An Empirical Study on Crosslingual Transfer in Probabilistic Topic Models

Shudong Hao
Bard College at Simon's Rock
Division of Science, Mathematics, and
Computing
shao@simons-rock.edu

Michael J. Paul
University of Colorado
Department of Information Science
mpaul@colorado.edu

Probabilistic topic modeling is a common first step in crosslingual tasks to enable knowledge transfer and extract multilingual features. Although many multilingual topic models have been developed, their assumptions about the training corpus are quite varied, and it is not clear how well the different models can be utilized under various training conditions. In this article, the knowledge transfer mechanisms behind different multilingual topic models are systematically studied, and through a broad set of experiments with four models on ten languages, we provide empirical insights that can inform the selection and future development of multilingual topic models.

1. Introduction

Popularized by Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), probabilistic topic models have been an important tool for analyzing large collections of texts (Blei 2012, 2018). Their simplicity and interpretability make topic models popular for many natural language processing tasks, such as discovery of document networks (Chen et al. 2013; Chang and Blei 2009) and authorship attribution (Seroussi, Zukerman, and Bohnert 2014).

Topic models take a corpus D as input, where each document $d \in D$ is usually represented as a sparse vector in a vocabulary space, and project these documents to a lower-dimensional **topic space**. In this sense, topic models are often used as a dimensionality reduction technique to extract representative and human-interpretable features.

Text collections, however, are often not in a single language, and thus there has been a need to generalize topic models from monolingual to multilingual settings. Given a corpus $D^{(1,...,L)}$ in languages $\ell \in \{1,...,L\}$, multilingual topic models learn topics in

each of the languages. From a human's view, each topic should be related to the same theme, even if the words are not in the same language (Figure 1(b)). From a machine's view, the word probabilities within a topic should be similar across languages, such that the low-dimensional representation of documents is not dependent on the language. In other words, the topic space in multilingual topic models is **language agnostic** (Figure 1(a)).

This article presents two major contributions to multilingual topic models. We first provide an alternative view of multilingual topic models by explicitly formulating a crosslingual knowledge transfer process during posterior inference (Section 3). Based on this analysis, we unify different multilingual topic models by defining a function called the **transfer operation**. This function provides an abstracted view of the knowledge transfer mechanism behind these models, while enabling further generalizations and improvements. Using this formulation, we analyze several existing multilingual topic models (Section 4).

Second, in our experiments we compare four representative models under different training conditions (Section 5). The models are trained and evaluated on ten languages from various language families to increase language diversity in the experiments. In particular, we include five languages with relatively high resources and five others with low resources. To quantitatively evaluate the models, we focus on topic quality in Section 5.3.1, and performance of downstream tasks using crosslingual document classification in Section 5.3.2. We investigate how sensitive the models are to different language resources (i.e., parallel/comparable corpus and dictionaries), and analyze what factors cause this difference (Sections 6 and 7).
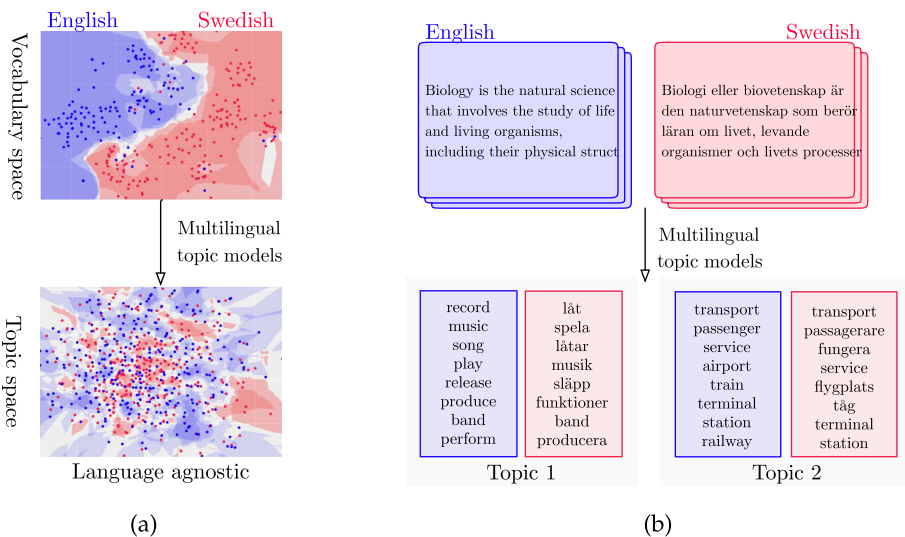
Figure 1
Overview of multilingual topic models. (a) Multilingual topic models project-language specific and high-dimensional features from the vocabulary space to a language-agnostic and low-dimensional topic space. This figure shows a t-SNE (Maaten and Hinton 2008) representation of a real data set. (b) Multilingual topic models produce theme-aligned topics for all languages. From a human's view, each topic contains different languages but the words are describing the same thing.

2. Background

We first review monolingual topic models, focusing on Latent Dirichlet Allocation, and then describe two families of multilingual extensions. Based on the types of supervision added to multilingual topic models, we separate the two model families into document-level and word-level supervision.

Topic models provide a high-level view of latent thematic structures in a corpus. Two main branches for topic models are non-probabilistic approaches such as Latent Semantic Analysis (LSA; Deerwester et al. 1990) and Non-Negative Matrix Factorization (Xu, Liu, and Gong 2003), and probabilistic ones such as Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan 2003) and probabilistic LSA (pLSA; Hofmann 1999). All these models were originally developed for monolingual data and later adapted to multilingual situations. Though there has been work to adapt non-probabilistic models, for example, based on "pseudo-bilingual" corpora approaches (Littman, Dumais, and Landauer 1998), most multilingual topic models that are trained on multilingual corpora are based on probabilistic models, especially LDA. Therefore, our work is focused on the probabilistic topic models, and in the following section we start by describing LDA.

2.1 Monolingual Topic Models

The most popular topic model is LDA, introduced by Blei, Ng, and Jordan (2003). This model assumes each document d is represented by a multinomial distribution θ_d over topics, and each "topic" k is a multinomial distribution $\phi^{(k)}$ over the vocabulary V. In the generative process, each θ and ϕ are generated from Dirichlet distributions parameterized by α and β, respectively. The hyperparameters for Dirichlet distributions can be asymmetric (Wallach, Mimno, and McCallum 2009), though in this work we use symmetric priors. Figure 2 shows the plate notation of LDA.

2.2 Multilingual Topic Models

We now describe a variety of multilingual topic models, organized into two families based on the type of supervision they use. Later, in Section 4, we focus on a subset of the models described here for deeper analysis using our knowledge transfer formulation, selecting the most general and representative models.

2.2.1 Document Level. The first model proposed to process multilingual corpora using LDA is the Polylingual Topic Model (PLTM; Mimno et al. 2009; Ni et al. 2009). This model extracts language-consistent topics from parallel or highly comparable multilingual corpora (for example, Wikipedia articles aligned across languages), assuming that document translations share the same topic distributions. This model has been
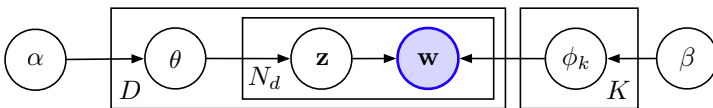
Figure 2
Plate notation of LDA. α and β are Dirichlet hyperparameters for θ and $\{\phi^{(k)}\}_{k=1}^{K}$. Topic assignments are denoted as **z**, and **w** denotes observed tokens.

extensively used and adapted in various ways for different crosslingual tasks (Krstovski and Smith 2011; Moens and Vulic 2013; Vulić and Moens 2014; Liu, Duh, and Matsumoto 2015; Krstovski and Smith 2016).

In the generative process, PLTM first generates language-specific topic-word distributions $\phi^{(\ell,k)} \sim \text{Dir}\left(\beta^{(\ell)}\right)$, for topics $k = 1, \ldots, K$ and languages $\ell = 1, \ldots, L$. Then, for each document tuple $\mathbf{d} = \left(d^{(1)}, \ldots, d^{(L)}\right)$, it generates a *tuple*-topic distribution $\theta_{\mathbf{d}} \sim \text{Dir}(\alpha)$. Every topic in this document tuple is generated from $\theta_{\mathbf{d}}$, and the word tokens in this document tuple are then generated from language-specific word distributions $\phi^{(\ell,k)}$ for each language. To apply PLTM, the corpus must be parallel or closely comparable to provide document-level supervision. We refer to this as the **document links** model (DOCLINK).

Models that transfer knowledge on the document level have many variants, including SOFTLINK (Hao and Paul 2018), comparable bilingual LDA (C-BILDA; Heyman, Vulic, and Moens 2016), the partially connected multilingual topic model (PCMLTM; Liu, Duh, and Matsumoto 2015), and multi-level hyperprior polylingual topic model (MLHPLTM; Krstovski, Smith, and Kurtz 2016). SOFTLINK generalizes DOCLINK by using a dictionary, so that documents can be linked based on overlap in their vocabulary, even if the corpus is not parallel or comparable. C-BILDA is a direct extension of DOCLINK that also models language-specific distributions to distinguish topics that are shared across languages from language-specific topics. PCMLTM adds an additional observed variable to indicate the absence of a language in a document tuple. MLHPLTM uses a hierarchy of hyperparameters to generate section-topic distributions. This model was motivated by applications to scientific research articles, where each section s has its own topic distribution $\theta^{(s)}$ shared by both languages.
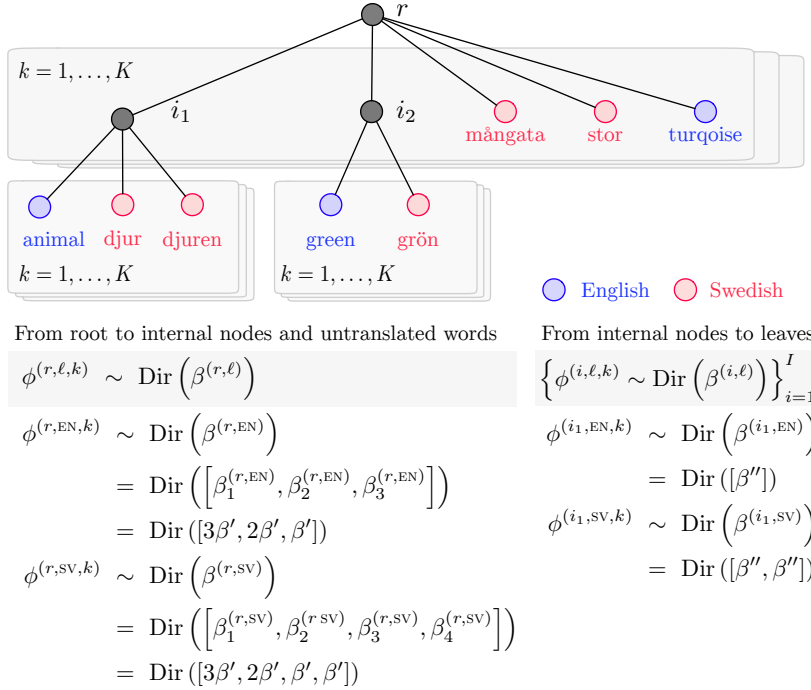
2.2.2 Word Level. Instead of document-level connections between languages, Boyd-Graber and Blei (2009) and Jagarlamudi and Daumé III (2010) proposed to model connections between languages through *words* using a multilingual dictionary and apply hyper-Dirichlet Type-I distributions (Andrzejewski, Zhu, and Craven 2009; Dennis III 1991). We refer to these approaches as the **vocabulary links** model (VOCLINK).

Specifically, VOCLINK uses a dictionary to create a tree structure where each internal node contains word translations, and words that are not translated are attached directly to the root of the tree r as leaves. In the generative process, for each language ℓ, VOCLINK first generates K multinomial distributions over all internal nodes and word types that are not translated, $\phi^{(r,\ell,k)} \sim \text{Dir}\left(\beta^{(r,\ell)}\right)$, where $\beta^{(r,\ell)}$ is a vector of Dirichlet prior from root r to internal nodes and untranslated words in language ℓ. Then, under each internal node i, for each language ℓ, VOCLINK generates a multinomial $\phi^{(i,\ell,k)} \sim \text{Dir}\left(\beta^{(i,\ell)}\right)$ over word types in language ℓ under the node i. Note that both $\beta^{(r,\ell)}$ and $\beta^{(i,\ell)}$ are vectors. In the first vector $\beta^{(r,\ell)}$, each cell is parameterized by scalar β' and scaled by the number of word translations under that internal node. For the second vector $\beta^{(i,\ell)}$, it is a symmetric hyperparameter where every cell uses the same scalar β''. See Figure 3 for an illustration.

Thus, to draw a word in language ℓ is equivalent to generating a *path* from the root to leaf nodes: $\left(r \to i, i \to w^{(\ell)}\right)$ or $\left(r \to w^{(\ell)}\right)$:

$$\text{Pr}\left(r \to i, i \to w^{(\ell)} | k\right) = \text{Pr}\left(i|k\right) \cdot \text{Pr}\left(w^{(\ell)} | k, i\right) \tag{1}$$

$$\text{Pr}\left(r \to w^{(\ell)} | k\right) = \text{Pr}\left(w^{(\ell)} | k\right) \tag{2}$$

Figure 3
An illustration of the tree structure used in word-level models. Hyperparameters $\beta^{(r,\ell)}$ and $\beta^{(i,\ell)}$ are both vectors, and β' and β'' are scalars. In the figure, i_1 has three translations, so the corresponding hyperparameter $\beta_1^{(r,\text{EN})} = \beta_1^{(r,\text{SV})} = 3\beta'$.

Document-topic distributions θ_d are generated in the same way as monolingual LDA, because no document translation is required.

The use of dictionaries to model similarities across topic-word distributions has been formulated in other ways as well. **PROBBILDA** (Ma and Nasukawa 2017) uses inverted indexing (Søgaard et al. 2015) to encode assumptions that word translations are generated from same distributions. PROBBILDA does not use tree structures in the parameters as in VOCLINK, but the general idea of sharing distributions among word translations is similar. Gutiérrez et al. (2016) use part-of-speech taggers to separate topic words (nouns) and perspective words (adjectives and verbs), developed for the application of detecting cultural differences, such as how different languages have different perspectives on the same topic. Topic words are modeled in the same way as in VOCLINK, whereas perspective words are modeled in a monolingual fashion.

3. Crosslingual Transfer in Probabilistic Topic Models

Conceptually, the term "knowledge transfer" indicates that there is a *process* of carrying information from a source to a destination. Using the representations of graphical models, the process can be visualized as the dependence of random variables. For example, $X \rightarrow Y$ implies that the generation of variable Y is conditioned on X, and thus the information of X is carried to Y. If X represents a probability distribution, the distribution of Y is informed by X, presenting a process of knowledge transfer, as we define it in this work.

