# Learning Multilingual Topics from Incomparable Corpora

**Shudong Hao**
Computer Science
University of Colorado
Boulder, CO, USA
shudong@colorado.edu

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

## Abstract

Multilingual topic models enable crosslingual tasks by extracting consistent topics from multilingual corpora. Most models require parallel or comparable training corpora, which limits their ability to generalize. In this paper, we first demystify the knowledge transfer mechanism behind multilingual topic models by defining an alternative but equivalent formulation. Based on this analysis, we then relax the assumption of training data required by most existing models, creating a model that only requires a dictionary for training. Experiments show that our new method effectively learns coherent multilingual topics from partially and fully *incomparable* corpora with limited amounts of dictionary resources.

## 1 Introduction

Multilingual topic models provide an overview of document structures in multilingual corpora, by learning language-specific versions of each topic (Figure 1). Their simplicity, efficiency and interpretability make models from this family popular for various crosslingual tasks, *e.g.,* feature extraction (Liu et al., 2015), cultural difference discovery (Shutova et al., 2017; Gutiérrez et al., 2016), translation detection (Krstovski et al., 2016; Krstovski and Smith, 2016), and others (Barrett et al., 2016; Agić et al., 2016; Hintz and Biemann, 2016).

Typical probabilistic multilingual topic models are based on Latent Dirichlet Allocation (LDA, Blei et al. (2003)), adding supervision on connections between languages. Most models achieve this by making strong assumptions on the training data—they either require a *parallel corpus* that has sentence-aligned documents in different languages (*e.g.,* EuroParl, Koehn (2005)), or a *comparable corpus* that has documents of similar content (*e.g.,* Wikipedia articles paired across languages). These training requirements limit the usage of such models: an adequately large parallel corpus is difficult to obtain, particularly for low-resource languages. For example, only 300 languages are available on Wikipedia,[1] and only 250 languages have more than 1,000 articles. Another common choice for parallel corpus in multilingual research, the Bible, is available in 2,530 languages (Agić et al., 2015).[2] However, studies show that its archaic themes and small corpus size (1,189 chapters) can limit performance (Hao et al., 2018; Moritz and Büchler, 2017). Therefore, the requirement of parallel/comparable corpora for multilingual topic models limits their usage in many situations.

Another line of research focuses on using multilingual dictionaries as supervision (Ma and Nasukawa, 2017; Gutiérrez et al., 2016; Liu et al., 2015; Jagarlamudi and Daumé III, 2010; Boyd-Graber and Blei, 2009). In contrast to parallel corpora, dictionaries are widely available and often easy to obtain. PANLEX, a free online dictionary database, for example, covers 5,700 languages and more than one billion dictionary entries (Kamholz et al., 2014; Baldwin et al., 2010).[3] Thus, a multilingual topic model built on a dictionary rather than a parallel corpus is potentially applicable to more languages.

A dictionary also allows for training on *incomparable* corpora—documents in different languages that are from different sources without direct connections—which have had less research on learning consistent topics. With a dictionary, a natural question is how to efficiently utilize the semantic information it

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[2] Reported by United Bible Societies at https://www.unitedbiblesocieties.org/
[3] https://panlex.org/

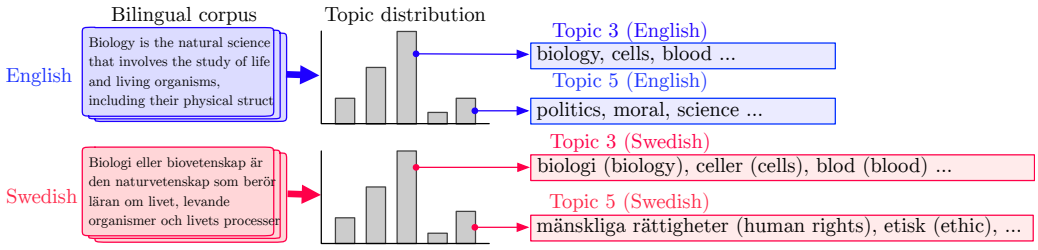



Figure 1: Multilingual topic models produce topics where each language has its own version.

carries so that a topic model can produce multilingually coherent topics. This work considers an alternative formulation of a dictionary-based topic model, one that borrows the structure of models used with comparable corpora, but uses a dictionary-based metric to learn connections between documents, instead of explicit connections from a comparable corpus. The main contributions of this work are:

- We summarize existing related work in Section 2 and propose a new formulation of multilingual topic models based on crosslingual transfer learning in Section 3. This new formulation explicitly shows the knowledge transfer mechanism during the generative process.
- Based on this new formulation, in Section 4 we generalize existing multilingual topic models and relax the assumptions of parallel/comparable datasets. Our approach requires only a dictionary, and is empirically shown to perform well even with only limited amounts of available entries.
- We evaluate our new model on five languages from different language families in Section 5. Our proposed model learns multilingually coherent topics and yields around a $25\%$ relative improvement in crosslingual classification performance.

## 2 Multilingual Topic Models

Multilingual topic models generate $K$ topics from a corpus consisting of multiple languages; each topic has a version specific to each language in the corpus (Figure 1). From a human's view, a coherent multilingual topic should talk about the same thing regardless of the language; from a machine's view, the success of multilingual topic models depends on the inferred topics being consistent across languages. For example, given an English-Swedish bilingual topic $\phi_k^{(\text{EN,SV})}$, the probability of an English word *island* and that of its translation in Swedish, *ö*, should be similar, *i.e.,* $\Pr\left(island_{\text{EN}}|\phi_k^{(\text{EN,SV})}\right) \approx \Pr\left(\ddot{o}_{\text{SV}}|\phi_k^{(\text{EN,SV})}\right)$. Most multilingual topic models extend LDA with one or both of two types of "link" information: document translations and word translations.

**Document Links.** The polylingual topic model (Mimno et al., 2009; Ni et al., 2009) assumes that during the generative process, a topic distribution $\theta_{\mathbf{d}}$ generates a tuple of comparable documents in different languages, *i.e.,* $\mathbf{d} = \left(d^{(\ell_1)}, \ldots, d^{(\ell_L)}\right)$ and each language $\ell$ has its own topic-word distributions, $\phi_k^{(\ell)}$. This model has been widely used (Vulić et al., 2013; Platt et al., 2010; Smet and Moens, 2009), but it requires a parallel/comparable corpus in order to link documents.

**Vocabulary Links.** Another type of model uses word translations (Jagarlamudi and Daumé III, 2010; Boyd-Graber and Blei, 2009) rather than linking documents. A multilingual dictionary is used to construct a tree structure where each internal node contains word translations, and applies hyper-Dirichlet type I distributions to generate words (Andrzejewski et al., 2009; Minka, 1999; Dennis III, 1991). For each topic $k$, a distribution from root $r$ to all the internal nodes $i$ is drawn by $\phi_{k,r} \sim \text{Dir}(\beta_r)$, and then a distribution from $i$ to a leaf node is drawn by $\phi_{k,i}^{(\ell)} \sim \text{Dir}\left(\beta_i^{(\ell)}\right)$. A word $w^{(\ell)}$ in language $\ell$ is drawn from a product of the two multinomial distributions by $w^{(\ell)} \sim \text{Mult}\left(\phi_{k,r} \cdot \phi_{k,i}^{(\ell)}\right)$.

