Analyzing Bayesian Crosslingual Transfer in Topic Models (Appendix)

Shudong Hao
Boulder, CO
shudonghao@gmail.com

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

A Notation

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 $k = 1, \dots, K$.
$\widehat{\varphi}^{(w)}$	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.

B Proofs

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. We first notice the triangle inequality:

$$\left|\widehat{\text{CVL}}^{(t)}(w_T, w_S) - \widehat{\text{CVL}}^{(t-1)}(w_T, w_S)\right| \tag{1}$$

$$= \left|\mathop{\mathbb{E}}_{x_S, x_T}\left[\widehat{\mathcal{L}}^{(t)}(x_T, w_S) + \widehat{\mathcal{L}}^{(t)}(x_S, w_T)\right] - \mathop{\mathbb{E}}_{x_S, x_T}\left[\widehat{\mathcal{L}}^{(t-1)}(x_T, w_S) + \widehat{\mathcal{L}}^{(t-1)}(x_S, w_T)\right]\right| \tag{2}$$

$$\begin{aligned} = &\left|\mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t)}(x_T, w_S)\right] + \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}^{(t)}(x_S, w_T)\right]\right. \\ &\left. - \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}}\left[\widehat{\mathcal{L}}^{(t-1)}(x_T, w_S)\right] - \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}}\left[\widehat{\mathcal{L}}^{(t-1)}(x_S, w_T)\right]\right| \end{aligned} \tag{3}$$

$$\begin{aligned}
\leq\quad & \left| \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\widehat{\mathcal{L}}^{(t)}(x_T, w_S) \right] - \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\widehat{\mathcal{L}}^{(t-1)}(x_T, w_S) \right] \right. \\
& \left. + \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}} \left[\widehat{\mathcal{L}}^{(t)}(x_S, w_T) \right] - \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}} \left[\widehat{\mathcal{L}}^{(t-1)}(x_S, w_T) \right] \right| \qquad (4)
\end{aligned}$$

$$\equiv \quad \left| \Delta \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\widehat{\mathcal{L}}(x_T, w_S) \right] + \Delta \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}} \left[\widehat{\mathcal{L}}(x_S, w_T) \right] \right| \qquad (5)$$

$$\leq \quad \left| \Delta \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\widehat{\mathcal{L}}(x_T, w_S) \right] \right| + \left| \Delta \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}} \left[\widehat{\mathcal{L}}(x_S, w_T) \right] \right| \qquad (6)$$

We look at the first term of Equation (6), 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 $\mathcal{P}^{(t)}_{x_T}$ as $t \to \infty$. Additionally, let $\mathcal{P}_{x_T, z_{x_S}}$ be the conditional probability for the token x_T being assigned to topic z_{x_S}:

$$\mathcal{P}_{x_T, z_{x_S}} \quad = \quad \Pr\left(k = z_{x_S}; w = w_{x_T}, \mathbf{z}_{-}, \mathbf{w}_{-}\right). \qquad (7)$$

