - H(\mathcal{P}_{\theta,x}),\end{aligned}$$

where w_S is the word translation to word w.

3 Off-Site Transfer

Off-site transfer refers to language transfer that happens while applying trained topic models to downstream crosslingual tasks such as document classification. Because transfer happens using the trained representations, the performance of off-site transfer heavily depends on that of on-site transfer. To analyze this problem, we focus on the task of crosslingual document classification.

In crosslingual document classification, a *document classifier*, h, is trained on documents from one language, and h is then applied to documents from another language. Specifically, after training bilingual topic models, we have K bilingual word distributions $\{\widehat{\phi}_k^{(S)}\}_{k=1}^K$ and $\{\widehat{\phi}_k^{(T)}\}_{k=1}^K$. These two distributions are used to infer document-topic distributions $\widehat{\theta}$ on unseen documents in the test corpus, and each document is represented by the inferred distributions. A document classifier is then trained on the $\widehat{\theta}$ vectors as features in source language S and tested on the target T.

We aim to show how the generalization risk on target languages T, denoted as $R_T(h)$, is related to the training risk on source languages S, $\widehat{R_S}(h)$. To differentiate the loss and classifiers in this section from those in Section 2, we use the term "risk" here, and h refers to the document classifiers, not the topic labeling process by the sampler.

3.1 Languages as Domains

Classic learning theory requires training and test sets to come from the same distribution $\mathcal{D}$, *i.e.,* $(\theta, y) \sim \mathcal{D}$, where θ is the document representation (features) and y the document label (Valiant, 1984). In practice, however, corpora in S and T may be sampled from different distributions, *i.e.,* $D^{(S)} = \{(\widehat{\theta}^{(d_S)}, y)\} \sim \widehat{\mathcal{D}}^{(S)}$ and $D^{(T)} = \{(\widehat{\theta}^{(d_T)}, y)\} \sim \widehat{\mathcal{D}}^{(T)}$. We refer to these distributions as *document spaces*. To relate $R_T(h)$ and $\widehat{R_S}(h)$, therefore, we have to take their distribution bias into consideration. This is often formulated as a problem of domain adaptation, and here we can formulate this such that each language is treated as a "domain".

We follow the seminal work by Ben-David et al. (2006), and define $\mathcal{H}$-distance as follows.

Definition 2 ($\mathcal{H}$-distance, Ben-David et al. (2006))**.** *Let $\mathcal{H}$ be a symmetric hypothesis space,* i.e., *for every hypothesis $h \in \mathcal{H}$ there exists its counterpart $1 - h \in \mathcal{H}$. We let $m = \left|D^{(S)}\right| + \left|D^{(T)}\right|$, the total size of test corpus. The $\mathcal{H}$-distance between $\widehat{\mathcal{D}}^{(S)}$ and $\widehat{\mathcal{D}}^{(T)}$ is defined as*

$$\begin{aligned}&\frac{1}{2}\widehat{d}_{\mathcal{H}}\left(\widehat{\mathcal{D}}^{(S)}, \widehat{\mathcal{D}}^{(T)}\right) \\ &= \max_{h\in\mathcal{H}} \frac{1}{m} \sum_{\ell\in\{S,T\}} \sum_{\mathbf{x}_d : h(\mathbf{x}_d)=\ell} \mathbb{1}\left\{\mathbf{x}_d \in D^{(\ell)}\right\},\end{aligned}$$

where $\mathbf{x}_d$ is the representation for document d, and $h(\mathbf{x}_d)$ outputs the language of this document.

This distance measures how identifiable the languages are based on their representations. If source and target languages are from entirely different distributions, a classifier can easily identify language-specific features, which could affect performance of the document classifiers.

With $\mathcal{H}$-distances, we have a measure of the "distance" between the two distributions $\widehat{\mathcal{D}}^{(S)}$ and $\widehat{\mathcal{D}}^{(T)}$. We state the following theorem from domain adaptation theory.

Theorem 4 (Ben-David et al. (2006))**.** *Let m be the corpus size of the source language,* i.e., *$m = \left|D^{(S)}\right|$, c the VC-dimension of document classifiers $h \in \mathcal{H}$, and $\widehat{d}_{\mathcal{H}}\left(\widehat{\mathcal{D}}^{(S)}, \widehat{\mathcal{D}}^{(T)}\right)$ the $\mathcal{H}$-distance between two languages in the document space. With probability at least $1 - \delta$, we have the following bound,*

$$\begin{aligned}R_T(h) \leq\ & \widehat{R}_S(h) + \widehat{d}_{\mathcal{H}}\left(\widehat{\mathcal{D}}^{(S)}, \widehat{\mathcal{D}}^{(T)}\right) + \widehat{\lambda} + \\ & \sqrt{\frac{4}{m}\left(c\log\frac{2em}{c} + \log\frac{4}{\delta}\right)}, \qquad (4)\end{aligned}$$

$$\widehat{\lambda} = \min_{h\in\mathcal{H}} \widehat{R}_S(h) + \widehat{R}_T(h). \qquad (5)$$

The term $\widehat{\lambda}$ in Theorem 4 defines a *joint risk*, *i.e.,* the training error on both source and target documents. This term usually cannot be estimated in practice since the labels for target documents are unavailable. However, we can still calculate this term for the purpose of analysis.

The theorem shows that the crosslingual classification risk is bounded by two critical components: the $\mathcal{H}$-distance, and the joint risk $\widehat{\lambda}$. Interestingly, these two quantities are based on the same set of features with different labeling rules: for $\mathcal{H}$-distance, the label for each instance is its language, while $\widehat{\lambda}$ uses the actual document label. Therefore, a better bound requires the consistency of features across languages, both in language and document labelings.

3.2 From Words to Documents

Since consistency of features depends on the document representations $\widehat{\theta}$, we need to trace back to the upstream training of topic models and show how the errors propagate to the formation of document representations. Thus, we first show the relations between $\widehat{\text{CVL}}$ and word representations $\widehat{\varphi}$ in the following lemma.

Lemma 1. *Given any bilingual word pair* (w_T, w_S)*, let* $\widehat{\varphi}^{(w)}$ *denote the distribution over topics of word type* w*. Then we have,*

$$1 - \widehat{\varphi}^{(w_T)\top} \cdot \widehat{\varphi}^{(w_S)} \quad \leq \quad \widehat{\text{CVL}}(w_T, w_S).$$

Proof. See Appendix. □

We need to connect the word representations $\widehat{\varphi}$, which are central to on-site transfer, to the document representations $\widehat{\theta}$, which are central to off-site transfer. To do this, we make an assumption that the inferred distribution over topics $\widehat{\theta}^{(d)}$ for each test document d is a weighted average over all word vectors, *i.e.,* $\widehat{\theta}^{(d)} \propto \sum_w f_w^d \cdot \widehat{\varphi}^{(w)}$, where f_w^d is the normalized frequency of word w in document d (Arora et al., 2013). When this assumption holds, we can bound the similarity of document representations $\widehat{\theta}^{(d_S)}$ and $\widehat{\theta}^{(d_T)}$ in terms of word representations and hence their $\widehat{\text{CVL}}$.