**Variations.** Many variations of these ideas have been proposed to deal with non-parallel corpus. Heyman et al. (2016) proposed C-BILDA, which distinguishes between shared and non-shared topics across languages, based on a document links model. The model, however, requires a comparable dataset that

provides document links between languages. A variation proposed by Ma and Nasukawa (2017) deals with non-parallel corpora. This model is essentially a modified version of Jagarlamudi and Daumé III (2010) and Boyd-Graber and Blei (2009), so we consider this work to be another vocabulary links model. Other models have been proposed for very specific situations that needs additional supervision. For example, Krstovski et al. (2016) requires scientific article section alignments, and Gutiérrez et al. (2016) requires Part-of-Speech (POS) taggers, which are not always available for all languages. Without POS taggers, this model is equivalent to vocabulary links. In our work, we focus on the standard document links and vocabulary links models, which are the most generalizable models.

## 3 Document Links: A Crosslingual Transfer Perspective

Before we introduce our new approach, we first present an alternative understanding of the document links model from the perspective of crosslingual transfer learning. In multilingual topic models, "knowledge" refers to word distributions for a topic in a language $\ell$, and we study how multilingual topic models transfer this knowledge from one language to another so that the model provides semantically coherent topics that are consistent across languages.

In the standard document links model, a "link" between a document $d_{\ell_1}$ in language $\ell_1$ and $d_{\ell_2}$ in $\ell_2$ indicates that they are translations or closely comparable. In this model, the topic assignments for both documents are independently generated from the same distribution, $\theta_{d_{\ell 1}, d_{\ell 2}}$. Thus, the joint likelihood of document links model is:

$$\Pr\left(\mathbf{w}_{d_{\ell 1}}, \mathbf{z}_{d_{\ell 1}}, \mathbf{w}_{d_{\ell 2}}, \mathbf{z}_{d_{\ell 2}} | \alpha, \beta\right), \tag{1}$$

where $\mathbf{w}_{d_\ell}$ and $\mathbf{z}_{d_\ell}$ are the word tokens and topic assignments of document $d_\ell$. We refer this formulation as the **joint generative model**, since the topics and words of $d_{\ell_1}$ and $d_{\ell_2}$ are generated *simultaneously*.

The simultaneousness of this model formulation, in which both languages generate topics jointly, masks the knowledge transfer process. To highlight this process, and to help us generalize the model in the next section, we define an alternative formulation in which $d_{\ell_1}$ and $d_{\ell_2}$ are generated *sequentially*.

Assume the topics of $d_{\ell_1}$ have already been generated from $\theta_{d_{\ell 1}} \sim \mathrm{Dir}(\alpha)$, and $\mathbf{n}_{d_{\ell 1}} \in \mathbb{N}^K$ is a vector of topic counts in $d_{\ell_1}$. In our alternative formulation, the generation of topics of $d_{\ell_2}$ depends on $d_{\ell_1}$ by $\theta_{d_{\ell 2}} \sim \mathrm{Dir}(\alpha + \mathbf{n}_{d_{\ell 1}})$, where the prior $\alpha + \mathbf{n}_{d_{\ell 1}}$ encourages the distribution $\theta_{d_{\ell 2}}$ to be similar to $\theta_{d_{\ell 1}}$. This formulation can go the other way, *i.e.,* generating $d_{\ell_2}$ first, and then $d_{\ell_1}$. The combined likelihood of this formulation is:

$$\Pr\left(\mathbf{w}_{d_{\ell 1}}, \mathbf{z}_{d_{\ell 1}} | \mathbf{w}_{d_{\ell 2}}, \mathbf{z}_{d_{\ell 2}}, \alpha, \beta\right) \cdot \Pr\left(\mathbf{w}_{d_{\ell 2}}, \mathbf{z}_{d_{\ell 2}} | \mathbf{w}_{d_{\ell 1}}, \mathbf{z}_{d_{\ell 1}}, \alpha, \beta\right), \tag{2}$$

and we refer to this formulation as the **conditional generative model**.

This alternative formulation explicitly shows the knowledge transfer process across languages by shaping the topic parameters for $\ell_2$ to be similar to that of the other language $\ell_1$, and vice versa. In this formulation, the likelihood of the conditional generative model is different from the joint generative model. In fact, this is an instance of *pseudolikelihood* (Besag, 1975; Leppä-aho et al., 2017), where the joint likelihood of the two documents is approximated as the product of each document's conditional likelihood given the other, *i.e.,* $\Pr(d_{\ell_1}, d_{\ell_2}) \approx \Pr(d_{\ell_1} | d_{\ell_2}) \cdot \Pr(d_{\ell_2} | d_{\ell_1})$. As Leppä-aho et al. (2017) suggests, pseudolikelihood is not a numerically accurate approximation to the joint likelihood; Theorem 1 below, however, states that this formulation yields exactly the same posterior estimations of $\theta$ and $\phi$.

**Theorem 1.** *The conditional generative model with document links yields the same posterior estimator to the joint generative model using collapsed Gibbs sampling.*

*Proof.* See Appendix. □

## 4 Generalizing Document Links

Obtaining parallel corpora for training the document links model is very demanding, particularly for low-resource languages. Therefore, as the second major contribution in this paper, we generalize the document links model using the formulation described above to require only a bilingual dictionary.

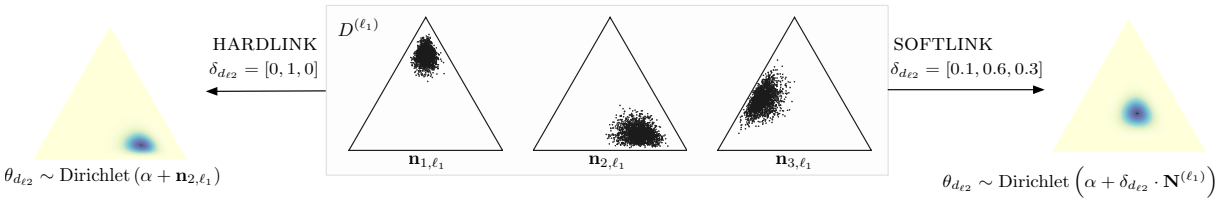



Figure 2: An illustration of how topic knowledge is transferred across languages through HARDLINK and SOFTLINK. To generate observations in $d_{\ell_2}$, both models uses topics in $\ell_1$ as prior knowledge to shape the Dirichlet prior for $d_{\ell_2}$. This transfer happens in HARDLINK by aligned documents in a comparable corpus, while SOFTLINK uses a generalized transfer distribution $\delta$.