In our study, "knowledge" can be loosely defined as K multinomial distributions over the vocabularies: $\{\phi^{(k)}\}_{k=1}^{K}$. Thus, to study the transfer mechanisms in topic models is to reveal how the models transfer $\{\phi^{(k)}\}_{k=1}^{K}$ from one language to another. To date, this transfer process has not been obvious in most models, because typical multilingual topic models assume the tokens in multiple languages are generated jointly.

In this section, we present a reformulation of these models that breaks down the co-generation assumption of current models and instead explicitly show the dependencies between languages. Starting with a simple example in Section 3.1, we show that our alternative formulation derives the same collapsed Gibbs sampler, and thus the same posterior distribution over samples, as in the original model. With this prerequisite, in Section 3.3 we introduce the **transfer operation**, which will be used to generalize and extend current multilingual topic models in Section 4.

3.1 Transfer Dependencies

We start with a simple graphical model, where $\theta \in \mathbb{R}_{+}^{K}$ is a K-dimensional categorical distribution, drawn from a Dirichlet parameterized by α, a symmetric hyperparameter (Figure 4(a)). Using θ, the model generates two variables, X and Y, and we use $\mathbf{x}$ and $\mathbf{y}$ to denote the generated observations. In the co-generation assumption, the variables X and Y are generated from the same θ at the same time, without dependencies between each other. Thus, we call this the joint model denoted as $\mathcal{G}^{(X,Y)}$ and the probability of the sample $(\mathbf{x}, \mathbf{y})$ is $\Pr\left(\mathbf{x}, \mathbf{y}; \alpha, \mathcal{G}^{(X,Y)}\right)$.

According to Bayes' theorem, there are two equivalent ways to expand the probability of $(\mathbf{x}, \mathbf{y})$:

$$\Pr\left(\mathbf{x}, \mathbf{y}; \alpha\right) = \Pr\left(\mathbf{x}|\mathbf{y}; \alpha\right) \cdot \Pr\left(\mathbf{y}; \alpha\right) \tag{3}$$

$$\Pr\left(\mathbf{x}, \mathbf{y}; \alpha\right) = \Pr\left(\mathbf{y}|\mathbf{x}; \alpha\right) \cdot \Pr\left(\mathbf{x}; \alpha\right) \tag{4}$$

where we notice that the generated sample is conditioned on another sample: $\Pr\left(\mathbf{x}|\mathbf{y}; \alpha\right)$ and $\Pr\left(\mathbf{y}|\mathbf{x}; \alpha\right)$, which fits into our concept of "transfer." We show both cases in Figures 4(b) and 4(c), and denote the graphical structures as $\mathcal{G}^{(Y|X)}$ and $\mathcal{G}^{(X|Y)}$, respectively, to show the dependencies between the two variables.

In this formulation, the model generates θ_x from Dirichlet (α) first and uses θ_x to generate the sample of $\mathbf{x}$. Using the histogram of $\mathbf{x}$ denoted as $\mathbf{n}_x = [n_{1|x}, n_{2|x}, \ldots, n_{K|x}]$

(a)   (b)   (c)
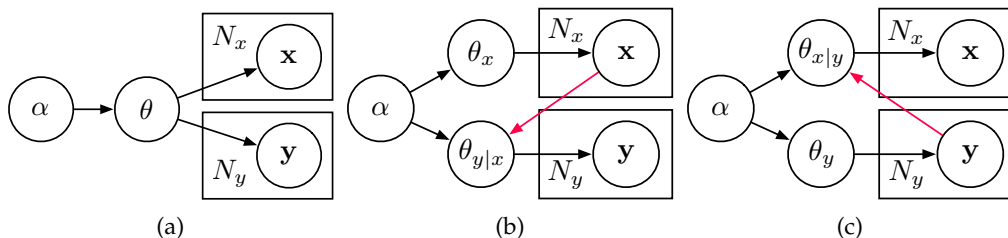
Figure 4
(a) The co-generation assumption generates $\mathbf{x}$ and $\mathbf{y}$ at the same time from the same θ. (b) To make the transfer process clear, we make the generation of $\mathbf{y}$ conditional on $\mathbf{x}$ and highlight the dependency in red. Because both x and y are exchangeable, the dependency can go the other way, as shown in (c).

where $n_{k|x}$ is the number of instances of X assigned to category k, together with hyper-parameter α, the model then generates a categorical distribution $\theta_{y|x} \sim \text{Dir}\,(\mathbf{n}_x + \alpha)$, from which the sample $\mathbf{y}$ is drawn.

This differs from the original joint model in that original parameter vector θ has been replaced with two variable-specific parameter vectors. The next section derives posterior inference with Gibbs sampling after integrating out the θ parameters, and we show that the sampler for each of two model formulations is equivalent and thus samples from an equivalent posterior distribution over $\mathbf{x}$ and $\mathbf{y}$.

3.2 Collapsed Gibbs Sampling

General approaches to infer posterior distributions over graphical model variables include Gibbs sampling, variational inference, and hybrid approaches (Kim, Voelker, and Saul 2013). We focus on collapsed Gibbs sampling (Griffiths and Steyvers 2004), which marginalizes out the parameters (θ in the example above) to focus on the variables of interest ($\mathbf{x}$ and $\mathbf{y}$ in the example).

Continuing with the example from the previous section, in each iteration of Gibbs sampling (a "sweep" of samples), the sampler goes through each example in the data, which can be viewed as sampling from the full posterior of a joint model $\mathcal{G}^{(X,Y)}$ as in Figure 5(a). Thus, when sampling an instance $x_i \in \mathbf{x}$, the collapsed conditional likelihood is

$$\Pr\left(x = k | \mathbf{x}_-, \mathbf{y}; \alpha\right) = \frac{\Pr(x = k, \mathbf{x}_-, \mathbf{y}; \alpha)}{\Pr(\mathbf{x}_-, \mathbf{y}; \alpha)} \tag{5}$$

$$= \frac{\Gamma\left(\alpha_k + n_{k|x} + n_{k|y}\right)}{\Gamma\left(N_x + N_y + \mathbf{1}^\top \alpha\right)} \cdot \frac{\Gamma\left(N_x + N_y^{(-i)} + \mathbf{1}^\top \alpha\right)}{\Gamma\left(\alpha_k + n_{k|x}^{(-i)} + n_{k|y}\right)} \tag{6}$$

$$= \frac{n_{k|x}^{(-i)} + n_{k|y} + \alpha_k}{N_x^{(-i)} + N_y + \mathbf{1}^\top \alpha} \tag{7}$$

where $\mathbf{x}_-$ is the set of tokens excluding the current one and $n_{k|x}^{(-i)}$ is the number of instances x assigned to category k except the current x_i. Note that in this equation, α is the hyperparameter for the Dirichlet prior, which gets added to the counts in the formula after integrating out the parameters θ.
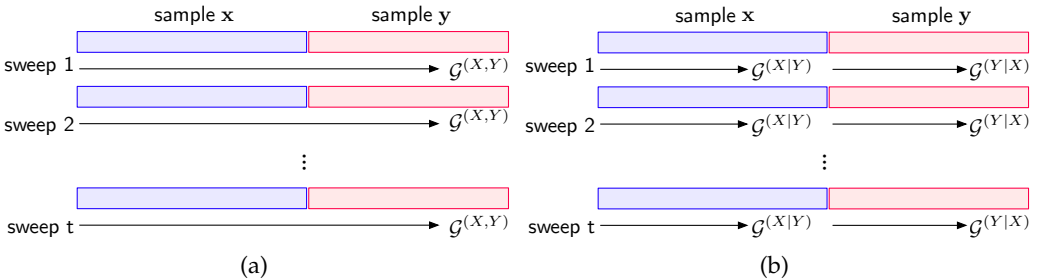
Figure 5
Sampling from a joint model $\mathcal{G}^{(X,Y)}$ (a) and two conditional models $\mathcal{G}^{(X|Y)}$ and $\mathcal{G}^{(Y|X)}$ (b) yields the same MAP estimates.

Using our formulation from the previous section, we can separate each sweep into two **subprocedures**, one for each variable. When sampling an instance of $x_i \in \mathbf{x}$, the histogram of sample $\mathbf{y}$ is fixed, and therefore it is sampling from the conditional model of $\mathcal{G}^{(X|Y)}$. Thus, the conditional likelihood is

$$\Pr\left(x = k | \mathbf{x}_-; \mathbf{y}, \alpha, \mathcal{G}^{(X|Y)}\right) = \frac{\Pr(x = k, \mathbf{x}_-; \mathbf{y}, \alpha)}{\Pr(\mathbf{x}_-; \mathbf{y}, \alpha)} \tag{8}$$

$$= \frac{\Gamma\left(n_{k|x} + (n_{k|y} + \alpha_k)\right)}{\Gamma\left(N_x + (N_y + \mathbf{1}^\top \alpha)\right)} \cdot \frac{\Gamma\left(N_x + (N_y^{(-i)} + \mathbf{1}^\top \alpha)\right)}{\Gamma\left(n_{k|x}^{(-i)} + (n_{k|y} + \alpha_k)\right)} \tag{9}$$

$$= \frac{n_{k|x}^{(-i)} + (n_{k|y} + \alpha_k)}{N_x^{(-i)} + (N_y + \mathbf{1}^\top \alpha)} \tag{10}$$

where the hyperparameter for variable X and category k becomes $n_{k|y} + \alpha_k$. Similarly, when sampling $y_i \in \mathbf{y}$ which is generated from the model $\mathcal{G}^{(Y|X)}$, the conditional likelihood is

$$\Pr\left(y = k | \mathbf{y}_-; \mathbf{x}, \alpha, \mathcal{G}^{(Y|X)}\right) = \frac{n_{k|y}^{(-i)} + (n_{k|x} + \alpha_k)}{N_y^{(-i)} + (N_x + \mathbf{1}^\top \alpha)} \tag{11}$$

with $n_{k|x} + \alpha_k$ as the hyperparameter for Y. This process is shown in Figure 5(b).

From the calculation perspective, although the meaning of Equations (7), (10), and (11) are different, their formulae are identical. This allows us to analyze similar models using the conditional formulation without changing the posterior estimation. A similar approach is the pseudo-likelihood approximation, where a joint model is reformulated as the combination of two conditional models, and the optimal parameters for the pseudo-likelihood function are the same as for the original joint likelihood function (Besag 1975; Koller and Friedman 2009; Leppä-aho et al. 2017).

3.3 Transfer Operation

Now that we have made the transfer process explicit and showed that this alternative formulation yields same collapsed posterior, we are able to describe a similar process in detail in the context of multilingual topic models.

If we treat X and Y in the previous example as two languages, and the samples $\mathbf{x}$ and $\mathbf{y}$ as either words, tokens, or documents from the two languages, we have a bilingual data set $(\mathbf{x}, \mathbf{y})$. Topic models have more complex graphical structures, where the examples (tokens) are organized within certain scopes (e.g., documents). To define the transfer process for a specific topic model, when generating samples in one language based on the transfer process of the model, we have to specify what examples we want to use from another language, how much, and where we want to use them. To this end, we define the **transfer operation**, which allows us to examine different models under a unified framework to compare them systematically.

Definition 1 (Transfer operation)
Let $\Omega \in \mathbb{R}^M$ be the target distribution of knowledge transfer with dimensionality M. A transfer operation on Ω from language ℓ_1 to ℓ_2 is defined as a function

$$h_\Omega : \mathbb{R}^{L_2 \times L_1} \times \mathbb{N}^{L_1 \times M} \times \mathbb{R}_+^{L_2 \times M} \mapsto \mathbb{R}^{L_2 \times M} \tag{12}$$

where L_1 and L_2 are the relevant dimensionalities for languages ℓ_1 and ℓ_2, respectively.

In this definition, the first argument of the transfer operation is where the two languages connect to each other, and can be defined as any bilingual supervision needed to enable transfer. The actual values of L_1 and L_2 depend on specific models. In an example of generating a document in language ℓ_2, L_1 is the number of documents in languages ℓ_1 and $L_2 = 1$, and $\delta \in \mathbb{R}^{L_1}$ could be an binary vector where $\delta_i = 1$ if document i is the translation to current document in ℓ_2, or zero otherwise. This is the core of crosslingual transfer through the transfer operation; later we will see that different multilingual topic models mostly only differ in the input of this argument, and designing this matrix is critical for an efficient knowledge transfer.

The second argument in the transfer operation is the sufficient statistics of the transfer source (ℓ_1 in the definition). After generating instances in language (ℓ_1), the statistics are organized into a matrix. The last argument is a prior distribution over the possible target distributions Ω.

The output of the transfer operation depends on and has the same dimensionality as the target distribution, which will be used as the prior to generate a multinomial distribution. Let Ω be the target distribution from which a topic of language ℓ_2 is generated: $z \sim \text{Multinomial}(\Omega)$. With a transfer operation, a topic is generated as follows:

$$\Omega \sim \text{Dirichlet}\left(h_\Omega\left(\delta, \mathbf{N}^{(\ell_1)}, \xi\right)\right) \tag{13}$$

$$z \sim \text{Multinomial}(\Omega) \tag{14}$$

where δ is bilingual supervision, $\mathbf{N}^{(\ell_1)}$ the generated sample of language ℓ_1, and ξ a prior distribution with the same dimensionality as Ω. See Figure 6 for an illustration.

In summary, this definition highlights three elements that are necessary to enable transfer:

(1) language transformations or supervision from the transfer source to destination;
(2) data statistics in the source; and
(3) a prior on the destination.

In the next section, we show how different topic models can be formulated with transfer operations, as well as how transfer operations can be used in the design of new models.