Theorem 5. *Let* $\widehat{\theta}^{(d_S)}$ *be the distribution over topics for document* d_S *(similarly for* d_T*),* $F(d_S, d_T) = \left(\sum_{w_S} {f_{w_S}^{d_S}}^2 \cdot \sum_{w_T} {f_{w_T}^{d_T}}^2\right)^{\frac{1}{2}}$ *where* f_w^d *is the* normalized *frequency of word* w *in document* d*, and* K *the number of topics. Then*

$$\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)} \leq F(d_S, d_T) \cdot \sqrt{K \cdot \sum_{w_S, w_T} \left(\widehat{\text{CVL}}(w_T, w_S) - 1\right)^2}.$$

Proof. See Appendix. □

This provides a spatial connection between document pairs and word pairs they have. Many kernalized classifiers such as support vector machines (SVM) explicitly use this inner product in the dual optimization objective (Platt, 1998). Since the inner product is directly related to the cosine similarity, Theorem 5 indicates that if two documents are spatially close, their inner product should be large, and thus the $\widehat{\text{CVL}}$ of all word pairs they share should be small. In an extreme case, if $\widehat{\text{CVL}}(w_T, w_S) = 1$ for all the bilingual word pairs appearing in document pair (d_S, d_T), then $\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)} = 0$, meaning the two documents are orthogonal and tend to be irrelevant topically.

With upstream training discussed in Section 2, we see that $\widehat{\text{CVL}}$ has an impact on the consistency of features across languages. A low $\widehat{\text{CVL}}$ indicates that the transfer from source to target is sufficient in two ways. First, languages share similar distributions, and therefore, it is harder to distinguish languages based on their distributions. Second, if there exists a latent mapping from a distribution to a label, it should produce similar labeling on both source and target data since they are similar. These two aspects correspond to the language $\mathcal{H}$-distance and joint risk $\widehat{\lambda}$ in Theorem 4.

4 Experiments

We experiment with five languages: Arabic (AR, Semitic), German (DE, Germanic), Spanish (ES, Romance), Russian (RU, Slavic), and Chinese (ZH, Sinitic). In the first two experiments, we pair each with English (EN, Germanic) and train PLTM on each language pair individually.

Training Data For each language pair, we use a subsample of 3,000 Wikipedia comparable documents, *i.e.,* 6,000 documents in total. We set $K = 50$, and train PLTM with default hyperparameters (McCallum, 2002). We run each experiment five times and average the results.

Test Data For experiments with document classification, we use Global Voices (GV) in all five language pairs as test sets. Each document in this dataset has a "categories" attribute that can be used

as the document label. In our classification experiments, we use *culture*, *technology*, and *education* as the labels to perform multiclass classification.

Evaluation To evaluate topic qualities, we use Crosslingual Normalized Pointwise Mutual Information (Hao et al., 2018, CNPMI), an intrinsic metric of crosslingual topic coherence. For any bilingual word pair (w_T, w_S),

$$\text{CNPMI}(w_T, w_S) = -\frac{\log \frac{\Pr(w_T, w_S)}{\Pr(w_T)\Pr(w_S)}}{\log \Pr(w_T, w_S)}, \quad (6)$$

where $\Pr(w_T, w_S)$ is the occurrence of w_T and w_S appearing in the same pair of comparable documents. We use 10,000 Wikipedia comparable document pairs outside PLTM training data for each language pair to calculate CNPMI scores. All datasets are publicly available at `http://opus.nlpl.eu/` (Tiedemann, 2012). Additional details of our datasets and experiment setup can be found in the appendix.

4.1 Sampling as Circular Validation

Our first experiment shows how $\widehat{\text{CVL}}$ changes over time during Gibbs sampling. According to the definition, the arguments of $\widehat{\text{CVL}}$ can include any bilingual word pairs; however, we suggest that it should be calculated specifically among word pairs that are expected to be related (and thus enable transfer). In our experiments, we select word pairs in the following way.

Recall that the output of a bilingual topic model is K topics, where each language has its own distribution. For each topic k, we can calculate $\widehat{\text{CVL}}(w_S, w_T)$ such that w_S and w_T belong to the same topic (*i.e.,* are in the top C most probable words in that topic), from the two languages, respectively. Using a cardinality C for each of the K topics, we have in total $C^2 \times K$ bilingual word pairs in the calculation of $\widehat{\text{CVL}}$.

At certain iterations, we collect the topic words as described above with cardinality $C = 5$, and calculate $\widehat{\text{CVL}}(w_T, w_S)$, $\text{CNPMI}(w_T, w_S)$, and the error term (the $\frac{1}{2}\sqrt{\cdots}$ term in Theorem 3) of all the bilingual word pairs. In the middle panel of Figure 2, $\widehat{\text{CVL}}$ over all word pairs from topic words is decreasing as sampling proceeds and becomes stable by the end of sampling. On the other hand, the correlations between CNPMI and $\widehat{\text{CVL}}$ are constantly decreasing. The negative correlations between $\widehat{\text{CVL}}$ and CNPMI implies that lower $\widehat{\text{CVL}}$ is associated with higher topic quality, since higher-quality topic has higher CNPMI but lower $\widehat{\text{CVL}}$.

4.2 What the PAC-Bayes Bound Shows

Theorem 3 provides insights into how knowledge is transferred during sampling and the factors that could affect this process. We analyze this bound from two aspects, the size of the training data (corresponding to $\frac{\ln n^\star}{n}$ term) and model assumptions (as in the crosslingual entropy terms).

4.2.1 Training Data: Downsampling

One factor that could affect $\widehat{\text{CVL}}$, according to Theorem 3, is the balance of tokens of a word pair. In an extreme case, if a word type w_S has only one token, while another word type w_T has a large number of tokens, the transfer from w_S to w_T is negligible. In this experiment, we will test if increasing the ratio term $\frac{\ln n^\star}{n}$ in the corpus lowers the performance of crosslingual transfer learning.

To this end, we specify a *sample rate* $\rho = 0.2, 0.4, 0.6, 0.8,$ and 1.0. For each word pair (w_T, w_S), we calculate n as in the ratio term $\frac{\ln n^\star}{n}$, and remove $(1-\rho)\cdot n$ tokens from the corpus (rounded to the nearest integer). Smaller ρ removes more tokens from the corpus and thus yields a larger ratio term on average.

We use a dictionary from Wiktionary to collect word pairs, where each word pair (w_S, w_T) is a translation pair. Figure 3 shows the results of downsampling using these two methods. Decreasing the sample rate ρ lowers the topic qualities. This implies that although PLTM can process comparable corpora, which need not be exact translations, one still needs to be careful about the token balance between linked document pairs.

For many low-resource languages, the target language corpus is much smaller than the source corpus, so the effect of this imbalance is important to be aware of. This is an important issue when choosing comparable documents, and Wikipedia is an illustrative example. Although one can collect comparable documents via Wikipedia's interlanguage links, articles under the same title but in different languages can have very large variations on document length, causing the imbalance of samples $\frac{\ln n^\star}{n}$, and thus potentially suboptimal performance of crosslingual training.