### 4.1 From Hard Links to Soft Links

Following the above discussion, we introduce our method assuming the directionality from language $\ell_1$ to $\ell_2$. We generalize the model above by rewriting the generation of the distribution $\theta_{d_{\ell 2}}$:

$$\theta_{d_{\ell 2}} \quad \sim \quad \text{Dirichlet}\left(\alpha + \delta_{d_{\ell 2}} \cdot \mathbf{N}^{(\ell_1)}\right), \tag{3}$$

where $\mathbf{N}^{(\ell_1)} \in \mathbb{N}^{|D^{(\ell_1)}| \times K}$ is the matrix of topic counts per document with $K$ topics in the corpus $D^{(\ell_1)}$ of language $\ell_1$. This is equivalent to the above document links model when $\delta_{d_{\ell 2}} \in \mathbb{R}^{|D^{(\ell_1)}|}$ is an indicator vector that has value 1 for the corresponding parallel document $d_{\ell_1} \in D^{(\ell_1)}$ and 0 elsewhere. We refer this as **hard links** (HARDLINK), where each document $d_{\ell_2} \in D^{(\ell_2)}$ is informed by exactly one document $d_{\ell_1}$, and this link is known *a priori* from a parallel corpus.

We create **soft links** (SOFTLINK) by relaxing the assumption that $\delta_{d_{\ell 2}}$ is an indicator vector, instead allowing $\delta_{d_{\ell 2}}$ to be any *distribution* over documents in $D^{(\ell_1)}$, a mixture of potentially multiple documents in language $\ell_1$ to inform parameters for a document $d_{\ell_2}$ in language $\ell_2$. We refer this distribution as the **transfer distribution**. The Dirichlet prior for document $d_{\ell_2}$ contains topic knowledge $\mathbf{N}^{(\ell_1)}$ transferred from corpus $D^{(\ell_1)}$, encouraging $\theta_{d_{\ell 2}}$ to be proportionally similar to documents in $D^{(\ell_1)}$. Figure 2 illustrates this process.

### 4.2 Defining the Transfer Distribution

The transfer distribution of document $d_{\ell_2}$ indicates how much knowledge should be transferred from every document $d_{\ell_1} \in D^{(\ell_1)}$. Intuitively, if $d_{\ell_1}$ and $d_{\ell_2}$ have a large amount of overlapping word translations, their topics should be similar as well. Therefore, we define the values of $\delta$ based on the similarity of document pairs using a bilingual dictionary. Specifically, for a document $d_{\ell_2} \in D^{(\ell_2)}$ , the transfer distribution of $d_{\ell_2}$, denoted as $\delta_{d_{\ell 2}}$, is a normalized vector of size $|D^{(\ell_1)}|$, *i.e.,* the size of corpus $D^{(\ell_1)}$. Each cell in $\delta_{d_{\ell 2}}$ corresponds to a document $d_{\ell_1} \in D^{(\ell_1)}$, defined as:

$$\left(\delta_{d_{\ell 2}}\right)_{d_{\ell 1}} \quad \propto \quad \frac{|\,\{w_{\ell_1}\} \cap \{w_{\ell_2}\}\,|}{|\,\{w_{\ell_1}\} \cup \{w_{\ell_2}\}\,|}, \qquad \forall\, w_{\ell_1} \in d_{\ell_1} w_{\ell_2} \in d_{\ell_2}, \tag{4}$$

where $\{w_\ell\}$ contains all the word types that appear in document $d_\ell$, and $\{w_{\ell_1}\} \cap \{w_{\ell_2}\}$ indicates all word pairs $(w_{\ell_1}, w_{\ell_2})$ that can be found in a dictionary as translations. In other words, $\left(\delta_{d_{\ell 2}}\right)_{d_{\ell 1}}$ is the proportion of words in the document pair $(d_{\ell_1}, d_{\ell_2})$ that are translations of each other.

In practice, a dense transfer distribution is computationally inefficient and is less meaningful than a sparse distribution, as it becomes approximately uniform due to the large size of the corpus. The transfer distribution should be more heavily concentrated on documents with higher word-level translation probabilities, while reducing the noise negatively transferred from those with low probabilities. To this end, we propose two approaches to help transfer distributions more efficiently focus on specific documents.

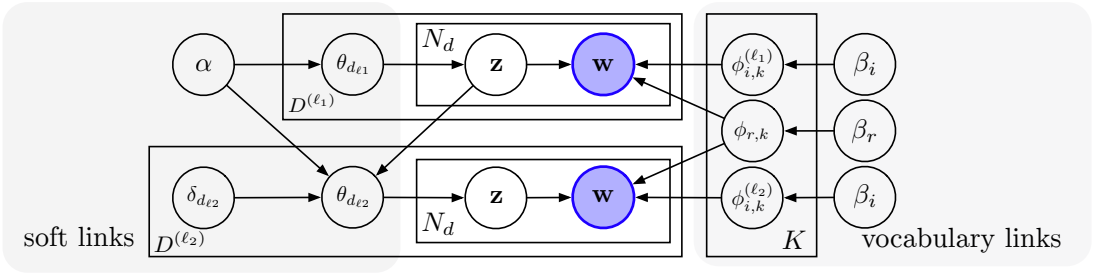



Figure 3: Plate notation of a multilingual topic model using soft links and vocabulary links.

#### 4.2.1 Static Focusing: a Threshold Method

The first method is to focus the distribution on the highest values such that values below a threshold are set to 0, while the remaining values are renormalized to sum to 1. The modified distribution is thus:

$$\left(\widetilde{\delta}_{d_{\ell 2}}\right)_{d_{\ell 1}} \quad \propto \quad \mathbb{1}\left\{\left(\delta_{d_{\ell 2}}\right)_{d_{\ell 1}} > \pi \cdot \max\left(\delta\right)\right\} \cdot \left(\delta_{d_{\ell 2}}\right)_{d_{\ell 1}} \tag{5}$$

where $\mathbb{1}$ is an indicator function, and $\pi \in [0, 1]$ is the **focal threshold**, a fixed parameter that adjusts the threshold. The threshold is defined with respect to the maximum value of $\delta$. A *corpus-wise* threshold chooses $\max(\delta)$ from all the $\delta_{d_{\ell 2}}$ in $D^{(\ell_2)}$ globally, while we also consider a *document-wise* threshold for each document, $\pi \cdot \max\left(\delta_{d_{\ell 2}}\right)$. We refer these two manners as the **selection scope**.

#### 4.2.2 Dynamic Focusing: an Annealing Method

Static focusing treats transfer distributions $\delta$ as fixed parameters during sampling, and it is difficult to decide how sparse a transfer distribution should be to achieve optimal performance. Therefore, we propose dynamic focusing, where we avoid choosing a specific focal threshold and selection scope. Specifically, we adjust the transfer distribution during inference dynamically, beginning with a dense transfer distribution and iteratively sharpening the distribution using deterministic annealing (Ueda and Nakano, 1994; Smith and Eisner, 2006; Paul and Dredze, 2015).