4. Representative Models

In this section, we describe four representative multilingual topic models in terms of the transfer operation formulation. These are also the models we will experiment on in Section 5. The plate notations of these models are shown in Figure 7, and we provide notations frequently used in these models in Table 1.
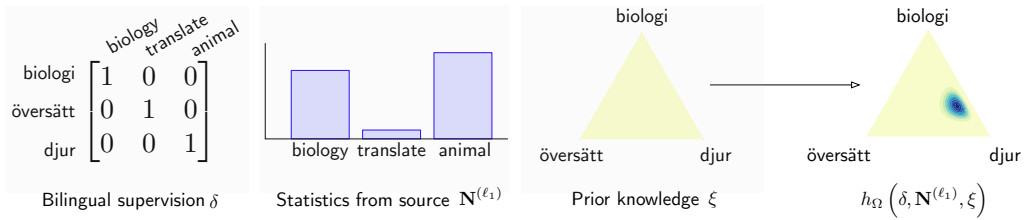
Figure 6
An illustration of a transfer operation on a 3-dimensional Dirichlet distribution. The first argument of h_Ω is a bilingual supervision δ, which is a 3×3 matrix, where $L_1 = L_2 = 3$, indicating word translations between two languages. The second argument $\mathbf{N}^{(\ell_1)}$ is the statistics (or histogram) from the sample in language ℓ_1, whose dimension is aligned with δ, and $M = 1$. With ξ as the prior knowledge (a symmetric hyperparameter), the result of h_Ω is then used as hyperparameters for the Dirichlet distribution.
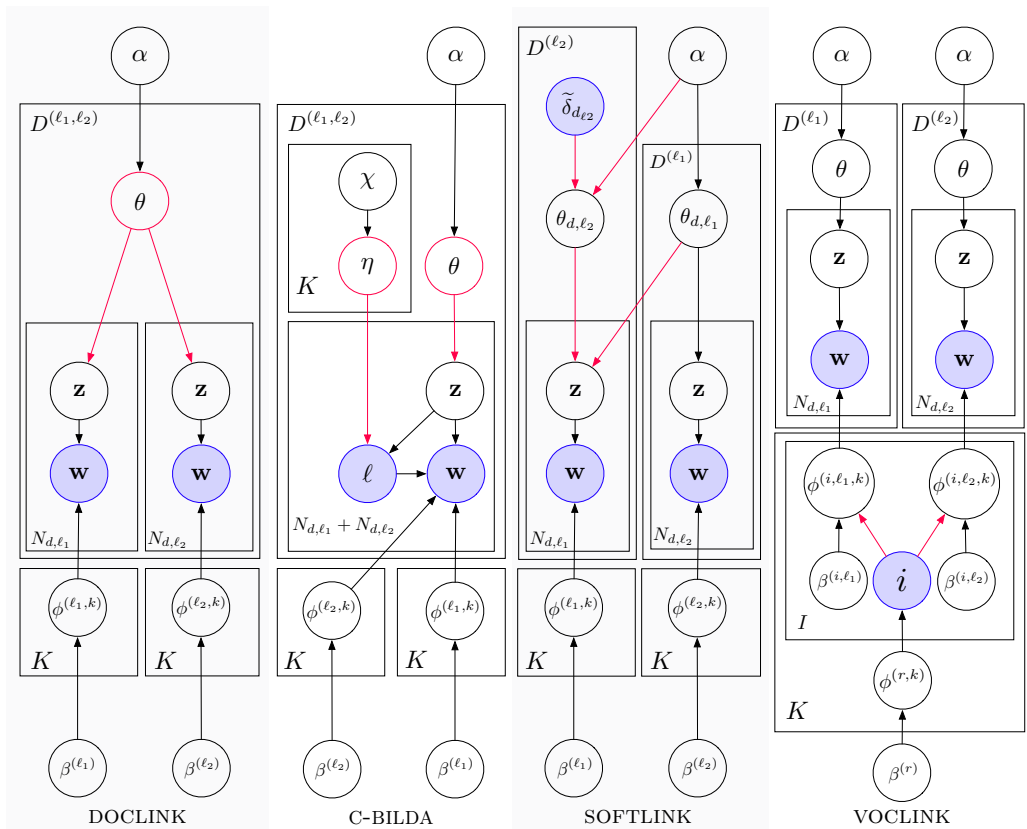
Figure 7
Plate notations of DOCLINK, C-BILDA, SOFTLINK, and VOCLINK (from left to right). We use red lines to make the knowledge transfer component clear. Note that in VOCLINK we assume every word is translated, so the plate notation does not include untranslated words.

Table 1
Notation table.

Notations	Descriptions
z	The topic assignment to a token.
$w^{(\ell)}$	A word type in language ℓ.
$V^{(\ell)}$	The size of vocabulary in language ℓ.
$D^{(\ell)}$	The size of corpus in language ℓ.
$D^{(\ell_1,\ell_2)}$	The number of document pairs in languages ℓ_1 and ℓ_2.
α	A symmetric Dirichlet prior vector of size K, where K is the number of topics, and each cell is denoted as α_k.
$\theta_{d,\ell}$	Multinomial distribution over topics for a document d in language ℓ.
$\beta^{(\ell)}$	A symmetric Dirichlet prior vector of size $V^{(\ell)}$, where $V^{(\ell)}$ is the size of vocabulary in language ℓ.
$\beta^{(r,\ell)}$	An asymmetric Dirichlet prior vector of size $I + V^{(\ell,-)}$, where I is the number of internal nodes in a Dirichlet tree, and $V^{(\ell,-)}$ the number of untranslated words in language ℓ. Each cell is denoted as $\beta_i^{(r,\ell)}$, indicating a scalar prior to a specific node i or an untranslated word type.
$\beta^{(i,\ell)}$	A symmetric Dirichlet prior vector of size $V_i^{(\ell)}$, where $V_i^{(\ell)}$ is the number of word types in language ℓ under internal node i.
$\phi^{(\ell,k)}$	Multinomial distribution over word types in language ℓ of topic k for topic k.
$\phi^{(r,\ell,k)}$	Multinomial distribution over internal nodes in a Dirichlet tree for topic k.
$\phi^{(i,\ell,k)}$	Multinomial distribution over all word types in language ℓ under internal node i for topic k.

4.1 Standard Models

Typical multilingual topic models are designed based on simple observations of multilingual data, such as parallel corpora and dictionaries. We focus on three popular models, and re-formulate them using the conditional generation assumption and the transfer operation we introduced in the previous sections.

4.1.1 DOCLINK. The document links model (DOCLINK) uses parallel/comparable data sets, so that each bilingual document pair shares the same distribution over topics. Assume the document d in language ℓ_1 is paired with d in language ℓ_2. Thus, the transfer target distribution is $\theta_{d,\ell_2} \in \mathbb{R}^K$ where K is the number of topics. For a document $d_{\ell 2}$, let $\delta \in \mathbb{N}_+^{D^{(\ell_1)}}$ be an indicator vector to indicate if a document $d_{\ell 1}$ is a translation or comparable document to $d_{\ell 2}$,

$$\delta_{d_{\ell 1}} = \mathbb{1}\left\{d_{\ell 2} \text{ and } d_{\ell 1} \text{ are translations}\right\} \tag{15}$$

where $D^{(\ell_1)}$ is the number of documents in language ℓ_1. Thus, the transfer operation for each document $d_{\ell 2}$ can be defined as

$$h_{\theta_{d,\ell_2}}\left(\delta, \mathbf{N}^{(\ell_1)}, \alpha\right) = \delta \cdot \mathbf{N}^{(\ell_1)} + \alpha \tag{16}$$

where $\mathbf{N}^{(\ell_1)} \in \mathbb{N}^{D^{(\ell_1)} \times K}$ is the sufficient statistics from language ℓ_1, and each cell n_{dk} is the count of topic k appearing in document d. We call this a "document-level" model, because the transfer target distribution is document-wise.

On the other hand, DOCLINK does not have any word-level knowledge, such as dictionaries, so the transfer operation on ϕ in DOCLINK is straightforward. For every topic $k = 1, \ldots, K$ and each word type w regardless of its language,

$$h_{\phi^{(\ell_2,k)}}\left(\mathbf{0}, \mathbf{N}^{(\ell_1)}, \beta^{(\ell_2)}\right) = \mathbf{0} \cdot \mathbf{N}^{(\ell_1)} + \beta^{(\ell_2)} = \beta^{(\ell_2)} \tag{17}$$

where $\beta^{(\ell_2)} \in \mathbb{R}_+^{V^{(\ell_2)}}$ is a symmetric Dirichlet prior for the topic-vocabulary distributions $\phi^{(\ell_2,k)}$, and $V^{(\ell_2)}$ is the size of vocabulary in language ℓ_2.

4.1.2 C-BILDA. As a variation of DOCLINK, C-BILDA has all of the components of DOCLINK and has the same transfer operations on θ and ϕ as in Equations (16) and (17), so this model is considered as a document-level model as well. Recall that C-BILDA additionally models topic-language distributions η.[1] For each document pair d and each topic k, a bivariate Bernoulli distribution over the two languages $\eta^{(k,d)} \in \mathbb{R}_+^2$ is drawn from a Beta distribution parameterized by $\left(\chi^{(d,\ell_1)}, \chi^{(d,\ell_2)}\right)$:

$$\eta^{(k,d)} \sim \text{Beta}\left(\chi^{(d,\ell_1)}, \chi^{(d,\ell_2)}\right) \tag{18}$$

$$\ell^{(k,m)} \sim \text{Bernoulli}\left(\eta^{(k,d)}\right) \tag{19}$$

where $\ell^{(k,m)}$ is the language of the m-th token assigned to topic k in the *entire* document pair d. Intuitively, $\eta_\ell^{(k,d)}$ is the probability of generating a token in language ℓ given the current document pair d and topic k.

Before diving into the specific definition of the transfer operation for this model, we need to take a closer look at the generative process of C-BILDA first, because in this model, language itself is a random variable as well. We describe the generative process in terms of the conditional formulation where one language is conditioned on the other. As usual, a monolingual model first generates documents in ℓ_1, and at this point each document pair d only has tokens in one language. Then for each document pair d, the conditional model additionally generates a number of topics z using the transfer operation on θ as defined in Equation (16). Instead of directly drawing a new word type in language ℓ_2 according to z, C-BILDA adds a step to generate a language ℓ' from $\eta^{(z,d)}$. Because the current token is supposed to be in language ℓ_2, if $\ell' \neq \ell_2$, this token is dropped, and the model keeps drawing the next topic z; otherwise, a word type is drawn from $\phi^{(z,\ell_2)}$ and attached to the document pair d. Once this process is over, each

---

[1] The original notation for topic-language distribution is δ (Heyman, Vulic, and Moens 2016). To avoid confusion in Equation (15), we change to η. We also follow the original paper where the model is for a bilingual case.
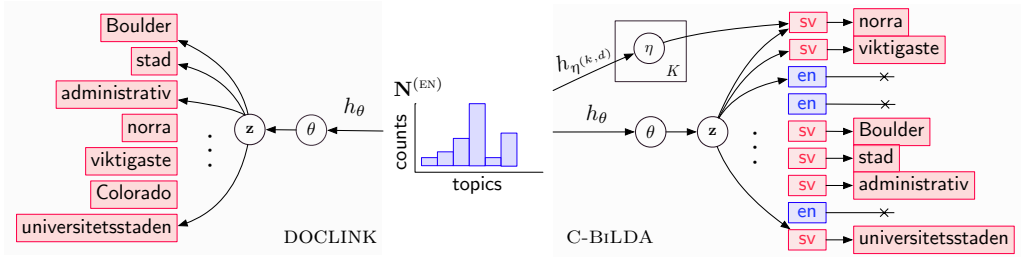
Figure 8
An illustration of difference between DOCLINK and C-BILDA in sequential generating process. DOCLINK uses a transfer operation on θ to generate topics and then word types in Swedish (SV). Additionally, C-BILDA uses a transfer operation on η to generate a language label according to a topic z. If the language generated is in Swedish, it draws a word type from the vocabulary; otherwise, the token is discarded.

document pair d contains tokens from two languages, and by separating the tokens based on their languages we can obtain the corresponding set of comparable document pairs. Conceptually, C-BILDA adds an additional "selector" in the generative process to decide if a topic should appear more in ℓ_2 based on topics in ℓ_1. We use Figure 8 as an illustration to show the difference between DOCLINK and C-BILDA.

It is clear that the generation of tokens in language ℓ_2 is affected by that of language ℓ_1; thus we define an additional transfer operation on $\eta^{(k,d)}$. The bilingual supervision δ is the same as Equation (15), which is a vector of dimension $D^{(\ell_1)}$ indicating document translations. We denote the statistics term $\mathbf{N}_k^{(\ell_1)} \in \mathbb{R}^{D^{(\ell_1)} \times 2}$, where each cell in the first column n_{dk} is the counts of topic k in document d, while the second column is a zero vector. Lastly, the prior term is also a two-dimensional vector $\chi^{(d)} = \left(\chi^{(d,\ell_1)}, \chi^{(d,\ell_1)}\right)$. Together, we have the transfer operation defined as

$$h_{\eta^{(k,d)}}\left(\delta, \mathbf{N}_k^{(\ell_1)}, \chi^{(d)}\right) = \delta \cdot \mathbf{N}_k^{(\ell_1)} + \chi^{(d)} \tag{20}$$

4.1.3 VOCLINK. Jagarlamudi and Daumé III (2010) and Boyd-Graber and Blei (2009) introduced another type of multilingual topic model, which uses a dictionary for word-level supervision instead of parallel/comparable documents as supervision, and we call this model VOCLINK.[2] Because no document-level supervision is used, the transfer operation on θ is simply defined as

$$h_{\theta_{d,\ell_2}}\left(\mathbf{0}, \mathbf{N}^{(\ell_1)}, \alpha\right) = \mathbf{0} \cdot \mathbf{N}^{(\ell_1)} + \alpha = \alpha \tag{21}$$

We now construct the transfer operation on the topic-word distribution φ based on the tree-structued priors in VOCLINK (Figure 3). Recall that each word $w^{(\ell)}$ is associated with at least one path, denoted as $\lambda_{w^{(\ell)}}$. If $w^{(\ell)}$ is translated, the path is $\lambda_{w^{(\ell)}} = \left(r \to i, i \to w^{(\ell)}\right)$ where r is the root and i an internal node; otherwise, the path is simply the edge from root to that word. Thus, on the first level of the tree, the Dirichlet

---

2 Although some models, as in Hu et al. (2014b), transfer knowledge at both document and word levels, in this analysis, we only focus on the word level where no transfer happens on the document level. The generalization simply involves using the same transfer operation on θ that is used in DOCLINK.

distribution $\phi^{(r,\ell_2,k)}$ is of dimension $I + V^{(\ell_2,-)}$, where I is the number of internal nodes (i.e., word translation entries), and $V^{(\ell_2,-)}$ are the untranslated word types in language ℓ_2. Let $\delta \in \mathbb{R}_+^{(I+V^{(\ell_2,-)}) \times V_1}$ be an indicator matrix where V_1 is the number of translated words in language ℓ_1, and each cell is

$$\delta_{i,w^{(\ell_1)}} = \mathbb{1} \left\{ w^{(\ell_1)} \text{ is under node } i \right\} \tag{22}$$

Given a topic k, the statistics argument $\mathbf{N}^{(\ell_1)} \in \mathbb{R}^{V_1}$ is a vector where each cell n_w is the count of word w assigned to topic k. Note that in the tree structure, the prior for Dirichlet is asymmetric and is scaled by the number of translations under each internal node. Thus, the transfer operation on $\phi^{(r,\ell_2,k)}$ is

$$h_{\phi^{(r,\ell_2,k)}} \left(\delta, \mathbf{N}^{(\ell_1)}, \beta^{(r,\ell_2)} \right) = \delta \cdot \mathbf{N}^{(\ell_1)} + \beta^{(r,\ell_2)} \tag{23}$$

Under each internal node, the Dirichlet is only related to specific languages, so no transfer happens, and the transfer operation on $\phi^{(i,\ell_2,k)}$ for an internal node i is simply $\beta^{(i,\ell_2)}$:

$$h_{\phi^{(i,\ell_2,k)}} \left(\mathbf{0}, \mathbf{N}^{(\ell_1)}, \beta^{(i,\ell_2)} \right) = \mathbf{0} \cdot \mathbf{N}^{(\ell_1)} + \beta^{(i,\ell_2)} = \beta^{(i,\ell_2)} \tag{24}$$

4.2 SOFTLINK: A Transfer Operation–Based Model

We have formulated three representative multilingual topic models by defining transfer operations for each model above. Our recent work, called SOFTLINK (Hao and Paul 2018), is explicitly designed according to the understanding of this transfer process. We present this model as a demonstration of how transfer operations can be used to build new multilingual topic models, which might not have an equivalent formulation using the standard co-generation model, by modifying the transfer operation.