4.2.2 Model Assumptions

Recall that the crosslingual entropy term can be decomposed into different levels, *e.g.,* document

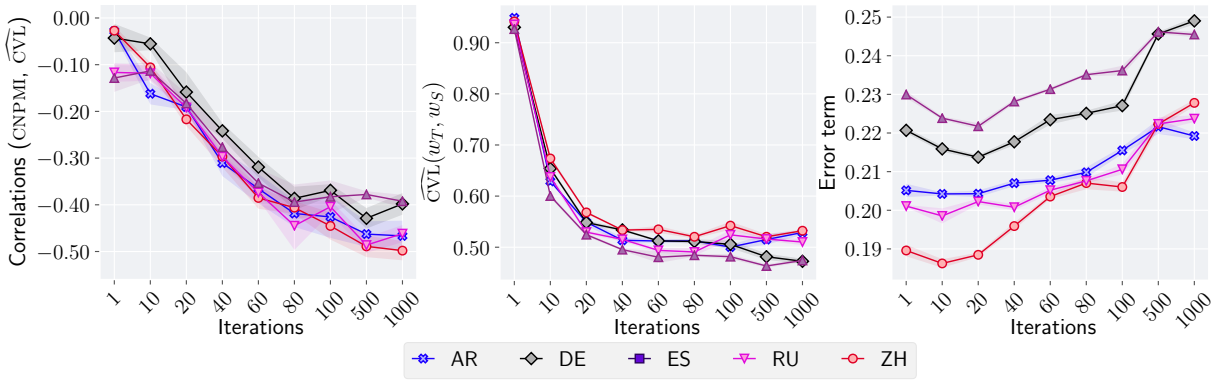

Figure 2: As Gibbs sampling progresses, $\widehat{\text{CVL}}$ of topic words drops, which leads to higher quality topics, and thus increases CNPMI. The left panel shows this negative correlation, and we use shades to indicate standard deviations across five chains.

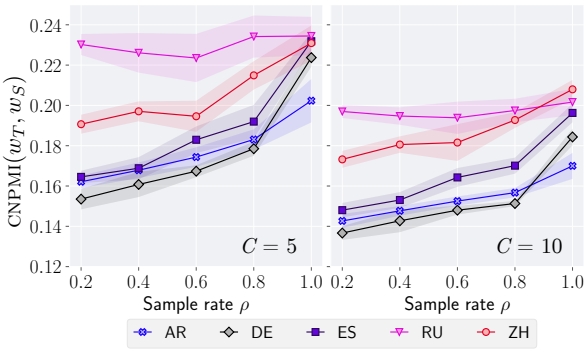

Figure 3: Increasing ρ results in smaller values of $\frac{\ln n^\star}{n}$ for translation pairs. Topic quality, evaluated by CNPMI, increases as well.

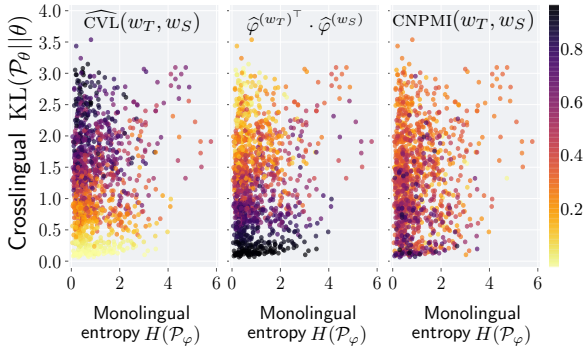

Figure 4: Each dot is a (EN,DE) word pair, and its color shows corresponding values of the indicated quantity. Best viewed in color.

level and word level, and we prefer a model with low crosslingual entropy but high monolingual entropy. In this experiment, we show how these two quantities affect the topic qualities, using English-German (EN-DE) documents as an example.

Given PLTM output in (EN,DE) and a cardinality $C = 5$, we collect $C^2 \times K$ bilingual word pairs as described in Section 4.1. For each word pair, we calculate three quantities: $\widehat{\text{CVL}}$, CNPMI, and the inner product of the word representations. In Figure 4, each dot is a word pair (w_S, w_T) colored by the values of these quantities. The word pair dots are positioned by their crosslingual and monolingual entropies.

We observe that $\widehat{\text{CVL}}$ decreases with crosslingual entropy on document level. The larger the crosslingual entropy, the harder it is to get a low $\widehat{\text{CVL}}$ because it needs larger monolingual entropy to decrease the bound, as shown in Section 2.4. On the other hand, the inner product of word pairs shows an opposite pattern of $\widehat{\text{CVL}}$, indicating a negative correlation (Lemma 1). In Figure 2 we see the correlation between CNPMI and $\widehat{\text{CVL}}$ is around -0.4 at the end of sampling, so there are fewer clear patterns for CNPMI in Figure 4. However, we also notice that the word pairs with higher CNPMI scores often appear at the bottom where crosslingual entropy is low while the monolingual entropy is high.

4.3 Downstream Task

We move on to crosslingual document classification as a downstream task. At various iterations of Gibbs sampling, we infer topics on the test sets for another 500 iterations and calculate the quantities shown in the Figure 5 (averaged over all languages), including the $\mathcal{H}$-distances for both training and test sets, and the joint risk $\widehat{\lambda}$.

We treat English as the source language and train support vector machines to obtain the best classifier $h^\star$ that fits the English documents. This classifier is then used to calculate the source and target risks $\widehat{R}_S(h^\star)$ and $\widehat{R}_T(h^\star)$. We also include $\frac{1}{2}\widehat{d}_{\mathcal{H}}(S, T)$, the $\mathcal{H}$-distance based on word rep-

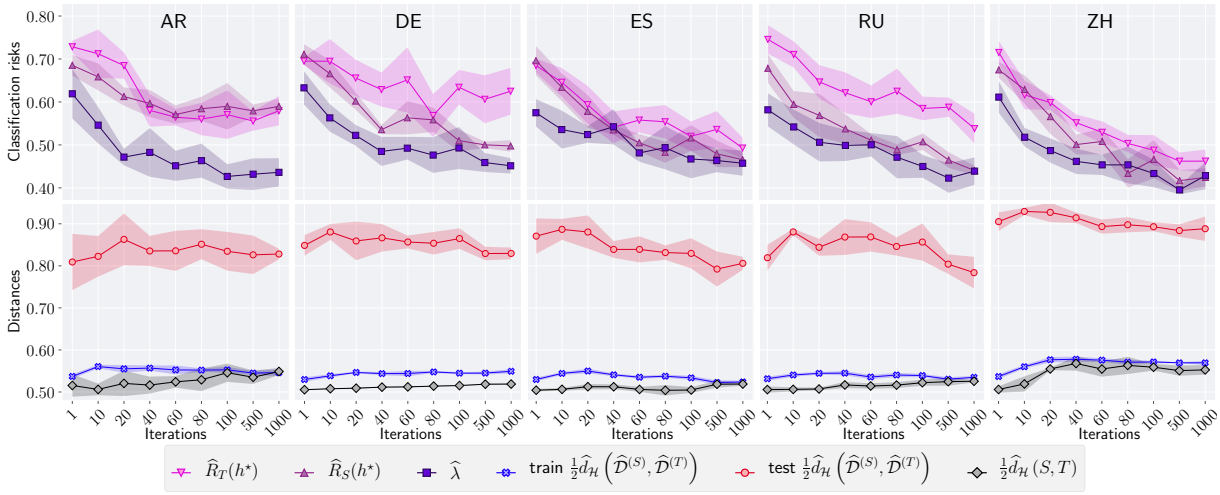

Figure 5: Gibbs sampling optimizes $\widehat{\text{CVL}}$, which decreases the joint risk $\widehat{\lambda}$ and $\mathcal{H}$-distances for test data.

resentations $\widehat{\varphi}$. As mentioned in Section 3.1, we train support vector machines to use languages as labels, and the accuracy score as the $\mathcal{H}$-distance.