Assume at iteration $t$, the transfer distribution for a document $d_{\ell_2}$ is denoted as $\delta_{d_{\ell 2}}^{(t)}$. Then at iteration $t'$, we anneal its transfer distribution by $\left(\delta_{d_{\ell 2}}^{(t')}\right)_{d_{\ell 1}} \propto \left(\delta_{d_{\ell 2}}^{(t)}\right)_{d_{\ell 1}}^{1/\tau}$ where $\tau$ is a fixed temperature, which we set to 0.9 in our experiments. We start with non-focused transfer distributions, and apply annealing at scheduling intervals during Gibbs sampling.

Designing an effective annealing schedule is critical. We propose two schedules below.

**Fixed Schedule.** The simplest schedule is to apply annealing for all transfer distributions every $I$ iterations. In our experiments, we set $I = 10$ (*i.e.,* $t' = t + 10$) and stop annealing after 400 iterations. A potential problem with fixed schedule is that it can "over-anneal" the transfer distributions, *i.e.,* all the mass converges to only one document.

**Adaptive Schedule.** A robust multilingual topic model should produce similar distributions over topics for a pair of word translations $c = (w_{\ell_1}, w_{\ell_2})$, where we call $c$ a concept. In other words, given a topic $k$, the probability of expressing a concept $i$ in language $\ell_1$ should be similar to language $\ell_2$. Thus, during iteration $t$, we calculate $\varphi_c^{(\ell,t)}$, the distribution over $K$ topics for each concept $c$ for each language $\ell$. Using $\varphi_c^{(\ell,t)}$ as features and its language $\ell$ as labels, we perform five-fold cross-validation by logistic regression for all concepts $c$. We define the average classification accuracy over the five folds as the **language identification score** (LIS). The lower the LIS, the better the model, since a high LIS means the inferred distributions are inconsistent enough to discriminate between languages. This idea is related to adversarial training between languages (Chen et al., 2016).

During Gibbs sampling, we calculate LIS after each iteration, and average LIS every $I$ iterations. We anneal all transfer distributions at iteration $t$ only if $\overline{\text{LIS}}_{t-I:t} > \overline{\text{LIS}}_{t-2I:t-I}$. That is, if the average LIS

score during iteration $t - I$ and $t$ has been increasing since iteration $t - 2I$ to $t - I$, we treat this as a warning sign of increased LIS and thus anneal the transfer distributions. As we sharpen $\delta$ by annealing, knowledge transfer between languages becomes more specific.

### 4.3 Modularity of Models

Multilingual topic models can include the different types of information we described in Section 2: document links, vocabulary links, or both (Hu et al., 2014), while a model with neither is equivalent to LDA. Document links can be either hard or soft, or a mix of both, as the only distinction is whether the transfer distribution is an indicator vector. A complete model with both soft document links and vocabulary links is shown in Figure 3. In Section 5, we experiment with a combination of SOFTLINK and VOCLINK.

## 5 Experiments

### 5.1 Data

We use five corpora in five languages from different language groups: Arabic (AR, Semitic), Spanish (ES, Romance), Farsi (FA, Indo-Iranian), Russian (RU, Slavic), and Chinese (ZH, Sinitic). Each language is paired with English (EN, Germanic), and we train multilingual topic models on these language pairs individually. All the corpora listed below are available at `http://opus.nlpl.eu/`. For preprocessing, we use stemmers to lemmatize and segment Chinese documents, and then remove stop words and the most frequent 100 word types for each language. Refer to the appendix for additional details.

**Training corpora.** As in many multilingual studies (Ruder et al., 2017), we use Wikipedia as our training corpus for multilingual topic models, and create two corpora, WIKI-INCO and WIKI-PACO for each language pair $(\text{EN}, \ell)$. For WIKI-INCO, we randomly select $2,000$ documents in each language without any connections, so that no documents are translations of each other (an *incomparable* corpus). We also create a *partially comparable* corpus, WIKI-PACO, which contains around $30\%$ comparable document pairs for each language pair.

**Test corpora.** We create two test corpora for each language pair $(\text{EN}, \ell)$ from TED Talks 2013 (TED) and Global Voices (GV), which provide categories for each document that can be used as classification labels. The first one, TED+TED, contains documents from TED in both languages, while the second one, TED+GV contains English documents from TED and non-English documents from GV. After training a topic model, we use $\phi^{(\text{EN})}$ and $\phi^{(\ell)}$ to infer topics from both languages. For TED+TED, we choose the five most frequent labels in TED as the label set (*technology*, *culture*, *science*, *global issues*, and *design*); for TED+GV, we replace *global issues* and *design* with *business* and *politics*, since the label set from GV does not include *global issues* and *design*.

**Dictionary.** We use Wiktionary to extract word translations for VOCLINK and to calculate transfer distribution values $\delta$ for SOFTLINK. The dictionary is available at `https://dumps.wikimedia.org/enwiktionary/`.

### 5.2 Inference Settings

For each compared model, we set the number of topics $K = 25$. We run the Gibbs samplers for $1,000$ training iterations and 500 iterations to infer topic distributions on test corpora. We set Dirichlet priors $\alpha = 0.1$ and $\beta = 0.01$ for HARDLINK and SOFTLINK. For VOCLINK, we set $\beta_r = 0.01$ for priors from root to internal nodes, and $\beta_i = 100$ from internal nodes $i$ to leaves, following Hu et al. (2014).

### 5.3 Evaluation Metrics

We evaluate each model in two ways. Experimental results below are averaged across all language pairs.

#### 5.3.1 Intrinsic Evaluation: Multilingual Topic Coherence

Typical topic model evaluations include intrinsic and extrinsic measurements. Intrinsic evaluation focuses on topic quality or coherence of the trained topics. The most widely-used metric for measuring

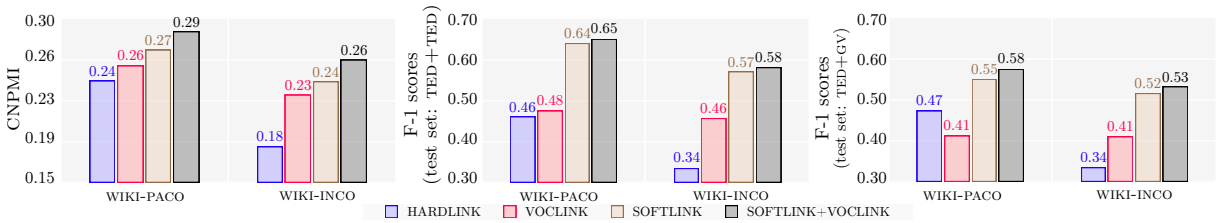



Figure 4: SOFTLINK consistently outperforms other models on both topic quality evaluation (CNPMI) and classification performance (F-1).

monolingual topic coherence is normalized pointwise mutual information (Lau et al., 2014; Newman et al., 2010). Hao et al. (2018) proposed crosslingual normalized pointwise mutual information (CNPMI) by extending this idea to multilingual settings, which correlates well with bilingual speakers' judgments on topic quality.