In DOCLINK, the supervision argument δ in the transfer operation is constructed using comparable data sets. This requirement, however, substantially limits the data that can be used. Moreover, the supervision δ is also limited by the data; if there is no translation available to a target document, δ is an all-zero vector, and the transfer operation defined in Equation (16) will cancel out all the available information $\mathbf{N}^{(\ell_1)}$ for the target document, which is an ineffective use of the resource. Unlike parallel corpora, dictionaries are widely available and often easy to obtain for many languages. Thus, the general idea of SOFTLINK is to use a dictionary to retrieve as much as possible information from ℓ_1 to construct δ in a way that links potentially comparable documents together, even if the corpus itself does not explicitly link together documents.

Specifically, for a document d_{ℓ_2}, instead of a pre-defined indicator vector, SOFTLINK defines δ as a probabilistic distribution over all documents in language ℓ_1:

$$\delta_{d_{\ell_1}} \propto \frac{\left| \left\{ w^{(\ell_1)} \right\} \cap \left\{ w^{(\ell_2)} \right\} \right|}{\left| \left\{ w^{(\ell_1)} \right\} \cup \left\{ w^{(\ell_2)} \right\} \right|} \tag{25}$$

where $\{w^{(\ell)}\}$ contains all the word types that appear in document d_ℓ, and $\{w^{(\ell_1)}\} \cap \{w^{(\ell_2)}\}$ indicates all word pairs $(w^{(\ell_1)}, w^{(\ell_2)})$ in a dictionary as translations. Thus, $\delta_{d_{\ell_1}}$ can be interpreted as the "probability" of d_{ℓ_1} being the translation of d_{ℓ_2}. We call δ the **transfer distribution**. See Figure 9 for an illustration.
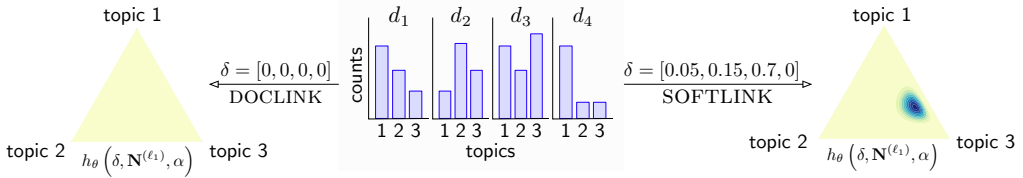
Figure 9
An example of how different inputs of transfer operation result in different Dirichlet priors through DOCLINK and SOFTLINK. The middle is a mini-corpus in language ℓ_1 and each document's topic histogram. When a document in ℓ_2 is not translation to any of those in ℓ_1, DOCLINK defines δ as an all-zero vector which leads to an uninformative symmetric prior. In contrast, SOFTLINK uses a dictionary to create δ as a distribution so that the topic histogram in each document in ℓ_1 can still be proportionally transferred.

In our initial work, we show that instead of a dense distribution, it is more efficient to make the transfer distributions sparse by thresholding,

$$\tilde{\delta}_{d_{\ell_1}} \propto \mathbb{1}\left\{\delta_{d_{\ell_1}} > \pi \cdot \max(\delta)\right\} \cdot \delta_{d_{\ell_1}} \tag{26}$$

where $\pi \in [0, 1]$ is a fixed threshold parameter. With the same definition of $\mathbf{N}^{(\ell_1)}$ and α in Equation (16) and δ defined as Equation (25), SOFTLINK completes the same transfer operations,

$$h_{\theta_{d,\ell_2}}\left(\delta, \mathbf{N}^{(\ell_1)}, \alpha\right) = \delta \cdot \mathbf{N}^{(\ell_1)} + \alpha \tag{27}$$

$$h_{\phi^{(\ell_2,k)}}\left(\mathbf{0}, \mathbf{N}^{(\ell_1)}, \beta^{(\ell_2)}\right) = \mathbf{0} \cdot \mathbf{N}^{(\ell_1)} + \beta^{(\ell_2)} = \beta^{(\ell_2)} \tag{28}$$

4.3 Summary: Transfer Levels and Transfer Models

We categorize transfer operations into two groups based on the target transfer distribution. *Document-level* operations transfer knowledge on distributions related to the entire document, such as θ in DOCLINK, C-BILDA, and SOFTLINK, and η in C-BILDA. *Word-level* operations transfer knowledge on those related to the entire vocabulary or specific word types, such as ϕ in VOCLINK.

When a model only has transfer operations on just one specific level, we also use the transfer level to refer the model. For example, DOCLINK, C-BILDA, and SOFTLINK are all document-level models, while VOCLINK is a word-level model. Those that transfer knowledge on multiple levels, such as Hu et al. (2014b), are called mixed-level models.

We summarize the transfer operation definitions for different models in Table 2, and add monolingual LDA as a reference to show how transfer operations are defined when no transfer takes place. We will experiment on the four multilingual models in Sections 4.1.1 through 4.2.

5. Experiment Settings

From discussions above, we are able to describe various multilingual topic models by defining different transfer operations, which explicitly represent the language transfer process. When designing and applying those transfer operations in practice, some

Table 2
Summary of transfer operations defined in the compared models, where we assume the direction of transfer is from ℓ_1 to ℓ_2.

Model	Document level	Word level	Parameters of h
LDA	α	$\beta^{(\ell_2)}$	—
DOCLINK	$\delta \cdot \mathbf{N}^{(\ell_1)} + \alpha$	$\beta^{(\ell_2)}$	δ: indicator vector;
C-BILDA	$\delta \cdot \mathbf{N}^{(\ell_1)} + \alpha,$ $\delta \cdot \mathbf{N}_k^{(\ell_1)} + \chi^{(d)}$	$\beta^{(\ell_2)}$	$\mathbf{N}^{(\ell_1)}$: doc-by-topic matrix; supervision: comparable documents;
SOFTLINK	$\delta \cdot \mathbf{N}^{(\ell_1)} + \alpha$	$\beta^{(\ell_2)}$	δ: transfer distribution; $\mathbf{N}^{(\ell_1)}$: doc-by-topic matrix; supervision: dictionary;
VOCLINK	α	$\delta \cdot \mathbf{N}^{(\ell_1)} + \beta^{(r,\ell_2)}$	δ: indicator vector; $\mathbf{N}^{(\ell_1)}$: node-by-word matrix; supervision: dictionary;

natural questions arise, such as which transfer operation is more effective in what type of situation, and how to design a model that is more generalizable regardless of availability of multilingual resources.

To study the model behaviors empirically, we train the four models described in the previous section—DOCLINK, C-BILDA, SOFTLINK, and VOCLINK—in ten languages. Considering the resources available, we separate the ten languages into two groups: high-resource languages (HIGHLAN) and low-resource languages (LOWLAN). For HIGHLAN, we have relatively abundant resources such as dictionary entries and document translations. We additionally use these languages to simulate the settings of LOWLAN by training multilingual topic models with different amounts of resources. For LOWLAN, we use all resources available to verify experiment results and conclusions from HIGHLAN.

5.1 Language Groups and Preprocessing

We separate the ten languages into two groups: HIGHLAN and LOWLAN. In this section, we describe the preprocessing details of these languages.

5.1.1 HIGHLAN. Languages in this group have a relatively large amount of resources, and have been widely experimented on in multilingual studies. Considering language diversity, we select representative languages from five different families: Arabic (AR, Semitic), German (DE, Germanic), Spanish (ES, Romance), Russian (RU, Slavic), and Chinese (ZH, Sinitic). We follow standard preprocessing procedures: We first use stemmers to process both documents and dictionaries (segmenter for Chinese), then we remove stopwords based on a fixed list and the most 100 frequent word types in the training corpus. The tools for preprocessing are listed in Table 3.

5.1.2 LOWLAN. Languages in this group have much fewer resources than those in HIGH-LAN, considered as low-resource languages. We similarly select five languages from different families: Amharic (AM, Afro-Asiatic), Aymara (AY, Aymaran), Macedonian (MK, Indo-European), Swahili (SW, Niger-Congo), and Tagalog (TL, Austronesian). Note

Table 3
List of source of stemmers and stopwords used in experiments for HIGHLAN.

Language	Family	Stemmer	Stopwords
EN	Germanic	SnowBallStemmer [3]	NLTK
DE	Germanic	SnowBallStemmer	NLTK
ES	Romance	SnowBallStemmer	NLTK
RU	Slavic	SnowBallStemmer	NLTK
AR	Semitic	Assem's Arabic Light Stemmer [4]	GitHub [5]
ZH	Sinitic	Jieba [6]	GitHub

that some of these are not strictly "low-resource" compared with many endangered languages. For the truly low-resource languages, it is very difficult to test the models with enough data, and, therefore, we choose languages that are understudied in natural language processing literature.

Preprocessing in this language group needs more consideration. Because they represent low-resource languages that most natural language processing tools are not available for, we do not use a fixed stopword list. Stemmers are also not available for these languages, so we do not apply stemming.

5.2 Training Sets and Model Configurations

There are many resources available for multilingual research, such as the European Parliament Proceedings parallel corpus (EUROPARL; Koehn 2005), the Bible, and Wikipedia. EuroParl provides a perfectly parallel corpus with precise translations, but it only contains 21 European languages, which limits its generalizability to most of the languages. The Bible, on the other hand, is also perfectly parallel and is widely available in 2,530 languages.[7] Its disadvantages, however, are that the contents are very limited (mostly about family and religion), the data set size is small (1,189 chapters), and many languages do not have digital format (Christodoulopoulos and Steedman 2015).

Compared with EUROPARL and the Bible, Wikipedia provides comparable documents in many languages with a large range of content, making it a very popular choice for many multilingual studies. In our experiments, we create ten bilingual Wikipedia corpora, each containing documents in one of the languages in either HIGHLAN or LOWLAN, paired with documents in English (EN). Though most multilingual topic models are not restricted to training bilingual corpora paired with English, this is a helpful way to focus our experiments and analysis.

We present the statistics of the training corpus of Wikipedia and the dictionary we use (from Wiktionary) in the experiments in Table 4. Note that we train topic models on

---

3 http://snowball.tartarus.org.
4 http://arabicstemmer.com.
5 https://github.com/6/stopwords-json.
6 https://github.com/fxsjy/jieba.
7 https://www.unitedbiblesocieties.org/.

Table 4
Statistics of training Wikipedia corpus and Wiktionary.

		English (EN)			Paired language			Wiktionary
		#docs	#tokens	#types	#docs	#tokens	#types	#entries
HighLan	AR	2,000	616,524	48,133	2,000	181,946	25,510	16,127
	DE	2,000	332,794	35,921	2,000	254,179	55,610	32,225
	ES	2,000	369,181	37,100	2,000	239,189	30,258	31,563
	RU	2,000	410,530	39,870	2,000	227,987	37,928	33,574
	ZH	2,000	392,745	38,217	2,000	168,804	44,228	23,276
LowLan	AM	2,000	3,589,268	161,879	2,000	251,708	65,368	4,588
	AY	2,000	1,758,811	84,064	2,000	169,439	24,136	1,982
	MK	2,000	1,777,081	100,767	2,000	489,953	87,329	6,895
	SW	2,000	2,513,838	143,691	2,000	353,038	46,359	15,257
	TL	2,000	2,017,643	261,919	2,000	232,891	41,618	6,552

bilingual pairs, where one of the languages is always English, so in the table we show statistics of English in every bilingual pair as well.

Lastly, we summarize the model configurations in Table 5. The goal of this study is to bring current multilingual topic models together, studying their corresponding strengths and limitations. To keep the experiments as comparable as possible, we use constant hyperparameters that are consistent across the models. For all models, we set the Dirichlet hyperparameter $\alpha_k = 0.1$ for each topic $k = 1, \ldots, K$. We run 1,000 Gibbs sampling iterations on the training set and 200 iterations on the test sets. The number of topics K is set to 20 by default for efficiency reasons.

Table 5
Model specifications.

Model	Hyperparameters
DOCLINK	We set β to be a symmetric vector where each cell $\beta_i = 0.01$ for all word types of all the languages, and use the MALLET implementation for training (McCallum 2002). To enable consistent comparison, we disable hyperparameter optimization provided in the package.
C-BILDA	Following the experiment results from Heyman, Vulic, and Moens (2016), we set $\chi = 2$ to make the results more competitive to DOCLINK. The rest of the settings are the same as for DOCLINK.
SOFTLINK	We use the document-wise thresholding approach for calculating the transfer distributions. The focus threshold is set to 0.8. The rest of the settings are the same as for DOCLINK.
VOCLINK	We set the scalar $\beta' = 0.01$ for hyperparameter $\beta^{(r,\ell)}$ from the root to both internal nodes or leaves. For those from internal nodes to leaves, we set $\beta'' = 100$, following the settings in Hu et al. (2014b).

5.3 Evaluation

We evaluate all models using both intrinsic and extrinsic metrics. Intrinsic evaluation is used to measure the topic quality or coherence learned from the training set, and extrinsic evaluation measures performance after applying the trained distributions to downstream crosslingual applications. For all the following experiments and tasks, we start by analyzing languages in HIGHLAN. Then we apply the analyzed results to LOWLAN.

