A Notation

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

B Proofs

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

**Proof.** We first notice the triangle inequality:

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

$$= \left| \mathbb{E}_{x_T \in S_{w_T}} \left[ \hat{L}^{(t)}(x_T, w_S) + \hat{L}^{(t-1)}(x_S, w_T) \right] - \mathbb{E}_{x_T \in S_{w_T}} \left[ \hat{L}^{(t-1)}(x_T, w_S) + \hat{L}^{(t-1)}(x_S, w_T) \right] \right|$$

$$= \left| \mathbb{E}_{x_T \in S_{w_T}} \left[ \hat{L}^{(t)}(x_T, w_S) \right] - \mathbb{E}_{x_T \in S_{w_T}} \left[ \hat{L}^{(t-1)}(x_T, w_S) \right] \right|$$

$$- \left| \mathbb{E}_{x_T \in S_{w_T}} \left[ \hat{L}^{(t)}(x_T, w_S) \right] - \mathbb{E}_{x_S \in S_{w_S}} \left[ \hat{L}^{(t)}(x_S, w_T) \right] \right|$$
We look at the first term of Equation (6), and the other term can be derived in the same way. We use $P_{xT}$ to denote the invariant distribution of the conditional $P_{xT}^{(t)}$ as $t \to \infty$. Additionally, let $P_{xT,z_w}$ be the conditional probability for the token $x_T$ being assigned to topic $z_w$: \[
P_{xT,z_w} = \Pr(k = z_w; w = w_{xT}, z_-, w_-).
\]

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

\[
\begin{align*}
\lim_{t \to \infty} \Delta \mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \hat{L}(x_T, w_S) & \leq \mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \left[ \hat{L}(x_T, w_S) - \hat{L}(x_T, w_S) \right] \\
& \quad + \mathbb{E}_{x_S \in \mathcal{S}_{w_S}} \left[ \hat{L}(x_S, w_T) - \hat{L}(x_S, w_T) \right] \\
& \equiv \Delta \mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \left[ \hat{L}(x_T, w_S) \right] + \Delta \mathbb{E}_{x_S \in \mathcal{S}_{w_S}} \left[ \hat{L}(x_S, w_T) \right] \\
& \quad \leq \Delta \mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \left[ \hat{L}(x_T, w_S) \right] + \Delta \mathbb{E}_{x_S \in \mathcal{S}_{w_S}} \left[ \hat{L}(x_S, w_T) \right] 
\end{align*}
\]

Therefore, we have

\[
\begin{align*}
\lim_{t \to \infty} \mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \left[ \hat{C}_{\mathcal{V}L}^{(t)}(w_T, w_S) - \hat{C}_{\mathcal{V}L}^{(t-1)}(w_T, w_S) \right] & \leq \lim_{t \to \infty} \Delta \mathbb{E}_{x_T \in \mathcal{S}_{w_T}} \left[ \hat{L}(x_T, w_S) \right] + \Delta \mathbb{E}_{x_S \in \mathcal{S}_{w_S}} \left[ \hat{L}(x_S, w_T) \right] = 0.
\end{align*}
\]
Theorem 3. Given a bilingual word pair \((w_T, w_S)\), with probability at least \(1 - \delta\), the following bound holds:

\[
CVL(w_T, w_S) \leq CVL(w_T, w_S) + \frac{1}{2} \sqrt{\frac{1}{n} \left( KL_{wT} + KL_{wS} + 2 \ln \frac{2}{\delta} \right) + \left( \ln n^* \right)},
\]

(18)

\[
n = \min \{ n_{wT}, n_{wS} \}, \quad n^* = \max \{ n_{wT}, n_{wS} \}.
\]

(19)

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

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

\[
L(x_T, w_S) \leq \hat{L}(x_T, w_S) + \sqrt{\frac{1}{2n_{wS}} \left( KL(P_{xT}||Q_{xT}) + \ln \frac{2\sqrt{n_{wS}}}{\delta} \right)}
\]

(20)

\[
= \hat{L}(x_T, w_S) + \sqrt{\frac{1}{2n_{wS}} \left( KL(P_{xT}||Q_{xT}) + \ln \frac{2}{\delta} + \frac{1}{2} \ln n_{wS} \right)}
\]

(21)

\[
= \hat{L}(x_T, w_S) + \epsilon(x_T, w_S).
\]

(22)

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

\[
L(x_S, w_T) \leq \hat{L}(x_S, w_T) + \sqrt{\frac{1}{2n_{wT}} \left( KL(P_{xS}||Q_{xS}) + \ln \frac{2}{\delta} + \frac{1}{2} \ln n_{wT} \right)}
\]

(23)

\[
= \hat{L}(x_S, w_T) + \epsilon(x_S, w_T).
\]

(24)

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,

\[
KL_{w_S} = KL(P_{x_T^\ast}||Q_{x_T^\ast}), \quad x_T^\ast = \arg \max_{x_T \in S_{w_T}} \epsilon(x_T, w_S);
\]