Given a bilingual topic $k$ in languages $\ell_1$ and $\ell_2$, and a parallel reference corpus $\mathcal{R}^{(\ell_1,\ell_2)}$, the CNPMI of topic $k$ is calculated as:

$$\text{CNPMI}(\ell_1, \ell_2, k) \quad = \quad \frac{1}{C^2} \sum_{i,j}^{C} \frac{1}{\log \Pr\left(w_i^{(\ell_1)}, w_j^{(\ell_2)}\right)} \cdot \log \frac{\Pr\left(w_i^{(\ell_1)}, w_j^{(\ell_2)}\right)}{\Pr\left(w_i^{(\ell_1)}\right) \Pr\left(w_j^{(\ell_2)}\right)} \tag{6}$$

where $C$ is the cardinality of a topic, *i.e.,* the $C$ most probable words in the topic-word distribution $\phi_k^{(\ell)}$. The co-occurrence probability of two words, $\Pr\left(w_i^{(\ell_1)}, w_j^{(\ell_2)}\right)$, is defined as the proportion of document pairs where both words appear. In the results below, we set $C = 20$, and average the CNPMI scores over $K = 25$ topics for each model output.

To calculate CNPMI scores, we use $10,000$ document pairs from a held-out portion of Wikipedia. CNPMI is an intrinsic evaluation, so it is only available for the training sets, WIKI-PACO and WIKI-INCO.

### 5.3.2 Extrinsic Evaluation: Crosslingual Classification

A successful multilingual topic model should provide informative features for crosslingual tasks. To show that our model is beneficial to downstream applications, we use crosslingual document classification to evaluate topic model performance. A high classification accuracy when testing on a different language from training indicates topic consistency across languages (Hermann and Blunsom, 2014; Klementiev et al., 2012; Smet et al., 2011).

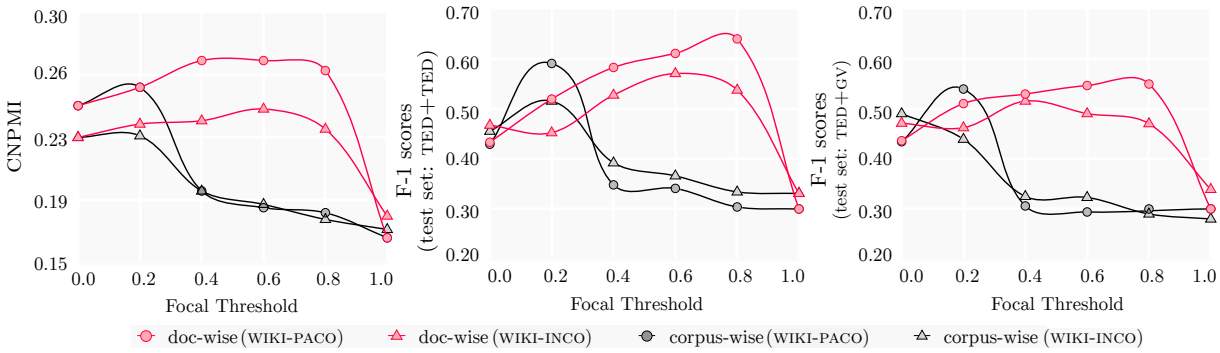
As in other studies on multilingual topic models, we first train topic models on a bilingual corpus $D^{(\ell_1,\ell_2)}$, and then use topic-word distributions $\phi^{(\ell_1)}$ and $\phi^{(\ell_2)}$ to infer document-topic distributions on unseen documents $D'^{(\ell_1)}$ and $D'^{(\ell_2)}$. Thus, a classifier is trained on $\theta_{d_{\ell 1}}$ with corresponding labels where $d_{\ell 1} \in D'^{(\ell_1)}$, and tested on $\theta_{d_{\ell 2}}$ where $d_{\ell 2} \in D'^{(\ell_2)}$, and vice versa. In our experiments, we use WIKI-PACO and WIKI-INCO to train topic models first, and then perform inference on either TED+TED (both English and non-English documents from TED) or TED+GV (English documents from TED and non-English from GV). For each language pair, we train multi-label classifiers using support vector machines (SVM) with five-fold cross-validation on documents in one language and test on the other. The F-1 scores reported below are micro-averaged over all labels.

## 5.4 Baseline Comparison

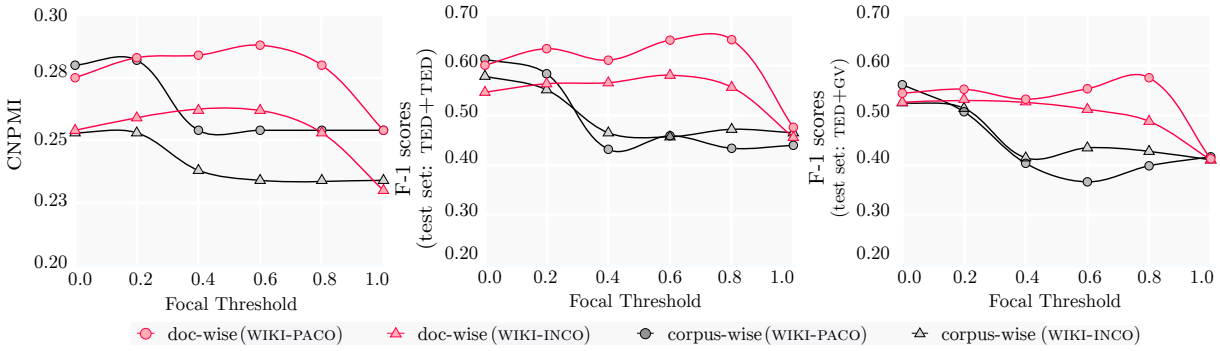
We first compare SOFTLINK with other models: HARDLINK, which is expected to do well on the partially comparable corpus (WIKI-PACO) but poorly on the incomparable corpus (WIKI-INCO), and VOCLINK. We additionally combine SOFTLINK+VOCLINK.

Figure 4 shows the performance (both intrinsic and extrinsic) of all models. For the SOFTLINK models, we used the optimal hyperparameter settings, but we compare other settings in Section 5.5.

When the training corpus is partially comparable (WIKI-PACO), all models can learn comparably coherent topics based on CNPMI scores, though the CNPMI of HARDLINK is lower than all other models.




(a) Training SOFTLINK model.




(b) Training SOFTLINK with VOCLINK.

Figure 5: CNPMI scores and F-1 scores of crosslingual classification under different focal thresholds and selection scope of the transfer distribution for SOFTLINK and SOFTLINK+VOCLINK (Section 4.2).

When the data is completely incomparable (WIKI-INCO), HARDLINK loses all connections between languages, so as expected its topics are least coherent. Similarly, when measuring classification performance, HARDLINK is comparable to VOCLINK on WIKI-PACO, but much worse on WIKI-INCO, where it loses all information. When the test set contains mostly parallel documents (TED+TED), the F-1 scores are higher, but when the test domain changes across languages (TED+GV), the performance drops.