We choose topic coherence (Hao, Boyd-Graber, and Paul 2018) and crosslingual document classification (Smet, Tang, and Moens 2011) as intrinsic and extrinsic evaluation tasks, respectively. The reason for choosing these two tasks is that they examine the models from different angles: Topic coherence looks at topic-word distributions, whereas classification focuses on document-topic distributions. Other evaluation tasks, such as word translation detection and crosslingual information retrieval, also utilize the trained distributions, but here we focus on a straightforward and representative task.

5.3.1 Intrinsic Evaluation: Topic Quality. Intrinsic evaluation refers to evaluating the learned model directly without applying it to any particular task; for topic models, this is usually based on the quality of the topics. Standard evaluation measures for monolingual models, such as perplexity (or held-out likelihood; Wallach et al. 2009) and Normalized Pointwise Mutual Information (NPMI, Lau, Newman, and Baldwin (2014)), could potentially be considered for crosslingual models. However, when evaluating multilingual topics, how words in different languages make sense together is also a critical criterion in addition to coherence within each of the languages.

In monolingual studies, Chang et al. (2009) show that held-out likelihood is not always positively correlated with human judgments of topics. Held-out likelihood is additionally suboptimal for multilingual topic models, because this measure is only calculated *within* each language, and the important crosslingual information is ignored.

Crosslingual Normalized Pointwise Mutual Information (CNPMI; Hao, Boyd-Graber, and Paul 2018) is a measure designed specifically for multilingual topic models. Extended from the widely used NPMI to measure topic quality in multilingual settings, CNPMI uses a parallel reference corpus to extract crosslingual coherence. CNPMI correlates well with bilingual speakers' judgments on topic quality and predictive performance in downstream applications. Therefore, we use CNPMI for intrinsic evaluations.

Definition 2 (Crosslingual Normalized Pointwise Mutual Information, CNPMI)
Let $\mathcal{W}_C^{(\ell_1,\ell_2)}$ be the set of top C words in a bilingual topic, and $\mathcal{R}^{(\ell_1,\ell_2)}$ a parallel reference corpus. The CNPMI of this topic is calculated as

$$\text{CNPMI}\left(\mathcal{W}_C^{(\ell_1,\ell_2)}\right) = -\frac{1}{C^2} \sum_{w_i,w_j \in \mathcal{W}_C^{(\ell_1,\ell_2)}} \frac{\log \frac{\Pr(w_i,w_j)}{\Pr(w_i)\Pr(w_j)}}{\log \Pr\left(w_i,w_j\right)} \tag{29}$$

where w_i and w_j are from languages ℓ_1 and ℓ_2, respectively. Let $\mathbf{d} = \left(d_{\ell_1}, d_{\ell_2}\right)$ be a pair of parallel documents from the reference corpus $\mathcal{R}^{(\ell_1,\ell_2)}$, whose size is denoted as $\left|\mathcal{R}^{(\ell_1,\ell_2)}\right|$. $\left|\{\mathbf{d} : w_i \in d_{\ell_1}, w_j \in d_{\ell_2}\}\right|$ is the number of parallel document pairs in which w_i and w_j

appear. The co-occurrence probability of a word pair and the probability of a single word are calculated as

$$\Pr\left(w_i, w_j\right) \triangleq \frac{\left|\left\{\mathbf{d} : w_i \in d_{\ell_1}, w_j \in d_{\ell_2}\right\}\right|}{\left|\mathcal{R}^{(\ell_1, \ell_2)}\right|} \tag{30}$$

$$\Pr\left(w_i\right) \triangleq \frac{\left|\left\{\mathbf{d} : w_i \in d_{\ell_1}\right\}\right|}{\left|\mathcal{R}^{(\ell_1, \ell_2)}\right|} \tag{31}$$

Intuitively, a coherent topic should contain words that make sense or fit in a specific context together. In the multilingual case, CNPMI measures how likely it is that a bilingual word pair appears in a similar context provided by the parallel reference corpus. We provide toy examples in Figure 10, where we show three bilingual topics. In Topic A, both languages are about "language," and all the bilingual word pairs have high probability of appearing in the same comparable document pairs. Thus Topic A is coherent *crosslingually*, and thus expected to have a high CNPMI score. Although we can identify the themes within each language in Topic B, that is, education in English and biology in Swahili, most of the bilingual word pairs do not make sense or appear in the same context, which gives us a low CNPMI score. The last topic is not coherent even within each language, so it has low CNPMI as well. Through this example, we see that CNPMI detects crosslingual coherence in multiple ways, unlike other intrinsic measures that might be adapted for crosslingual models.

In our experiments, we use $10,000$ linked Wikipedia article pairs for each language pair (EN, ℓ) ($20,000$ in total) as the reference corpus, and set $C = 10$ by default. Note that HIGHLAN has more Wikipedia articles, and we make sure the articles used for evaluating CNPMI scores do not appear in the training set. However, for LOWLAN, because the number of linked Wikipedia articles is extremely limited, we use all the available pairs to evaluate CNPMI scores. The statistics are shown in Table 6.

5.3.2 Extrinsic Evaluation: Crosslingual Classification. Crosslingual document classification is the most common downstream application for multilingual topic models (Smet, Tang,

Topic A (English-Amharic)		Topic B (English-Swahili)		Topic C (English-Macedonian)	
CNPMI = 0.3632		CNPMI = −0.0094		CNPMI = 0.0643	
word	ቋንቋ (language)	degree	bakteria (bacteria)	food	крава (cow)
dialect	ቃላት (words)	science	asidi (acid)	secret	книга (book)
vowel	ፊደል (letter)	professor	spishi (species)	under	безбедност (security)
latin	ጽሕፈት (writing)	award	amino (amino)	bridge	универзумот (universe)
spoken	አልፋቤት (alphabet)	bachelor	seli (cells)	pills	започнете (start)
letter	ድምፅ (audio)	program	aina (type)	diploma	списание (magazine)
Arabic	እንግሊዝኛ (English)	academic	bata (duck)	frogs	дрвја (trees)
speaker	ልሳናት (tongues)	institute	maji (water)	lights	завеса (curtain)
verb	ምልክት (signal)	student	wanyama (animals)	lie	чудо (miracle)
linguist	ሰዎች (people)	chemistry	protini (protein)	donuts	витамин (vitamin)

Figure 10
CNPMI measures how likely a bilingual word pair appears in a similar context in two languages, provided by a reference corpus. Topic A has a high CNPMI score because both languages are talking about the same theme. Both Topic B and Topic C are incoherent multilingual topics, although Topic B is coherent within each language.

Table 6
Statistics of Wikipedia corpus for topic coherence evaluation (CNPMI).

		English			Paired language		
		#docs	#tokens	#types	#docs	#tokens	#types
HIGHLAN	AR	10,000	3,597,322	128,926	10,000	996,801	64,197
	DE	10,000	2,155,680	103,812	10,000	1,459,015	166,763
	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
LOWLAN	AM	4,316	9,632,700	269,772	4,316	403,158	91,295
	AY	4,187	5,231,260	167,531	4,187	280,194	32,424
	MK	10,000	11,080,304	301,026	10,000	3,175,182	245,687
	SW	10,000	13,931,839	341,231	10,000	1,755,514	134,152
	TL	6,471	7,720,517	645,534	6,471	1,124,049	83,967

and Moens 2011; Vulić et al. 2015; Heyman, Vulic, and Moens 2016). Typically, a model is trained on a multilingual training set $D^{(\ell_1, \ell_2)}$ in languages ℓ_1 and ℓ_2. Using the trained topic-vocabulary distributions ϕ, the model infers topics in test sets $D'^{(\ell_1)}$ and $D'^{(\ell_2)}$.

In multilingual topic models, document-topic distributions θ can be used as features for classification, where the $\widehat{\theta}_{d,\ell_1}$ vectors in language ℓ_1 train a classifier tested by the $\widehat{\theta}_{d,\ell_2}$ vectors in language ℓ_2. A better classification performance indicates more consistent features across languages. See Figure 11 for an illustration. In our experiments, we use a linear support vector machine to train multilabel classifiers with five-fold cross-validation. Then, we use micro-averaged F-1 scores to evaluate and compare performance across different models.

For crosslingual classification, we also require held-out test data with labels or annotations. In our experiments, we construct test sets from two sources: TED Talks 2013 (TED) and Global Voices (GV). TED contains parallel documents in all languages in HIGHLAN, whereas GV contains all languages from both HIGHLAN and LOWLAN.
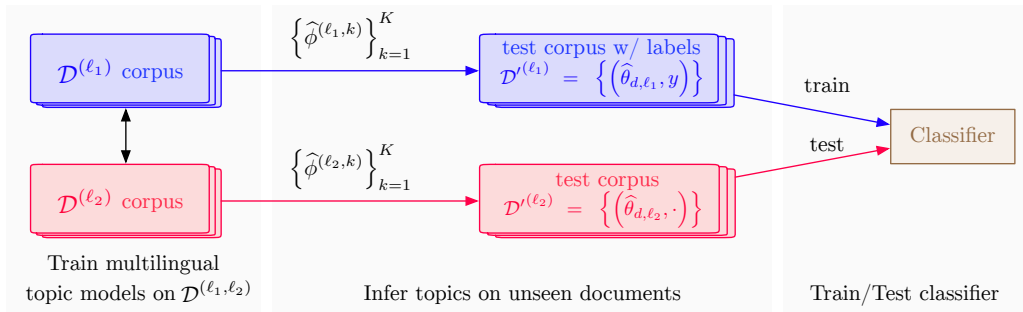
Figure 11
An illustration of crosslingual document classification. After training multilingual topic models, the topics, $\{\widehat{\phi}^{(\ell,k)}\}$ are used to infer document-topic distributions $\widehat{\theta}$ of unseen documents in both languages. A classifier is trained with the inferred distributions $\widehat{\theta}_{d,\ell_1}$ as features and the labels y in language ℓ_1, and predicts labels in language ℓ_2.

Using the two multilingual sources, we create two types of test sets for HIGHLAN—TED + TED and TED + GV, and only one type for LOWLAN—TED+GV. In **TED+TED**, we infer document-topic distributions on documents from TED in English and the paired language. This only applies to HIGHLAN, because TED do not have documents in LOWLAN. In **TED+GV**, we infer topics on English documents from TED, and infer topics on documents from GV in the paired language (both HIGHLAN and LOWLAN). The two types of test sets also represent different application situations. TED + TED implies that the test documents in both languages are parallel and come from the same source, whereas TED + GV represents how the topic model performs when the two languages have different data sources.

Both corpora are retrieved from `http://opus.nlpl.eu/` (Tiedemann 2012). The labels, however, are manually retrieved from `http://ted.com/` and `http://globalvoices.org/`. In TED corpus, each document is a transcript of a talk and is assigned to multiple categories on the Web page, such as "technology," "arts," and so forth. We collect all categories for the entire TED corpus, and use the three most frequent categories—*technology*, *culture*, *science*—as document labels. Similarly, in GV corpus, each document is a news story, and has been labeled with multiple categories on the Web page of the story. Because in TED + GV, the two sets are from different sources, and training and testing is only possible when both sets share the same labels, we apply the same three labels from TED to GV as well. This processing requires minor mappings, for example, from "arts-culture" in GV to "culture" in TED. The data statistics are presented in Table 7.

6. Document-Level Transfer and Its Limitations

We first explore the empirical characteristics of document-level transfer, using DOC-LINK, C-BILDA, and SOFTLINK.

Multilingual corpora can be loosely categorized into three types: parallel, comparable, and incomparable. A parallel corpus contains exact document translations

Table 7
Statistics of TED Talks 2013 (TED) and Global Voices (GV) corpus.

		Corpus statistics			Label distributions		
		#docs	#types	#tokens	*#technology*	*culture*	*science*
TED	AR	1,112	1,066,754	15,124	384	304	290
	DE	1,063	774,734	19,826	364	289	276
	ES	1,152	933,376	13,088	401	312	295
	RU	1,010	831,873	17,020	346	275	261
	ZH	1,123	1,032,708	19,594	386	315	290
GV (HIGHLAN)	AR	2,000	325,879	13,072	510	489	33
	DE	1,481	269,470	16,031	346	344	42
	ES	2,000	367,631	11,104	457	387	38
	RU	2,000	488,878	16,157	516	369	62
	ZH	2,000	528,370	18,194	499	366	56
GV (LOWLAN)	AM	39	10,589	4,047	3	3	1
	AY	674	66,076	4,939	76	100	46
	MK	1,992	388,713	29,022	343	426	182
	SW	1,383	359,066	14,072	137	110	71
	TL	254	26,072	6,138	32	67	19

across languages, of which EUROPARL and the Bible, discussed before, are examples. A comparable corpus contains document pairs (in the bilingual case), where each document in one language has a related counterpart in the other language. However, these document pairs are not exact translations of each other, and they can only be connected through a loosely defined "theme." Wikipedia is an example, where document pairs are linked by article titles. Incomparable corpora contain potentially unrelated documents across languages, with no explicit indicators of document pairs.

With different levels of comparability comes different availabilities of such corpora: It is much harder to find parallel corpora in low-resource languages. Therefore, we first focus on HIGHLAN, and use Wikipedia to simulate the low-resource situation in Section 6.1, where we find that DOCLINK and C-BILDA are very sensitive to the training corpus, and thus might not be the best option when it comes to low-resource languages. We then examine LOWLAN in Section 6.2.

6.1 Sensitivity to Training Corpus

We first vary the comparability of the training corpus and study how different models behave under different situations. All models are potentially affected by the comparability of the training set, although only DOCLINK and C-BILDA explicitly rely on this information to define transfer operations. This experiment shows that models transferring knowledge on the document level (DOCLINK and C-BILDA) are very sensitive to the training set, but can be almost entirely insensitive with appropriate modifications to the transfer operation as in SOFTLINK.

6.1.1 Experiment Settings. For each language pair (EN, ℓ), we construct a random subsample of $2,000$ documents from Wikipedia in each language ($4,000$ in total). To vary the comparability, we vary the proportion of linked Wikipedia articles between the two languages, from 0.0, 0.01, 0.05, 0.1, 0.2, 0.4, 0.8, to 1. When the percentage is zero, the bilingual corpus is entirely **incomparable**, that is, no document-level translations can be found in another language, and DOCLINK and C-BILDA degrade into monolingual LDAs. The indicator matrix used by transfer operations in Section 4.1.1 is a zero matrix $\delta = \mathbf{0}$. When the percentage is one, meaning each document from one language is linked to one document from another language, the corpus is considered fully comparable, and δ is an identity matrix $\mathbf{1}$. Any number between 0 and 1 makes the corpus **partially comparable** to different degrees. The CNPMI and crosslingual classification results are shown in Figure 12, and the shades indicate the standard deviations across five Gibbs sampling chains. For VOCLINK and SOFTLINK, we use all the dictionary entries.