Another assumption we made is once the source language is converged, we keep the states of it fixed. That is, $z^{(t)}_{x_S} = z^{(t-1)}_{x_S}$, 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| && (8) \\
= \quad & \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] - \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\widehat{\mathcal{L}}^{(t-1)}(x_T, w_S) \right] \right| && (9) \\
= \quad & \lim_{t\to\infty} \left| \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\frac{1}{n_{w_S}} \sum_{x_S \in \mathcal{S}_{w_S}} \mathbb{E}_{h \sim \mathcal{P}^{(t)}_{x_T}} \mathbb{1}\left\{ h(x_S) \neq z^{(t)}_{x_S} \right\} \right] \right. \\
& \left. - \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\frac{1}{n_{w_S}} \sum_{x_S \in \mathcal{S}_{w_S}} \mathbb{E}_{h \sim \mathcal{P}^{(t-1)}_{x_T}} \mathbb{1}\left\{ h(x_S) \neq z^{(t-1)}_{x_S} \right\} \right] \right| && (10) \\
= \quad & \lim_{t\to\infty} \frac{1}{n_{w_S}} \sum_{x_S \in \mathcal{S}_{w_S}} \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\left| \mathbb{E}_{h \sim \mathcal{P}^{(t)}_{x_T}} \mathbb{1}\left\{ h(x_S) \neq z^{(t)}_{x_S} \right\} - \mathbb{E}_{h \sim \mathcal{P}^{(t-1)}_{x_T}} \mathbb{1}\left\{ h(x_S) \neq z^{(t-1)}_{x_S} \right\} \right| \right] && (11) \\
= \quad & \lim_{t\to\infty} \frac{1}{n_{w_S}} \sum_{x_S \in \mathcal{S}_{w_S}} \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\left| \mathbb{E}_{h \sim \mathcal{P}^{(t)}_{x_T}} \mathbb{1}\left\{ h(x_S) \neq z_{x_S} \right\} - \mathbb{E}_{h \sim \mathcal{P}^{(t-1)}_{x_T}} \mathbb{1}\left\{ h(x_S) \neq z_{x_S} \right\} \right| \right] && (12) \\
= \quad & \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}^{(t)}_{x_T, z_{x_S}}\right) - \left(1 - \mathcal{P}^{(t-1)}_{x_T, z_{x_S}}\right) \right| \right] && (13) \\
= \quad & \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}^{(t-1)}_{x_T, z_{x_S}} - \mathcal{P}^{(t)}_{x_T, z_{x_S}} \right| \right] && (14) \\
= \quad & \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] \quad = \quad 0. && (15)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \lim_{t\to\infty} \left| \widehat{\mathrm{CVL}}^{(t)}(w_T, w_S) - \widehat{\mathrm{CVL}}^{(t-1)}(w_T, w_S) \right| && (16) \\
\leq \quad & \lim_{t\to\infty} \left| \Delta \mathop{\mathbb{E}}_{x_T \in \mathcal{S}_{w_T}} \left[\widehat{\mathcal{L}}(x_T, w_S) \right] \right| + \left| \Delta \mathop{\mathbb{E}}_{x_S \in \mathcal{S}_{w_S}} \left[\widehat{\mathcal{L}}(x_S, w_T) \right] \right| \quad = \quad 0. && (17)
\end{aligned}$$

$\square$

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

$$\textsc{cvl}(w_T, w_S) \leq \widehat{\textsc{cvl}}(w_T, w_S) + \frac{1}{2}\sqrt{\frac{1}{n}\left(\mathrm{KL}_{w_T} + \mathrm{KL}_{w_S} + 2\ln\frac{2}{\delta}\right) + \left(\frac{\ln n^\star}{n}\right)}, \tag{18}$$

$$n = \min\left\{n_{w_T}, n_{w_S}\right\}, \quad n^\star = \max\left\{n_{w_T}, n_{w_S}\right\}. \tag{19}$$

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 $1-\delta$,

$$\begin{aligned}
\mathcal{L}(x_T, w_S) &\leq \widehat{\mathcal{L}}(x_T, w_S) + \sqrt{\frac{1}{2n_{w_S}}\left(\mathrm{KL}\left(\mathcal{P}_{x_T}||Q_{x_T}\right) + \ln\frac{2\sqrt{n_{w_S}}}{\delta}\right)} && (20)\\
&= \widehat{\mathcal{L}}(x_T, w_S) + \sqrt{\frac{1}{2n_{w_S}}\left(\mathrm{KL}\left(\mathcal{P}_{x_T}||Q_{x_T}\right) + \ln\frac{2}{\delta} + \frac{1}{2}\ln n_{w_S}\right)} && (21)\\
&\equiv \widehat{\mathcal{L}}(x_T, w_S) + \epsilon(x_T, w_S). && (22)
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{L}(x_S, w_T) &\leq \widehat{\mathcal{L}}(x_S, w_T) + \sqrt{\frac{1}{2n_{w_T}}\left(\mathrm{KL}\left(\mathcal{P}_{x_S}||Q_{x_S}\right) + \ln\frac{2}{\delta} + \frac{1}{2}\ln n_{w_T}\right)} && (23)\\
&\equiv \widehat{\mathcal{L}}(x_S, w_T) + \epsilon(x_S, w_T). && (24)
\end{aligned}$$