The classification risks, such as $\widehat{R}_S(h^\star)$, $\widehat{R}_T(h^\star)$, and $\widehat{\lambda}$, are decreasing as expected (upper row in Figure 5), which shows very similar trends as $\widehat{\text{CVL}}$ in Figure 2. On the other hand, we notice that the $\mathcal{H}$-distances of training documents and vocabularies, $\frac{1}{2}\widehat{d}_{\mathcal{H}}\left(\widehat{\mathcal{D}}^{(S)}, \widehat{\mathcal{D}}^{(T)}\right)$ and $\frac{1}{2}\widehat{d}_{\mathcal{H}}(S,T)$, stabilize around 0.5 to 0.6, meaning it is difficult to differentiate the languages based on their representations. Interestingly, the $\mathcal{H}$-distances of test documents are at a less ideal value, although they are slightly decreasing in most of the languages except AR. However, recall that the target risk also depends on other factors than $\mathcal{H}$-distance (Theorem 4), and we use Figure 6 to illustrate this point.

We further explore the relationship between the predictability of languages vs document classes in Figure 6. We collect documents correctly classified for both document class and language labels, from which we randomly choose 200 documents for each language, and use $\widehat{\theta}$ to plot t-SNE scatterplots. Note that the two plots are from the same set of documents, and so the spatial relations between any two points are fixed, but we color them with different labelings. Although the classifier can identify the languages (right panel), the features are still consistent, because on the left panel, the decision boundary changes its direction and also successfully classifies the documents based on actual label class. This illustrates why a single $\mathcal{H}$-distance does not necessarily mean inconsistent features across languages and high target risks.

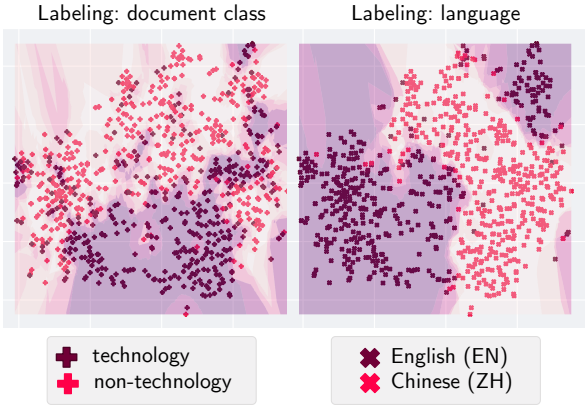

Figure 6: Although the classifier identifies the languages (right), the features are still consistent based on actual document class (left).

5 Conclusions and Future Directions

This study gives new insights into crosslingual transfer learning in multilingual topic models. By formulating the inference process as a circular validation, we derive a PAC-Bayesian theorem to show the factors that affect the success of crosslingual learning. We also connect topic model learning with downstream crosslingual tasks to show how errors propagate.

As the first step toward more theoretically justified crosslingual transfer learning, our study suggests considerations for constructing crosslingual transfer models in general. For example, an effective model should strengthen crosslingual transfer while minimizing non-transferred components, use a balanced dataset or specific optimization algorithms for low-resource languages, and support evaluation metrics that relate to CVL.

References

Sanjeev Arora, Rong Ge, Yonatan Halpern, David M. Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A Practical Algorithm for Topic Modeling with Provable Guarantees. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 280–288.

Maria Barrett, Frank Keller, and Anders Søgaard. 2016. Cross-lingual Transfer of Correlations between Parts of Speech and Gaze Features. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1330–1339.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 137–144.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Lorenzo Bruzzone and Mattia Marconcini. 2010. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787.

Manaal Faruqui and Chris Dyer. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 462–471.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. 2012. PAC-Bayesian Learning and Domain Adaptation. *CoRR*, abs/1212.2340.

Thomas L Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

E. Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting Cross-cultural Differences Using a Multilingual Topic Model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Shudong Hao, Jordan L. Boyd-Graber, and Michael J. Paul. 2018. Lessons from the Bible on Modern Topics: Low-Resource Multilingual Topic Model Evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1090–1100.

Shudong Hao and Michael J. Paul. 2018. Learning Multilingual Topics from Incomparable Corpora. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2595–2609.

Geert Heyman, Ivan Vulic, and Marie-Francine Moens. 2016. C-BiLDA: Extracting Cross-lingual Topics from Non-parallel Texts by Distinguishing Shared from Unshared Content. *Data Mining and Knowledge Discovery*, 30(5):1299–1323.

Yuening Hu, Jordan L. Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive Topic Modeling. *Machine Learning*, 95(3):423–469.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting Multilingual Topics from Unaligned Comparable Corpora. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, pages 444–456.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing Crosslingual Distributed Representations of Words. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 1459–1474.

Tengfei Ma and Tetsuya Nasukawa. 2017. Inverted Bilingual Topic Models for Lexicon Extraction from Non-parallel Data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4075–4081.

David A. McAllester. 1999. PAC-Bayesian Model Averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, Santa Cruz, CA, USA, July 7-9, 1999*, pages 164–170.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 880–889.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining Multilingual Topics from Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 1155–1156.

John Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2018. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, abs/1706.04902.

Ekaterina Shutova, Lin Sun, E. Dario Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual Metaphor Processing: Experiments with Semi-Supervised and Unsupervised Learning. *Computational Linguistics*, 43(1):71–123.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language Linking of News Stories on the Web Using Interlingual Topic Modelling. In *Proceedings of the 2nd ACM Workshop on Social Web Search and Mining, CIKM-SWSM 2009, Hong Kong, China, November 2, 2009*, pages 57–64.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218.

Leslie G. Valiant. 1984. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142.

Erheng Zhong, Wei Fan, Qiang Yang, Olivier Verscheure, and Jiangtao Ren. 2010. Cross Validation Framework to Choose amongst Models and Datasets for Transfer Learning. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, pages 547–562.

Appendix A Notation

See Table 1.

Notation	Description
S, T	Source and target languages. They are interchangeable during Gibbs sampling. For example, when training English and German, English can be either source or target.
w_ℓ	A **word type** of language ℓ.
x_ℓ	An individual **token** of language ℓ.
z_{x_ℓ}	The topic assignment of token x_ℓ.
$\mathcal{S}_{w_\ell}$	The sample of word type w_ℓ, the set containing all the tokens x_ℓ that are of this word type.
$\mathcal{P}_{x_\ell}, \mathcal{P}_{x_\ell,k}$	$\mathcal{P}_{x_\ell}$ denotes the conditional distribution over all topics for token x_ℓ. The conditional probability of sampling a topic k from $\mathcal{P}_{x_\ell}$ is denoted as $\mathcal{P}_{x_\ell,k}$.
$D^{(\ell)}$	The set of documents in language ℓ. This usually refers to the test corpus.
$\widehat{\mathcal{D}}^{(\ell)}$	The array of document representations from the corpus $D^{(\ell)}$ and their document labels.
$\widehat{\phi}_k^{(\ell)}$	The empirical distribution over vocabulary of language ℓ for topic $k = 1, \ldots, K$.
$\widehat{\varphi}^{(w)}$	The word representation, *i.e.,* the empirical distribution over K topics for a word type w. This can be obtained by re-normalizing $\widehat{\phi}_k^{(\ell)}$.
$\widehat{\theta}^{(d)}$	The document representation, *i.e.,* the empirical distribution over K topics for a document d.