6.1.2 Results. In terms of topic coherence (CNPMI), both DOCLINK and C-BILDA have competitive performance on CNPMI, and achieve full potential when the corpus is fully comparable. As expected, models transferring knowledge at the document level (DOCLINK and C-BILDA) are very sensitive to the training corpus: The more aligned the corpus is, the better topics the model learns. For the word-level model, VOCLINK roughly stays at the same performance level, which is also expected, because this model does not use linked documents as supervision. However, its performance on Russian is surprisingly low compared with other languages and models. In the next section, we will look closer at this problem by investigating the impact of dictionaries.

It is notable that SOFTLINK, a document-level model, is also insensitive to the training corpus and outperforms other models most of the time. Recall that on the document level, SOFTLINK defines transfer operation on document-topic distributions θ,
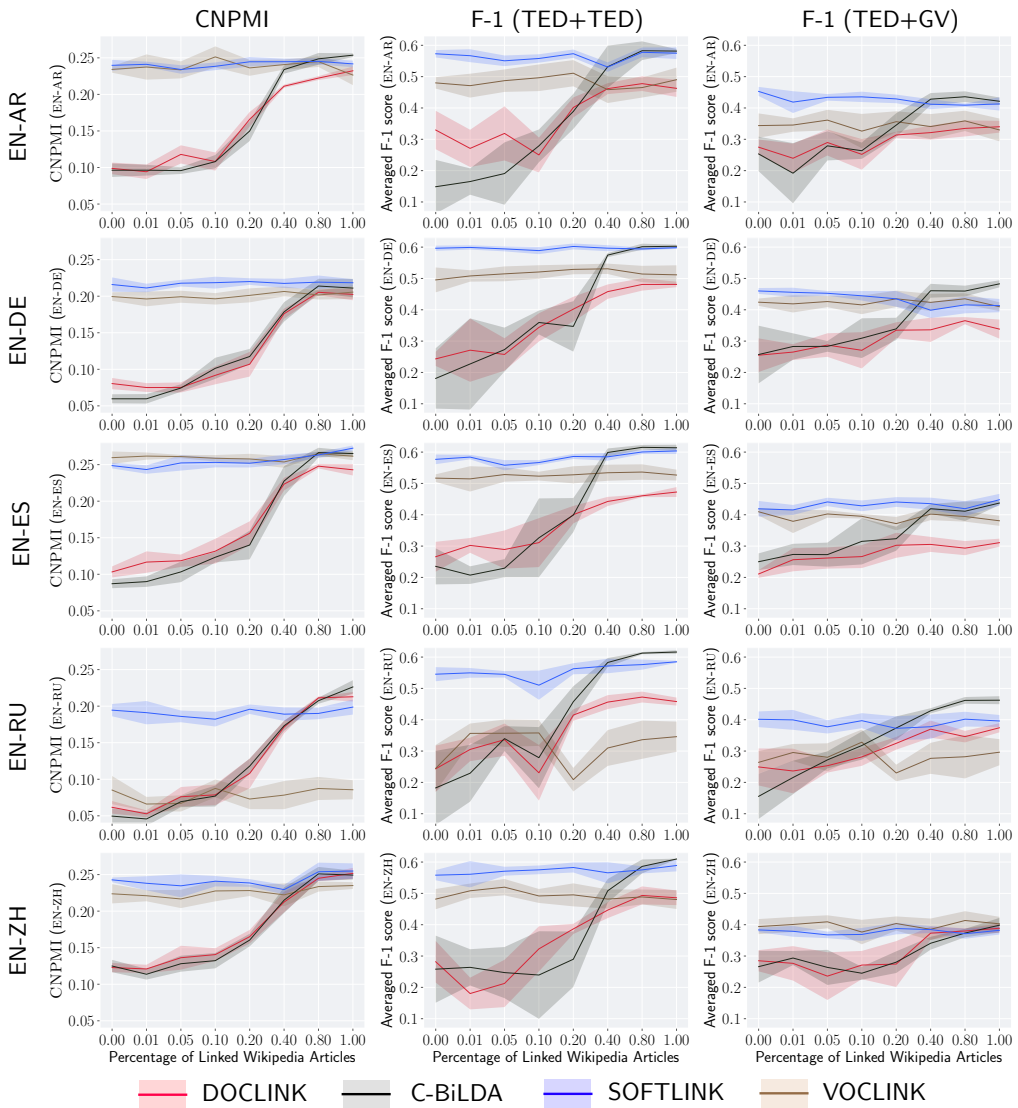
Figure 12
Both SOFTLINK and VOCLINK stay at a stable performance level of either CNPMI or F-1 scores, whereas DOCLINK and C-BILDA expectedly have better performance as there are more linked Wikipedia articles.

similarly to DOCLINK and C-BILDA, but using dictionary resources. This implies that good design of the supervision δ in the transfer operation could lead to a more stable performance across different training situations.

When it comes to the classification task, the F-1 scores of DOCLINK and C-BILDA have very large variations, and the increasing trend of F-1 scores is less obvious than with CNPMI. This is especially true when the percentage of linked documents is very small. For one, when the percentage is small, the transfer on the document level is less constrained, leaving the projection of two languages into the same topic space less predictive. The evaluation scope of CNPMI is actually much smaller and more

concentrated than classification, because it only focuses on the top *C* words, which does not lead to large variations.

One consistent result we notice is that SOFTLINK still performs well on classification with very small variations and stable F-1 scores, which again benefits from the definition of transfer operation in SOFTLINK. When transferring topics to another language, SOFTLINK uses dictionary constraints as in VOCLINK, but instead of a simple one-on-one word type mapping, it expands the transfer scope to the entire document. Additionally, SOFTLINK distributionally transfers knowledge from the entire corpus in another language, which actually reinforces the transfer efficiency without relying on direct supervision at the document level.

6.2 Performance on LOWLAN

In this section, we take a look at languages in LOWLAN. For SOFTLINK and VOCLINK, we use all dictionary entries to train languages in LOWLAN, because the sizes of dictionaries in these languages are already very small. We again use a subsample of $2,000$ Wikipedia document pairs with English to make the results comparable with HIGHLAN. In Figure 13(a), we also present results of models for HIGHLAN using fully comparable training corpora and full dictionaries for direct comparison of the effect of language resources.

In most cases, transfer on document level (particularly C-BILDA) performs better than on word levels, in both HIGHLAN and LOWLAN. Considering the number of dictionary entries available from Table 4, it is reasonable to suspect that the dictionary is a major factor affecting the performance of word-level transfer.

On the other hand, although SOFTLINK does not model vocabularies directly as in VOCLINK, transferring knowledge at the document level with a limited dictionary still yields competitive CNPMI scores. Therefore, in this experiment on LOWLAN, we see that with the same lexicon resource, it is generally more efficient to transfer knowledge at the document level. We will also explore this in detail in Section 7.

We also present a comparison of micro-averaged F-1 scores between HIGHLAN and LOWLAN in Figure 13(b). The test set used for this comparison is TED + GV, since TED does not have articles available in LOWLAN. Also, languages such as Amharic (AM) have fewer than 50 GV articles available, which is an extremely small number for training a robust classifier, so in these experiments, we only train classifiers on English (TED articles) and test them on languages in HIGHLAN and LOWLAN (GV articles).

Similarly, the classification results are generally better in document-level transfer, and both C-BILDA and SOFTLINK give similar scores. However, it is worth noting that VOCLINK has very large variations in all languages, and the F-1 scores are very low. This again suggests that transferring knowledge on the word level is less effective, and in Section 7 we study in detail why this is the case.

7. Word-Level Transfer and Its Limitations

In the previous section, we compared different multilingual topic models with a focus on document-level models. We draw conclusions that DOCLINK and C-BILDA are very sensitive to the training corpus, which is natural due to their definition of supervision as a one-to-one document pair mapping. On the other hand, the word-level model VOCLINK in general has lower performance, especially with LOWLAN, even if the corpus is entirely comparable.
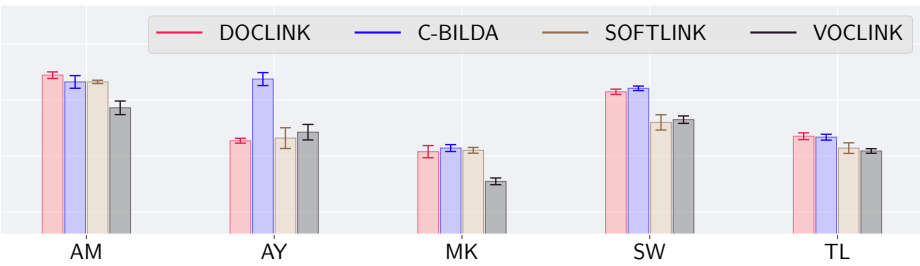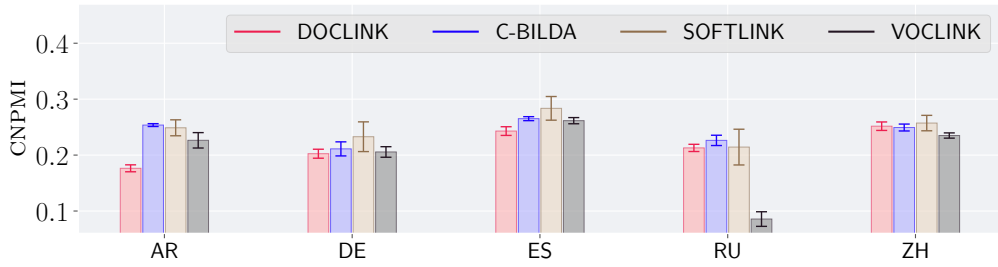
(a) score comparison of different models and languages with cardinality $C = 10$.
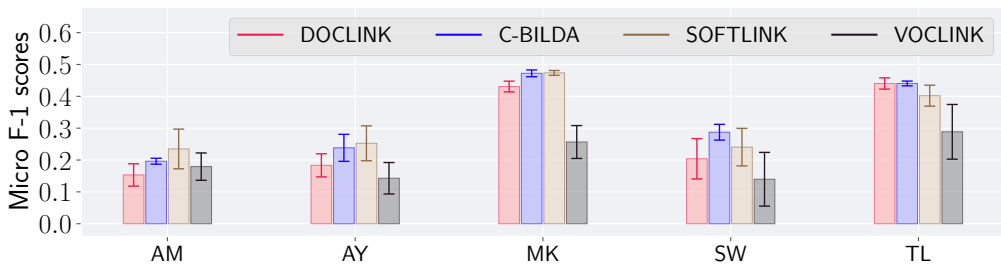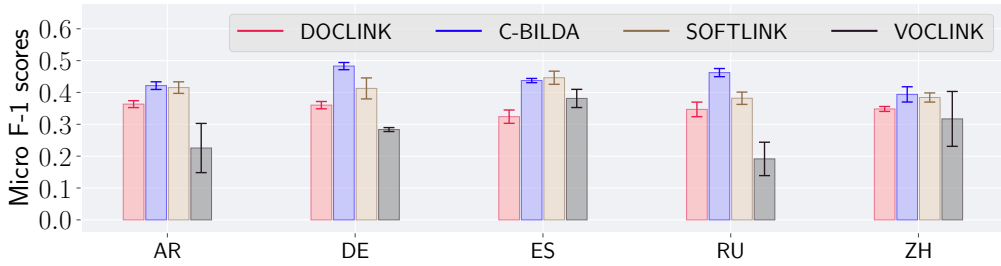
(b) Micro-averaged F-1 scores of different models and languages on TED+GV corpora.

Figure 13
Topic quality evaluation and classification performance on both HIGHLAN and LOWLAN. We notice that VOCLINK has lower CNPMI and F-1 scores in general, with large standard deviations. C-BILDA, on the other hand, outperforms other models in most of the languages.

One interesting result we observed from the previous section is that SOFTLINK and VOCLINK use the same dictionary resource while transferring topics on different levels, and SOFTLINK generally has better performance than VOCLINK. Therefore, in this section, we explore the characteristics of the word-level model VOCLINK and compare it with SOFTLINK to study why it does not use the same dictionary resource as effectively.

To this end, we first vary the amount of dictionary entries available and compare how SOFTLINK and VOCLINK perform (Section 7.1). Based on the results, we analyze word-level transfer from three different angles: dictionary usage (Section 7.2) as an intuitive explanation of the models, topic analysis (Section 7.3) from a more qualitative perspective, and comparing transfer strength (Section 7.4) as a quantitative analysis.

7.1 Sensitivity to Dictionaries

Word-level models such as VOCLINK use a dictionary as supervision, and thus will naturally be affected by the dictionary used. Although SOFTLINK transfers knowledge on the document level, it uses the dictionary to calculate the transfer distributions used in its document-level transfer operation. In this section, we focus on the comparison of SOFTLINK and VOCLINK.

7.1.1 Sampling the Dictionary Resource. The dictionary is the essential part of SOFTLINK and VOCLINK and is used in different ways to define transfer operations. The availability of dictionaries, however, varies among different languages. From Table 4, we notice that for LOWLAN the number of available dictionary entries is very limited, which suggests it could be a major factor affecting the performance of word-level topic models. Therefore, in this experiment, we sample different numbers of dictionary entries in HIGHLAN to study how this alters performance of SOFTLINK and VOCLINK.

Given a bilingual dictionary, we add only a proportion of entries in it to SOFTLINK and VOCLINK. As in the previous experiments varying the proportion of document links, we change the proportion from 0, 0.01, 0.05, 0.1, 0.2, 0.4, 0.8, to 1.0. When the proportion is 0, both SOFTLINK and VOCLINK become monolingual LDA and no transfer happens; when the proportion is 1, both models reach their highest potential with all the dictionary entries available.

We also sample the dictionary in two manners: random- and frequency-based. In random-based, the entries are randomly chosen from the dictionary, and the five chains have different entries added to the models. In frequency-based, we select the most frequent word types from the training corpus.

Figure 14 shows a detailed comparison among different evaluations and languages. As expected, adding more dictionary entries helps both SOFTLINK and VOCLINK, with increasing CNPMI scores and F-1 scores in general. However, we notice that adding more dictionary entries can boost SOFTLINK's performance very quickly, whereas the increase in VOCLINK's CNPMI scores is slower. Similar trends can be observed in the classification task as well, where adding more words does not necessarily increase VOCLINK's F-1 scores, and the variations are very high.