(25)

\[
KL_{w_T} = KL(P_{x_S^\ast}||Q_{x_S^\ast}), \quad x_S^\ast = \arg \max_{x_S \in S_{w_S}} \epsilon(x_S, w_T).
\]

(26)

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

\[
CVL(w_T, w_S) \leq \frac{1}{2} \mathbb{E}_{x_T, x_T} [L(x_T, w_S) + L(x_S, w_T)]
\]

(27)

\[
= \frac{1}{2} \mathbb{E}_{x_T, x_T} [\hat{L}(x_T, w_S) + \hat{L}(x_S, w_T)]
\]

(28)

\[
\leq \frac{1}{2} \left( \mathbb{E}_{x_T \in S_{w_T}} \hat{L}(x_T, w_S) + \mathbb{E}_{x_S \in S_{w_S}} \hat{L}(x_S, w_T) \right)
\]

(29)

\[
+ \frac{1}{2} \left( \mathbb{E}_{x_T \in S_{w_T}} \epsilon(x_T, w_S) + \mathbb{E}_{x_S \in S_{w_S}} \epsilon(x_S, w_T) \right)
\]

(30)

\[
= \hat{CVL}(w_T, w_S) + \frac{1}{2} \left( \mathbb{E}_{x_T \in S_{w_T}} \epsilon(x_T, w_S) + \mathbb{E}_{x_S \in S_{w_S}} \epsilon(x_S, w_T) \right)
\]

(31)

\[
\leq \hat{CVL}(w_T, w_S) + \frac{1}{2} \left( \epsilon(x_T^\ast, w_S) + \epsilon(x_S^\ast, w_T) \right)
\]

(32)

\[
\leq \hat{CVL}(w_T, w_S) + \frac{1}{2} \left( \epsilon(x_T^\ast, w_S) + \epsilon(x_S^\ast, w_T) \right)
\]

(33)

\[
\leq \hat{CVL}(w_T, w_S)
\]

(34)
\[ \frac{1}{2} \left( \sqrt{\frac{1}{2n_{wT}} \left( \text{KL}_{wT} + \frac{2}{\delta} + \frac{1}{2} \ln n_{wT} \right)} \right) + \frac{1}{2} \left( \sqrt{\frac{1}{2n_{wS}} \left( \text{KL}_{wS} + \frac{2}{\delta} + \frac{1}{2} \ln n_{wS} \right)} \right) \]

\[ \leq \text{CVL}(w_T, w_S) + \frac{1}{2} \sqrt{\frac{1}{n} \left( \text{KL}_{wT} + \text{KL}_{wS} + 2 \ln \frac{2}{\delta} + \left( \frac{\ln (n_{wT} \cdot n_{wS})}{2n} \right) \right)} \]

\[ \leq \text{CVL}(w_T, w_S) + \frac{1}{2} \sqrt{\frac{1}{n} \left( \text{KL}_{wT} + \text{KL}_{wS} + 2 \ln \frac{2}{\delta} + \left( \frac{\ln n^*}{n} \right) \right)} \]

which gives us the result.

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

\[ 1 - \hat{\varphi}(w_T)^\top \cdot \hat{\varphi}(w_S) \leq \text{CVL}(w_T, w_S). \]

Proof. We expand the equation of \(\text{CVL}\) as follows,

\[ \text{CVL}(w_T, w_S) \]

\[ = \frac{1}{2} \mathbb{E}_{x_T, x_S} \left[ \hat{L}(x_T, w_S) + \hat{L}(x_S, w_T) \right] \]

\[ = \frac{1}{2} \left( \mathbb{E}_{x_T} \left[ \hat{L}(x_T, w_S) \right] + \mathbb{E}_{x_S} \left[ \hat{L}(x_S, w_T) \right] \right) \]

\[ = \frac{1}{2} \left( \sum_{x_T \in S_{wT}} \sum_{x_S \in S_{wS}} \mathbb{E}_{h \sim P_{x_T}} \left[ \mathbb{I} \{ h(x_S) \neq z_{x_T} \} \right] \right) \]

\[ = \frac{1}{2} \left( \sum_{x_T \in S_{wT}} \sum_{x_S \in S_{wS}} \mathbb{E}_{h \sim P_{x_T}} \left[ \mathbb{I} \{ h(x_T) \neq z_{x_T} \} \right] \right) \]

\[ = 1 - \frac{1}{2} \left( \sum_{x_T \in S_{wT}} \sum_{x_S \in S_{wS}} \frac{1 - P_{x_T, z_{x_S}}}{n_{wT} \cdot n_{wS}} + \sum_{x_S \in S_{wS}} \sum_{x_T \in S_{wT}} \frac{1 - P_{x_T, z_{x_S}}}{n_{wS} \cdot n_{wT}} \right) \]

\[ = 1 - \frac{1}{2} \left( \sum_{x_T \in S_{wT}} \sum_{x_S \in S_{wS}} \frac{P_{x_T, z_{x_S}}}{n_{wT} \cdot n_{wS}} + \sum_{x_S \in S_{wS}} \sum_{x_T \in S_{wT}} \frac{P_{x_T, z_{x_S}}}{n_{wS} \cdot n_{wT}} \right) \]