Table 1: Notation table.

Appendix B Proofs

Theorem 1. *Let* $\widehat{\mathrm{CVL}}^{(t)}(w_T, w_S)$ *be the empirical circular validation loss of any bilingual word pair at iteration* t *of Gibbs sampling. Then* $\widehat{\mathrm{CVL}}^{(t)}(w_T, w_S)$ *converges as* $t \to \infty$.

Proof. We first notice the triangle inequality:

$$
\begin{aligned}
&\left|\widehat{\mathrm{CVL}}^{(t)}(w_T, w_S) - \widehat{\mathrm{CVL}}^{(t-1)}(w_T, w_S)\right| \\
=\ &\Bigg|\mathop{\mathbb{E}}_{x_S, x_T}\left[\widehat{\mathcal{L}}^{(t)}(x_T, w_S) + \widehat{\mathcal{L}}^{(t)}(x_S, w_T)\right] \\
&- \mathop{\mathbb{E}}_{x_S, x_T}\left[\widehat{\mathcal{L}}^{(t-1)}(x_T, w_S) + \widehat{\mathcal{L}}^{(t-1)}(x_S, w_T)\right]\Bigg| \\
=\ &\Bigg|\mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t)}(x_T, w_S)\right] + \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}^{(t)}(x_S, w_T)\right] \\
&- \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t-1)}(x_T, w_S)\right] \\
&- \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}^{(t-1)}(x_S, w_T)\right]\Bigg| \\
\leq\ &\Bigg|\mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t)}(x_T, w_S)\right] - \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t-1)}(x_T, w_S)\right] \\
&+ \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}^{(t)}(x_S, w_T)\right] - \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}^{(t-1)}(x_S, w_T)\right]\Bigg| \\
\equiv\ &\left|\Delta \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}(x_T, w_S)\right] + \Delta \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}(x_S, w_T)\right]\right| \\
\leq\ &\left|\Delta \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}(x_T, w_S)\right]\right| + \left|\Delta \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}(x_S, w_T)\right]\right|.
\end{aligned}
$$

We look at the first term of the last equation, and the other term can be derived in the same way. We use $\mathcal{P}_{x_T}$ to denote the invariant distribution of the conditional $\mathcal{P}_{x_T}^{(t)}$ as $t \to \infty$. Additionally, let $\mathcal{P}_{x_T, z_{x_S}}$ be the conditional probability for the to-

ken x_T being assigned to topic z_{x_S}:

$$\mathcal{P}_{x_T,z_{x_S}} \;=\; \Pr\left(k = z_{x_S}; w = w_{x_T}, \mathbf{z}_{-}, \mathbf{w}_{-}\right).$$

Another assumption we made is once the source language is converged, we keep the states of it fixed. That is, $z_{x_S}^{(t)} = z_{x_S}^{(t-1)}$, and only sample the target language. Taking the difference between the expectation at iterations t and $t-1$, we have

$$\begin{aligned}
&\lim_{t\to\infty}\left|\Delta \mathop{\mathbb{E}}_{x_T\in\mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}(x_T,w_S)\right]\right| \\
=\;&\lim_{t\to\infty}\left|\mathop{\mathbb{E}}_{x_T\in\mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t)}(x_T,w_S)\right]\right. \\
&\left.-\mathop{\mathbb{E}}_{x_T\in\mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t-1)}(x_T,w_S)\right]\right| \\
=\;&\lim_{t\to\infty}\left|\mathop{\mathbb{E}}_{x_T}\left[\frac{1}{n_{w_S}}\sum_{x_S}\mathbb{E}_{h\sim\mathcal{P}_{x_T}^{(t)}}\mathbb{1}\left\{h(x_S)\neq z_{x_S}^{(t)}\right\}\right]\right. \\
&\left.-\mathop{\mathbb{E}}_{x_T}\left[\frac{1}{n_{w_S}}\sum_{x_S}\mathbb{E}_{h\sim\mathcal{P}_{x_T}^{(t-1)}}\mathbb{1}\left\{h(x_S)\neq z_{x_S}^{(t-1)}\right\}\right]\right| \\
=\;&\lim_{t\to\infty}\frac{1}{n_{w_S}}\sum_{x_S}\mathop{\mathbb{E}}_{x_T}\left[\left|\mathbb{E}_{h\sim\mathcal{P}_{x_T}^{(t)}}\mathbb{1}\left\{h(x_S)\neq z_{x_S}^{(t)}\right\}\right.\right. \\
&\left.\left.-\mathbb{E}_{h\sim\mathcal{P}_{x_T}^{(t-1)}}\mathbb{1}\left\{h(x_S)\neq z_{x_S}^{(t-1)}\right\}\right|\right] \\
=\;&\lim_{t\to\infty}\frac{1}{n_{w_S}}\sum_{x_S}\mathop{\mathbb{E}}_{x_T}\left[\left|\mathbb{E}_{h\sim\mathcal{P}_{x_T}^{(t)}}\mathbb{1}\left\{h(x_S)\neq z_{x_S}\right\}\right.\right. \\
&\left.\left.-\mathbb{E}_{h\sim\mathcal{P}_{x_T}^{(t-1)}}\mathbb{1}\left\{h(x_S)\neq z_{x_S}\right\}\right|\right] \\
=\;&\lim_{t\to\infty}\frac{1}{n_{w_S}}\sum_{x_S\in\mathcal{S}_{w_S}}\mathbb{E}_{x_T\in\mathcal{S}_{w_T}}\left[\left|\left(1-\mathcal{P}_{x_T,z_{x_S}}^{(t)}\right)\right.\right. \\
&\left.\left.-\left(1-\mathcal{P}_{x_T,z_{x_S}}^{(t-1)}\right)\right|\right] \\
=\;&\lim_{t\to\infty}\frac{1}{n_{w_S}}\sum_{x_S\in\mathcal{S}_{w_S}}\mathbb{E}_{x_T\in\mathcal{S}_{w_T}}\left[\left|\mathcal{P}_{x_T,z_{x_S}}^{(t-1)}-\mathcal{P}_{x_T,z_{x_S}}^{(t)}\right|\right] \\
=\;&\lim_{t\to\infty}\frac{1}{n_{w_S}}\sum_{x_S\in\mathcal{S}_{w_S}}\mathbb{E}_{x_T\in\mathcal{S}_{w_T}}\left[\left|\mathcal{P}_{x_T,z_{x_S}}-\mathcal{P}_{x_T,z_{x_S}}\right|\right] \\
=\;&0.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\lim_{t\to\infty}\left|\widehat{\mathrm{CVL}}^{(t)}(w_T,w_S)-\widehat{\mathrm{CVL}}^{(t-1)}(w_T,w_S)\right| \\
\leq\;&\lim_{t\to\infty}\left|\Delta\mathop{\mathbb{E}}_{x_T\in\mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}(x_T,w_S)\right]\right| \\
&+\left|\Delta\mathop{\mathbb{E}}_{x_S\in\mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}(x_S,w_T)\right]\right| \\
=\;&0.
\end{aligned}$$