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,

$$\begin{aligned}
\mathrm{KL}_{w_S} &= \mathrm{KL}\left(\mathcal{P}_{x_T^\star}||Q_{x_T^\star}\right), \quad x_T^\star = \operatorname*{arg\,max}_{x_T \in \mathcal{S}_{w_T}} \epsilon(x_T, w_S); && (25)\\
\mathrm{KL}_{w_T} &= \mathrm{KL}\left(\mathcal{P}_{x_S^\star}||Q_{x_S^\star}\right), \quad x_S^\star = \operatorname*{arg\,max}_{x_S \in \mathcal{S}_{w_S}} \epsilon(x_S, w_T). && (26)
\end{aligned}$$

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

$$\begin{aligned}
& \textsc{cvl}(w_T, w_S) && (27)\\
=\ & \frac{1}{2}\mathop{\mathbb{E}}_{x_S, x_T}\left[\mathcal{L}(x_T, w_S) + \mathcal{L}(x_S, w_T)\right] && (28)\\
=\ & \frac{1}{2}\left(\mathbb{E}_{x_T}\mathcal{L}(x_T, w_S) + \mathbb{E}_{x_S}\mathcal{L}(x_S, w_T)\right) && (29)\\
\leq\ & \frac{1}{2}\left(\mathbb{E}_{x_T \in \mathcal{S}_{w_T}}\widehat{\mathcal{L}}(x_T, w_S) + \mathbb{E}_{x_S \in \mathcal{S}_{w_S}}\widehat{\mathcal{L}}(x_S, w_T)\right) && (30)\\
& + \frac{1}{2}\left(\mathbb{E}_{x_T \in \mathcal{S}_{w_T}}\epsilon(x_T, w_S) + \mathbb{E}_{x_S \in \mathcal{S}_{w_S}}\epsilon(x_S, w_T)\right) && (31)\\
=\ & \widehat{\textsc{cvl}}(w_T, w_S) + \frac{1}{2}\left(\mathbb{E}_{x_T \in \mathcal{S}_{w_T}}\epsilon(x_T, w_S) + \mathbb{E}_{x_S \in \mathcal{S}_{w_S}}\epsilon(x_S, w_T)\right) && (32)\\
\leq\ & \widehat{\textsc{cvl}}(w_T, w_S) + \frac{1}{2}\left(\epsilon(x_T^\star, w_S) + \epsilon(x_S^\star, w_T)\right) && (33)\\
\leq\ & \widehat{\textsc{cvl}}(w_T, w_S) && (34)
\end{aligned}$$

$$+\frac{1}{2}\left(\sqrt{\frac{1}{2n_{w_T}}\left(\mathrm{KL}_{w_T}+\ln\frac{2}{\delta}+\frac{1}{2}\ln n_{w_T}\right)}\right. \tag{35}$$

$$\left.+\sqrt{\frac{1}{2n_{w_S}}\left(\mathrm{KL}_{w_S}+\ln\frac{2}{\delta}+\frac{1}{2}\ln n_{w_S}\right)}\right) \tag{36}$$

$$\leq \widehat{\mathrm{CVL}}(w_T,w_S)+\frac{1}{2}\sqrt{\frac{1}{n}\left(\mathrm{KL}_{w_T}+\mathrm{KL}_{w_S}+2\ln\frac{2}{\delta}\right)+\left(\frac{\ln\left(n_{w_T}\cdot n_{w_S}\right)}{2n}\right)} \tag{37}$$

$$\leq \widehat{\mathrm{CVL}}(w_T,w_S)+\frac{1}{2}\sqrt{\frac{1}{n}\left(\mathrm{KL}_{w_T}+\mathrm{KL}_{w_S}+2\ln\frac{2}{\delta}\right)+\left(\frac{\ln n^{\star}}{n}\right)}, \tag{38}$$

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,*

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

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

$$\widehat{\mathrm{CVL}}(w_T,w_S) \tag{39}$$

$$=\frac{1}{2}\mathop{\mathbb{E}}_{x_S,x_T}\left[\widehat{\mathcal{L}}(x_T,w_S)+\widehat{\mathcal{L}}(x_S,w_T)\right] \tag{40}$$

$$=\frac{1}{2}\left(\mathbb{E}_{x_T}\left[\widehat{\mathcal{L}}(x_T,w_S)\right]+\mathbb{E}_{x_S}\left[\widehat{\mathcal{L}}(x_S,w_T)\right]\right) \tag{41}$$