\[ = 1 - \frac{1}{2} \sum_{k=1}^{K} \left( \hat{\varphi}_k(w_S) \cdot \frac{n_{k|w_T}}{n_{wT}} + \hat{\varphi}_k(w_T) \cdot \frac{n_{k|w_S}}{n_{wS}} \right) \]

\[ = 1 - \frac{1}{2} \sum_{k=1}^{K} \left( \hat{\varphi}_k(w_T) \cdot \hat{\varphi}_k(w_S) \right) \]

which concludes the proof.
Theorem 5. Let $\hat{\theta}^{(ds)}$ be the distribution over topics for document $d_S$ (similarly for $d_T$), $F(d_S, d_T) = \left( \sum_{w_S} f^{d_S}_{w_S} \cdot \sum_{w_T} f^{d_T}_{w_T} \right)^{\frac{1}{2}}$ where $f^d_w$ is the normalized frequency of word $w$ in document $d$, and $K$ the number of topics. Then

$$\hat{\theta}^{(ds)} \cdot \hat{\theta}^{(dt)} \leq F(d_S, d_T) \cdot \sqrt{K \cdot \sum_{w_S, w_T} (\overline{CVL}(w_S, w_T) - 1)^2}.$$

Proof. We first expand the inner product of $\hat{\theta}^{(ds)} \cdot \hat{\theta}^{(dt)}$ as follows,

$$\hat{\theta}^{(ds)} \cdot \hat{\theta}^{(dt)} = \sum_{k=1}^{K} \hat{\theta}_k^{(ds)} \cdot \hat{\theta}_k^{(dt)} \quad (51)$$

$$= \sum_{k=1}^{K} \left( \left( \sum_{w_S \in V(S)} f^{d_S}_{w_S} \cdot \hat{\varphi}^{(ws)}_k \right) \cdot \left( \sum_{w_T \in V(T)} f^{d_T}_{w_T} \cdot \hat{\varphi}^{(wt)}_k \right) \right) \quad (52)$$

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

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

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.

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

$$= \sum_{k=1}^{K} \left( \sum_{w_S, w_T} \hat{\varphi}^{(ws)}_k \cdot \hat{\varphi}^{(wt)}_k \right)^{\frac{1}{2}} \quad (56)$$

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

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

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

$$= \sqrt{K} \cdot \left( \sum_{w_S, w_T} \left( \hat{\varphi}^{(wt)} \cdot \hat{\varphi}^{(ws)} \right)^{\frac{1}{2}} \right). \quad (60)$$

Thus, we have the following inequality:

$$\hat{\theta}^{(ds)} \cdot \hat{\theta}^{(dt)} \leq F(d_S, d_T) \cdot \sqrt{K} \cdot \left( \sum_{w_S, w_T} \left( \hat{\varphi}^{(wt)} \cdot \hat{\varphi}^{(ws)} \right)^{\frac{1}{2}} \right)^{\frac{1}{2}}. \quad (61)$$
Plug in Lemma 1, we see that
\[
\hat{\theta}^\top(d_S) \cdot \hat{\theta}^\top(d_T) \leq F(d_S, d_T) \cdot \sqrt{K} \cdot \left( \sum_{w_S, w_T} (\tilde{CVL}(w_T, w_S) - 1)^2 \right)^{\frac{1}{2}}.
\] (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.

<table>
<thead>
<tr>
<th>Language</th>
<th>Family</th>
<th>Stemmer</th>
<th>Stopwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Semitic</td>
<td>Assem’s Arabic Light Stemmer</td>
<td>GitHub</td>
</tr>
<tr>
<td>DE</td>
<td>Germanic</td>
<td>SnowBallStemmer</td>
<td>NLTK</td>
</tr>
<tr>
<td>EN</td>
<td>Germanic</td>
<td>SnowBallStemmer</td>
<td>NLTK</td>
</tr>
<tr>
<td>ES</td>
<td>Romance</td>
<td>SnowBallStemmer</td>
<td>NLTK</td>
</tr>
<tr>
<td>RU</td>
<td>Slavic</td>
<td>SnowBallStemmer</td>
<td>NLTK</td>
</tr>
<tr>
<td>ZH</td>
<td>Sinitic</td>
<td>Jieba</td>
<td>GitHub</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Paired language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#docs</td>
<td>#token</td>
</tr>
<tr>
<td>AR</td>
<td>3,000</td>
<td>724,362</td>
</tr>
<tr>
<td>DE</td>
<td>3,000</td>
<td>409,381</td>
</tr>
<tr>
<td>ES</td>
<td>3,000</td>
<td>451,115</td>
</tr>
<tr>
<td>RU</td>
<td>3,000</td>
<td>480,715</td>
</tr>
<tr>
<td>ZH</td>
<td>3,000</td>
<td>480,142</td>
</tr>
</tbody>
</table>

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.

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

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

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

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

#### 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