□

Theorem 3. *Given a bilingual word pair (w_T, w_S), with probability at least $1-\delta$, the following bound holds:*

$$\begin{aligned}
&\mathrm{CVL}(w_T,w_S) \;\leq\; \widehat{\mathrm{CVL}}(w_T,w_S) \;+ \\
&\frac{1}{2}\sqrt{\frac{1}{n}\left(\mathrm{KL}_{w_T}+\mathrm{KL}_{w_S}+2\ln\frac{2}{\delta}\right)+\frac{\ln n^\star}{n}}, \\
&n=\min\left\{n_{w_T},n_{w_S}\right\},\quad n^\star=\max\left\{n_{w_T},n_{w_S}\right\}.
\end{aligned}$$

For brevity we use KL_w to denote $\mathrm{KL}(\mathcal{P}_x||Q_x)$, where $\mathcal{P}_x$ is the conditional distribution from Gibbs sampling of token x with word type w that gives highest loss $\widehat{\mathcal{L}}(x,w)$, and Q_x a prior.

Proof. From Theorem 2, for target language, with probability at least $1-\delta$,

$$\begin{aligned}
&\mathcal{L}(x_T,w_S) \\
\leq\;&\widehat{\mathcal{L}}(x_T,w_S)+\sqrt{\frac{\mathrm{KL}\left(\mathcal{P}_{x_T}||Q_{x_T}\right)+\ln\frac{2\sqrt{n_{w_S}}}{\delta}}{2n_{w_S}}} \\
=\;&\widehat{\mathcal{L}}(x_T,w_S)+\sqrt{\frac{\mathrm{KL}\left(\mathcal{P}_{x_T}||Q_{x_T}\right)+\ln\frac{2}{\delta}+\frac{\ln n_{w_S}}{2n_{w_S}}}{2}} \\
\equiv\;&\widehat{\mathcal{L}}(x_T,w_S)+\epsilon(x_T,w_S).
\end{aligned}$$

For the source language, similarly, with probability at least $1-\delta$,

$$\begin{aligned}
&\mathcal{L}(x_S,w_T) \\
\leq\;&\widehat{\mathcal{L}}(x_S,w_T)+\sqrt{\frac{\mathrm{KL}\left(\mathcal{P}_{x_S}||Q_{x_S}\right)+\ln\frac{2}{\delta}+\frac{\ln n_{w_T}}{2}}{2n_{w_T}}} \\
\equiv\;&\widehat{\mathcal{L}}(x_S,w_T)+\epsilon(x_S,w_T).
\end{aligned}$$

Given a word type w_T, we notice that only the KL-divergence term in $\epsilon(x_T,w_S)$ varies among different tokens x_T. Thus, we use KL_{w_S} and KL_{w_T} to denote the maximal values of KL-divergence over all the tokens,

$$\begin{aligned}
\mathrm{KL}_{w_S} \;&=\; \mathrm{KL}\left(\mathcal{P}_{x_T^\star}||Q_{x_T^\star}\right), \\
x_T^\star \;&=\; \mathop{\arg\max}_{x_T\in\mathcal{S}_{w_T}}\epsilon(x_T,w_S); \\
\mathrm{KL}_{w_T} \;&=\; \mathrm{KL}\left(\mathcal{P}_{x_S^\star}||Q_{x_S^\star}\right), \\
x_S^\star \;&=\; \mathop{\arg\max}_{x_S\in\mathcal{S}_{w_S}}\epsilon(x_S,w_T).
\end{aligned}$$

Let $n = \min\{n_{w_T}, n_{w_S}\}$, and $n^\star = \max\{n_{w_T}, n_{w_S}\}$. Due to the fact that $\sqrt{x}+\sqrt{y} \leq \frac{2}{\sqrt{2}}\sqrt{x+y}$ for $x, y > 0$, we have

$$
\begin{aligned}
&\mathrm{CVL}(w_T, w_S) \\
=\ & \frac{1}{2} \mathop{\mathbb{E}}_{x_S, x_T} \left[\mathcal{L}(x_T, w_S) + \mathcal{L}(x_S, w_T)\right] \\
=\ & \frac{1}{2} \left(\mathbb{E}_{x_T} \mathcal{L}(x_T, w_S) + \mathbb{E}_{x_S} \mathcal{L}(x_S, w_T)\right) \\
\leq\ & \frac{1}{2} \left(\mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \widehat{\mathcal{L}}(x_T, w_S) + \mathbb{E}_{x_S \in \mathcal{S}_{w_S}} \widehat{\mathcal{L}}(x_S, w_T)\right) \\
& + \frac{1}{2} \left(\mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \epsilon(x_T, w_S) + \mathbb{E}_{x_S \in \mathcal{S}_{w_S}} \epsilon(x_S, w_T)\right) \\
=\ & \widehat{\mathrm{CVL}}(w_T, w_S) \\
& + \frac{1}{2} \left(\mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \epsilon(x_T, w_S) + \mathbb{E}_{x_S \in \mathcal{S}_{w_S}} \epsilon(x_S, w_T)\right) \\
\leq\ & \widehat{\mathrm{CVL}}(w_T, w_S) + \frac{1}{2} \left(\epsilon(x_T^\star, w_S) + \epsilon(x_S^\star, w_T)\right) \\
\leq\ & \widehat{\mathrm{CVL}}(w_T, w_S) \\
& + \frac{1}{2} \left(\sqrt{\frac{1}{2n_{w_T}} \left(\mathrm{KL}_{w_T} + \ln \frac{2}{\delta} + \frac{1}{2} \ln n_{w_T}\right)}\right. \\
& \left. + \sqrt{\frac{1}{2n_{w_S}} \left(\mathrm{KL}_{w_S} + \ln \frac{2}{\delta} + \frac{1}{2} \ln n_{w_S}\right)}\right) \\
\leq\ & \widehat{\mathrm{CVL}}(w_T, w_S) \\
& + \frac{1}{2} \sqrt{\frac{\mathrm{KL}_{w_T} + \mathrm{KL}_{w_S} + 2 \ln \frac{2}{\delta}}{n} + \left(\frac{\ln \left(n_{w_T} \cdot n_{w_S}\right)}{2n}\right)} \\
\leq\ & \widehat{\mathrm{CVL}}(w_T, w_S) \\
& + \frac{1}{2} \sqrt{\frac{\mathrm{KL}_{w_T} + \mathrm{KL}_{w_S} + 2 \ln \frac{2}{\delta}}{n} + \left(\frac{\ln n^\star}{n}\right)},
\end{aligned}
$$

which gives us the result. □

Lemma 1. *Given any bilingual word pair (w_T, w_S), let $\widehat{\varphi}^{(w)}$ denote the distribution over topics of word type w. Then we have,*

$$
1 - \widehat{\varphi}^{(w_T)\top} \cdot \widehat{\varphi}^{(w_S)} \ \leq\ \widehat{\mathrm{CVL}}(w_T, w_S).
$$