This comparison provides an interesting insight to increasing lexical resources efficiently. In some applications, especially related to low-resource languages, the number of available lexicon resources is very small, and one way to solve this problem is to incorporate human feedback, such as interactive topic modeling proposed by Hu et al. (2014a). In our case, a native speaker of the low-resource language could provide word translations that could be incorporated into topic models. Because of limited time and financial budget, however, it is impossible to translate all the word types that appear in the corpus, so the challenge is how to boost the performance of the target task as much as possible with less effort from humans. In this comparison, we see that if the target task is to train coherent multilingual topics, training SOFTLINK is a more efficient way than VOCLINK.
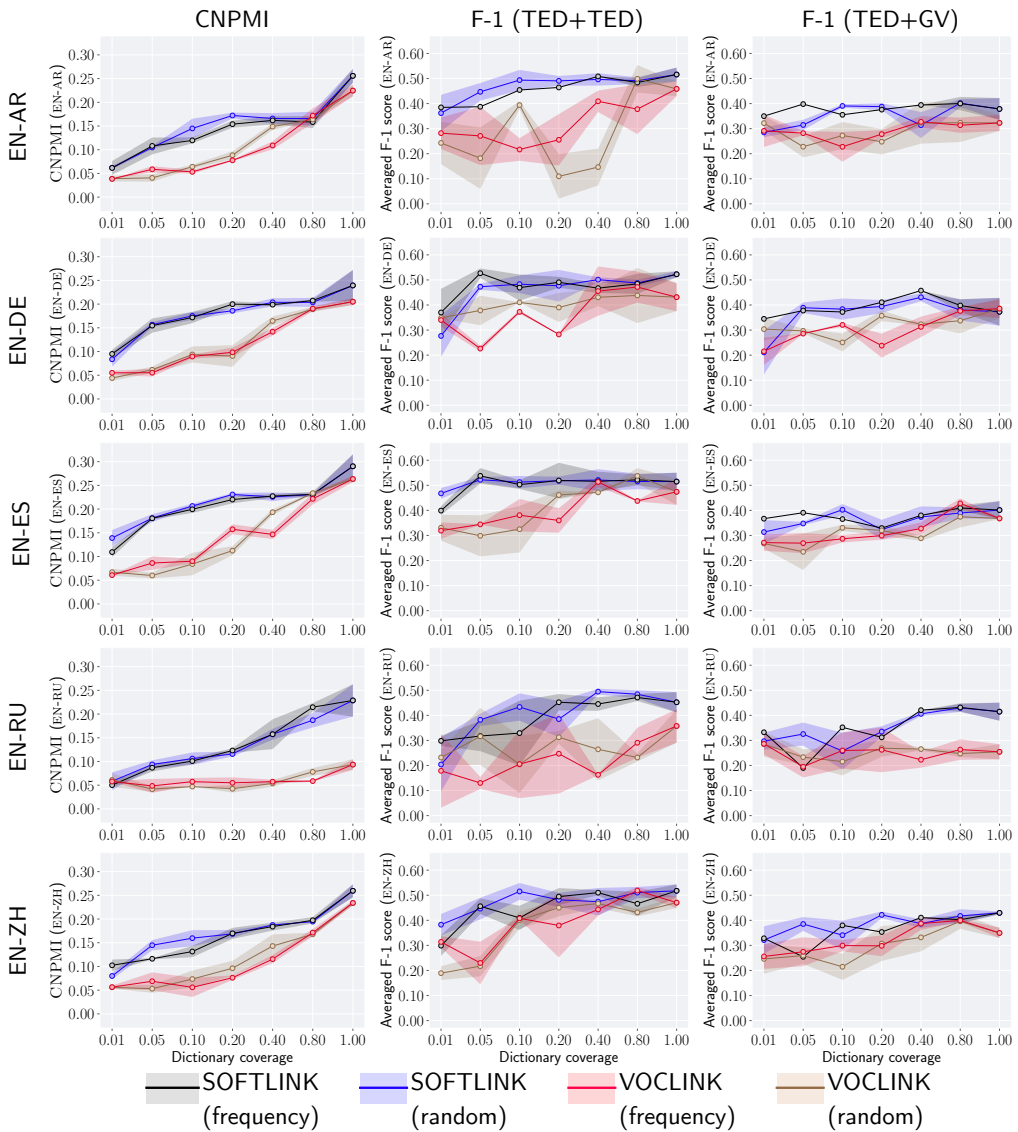
Figure 14
SOFTLINK produces better topics and is more capable of crosslingual classification tasks than
VOCLINK when the number of dictionary entries is very limited.

7.1.2 Varying Comparability of the Corpus. For SOFTLINK and VOCLINK, the dictionary is
only one aspect of the training situation. As discussed in our document-level experi-
ments, the training corpus is also an important factor that could affect the performance
of all topic models. Although corpus comparability is not an explicit requirement of
SOFTLINK and VOCLINK, the comparability of the corpus might affect the coverage pro-
vided by the dictionary or affect performance in other ways. In SOFTLINK, comparability
could also affect the transfer operator's ability to find similar documents to link to. In
this section, we study the relationship between dictionary coverage and comparability
of the training corpus.

Similar to the previous section, we vary the dictionary coverage from 0.01, 0.05, 0.1, 0.2, 0.4, 0.8, to 1, using the frequency-based method as in the last experiment. We also vary the number of linked Wikipedia articles from 0, 0.01, 0.05, 0.1, 0.2, 0.4, 0.8, to 1. We present CNPMI scores in Figure 15(a), where the results are averaged over all five languages in HIGHLAN. It is clear that SOFTLINK outperforms VOCLINK, regardless of training corpus and dictionary size. This implies that SOFTLINK could potentially learn coherent multilingual topics even when the training conditions are unfavorable: for example, when the training corpus is incomparable and there is only a small number of dictionary entries.

SOFTLINK (Average CNPMI scores)

Dictionary Coverage	0.0	0.01	0.05	0.1	0.2	0.4	0.8	1.0
1.0	.167	.170	.169	.168	.169	.178	.179	.166
0.8	.165	.157	.158	.170	.171	.166	.174	.184
0.4	.146	.155	.141	.152	.147	.159	.161	.167
0.2	.138	.151	.140	.137	.147	.141	.147	.149
0.1	.121	.122	.129	.128	.132	.133	.129	.133
0.05	.096	.105	.101	.108	.106	.099	.110	.107
0.01	.079	.072	.073	.079	.068	.067	.073	.077

VOCLINK (Average CNPMI scores)

Dictionary Coverage	0.0	0.01	0.05	0.1	0.2	0.4	0.8	1.0
1.0	.152	.157	.160	.162	.156	.165	.158	.155
0.8	.153	.153	.149	.154	.147	.153	.147	.165
0.4	.109	.114	.120	.102	.115	.113	.116	.117
0.2	.077	.079	.077	.074	.084	.079	.081	.085
0.1	.056	.058	.054	.054	.056	.061	.060	.066
0.05	.048	.039	.044	.051	.045	.046	.054	.052
0.01	.043	.040	.042	.042	.042	.044	.043	.042

Percentage of Linked Wikipedia Articles

(a) Average CNPMI scores on multilingual topic coherence.

SOFTLINK (Test set: TED + TED)

Dictionary Coverage	0.0	0.01	0.05	0.1	0.2	0.4	0.8	1.0
1.0	.482	.487	.467	.452	.515	.473	.480	.521
0.8	.477	.495	.468	.456	.501	.482	.474	.457
0.4	.407	.484	.465	.513	.470	.465	.460	.461
0.2	.464	.455	.478	.445	.465	.478	.478	.445
0.1	.405	.469	.428	.473	.417	.424	.426	.455
0.05	.410	.427	.442	.409	.425	.424	.448	.460
0.01	.386	.352	.374	.410	.344	.390	.383	.359

VOCLINK (Test set: TED + TED)

Dictionary Coverage	0.0	0.01	0.05	0.1	0.2	0.4	0.8	1.0
1.0	.422	.459	.462	.465	.449	.468	.476	.413
0.8	.427	.470	.431	.362	.373	.470	.400	.387
0.4	.326	.438	.366	.352	.346	.367	.369	.320
0.2	.261	.317	.302	.403	.302	.347	.249	.303
0.1	.294	.191	.239	.250	.313	.216	.353	.327
0.05	.262	.249	.244	.195	.312	.309	.165	.167
0.01	.095	.251	.320	.326	.216	.212	.284	.220

SOFTLINK (Test set: TED + GV)

Dictionary Coverage	0.0	0.01	0.05	0.1	0.2	0.4	0.8	1.0
1.0	.384	.381	.391	.410	.406	.336	.382	.400
0.8	.373	.372	.348	.387	.393	.373	.398	.399
0.4	.393	.425	.373	.390	.364	.397	.349	.354
0.2	.381	.360	.401	.342	.365	.357	.374	.365
0.1	.335	.331	.336	.361	.350	.319	.342	.371
0.05	.372	.373	.351	.359	.366	.333	.332	.340
0.01	.304	.334	.290	.349	.322	.286	.286	.286

VOCLINK (Test set: TED + GV)

Dictionary Coverage	0.0	0.01	0.05	0.1	0.2	0.4	0.8	1.0
1.0	.334	.331	.354	.335	.352	.373	.367	.365
0.8	.270	.347	.301	.339	.314	.324	.319	.357
0.4	.288	.331	.321	.326	.307	.302	.294	.301
0.2	.290	.298	.221	.293	.279	.285	.217	.257
0.1	.273	.272	.266	.230	.267	.242	.251	.203
0.05	.270	.228	.290	.261	.317	.281	.184	.261
0.01	.207	.211	.256	.263	.212	.213	.263	.214

Percentage of Linked Wikipedia Articles

(b) Multilabel crosslingual document classification F-1 scores in HIGHLAN.

Figure 15
Adding more dictionary entries has a higher impact on word-level model VOCLINK. SOFTLINK learns better quality topics than VOCLINK. SOFTLINK also generally performs better on classification.

The results of crosslingual classification are shown in Figure 15(b). When the test sets are from the same source (TED + TED), SOFTLINK utilizes the dictionary more efficiently and performs better than VOCLINK. In particular, F-1 scores of SOFTLINK using only 20% of dictionary entries is already outperforming VOCLINK using the full dictionary. A similar comparison can also be drawn when the test sets are from different sources such as TED + GV.

7.1.3 Discussion. From the results so far, it is empirically clear that transferring knowledge on the word level tends to be less efficient than the document level. This is arguably counter-intuitive. Recall that the goal of multilingual topic models is to let semantically related words and translations have similar distributions over topics. The word-level model VOCLINK directly uses this information—dictionary entries—to define transfer operations, yet its CNPMI scores are lower. In the following sections, therefore, we try to explain this apparent contradiction. We first analyze the dictionary usage of VOCLINK (Section 7.2), and then lead our discussion on the transfer strength comparisons between document and word levels for all models (Sections 7.3 and 7.4).

7.2 Dictionary Usage

In practice, the assumption of VOCLINK is also often weakened by another important factor: the presence of word translations in the training corpus. Given a word pair $\left(w^{(\ell_1)}, w^{(\ell_2)}\right)$, the assumption of VOCLINK is valid only when both words appear in the training corpus in their respective languages. If $w^{(\ell_2)}$ is not in $D^{(\ell_2)}$, $w^{(\ell_1)}$ will be treated as an untranslated word instead. Figure 16 shows an example of how tree structures in VOCLINK are affected by the corpus and the dictionary.

In Figure 17, we present the statistics of word types from different sources on a logarithmic scale. "Dictionary" is the number of word types that appeared in the original dictionary as shown in the last column of Table 4, and we use the same preprocessing to the dictionary as to the training corpus to make sure the quantities are comparable. "Training set" is the number of word types that appeared in the training
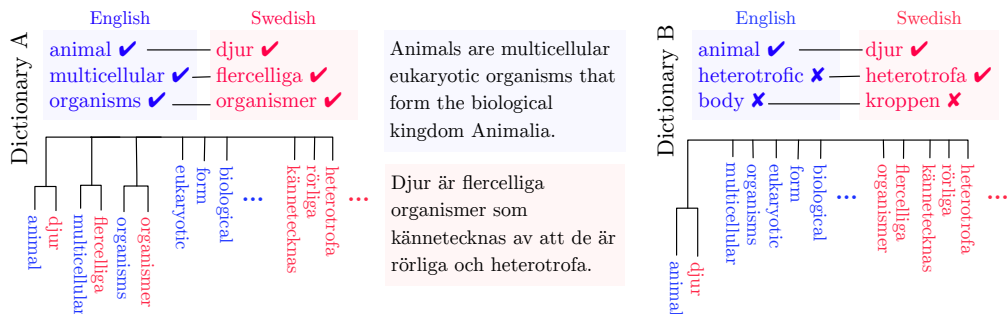
Figure 16
The dictionary used by VOCLINK is affected by its overlap with the corpus. In this example, the three entries in Dictionary A can all be found in the corpus, so the tree structure has all of them. However, only one entry in Dictionary B can be found in the corpus. Although the Swedish word "heterotrofa" is also in the dictionary, its English translation cannot be found in the corpus, so Dictionary B ends up a tree with only one entry.
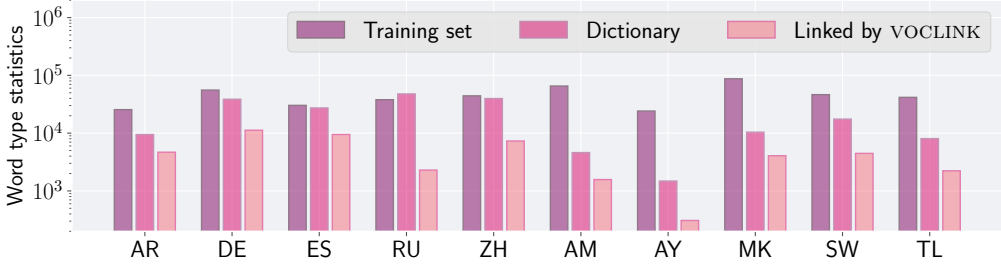
Figure 17
The number of word types that are linked in VOCLINK is far less than the original dictionary and even than that of word types in the training sets.

set, and "Linked by VOCLINK" is the number of word types that are actually used in VOCLINK, that is, the number of non-zero entries in δ in the transfer operation.

Note that even when we use the complete dictionary to create the tree structure in VOCLINK, in LOWLAN, there are far more word types in the training set than those in the dictionary. In other words, the supervision matrix δ used by $h_{\phi^{(r,k)}}$ is never actually full rank, and thus, the full potential of VOCLINK is very difficult to achieve due to the properties of the training corpus. This situation is as if the document-level model DOCLINK had only half of the linked documents in the training corpus.

On the other hand, we notice that in HIGHLAN, the number of word types in the dictionary is usually comparable to that of the training set (except in AR). For LOWLAN, however, the situation is quite the contrary: There are more word types in the training set than in the dictionary. Thus, the availability of sufficient dictionary entries is especially a problem for LOWLAN.