$$=\frac{1}{2}\left(\frac{\sum_{x_T\in\mathcal{S}_{w_T}}\sum_{x_S\in\mathcal{S}_{w_S}}\mathbb{E}_{h\sim\mathcal{P}_{x_T}}\left[\mathbb{1}\left\{h(x_S)\neq z_{x_S}\right\}\right]}{n_{w_T}\cdot n_{w_S}}\right. \tag{42}$$

$$\left.+\frac{\sum_{x_S\in\mathcal{S}_{w_S}}\sum_{x_T\in\mathcal{S}_{w_T}}\mathbb{E}_{h\sim\mathcal{P}_{x_S}}\left[\mathbb{1}\left\{h(x_T)\neq z_{x_T}\right\}\right]}{n_{w_S}\cdot n_{w_T}}\right) \tag{43}$$

$$=\frac{1}{2}\left(\frac{\sum_{x_T\in\mathcal{S}_{w_T}}\sum_{x_S\in\mathcal{S}_{w_S}}\left(1-\mathcal{P}_{x_T,z_{x_S}}\right)}{n_{w_T}\cdot n_{w_S}}+\frac{\sum_{x_S\in\mathcal{S}_{w_S}}\sum_{x_T\in\mathcal{S}_{w_T}}\left(1-\mathcal{P}_{x_S,z_{x_T}}\right)}{n_{w_S}\cdot n_{w_T}}\right) \tag{44}$$

$$=1-\frac{1}{2}\left(\frac{\sum_{x_T\in\mathcal{S}_{w_T}}\sum_{x_S\in\mathcal{S}_{w_S}}\mathcal{P}_{x_T,z_{x_S}}}{n_{w_T}\cdot n_{w_S}}+\frac{\sum_{x_S\in\mathcal{S}_{w_S}}\sum_{x_T\in\mathcal{S}_{w_T}}\mathcal{P}_{x_S,z_{x_T}}}{n_{w_S}\cdot n_{w_T}}\right) \tag{45}$$

$$=1-\frac{1}{2}\sum_{k=1}^{K}\left(\frac{n_{k|w_S}\cdot\sum_{x_T\in\mathcal{S}_{w_T}}\mathcal{P}_{x_T,k}}{n_{w_T}\cdot n_{w_S}}+\frac{n_{k|w_T}\cdot\sum_{x_S\in\mathcal{S}_{w_S}}\mathcal{P}_{x_S,z_{x_T}}}{n_{w_S}\cdot n_{w_T}}\right) \tag{46}$$

$$=1-\frac{1}{2}\sum_{k=1}^{K}\left(\widehat{\varphi}_k^{(w_S)}\cdot\frac{\sum_{x_T\in\mathcal{S}_{w_T}}\mathcal{P}_{x_T,k}}{n_{w_T}}+\widehat{\varphi}_k^{(w_T)}\cdot\frac{\sum_{x_S\in\mathcal{S}_{w_S}}\mathcal{P}_{x_S,z_{x_T}}}{n_{w_S}}\right) \tag{47}$$

$$\geq 1-\frac{1}{2}\sum_{k=1}^{K}\left(\widehat{\varphi}_k^{(w_S)}\cdot\frac{n_{k|w_T}}{n_{w_T}}+\widehat{\varphi}_k^{(w_T)}\cdot\frac{n_{k|w_S}}{n_{w_S}}\right) \tag{48}$$

$$=1-\frac{1}{2}\sum_{k=1}^{K}\left(\widehat{\varphi}_k^{(w_S)}\cdot\widehat{\varphi}_k^{(w_T)}+\widehat{\varphi}_k^{(w_T)}\cdot\widehat{\varphi}_k^{(w_S)}\right) \tag{49}$$

$$=1-\widehat{\varphi}^{(w_S)\top}\cdot\widehat{\varphi}^{(w_T)} \tag{50}$$

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*

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

Proof. We first expand the inner product of $\widehat{\theta_S}^\top \cdot \widehat{\theta_T}$ as follows,

$$\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)} = \sum_{k=1}^{K} \widehat{\theta}_k^{(d_S)\top} \cdot \widehat{\theta}_k^{(d_T)} \tag{51}$$

$$= \sum_{k=1}^{K} \left(\left(\sum_{w_S \in V^{(S)}} f_{w_S}^{d_S} \cdot \widehat{\varphi}_k^{(w_S)} \right) \cdot \left(\sum_{w_T \in V^{(T)}} f_{w_T}^{d_T} \cdot \widehat{\varphi}_k^{(w_T)} \right) \right) \tag{52}$$