Proof. We expand the equation of $\widehat{\mathrm{CVL}}$ as follows,

$$
\begin{aligned}
&\widehat{\mathrm{CVL}}(w_T, w_S) \\
=\ & \frac{1}{2} \mathop{\mathbb{E}}_{x_S, x_T} \left[\widehat{\mathcal{L}}(x_T, w_S) + \widehat{\mathcal{L}}(x_S, w_T)\right] \\
=\ & \frac{1}{2} \left(\mathbb{E}_{x_T} \left[\widehat{\mathcal{L}}(x_T, w_S)\right] + \mathbb{E}_{x_S} \left[\widehat{\mathcal{L}}(x_S, w_T)\right]\right) \\
=\ & \frac{1}{2} \left(\frac{\sum_{x_T \in \mathcal{S}_{w_T}} \sum_{x_S \in \mathcal{S}_{w_S}} \mathbb{E}_{h \sim \mathcal{P}_{x_T}} \left[\mathbb{1}\left\{h(x_S) \neq z_{x_S}\right\}\right]}{n_{w_T} \cdot n_{w_S}}\right. \\
& \left. + \frac{\sum_{x_S \in \mathcal{S}_{w_S}} \sum_{x_T \in \mathcal{S}_{w_T}} \mathbb{E}_{h \sim \mathcal{P}_{x_S}} \left[\mathbb{1}\left\{h(x_T) \neq z_{x_T}\right\}\right]}{n_{w_S} \cdot n_{w_T}}\right) \\
=\ & \frac{1}{2} \left(\frac{\sum_{x_T \in \mathcal{S}_{w_T}} \sum_{x_S \in \mathcal{S}_{w_S}} \left(1 - \mathcal{P}_{x_T, z_{x_S}}\right)}{n_{w_T} \cdot n_{w_S}}\right. \\
& \left. + \frac{\sum_{x_S \in \mathcal{S}_{w_S}} \sum_{x_T \in \mathcal{S}_{w_T}} \left(1 - \mathcal{P}_{x_S, z_{x_T}}\right)}{n_{w_S} \cdot n_{w_T}}\right) \\
=\ & 1 - \frac{1}{2} \left(\frac{\sum_{x_T \in \mathcal{S}_{w_T}} \sum_{x_S \in \mathcal{S}_{w_S}} \mathcal{P}_{x_T, z_{x_S}}}{n_{w_T} \cdot n_{w_S}}\right. \\
& \left. + \frac{\sum_{x_S \in \mathcal{S}_{w_S}} \sum_{x_T \in \mathcal{S}_{w_T}} \mathcal{P}_{x_S, z_{x_T}}}{n_{w_S} \cdot n_{w_T}}\right) \\
=\ & 1 - \frac{1}{2} \sum_{k=1}^{K} \left(\frac{n_{k|w_S} \cdot \sum_{x_T \in \mathcal{S}_{w_T}} \mathcal{P}_{x_T, k}}{n_{w_T} \cdot n_{w_S}}\right. \\
& \left. + \frac{n_{k|w_T} \cdot \sum_{x_S \in \mathcal{S}_{w_S}} \mathcal{P}_{x_S, z_{x_T}}}{n_{w_S} \cdot n_{w_T}}\right) \\
=\ & 1 - \frac{1}{2} \sum_{k=1}^{K} \left(\widehat{\varphi}_k^{(w_S)} \cdot \frac{\sum_{x_T \in \mathcal{S}_{w_T}} \mathcal{P}_{x_T, k}}{n_{w_T}}\right. \\
& \left. + \widehat{\varphi}_k^{(w_T)} \cdot \frac{\sum_{x_S \in \mathcal{S}_{w_S}} \mathcal{P}_{x_S, z_{x_T}}}{n_{w_S}}\right) \\
\geq\ & 1 - \frac{1}{2} \sum_{k=1}^{K} \left(\widehat{\varphi}_k^{(w_S)} \cdot \frac{n_{k|w_T}}{n_{w_T}} + \widehat{\varphi}_k^{(w_T)} \cdot \frac{n_{k|w_S}}{n_{w_S}}\right) \\
=\ & 1 - \frac{1}{2} \sum_{k=1}^{K} \left(\widehat{\varphi}_k^{(w_S)} \cdot \widehat{\varphi}_k^{(w_T)} + \widehat{\varphi}_k^{(w_T)} \cdot \widehat{\varphi}_k^{(w_S)}\right) \\
=\ & 1 - \widehat{\varphi}^{(w_S)\top} \cdot \widehat{\varphi}^{(w_T)}
\end{aligned}
$$

which concludes the proof. □

Theorem 5. *Let $\widehat{\theta}^{(d_S)}$ be the distribution over topics for document d_S (similarly for d_T), $F(d_S, d_T) = \left(\sum_{w_S} {f_{w_S}^{d_S}}^2 \cdot \sum_{w_T} {f_{w_T}^{d_T}}^2\right)^{\frac{1}{2}}$ where f_w^d is the* normalized *frequency of word w in document d, and K the number of topics. Then*

$$
\begin{aligned}
\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)} \ \leq\ & F(d_S, d_T) \\
& \cdot \sqrt{K \cdot \sum_{w_S, w_T} \left(\widehat{\mathrm{CVL}}(w_T, w_S) - 1\right)^2}.
\end{aligned}
$$

Proof. We first expand the inner product of $\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)}$ as follows,

$$
\begin{aligned}
&\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)} \\
=\ & \sum_{k=1}^{K} \widehat{\theta}_k^{(d_S)} \cdot \widehat{\theta}_k^{(d_T)} \\
=\ & \sum_{k=1}^{K} \left(\left(\sum_{w_S \in V^{(S)}} f_{w_S}^{d_S} \cdot \widehat{\varphi}_k^{(w_S)}\right)\right. \\
& \left. \cdot \left(\sum_{w_T \in V^{(T)}} f_{w_T}^{d_T} \cdot \widehat{\varphi}_k^{(w_T)}\right)\right) \\
\leq\ & F(d_S, d_T) \cdot \sum_{k=1}^{K} \left(\left(\sum_{w_S \in V^{(S)}} {\widehat{\varphi}_k^{(w_S)}}^2\right)^{\frac{1}{2}} \cdot\right. \\
& \left. \left(\sum_{w_T \in V^{(T)}} {\widehat{\varphi}_k^{(w_T)}}^2\right)^{\frac{1}{2}}\right), \\
& F(d_S, d_T) \\
=\ & \left(\sum_{w_S \in V^{(S)}} {f_{w_S}^{d_S}}^2\right)^{\frac{1}{2}} \cdot \left(\sum_{w_T \in V^{(T)}} {f_{w_T}^{d_T}}^2\right)^{\frac{1}{2}},
\end{aligned}
$$

where $F(d_S, d_T)$ is a constant independent of topic k, and the last inequality due to Hölder's. We then focus on the topic-dependent part of the last inequality.