On the other hand, SOFTLINK consistently outperforms other models regardless of training and test sets. It seems that SOFTLINK benefits from learning new connections between documents, even when part of the corpus contains direct links for training HARDLINK. It is also interesting that SOFTLINK uses the same dictionary resource as VOCLINK, but has a relative performance increase around 25%. It seems SOFTLINK can more efficiently utilize lexical information in a dictionary. We explore this relationship more in Section 5.6.

Finally, we observe that combining SOFTLINK+VOCLINK provides a performance boost over SOFTLINK in all cases, though the increase is small.

### 5.5 Comparison of Focusing Methods

We have shown that, when optimized, SOFTLINK can better utilize dictionary resources and outperform other models. We now focus on different training configurations for SOFTLINK, specifically, different methods of focusing the transfer distribution (Section 4.2).

Figure 5 shows how F-1 and CNPMI scores change with different static focusing methods. We vary the focal threshold and selection scope (*i.e.,* doc-wise or corpus-wise) for transfer distributions. As we increase the focal threshold $\pi$, more documents are zeroed out in the transfer distributions. When $\pi = 0.6$ or $0.8$, the transfer distributions are very sparse, and we notice that document-wise selection achieves the best performance. In the extreme case that $\pi = 1$, the transfer distributions are all zero, so SOFTLINK loses its connections between $\ell_1$ and $\ell_2$, and thus degrades to monolingual LDA. When

| | F-1 scores (TED+TED) | | F-1 scores (TED+GV) | | CNPMI | |
|---|---|---|---|---|---|---|
| | LIS | Fixed | LIS | Fixed | LIS | Fixed |
| WIKI-PACO | 0.627 | 0.638 | 0.551 | 0.534 | 0.256 | 0.258 |
| WIKI-INCO | 0.551 | 0.526 | 0.475 | 0.470 | 0.220 | 0.217 |

(a) Training SOFTLINK model.

| | F-1 scores (TED+TED) | | F-1 scores (TED+GV) | | CNPMI | |
|---|---|---|---|---|---|---|
| | LIS | Fixed | LIS | Fixed | LIS | Fixed |
| WIKI-PACO | 0.640 | 0.647 | 0.557 | 0.543 | 0.261 | 0.266 |
| WIKI-INCO | 0.546 | 0.517 | 0.459 | 0.465 | 0.242 | 0.233 |

(b) Training SOFTLINK with VOCLINK.

Table 1: Dynamically focusing transfer distributions in SOFTLINK yields competitive results on classification and topic quality evaluation. There is no significant difference between Fixed and LIS schedules.

training SOFTLINK with VOCLINK, the change of CNPMI and F-1 scores are less obvious as we increase focal threshold, since increasing focal threshold only has an impact on the SOFTLINK component of the model. When the focal threshold is higher, fewer soft links are active, so the model is closer to a plain VOCLINK model.

Interestingly, when focal threshold $\pi$ changes from $0.2$ to $0.4$, F-1 scores of corpus-wise selection scope trained on SOFTLINK drops drastically, in contrast to document-wise. This is because using corpus-wise selection could set a large portion of transfer distributions to zero, and only a small number of documents have non-zero transfer distributions. Since corpus-wise selection relies on the entire training corpus, it must be used with caution.

We find that using annealing to dynamically focus the distributions works well and is competitive with static focusing (Table 1). Annealing does better than the majority of settings of static focusing, though is worse than optimally-tuned focusing. We do not observe a significant difference between the two annealing schedules. When combining SOFTLINK and VOCLINK, the patterns are similar to that of SOFTLINK only.

### 5.6 Sensitivity to Dictionary Size

Both VOCLINK and SOFTLINK use the same dictionary resource, yet SOFTLINK produces better features for downstream tasks. To understand this behavior better, we experiment with different dictionary sizes to understand how well the models are utilizing the resource.

In Figure 6, we use different proportions ($20\%$, $40\%$, ..., $80\%$) of the dictionary to train SOFTLINK and VOCLINK.[4] We observe that the performance of VOCLINK (both F-1 and CNPMI) increases almost linearly with the dictionary size. In contrast, SOFTLINK is already at its best performance with only 20% of the available dictionary entries. This is further confirmation that SOFTLINK is using this resource in a more efficient way.

In VOCLINK, knowledge transfer happens through internal nodes of the word distribution priors, *i.e.,* word translations pairs, and words without translations are directly connected to the Dirichlet tree's root. If the dictionary cannot cover all the word types appeared in the training set, VOCLINK will have a set of word types in $\ell_1$ that cannot transfer enough topic knowledge to $\ell_2$ and vice versa. The fewer entries the dictionary provides, the more VOCLINK degrades to monolingual LDA. In contrast, SOFTLINK can potentially transfer knowledge from the whole corpus. For SOFTLINK, the dictionary is not used directly for modeling, rather it is only used for linking documents. Thus, knowledge transfer does not heavily rely on the number of entries in the dictionary.

### 5.7 Discussion

The sensitivity to dictionary size is an important factor to be considered in practice. For low-resource languages, a dictionary is easier to obtain than a large parallel corpus (Section 1). Models that rely on dictionaries such as VOCLINK and SOFTLINK are therefore more applicable to low-resource languages

[4]We use document-wise selection scope and focal threshold $\pi = 0.6$ for training SOFTLINK; same as in Section 5.4.

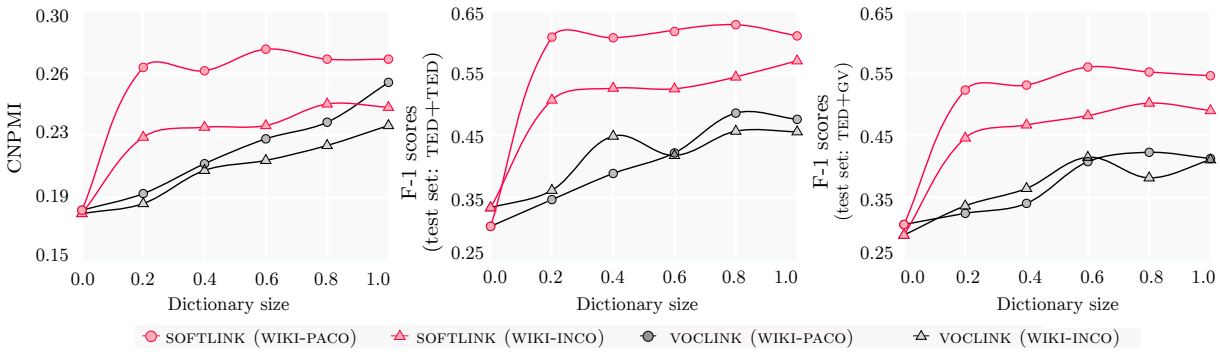



Figure 6: Performance of VOCLINK continues increasing when more dictionary entries are added, while SOFTLINK performance mostly stabilizes after using only 20% of available dictionary entries.

than HARDLINK. However, there are also large variations in dictionary size among languages. For example, in Wiktionary, 57 languages have fewer than 1,000 entries, while 77 languages have more than 100,000 entries. For truly low-resource languages, dictionary size could be a limiting factor. Since SOFTLINK can outperform VOCLINK with only a limited amount of lexical information, it may be able to transfer knowledge to low-resource languages more effectively than other approaches.