$$\leq F(d_S, d_T) \cdot \sum_{k=1}^{K} \left(\left(\sum_{w_S \in V^{(S)}} \widehat{\varphi}_k^{(w_S)2} \right)^{\frac{1}{2}} \cdot \left(\sum_{w_T \in V^{(T)}} \widehat{\varphi}_k^{(w_T)2} \right)^{\frac{1}{2}} \right), \tag{53}$$

$$F(d_S, d_T) = \left(\sum_{w_S \in V^{(S)}} {f_{w_S}^{d_S}}^2 \right)^{\frac{1}{2}} \cdot \left(\sum_{w_T \in V^{(T)}} {f_{w_T}^{d_T}}^2 \right)^{\frac{1}{2}}, \tag{54}$$

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.

$$\sum_{k=1}^{K} \left(\left(\sum_{w_S \in V^{(S)}} \widehat{\varphi}_k^{(w_S)2} \right)^{\frac{1}{2}} \cdot \left(\sum_{w_T \in V^{(T)}} \widehat{\varphi}_k^{(w_T)2} \right)^{\frac{1}{2}} \right) \tag{55}$$

$$= \sum_{k=1}^{K} \left(\sum_{w_S, w_T} \left(\widehat{\varphi}_k^{(w_S)} \cdot \widehat{\varphi}_k^{(w_T)} \right)^2 \right)^{\frac{1}{2}} \tag{56}$$

$$\leq \sqrt{K} \cdot \left(\sum_{k=1}^{K} \sum_{w_S, w_T} \left(\widehat{\varphi}_k^{(w_S)} \cdot \widehat{\varphi}_k^{(w_T)} \right)^2 \right)^{\frac{1}{2}} \tag{57}$$

$$= \sqrt{K} \cdot \left(\sum_{w_S, w_T} \sum_{k=1}^{K} \left(\widehat{\varphi}_k^{(w_S)} \cdot \widehat{\varphi}_k^{(w_T)} \right)^2 \right)^{\frac{1}{2}} \tag{58}$$

$$\leq \sqrt{K} \cdot \left(\sum_{w_S, w_T} \left(\sum_{k=1}^{K} \widehat{\varphi}_k^{(w_S)} \cdot \widehat{\varphi}_k^{(w_T)} \right)^2 \right)^{\frac{1}{2}} \tag{59}$$

$$= \sqrt{K} \cdot \left(\sum_{w_S, w_T} \left(\widehat{\varphi}^{(w_T)\top} \cdot \widehat{\varphi}^{(w_S)} \right)^2 \right)^{\frac{1}{2}}. \tag{60}$$

Thus, we have the following inequality:

$$\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)} \leq F(d_S, d_T) \cdot \sqrt{K} \cdot \left(\sum_{w_S, w_T} \left(\widehat{\varphi}^{(w_T)\top} \cdot \widehat{\varphi}^{(w_S)} \right)^2 \right)^{\frac{1}{2}}. \tag{61}$$

Plug in Lemma 1, we see that

$$\widehat{\theta}^{(d_S)\top} \cdot \widehat{\theta}^{(d_T)} \quad \leq \quad F(d_S, d_T) \cdot \sqrt{K} \cdot \left(\sum_{w_S, w_T} \left(\widehat{\text{CVL}}(w_T, w_S) - 1 \right)^2 \right)^{\frac{1}{2}} . \tag{62}$$

□

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 1 lists the source of the stemmers and stopwords.

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 1: List of source of stemmers and stopwords used in experiments.

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 2. The numbers are calculated after stemming and lemmatization.

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

Table 2: Statistics of the Wikipedia training corpus.

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 3. This Wikipedia corpus does not overlap with the training set.

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

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

Table 3: Statistics of the Wikipedia corpus for topic coherence evaluation (CNPMI).

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

Table 4: Statistics of the Global Voices (GV) corpus.

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 4. 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/`.

D Topic Model Configurations

For each experiment, we run five chains of Gibbs sampling using the Polylingual Topic Model implemented in MALLET (McCallum, 2002; Mimno et al., 2009), 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.

References

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.

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.