$$\begin{aligned}
&\sum_{k=1}^{K}\left(\left(\sum_{w_S\in V^{(S)}}\widehat{\varphi}_k^{(w_S)^2}\right)^{\frac{1}{2}}\cdot\left(\sum_{w_T\in V^{(T)}}\widehat{\varphi}_k^{(w_T)^2}\right)^{\frac{1}{2}}\right)\\
=&\sum_{k=1}^{K}\left(\sum_{w_S,w_T}\left(\widehat{\varphi}_k^{(w_S)}\cdot\widehat{\varphi}_k^{(w_T)}\right)^2\right)^{\frac{1}{2}}\\
\leq&\sqrt{K}\cdot\left(\sum_{k=1}^{K}\sum_{w_S,w_T}\left(\widehat{\varphi}_k^{(w_S)}\cdot\widehat{\varphi}_k^{(w_T)}\right)^2\right)^{\frac{1}{2}}\\
=&\sqrt{K}\cdot\left(\sum_{w_S,w_T}\sum_{k=1}^{K}\left(\widehat{\varphi}_k^{(w_S)}\cdot\widehat{\varphi}_k^{(w_T)}\right)^2\right)^{\frac{1}{2}}\\
\leq&\sqrt{K}\cdot\left(\sum_{w_S,w_T}\left(\sum_{k=1}^{K}\widehat{\varphi}_k^{(w_S)}\cdot\widehat{\varphi}_k^{(w_T)}\right)^2\right)^{\frac{1}{2}}\\
=&\sqrt{K}\cdot\left(\sum_{w_S,w_T}\left(\widehat{\varphi}^{(w_T)\top}\cdot\widehat{\varphi}^{(w_S)}\right)^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Thus, we have the following inequality:

$$\begin{aligned}
\widehat{\theta}^{(d_S)\top}\cdot\widehat{\theta}^{(d_T)}\leq&\ F(d_S,d_T)\cdot\sqrt{K}\\
&\cdot\left(\sum_{w_S,w_T}\left(\widehat{\varphi}^{(w_T)\top}\cdot\widehat{\varphi}^{(w_S)}\right)^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Plug in Lemma 1, we see that

$$\begin{aligned}
\widehat{\theta}^{(d_S)\top}\cdot\widehat{\theta}^{(d_T)}\leq&\ F(d_S,d_T)\cdot\sqrt{K}\\
&\cdot\left(\sum_{w_S,w_T}\left(\widehat{\mathrm{CVL}}(w_T,w_S)-1\right)^2\right)^{\frac{1}{2}}.
\end{aligned}$$

□

Appendix C Dataset Details

C.1 Pre-processing

For all the languages, we use existing stemmers to stem words in the corpora and the entries in Wiktionary. Since Chinese does not have stemmers, we loosely use "stem" to refer to "segment" Chinese sentences into words. We also use fixed stopword lists to filter out stop words. Table 2 lists the source of the stemmers and stopwords.

[1] http://snowball.tartarus.org;
[2] http://arabicstemmer.com;
[3] https://github.com/6/stopwords-json;
[4] https://github.com/fxsjy/jieba.

C.2 Training Sets

Our training set is a comparable corpus from Wikipedia. For each Wikipedia article page, there exists an interlingual link to view the article in another language. This interlingual link provides the same article in different languages and is commonly used to create comparable corpora in multilingual studies. We show the statistics of this training corpus in Table 3. The numbers are calculated after stemming and lemmatization.

C.3 Test Sets

C.3.1 Topic Coherence Evaluation Sets

Topic coherence evaluation for multilingual topic models was proposed by Hao et al. (2018), where a comparable corpus is used to calculate bilingual word pair co-occurrence and CNPMI scores. We use a Wikipedia corpus to calculate this score, and the statistics are shown in Table 4. This Wikipedia corpus does not overlap with the training set.

C.3.2 Unseen Document Inference

We use the Global Voices (GV) corpus to create test sets, which can be retrieved from the website https://globalvoices.org directly, or from the OPUS collection at http://opus.nlpl.eu/GlobalVoices.php. We show the statistics in Table 5. After the column showing number of documents, we also include the statistics of specific labels. The multiclass labels are mutual exclusive, and each document has only one label.

Note that although all the language pairs share the same set of English test documents, the document representations are inferred from different topic models trained specifically for that language pair. Thus, the document representations for the same English document are different across different language pairs.

Lastly, the number of word types is based on the training set and after stemming and lemmatization. When a word type in the test set does not appear in the training set, we ignore this type.

C.3.3 Wiktionary

In downsampling experiments (Section 4.2), we use English Wiktionary to create bilingual dictionaries, which can be downloaded at https://dumps.wikimedia.org/enwiktionary/.

Language	Family	Stemmer	Stopwords
AR	Semitic	`Assem's Arabic Light Stemmer` [1]	GitHub [2]
DE	Germanic	`SnowBallStemmer` [3]	NLTK
EN	Germanic	`SnowBallStemmer`	NLTK
ES	Romance	`SnowBallStemmer`	NLTK
RU	Slavic	`SnowBallStemmer`	NLTK
ZH	Sinitic	`Jieba` [4]	GitHub

Table 2: List of source of stemmers and stopwords used in experiments.

	English		
Language	#docs	#token	#types
AR	3,000	724,362	203,024
DE	3,000	409,381	125,071
ES	3,000	451,115	134,241
RU	3,000	480,715	142,549
ZH	3,000	480,142	141,679
	Paired language		
Language	#docs	#token	#types
AR	3,000	223,937	61,267
DE	3,000	285,745	125,169
ES	3,000	276,188	95,682
RU	3,000	276,462	96,568
ZH	3,000	233,773	66,275

Table 3: Statistics of the Wikipedia training corpus.

Appendix D Topic Model Configurations

For each experiment, we run five chains of Gibbs sampling using the Polylingual Topic Model implemented in MALLET, [5] and take the average over all chains. Each chain has 1,000 iterations, and we do not set a burn-in period. We set the topic number $K = 50$. Other hyperparameters are $\alpha = \frac{50}{K} = 1$ and $\beta = 0.01$ which are the default settings. We do not enable hyperparameter optimization procedures.

[5] http://mallet.cs.umass.edu/topics-polylingual.php.

	English		
Language	#docs	#token	#types
AR	10,000	3,092,721	143,504
DE	10,000	2,779,963	146,757
ES	10,000	3,021,732	149,423
RU	10,000	3,016,795	154,442
ZH	10,000	1,982,452	112,174
	Paired language		
Language	#docs	#token	#types
AR	10,000	1,477,312	181,734
DE	10,000	1,702,101	227,205
ES	10,000	1,737,312	142,086
RU	10,000	2,299,332	284,447
ZH	10,000	1,335,922	144,936

Table 4: Statistics of the Wikipedia corpus for topic coherence evaluation (CNPMI).

Language	#docs	#token	#types
EN	11,012	3,838,582	104,164
AR	1,086	314,918	53,030
DE	773	334,611	38,702
ES	7,470	3,454,304	110,134
RU	1,035	454,380	67,202
ZH	1,590	804,720	61,319
	#tech.	*#culture*	*#edu.*
EN	4,384	4,679	1,949
AR	457	430	199
DE	315	294	164
ES	2,961	3,121	1,388
RU	362	456	217
ZH	619	622	349

Table 5: Statistics of the Global Voices (GV) corpus.