In summary, SOFTLINK relaxes and generalizes HARDLINK to be adaptable to more situations, while using dictionary information more efficiently than VOCLINK.

## 6 Conclusions and Future Work

We have described a new formulation for multilingual topic models which explicitly shows the knowledge transfer process across languages. Based on this analysis, we proposed a new multilingual topic model that can learn multilingually coherent topics and provide consistent topic features for crosslingual tasks. Unlike existing models, our approach is flexible and adaptable to incomparable corpora with only a dictionary, which is beneficial in many situations, in particular low-resource settings.

There are many possible directions following this work. First, our formulation of the knowledge transfer process enables future work focusing on how to develop more efficient algorithms that transfer knowledge with minimal supervision. Second, for SOFTLINK we plan to explore more about characteristics of languages that can lead to better formulations and learning of the transfer distributions.

## References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If All You Have is a bit of the Bible: Learning POS Taggers for Truly Low-Resource Languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 268–272.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual Projection for Parsing Truly Low-Resource Languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 25–32.

Timothy Baldwin, Jonathan Pool, and Susan M. Colowick. 2010. PanLex and LEXTRACT: Translating All Words of All Languages of the World. In *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China*, pages 37–40.

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.

Julian Besag. 1975. Statistical Analysis of Non-lattice Data. *The statistician*, pages 179–195.

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

Jordan L. Boyd-Graber and David M. Blei. 2009. Multilingual Topic Models for Unaligned Text. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 75–82.

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie. 2016. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. *CoRR*, abs/1606.01614.

Samuel Y. Dennis III. 1991. On the Hyper-Dirichlet Type 1 and Hyper-Liouville Distributions. *Communications in Statistics — Theory and Methods*, 20(12):4069–4081.

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: Adapting Topic Model Evaluation to Multilingual and Low-Resource Settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1090–1100.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 58–68.

Geert Heyman, Ivan Vulić, 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.

Gerold Hintz and Chris Biemann. 2016. Language Transfer Learning for Supervised Lexical Substitution. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, pages 118–129.

Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan L. Boyd-Graber. 2014. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1166–1176.

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.

David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. PanLex: Building a Resource for Panlingual Lexical Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 3145–3150.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit*.

Kriste Krstovski and David A. Smith. 2016. Bootstrapping Translation Detection and Sentence Extraction from Comparable Corpora. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1127–1132.

Kriste Krstovski, David A. Smith, and Michael J. Kurtz. 2016. Online Multilingual Topic Models with Multi-Level Hyperpriors. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 454–459.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 530–539.

Janne Leppä-aho, Johan Pensar, Teemu Roos, and Jukka Corander. 2017. Learning Gaussian Graphical Models with Fractional Marginal Pseudo-Likelihood. *International Journal of Approximate Reasoning*, 83:21–42.

Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2015. Multilingual Topic Models for Bilingual Dictionary Extraction. *ACM Transactions on Asian & Low-Resource Langanguage Information Processing.*, 14(3):11:1–11:22.

Tengfei Ma and Tetsuya Nasukawa. 2017. Inverted Bilingual Topic Models for Lexicon Extraction from Nonparallel 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 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.

Thomas Minka. 1999. The Dirichlet-tree Distribution.

Maria Moritz and Marco Büchler. 2017. Ambiguity in Semantically Related Word Substitutions: an Investigation in Historical Bible Translations. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Gothenburg, Sweden, May 22, 2017*, pages 18–23.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 100–108.

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.

Michael J. Paul and Mark Dredze. 2015. SPRITE: Generalizing Topic Models with Structured Priors. *Transactions of the Association for Computational Linguistics*, 3:43–57.

John C. Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual Document Representations from Discriminative Projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 251–261.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A Survey of Cross-lingual Word Embedding Models. *CoRR*, 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.

Wim De Smet, Jie Tang, and Marie-Francine Moens. 2011. Knowledge Transfer across Multilingual Corpora via Latent Topics. In *Advances in Knowledge Discovery and Data Mining - 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part I*, pages 549–560.

Noah A. Smith and Jason Eisner. 2006. Annealing Structural Bias in Multilingual Weighted Grammar Induction. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*.

Naonori Ueda and Ryohei Nakano. 1994. Deterministic Annealing Variant of the EM Algorithm. In *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 545–552.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2013. Cross-language Information Retrieval Models Based on Latent Topic Models Trained with Document-aligned Comparable Corpora. *Information Retrieval*, 16(3):331–368.

| Language | Family | Stemmer | Stopwords |
|---|---|---|---|
| EN | Germanic | `SnowBallStemmer` [1] | NLTK |
| ES | Romance | `SnowBallStemmer` | NLTK |
| RU | Slavic | `SnowBallStemmer` | NLTK |
| AR | Semitic | `Assem's Arabic Light Stemmer` [2] | GitHub [3] |
| FA | Indo-Iranian | `Hazm` [4] | GitHub |
| ZH | Sinitic | `Jieba` [5] | GitHub |

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

## Appendix A Pseudolikelihood

**Theorem 1.** *The conditional generative model with document links yields the same posterior estimator to the joint generative model using collapsed Gibbs sampling.*

*Proof.* Suppose the document links model is sampling topic of the $m$-th token in document $d_{\ell_2}$. The sampler calculates the conditional topic distribution, and then draw a topic assignment. Using collapsed Gibbs sampling, we calculate the conditional probability of a topic $k$:

$$\Pr\left(z_{d_{\ell 2},m}=k|\mathbf{z}_{d_{\ell 2},-},\mathbf{w}_{d_{\ell 2}};\mathbf{n}_{d_{\ell 1}},\alpha,\beta\right)=\frac{\Pr\left(z_{d_{\ell 2},m}=k,\mathbf{z}_{d_{\ell 2},-},\mathbf{w}_{d_{\ell 2}};\mathbf{n}_{d_{\ell 1}},\alpha,\beta\right)}{\Pr\left(\mathbf{z}_{d_{\ell 2},-},\mathbf{w}_{d_{\ell 2}};\mathbf{n}_{d_{\ell 1}},\alpha,\beta\right)}$$

$$=\frac{\Pr\left(z_{d_{\ell 2},m}=k,\mathbf{z}_{d_{\ell 2},-},\mathbf{w}_{d_{\ell 2}};\mathbf{n}_{d_{\ell 1}},\alpha\right)}{\Pr\left(\mathbf{z}_{d_{\ell 2},-};\mathbf{n}_{d_{\ell 1}},\alpha\right)}\cdot\frac{\Pr\left(\mathbf{w}_{d_{\ell 2}}|z_{d_{\ell 2},n}=k,\mathbf{z}_{d_{\ell 2},-};\beta\right)}{\Pr\left(\mathbf{w}_{d_{\ell 2}}|\mathbf{z}_{d_{\ell 2},-},\beta\right)}$$

$$=\frac{\frac{\prod_{k'\neq k}\Gamma\left(n_{k'|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1},k'}+\alpha\right)\cdot\Gamma\left(n_{k|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1},k}+\alpha+1\right)}{\Gamma\left(n_{\cdot|d}+\mathbf{n}_{d_{\ell 1}}+\alpha+1\right)}}{\frac{\prod_{k}\Gamma\left(n_{k|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1},k}+\alpha\right)}{\Gamma\left(n_{\cdot|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1}}+K\alpha\right)}}\cdot\frac{\frac{\prod_{w\neq w_{d_{\ell 2},m}}\Gamma\left(n_{w|k}+\beta\right)\cdot\Gamma\left(n_{w_{d_{\ell 2},m}|k}+\beta+1\right)}{\Gamma\left(n_{\cdot|k}+V^{(\ell_2)}\beta+1\right)}}{\frac{\prod_{w}\Gamma\left(n_{w|k}+\beta\right)}{\Gamma\left(n_{\cdot|k}+V^{(\ell_2)}\beta\right)}}$$

$$=\frac{\Gamma\left(n_{k|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1},k}+\alpha+1\right)}{\Gamma\left(n_{k|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1},k}+\alpha\right)}\cdot\frac{\Gamma\left(n_{\cdot|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1}}+K\alpha\right)}{\Gamma\left(n_{\cdot|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1}}+\alpha+1\right)}\cdot\frac{\Gamma\left(n_{w_{d_{\ell 2},m}|k}+\beta\right)}{\Gamma\left(n_{w_{d_{\ell 2},m}|k}+\beta+1\right)}\cdot\frac{\Gamma\left(n_{\cdot|k}+V^{(\ell_2)}\beta\right)}{\Gamma\left(n_{\cdot|k}+V^{(\ell_2)}\beta+1\right)}$$

$$=\frac{n_{k|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1},k}+\alpha}{n_{\cdot|d_{\ell 2}}+\mathbf{n}_{d_{\ell 1}}+K\alpha}\cdot\frac{n_{w_{d_{\ell 2},m}|k}+\beta}{n_{\cdot|k}+V^{(\ell_2)}\beta},$$

where $\mathbf{z}_{d_{\ell 2},-}$ is all the topic assignments in $d_{\ell_2}$ except the current one, $n_{\cdot|d_{\ell 2}}$ the number of tokens in $d_{\ell_2}$, $n_{k|d_{\ell 2}}$ the number of tokens assigned to topic $k$ in $d_{\ell_2}$, $n_{\cdot|k}$ the number of tokens assigned to topic $k$, $n_{w|k}$ the number of word type $w$ assigned to topic $k$, and $V^{(\ell_2)}$ the vocabulary size of language $\ell_2$. The roles of $\ell_1$ and $\ell_2$ are interchangeable, so both languages use the same conditional distributions. The last equation of the derivation above gives identical posterior estimation in the original model. Thus, the alternative formulation, despite not a numerically accurate likelihood approximation, does not make a difference for parameter estimation. □

## Appendix B Dataset Processing Details

### B.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/sobhe/hazm`;
[5] `https://github.com/fxsjy/jieba`.

| | | WIKI-PACO | WIKI-INCO | TED | GV | Wikipedia (for CNPMI) | Wiktionary |
|---|---|---|---|---|---|---|---|
| AR | #docs | 2,000 | 2,000 | 1,112 | 2,000 | 8,862 | 16,127 |
| | #tokens | 1,075,691 | 293,640 | 1,521,334 | 466,859 | 79,740 | |
| | #types | 32,843 | 19,900 | 44,982 | 32,468 | 1,533,261 | |
| ES | #docs | 2,000 | 2,000 | 1,152 | 2,000 | 9,325 | 31,563 |
| | #tokens | 475,234 | 237,561 | 1,228,469 | 493,327 | 1,763,897 | |
| | #types | 35,069 | 27,465 | 30,247 | 28,471 | 91,428 | |
| FA | #docs | 2,000 | 2,000 | 687 | 401 | 9,669 | 14,952 |
| | #tokens | 415,620 | 91,623 | 1,415,263 | 89,414 | 940,672 | |
| | #types | 18,316 | 9,987 | 36,670 | 9,447 | 46,995 | |
| RU | #docs | 2,000 | 2,000 | 1,010 | 2,000 | 9,837 | 33,574 |
| | #tokens | 4,368,563 | 766,887 | 1,133,098 | 679,217 | 2,356,994 | |
| | #types | 51,740 | 24,341 | 44,577 | 47,395 | 134,424 | |
| ZH | #docs | 2,000 | 2,000 | 1,123 | 2,000 | 8,222 | 23,276 |
| | #tokens | 3,095,977 | 303,634 | 1,428,532 | 745,307 | 1,338,116 | |
| | #types | 59,431 | 30,481 | 71,906 | 69,872 | 144,765 | |

Table 3: Statistics of corpora and dictionary in the five languages used in the experiments.

| Languages | AR | ES | FA | RU | ZH |
|---|---|---|---|---|---|
| Proportion | 12.2% | 9.35 % | 50.85 % | 50.20 % | 17.90 % |

Table 4: Proportions of linked document pairs in corpus WIKI-PACO.

### B.2 Data Source

We list the statistics in Table 3.

**Wikipedia (WIKI-PACO, and WIKI-INCO).** For training multilingual topic models, the dataset Wikipedia can be downloaded at `http://opus.nlpl.eu/TED2013.php`. For each language pair (EN, $\ell$), we create WIKI-INCO, a completely incomparable corpus, where $2,000$ EN documents and $2,000$ non-English documents are randomly chosen but do not contain document-level translations to each other.

We also create WIKI-PACO, a partially comparable corpus. Each language has different proportions of comparable document pairs. See Table 4.

**TED Talks 2013 (TED).** TED Talks 2013 contains mostly parallel documents, and can be obtained from OPUS: `http://opus.nlpl.eu/TED2013.php`. Note that not all English documents have translations to another language, which is slightly different from the original assumptions in polylingual topic models.

The classification labels can be obtained from the documents. Each document has several "categories" that can be regarded as labels. Thus, we retrieve those labels, and choose the most frequent five labels for classification: *technology*, *culture*, *science*, *global issues*, and *design*.

**Global Voices (GV).** Global Voices can be obtained from OPUS as well: `http://opus.nlpl.eu/GlobalVoices.php`. Global Voices corpus has a large number of documents, so for efficiency, we randomly choose a sample of at most 2,000 documents for each language.

There's no label information from the corpus itself. However, the labels can be retrieved from the webpage of each document, at `https://globalvoices.org`. To make sure Global Voices have the same label set to TED Talks, we changed the label set to: *technology*, *culture*, *science*, *business*, and *politics*.

**Wiktionary.** We use English Wiktionary to create bilingual dictionaries, which can be downloaded at `https://dumps.wikimedia.org/enwiktionary/`.