We conclude from Figure 15(a) that adding more dictionary entries will slowly improve VOCLINK, but even when there are enough dictionary items, due to model assumptions, VOCLINK will not achieve its full potential unless every word in the training corpus is in the dictionary. A possible solution is to first extract word alignments from parallel corpora, and then create a tree structure using those word alignments, as experimented in Hu et al. (2014b). However, when parallel corpora are available, we have shown that document-level models such as DOCLINK work better anyway, and the accuracy of word aligners is another possible limitation to consider.

7.3 Topic Analysis

Whereas VOCLINK uses a dictionary to directly model word translations, SOFTLINK uses the same dictionary to define the supervision in transfer operation differently on the document level. Experiments show that transferring knowledge on the document level with a dictionary (i.e., SOFTLINK) is more efficient, resulting in stable and low-variance topic qualities in various training situations. A natural question is why the same resource results in different performance on different levels of transfer operations. To answer this question from another angle, we further look into the actual topics trained from SOFTLINK and VOCLINK in this section. The general idea is to look into the same topic output from SOFTLINK and VOCLINK and see what topic words they have in common (denoted as $\mathcal{W}^+$), and what words they have exclusively, denoted as $\mathcal{W}^{-,\text{SOFT}}$ and $\mathcal{W}^{-,\text{VOC}}$ for SOFTLINK and VOCLINK, respectively. The words in $\mathcal{W}^{-,\text{VOC}}$ are

those with lower topic coherence and are thus the key to understanding the suboptimal performance of VOCLINK.

7.3.1 Aligning Topics. To this end, the first step is to align possible topics between VOCLINK and SOFTLINK, since the initialization of Gibbs samplers is random. Let $\{\mathcal{W}_k^{\text{VOC}}\}_{k=1}^K$ and $\{\mathcal{W}_k^{\text{SOFT}}\}_{k=1}^K$ be the K topics learned by VOCLINK and SOFTLINK respectively, from the same training conditions. For each topic pair (k, k') we calculate the Jaccard index $\mathcal{W}_k^{\text{VOC}}$ and $\mathcal{W}_{k'}^{\text{SOFT}}$, one for each language, and use the average over the two languages as the matching score $m_{k,k'}$ of the topic pair:

$$m_{k,k'} = \frac{1}{2}\left(J\left(\mathcal{W}_{k,\ell_1}^{\text{VOC}}, \mathcal{W}_{k',\ell_1}^{\text{SOFT}}\right) + J\left(\mathcal{W}_{k,\ell_2}^{\text{VOC}}, \mathcal{W}_{k',\ell_2}^{\text{SOFT}}\right)\right) \tag{32}$$

where $J(X, Y)$ is the Jaccard index between sets X and Y. Thus, there are K^2 matching scores with a number of topics K. We set a threshold of 0.8, so that a matching score is valid only when it is greater than $0.8 \cdot \max m_{k,k'}$ over all the K^2 scores. For each topic k, if its matching score is valid, we align $\mathcal{W}_k^{\text{VOC}}$ with $\mathcal{W}_{k'}^{\text{SOFT}}$, and treat them as potentially the same topic. When multiple matching scores are valid, we use the topic with the highest score and ignore the rest.

7.3.2 Comparing Document Frequency. Using the approximate alignment algorithm we described above, we are now able to compare each aligned topic pair between VOCLINK and SOFTLINK.

For a word type w, we define the document frequency as the percentage of documents where w appears. A low document frequency of word w implies that w only appears in a small number of documents. For every aligned topic pair $(\mathcal{W}_i, \mathcal{W}_j)$ where $\mathcal{W}_i$ and $\mathcal{W}_j$ are topic word sets from SOFTLINK and VOCLINK, respectively, we have three sets of topic words derived from this pair:

$$\mathcal{W}^+ \quad = \mathcal{W}_i \cap \mathcal{W}_j \tag{33}$$

$$\mathcal{W}^{-,\text{VOC}} = \mathcal{W}_i \setminus \mathcal{W}_j \tag{34}$$

$$\mathcal{W}^{-,\text{SOFT}} = \mathcal{W}_j \setminus \mathcal{W}_i \tag{35}$$

Then we calculate the average document frequencies over all the words in each of the sets, and we show the results in Figure 18.

We observe that the average document frequencies over words in $\mathcal{W}^{-,\text{VOC}}$ are consistently lower in every language, whereas those in $\mathcal{W}^+$ are higher. This implies that VOCLINK tends to give rare words higher probability in the topic-word distributions. In other words, VOCLINK gives high probabilities to words that only appear in specific contexts, such as named entities. Thus, when evaluating topics using a reference corpus, the co-occurrence of such words with other words is relatively low due to lack of that specific context in the reference corpus.

We show an example of an aligned topic in Figure 19. In this example, we see that although both VOCLINK and SOFTLINK can discover semantically coherent words shown in $\mathcal{W}^+$, VOCLINK focuses more on words that only appear in specific contexts: There are many words (mostly named entities) in $\mathcal{W}^{-,\text{VOC}}$ that only appear in one document. Due to lack of this very specific context in the reference corpora, the
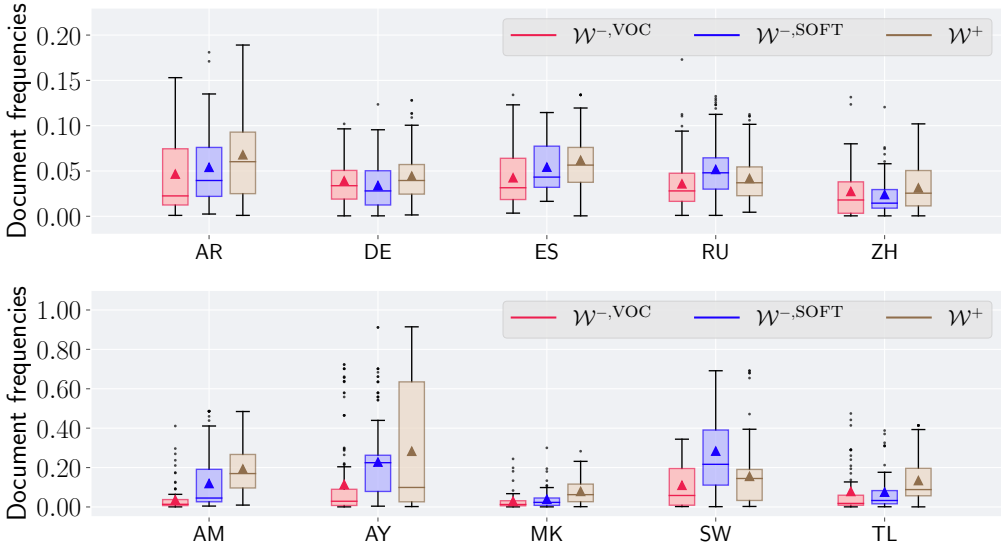
Figure 18
Average document frequencies of $\mathcal{W}^{-,\text{VOC}}$ are generally lower than $\mathcal{W}^{-,\text{SOFT}}$ and $\mathcal{W}^+$, shown in the triangle markers.

Figure 19
An example of real data showing the topic words of SOFTLINK and VOCLINK. Words that appear in both models are in $\mathcal{W}^+$; words that only appear in SOFTLINK or VOCLINK are included in $\mathcal{W}^{-,\text{SOFT}}$ or $\mathcal{W}^{-,\text{VOC}}$, respectively.

co-occurrence of these words with other more general words is likely to be zero, resulting in lower CNPMI.

7.4 Comparing Transfer Strength

While we have looked at the topics to explain what kind of words produced by VOC-LINK make the model's performance lower than SOFTLINK, in this section, we try to explain why this happens by analyzing their transfer operations. Recall that VOCLINK defines transfer operations on topic-node distributions $\{\phi_{k,r}\}_{k=1}^K$ (Equation (23)), while SOFTLINK defines transfer on document-topic distributions θ. The differences between transfer levels with the same resources leads to a suspicion that document level has a "stronger" transfer power.

The first question is to understand how this transfer operation actually functions in the training of topic models. During Gibbs' sampling of monolingual LDA, the

conditional distribution for a token, denoted as $\mathcal{P}$, is calculated by conditioning on all the other tokens and their topics, and can be factorized into two conditionals: document-level $\mathcal{P}_\theta$ and word-level $\mathcal{P}_\phi$. Let the current token be of word type w, and $\mathbf{w}_-$ and $\mathbf{z}_-$ all the other words and their current topic assignments in the corpus. The conditional is then

$$\mathcal{P}_k = \mathrm{Pr}\left(z = k | w, \mathbf{w}_-, \mathbf{z}_-\right) \tag{36}$$

$$\propto \left(n_{k|d} + \alpha_k\right) \cdot \frac{n_{w|k} + \beta_w}{n_{\cdot|k} + \mathbf{1}^\top \beta} \tag{37}$$

$$= \mathcal{P}_{\theta k} \cdot \mathcal{P}_{\phi k} \tag{38}$$

where $n_{k|d}$ is the number of topic k in document d, $n_{w|k}$ the number of word type w in topic k, $n_{\cdot|k}$ the number of tokens assigned to topic k, and $\mathbf{1}$ an all-one vector. In this equation, the final conditional distribution can be treated as a "vote" from the two conditionals: $\mathcal{P}_\theta$ and $\mathcal{P}_\phi$ (Yuan et al. 2015). If $\mathcal{P}_\phi$ is a uniform distribution, then $\mathcal{P} = \mathcal{P}_\theta$, meaning the conditional on document $\mathcal{P}_\theta$ dominates the decision of choosing a topic, while the conditional on word $\mathcal{P}_\phi$ is uninformative.

We apply this similar idea to multilingual topic models. For a token in language ℓ_2, we let w be its word type, and $\mathcal{P}$ can also generally be factorized to two individual conditionals,

$$\mathcal{P}_k = \mathrm{Pr}\left(z = k | w, \mathbf{w}_-, \mathbf{z}_-\right) \tag{39}$$

$$\propto \underbrace{\left[n_{k|d} + h_\theta\left(\delta, \mathbf{N}^{(\ell_1)}, \alpha\right)_k\right]}_{\mathcal{P}_{\mathrm{DOC},k}} \cdot \underbrace{\frac{n_{w|k} + h_\phi\left(\delta', \mathbf{N}^{(\ell_1)}, \beta\right)_w}{n_{\cdot|k} + \mathbf{1}^\top h_\phi\left(\delta', \mathbf{N}^{(\ell_1)}, \beta\right)}}_{\mathcal{P}_{\mathrm{VOC},k}} \tag{40}$$

$$= \mathcal{P}_{\mathrm{DOC},k} \cdot \mathcal{P}_{\mathrm{VOC},k} \tag{41}$$

where the transfer operation is clearly incorporated into the calculation of the conditional, and $\mathcal{P}_{\mathrm{DOC}}$ and $\mathcal{P}_{\mathrm{VOC}}$ are conditional distributions on document and word levels, respectively. Thus, it is easy to see how transfers on different levels contribute to the decision of a topic. This is also where our comparison of "transfer strength" starts.

To apply this idea, for each token, we first obtain three distributions described before: $\mathcal{P}$, $\mathcal{P}_{\mathrm{DOC}}$, and $\mathcal{P}_{\mathrm{VOC}}$. Then we calculate cosine similarities $\cos\left(\mathcal{P}_{\mathrm{DOC}}, \mathcal{P}\right)$ and $\cos\left(\mathcal{P}_{\mathrm{VOC}}, \mathcal{P}\right)$. If $r = \frac{\cos(\mathcal{P}_{\mathrm{DOC}}, \mathcal{P})}{\cos(\mathcal{P}_{\mathrm{VOC}}, \mathcal{P})} > 1$, we know that $\mathcal{P}_{\mathrm{DOC}}$ is dominant and helps shape the conditional distribution $\mathcal{P}$; in other words, the document level transfer is stronger. We calculate the ratio of similarities $r = \frac{\cos(\mathcal{P}_{\mathrm{DOC}}, \mathcal{P})}{\cos(\mathcal{P}_{\mathrm{VOC}}, \mathcal{P})}$ for all the tokens in every model, and take the model-wise average over all the tokens (Figure 20). The most balanced situation is $r = 1$, meaning transfers on both word and document levels are contributing equally to the conditional distributions.

From the results, we notice that both DOCLINK and C-BILDA have stronger transfer strength on the document level, which means that the transfer operations on the document levels are actually informing the decision of a token's topic. However, we also notice that VOCLINK has very comparable transfer strength to DOCLINK and C-BILDA, which makes less sense, because VOCLINK defines transfer operations on the word level. This implies that transferring knowledge on the word level is weaker. This also explains
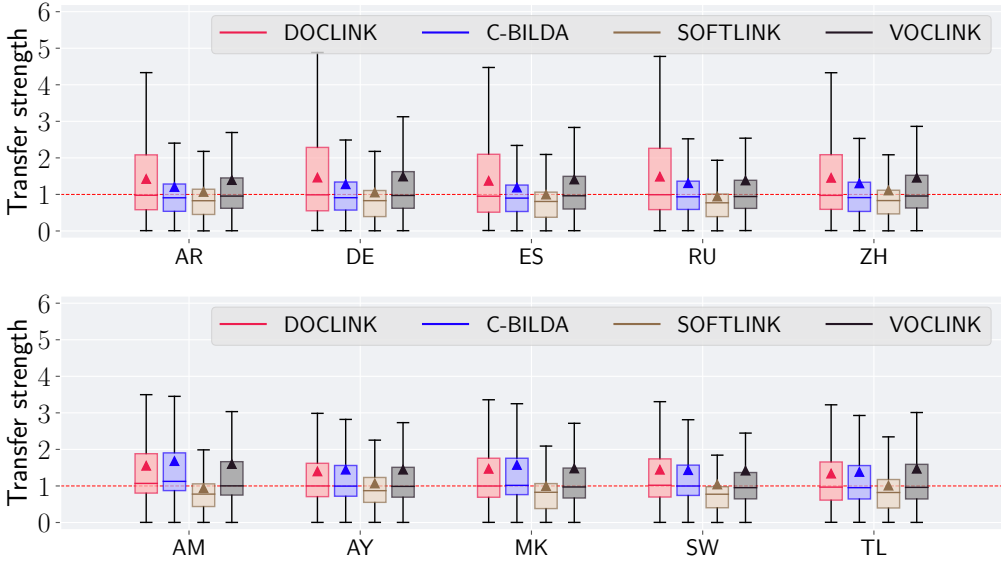
Figure 20
Comparisons of transfer strength. A value of one (shown in red dotted line) means an equal balance of transfer between document and word levels. We notice SOFTLINK has the most balanced transfer strength, whereas VOCLINK has stronger transfer at the document level although its transfer operation is defined on the word level.

why, in the previous section, VOCLINK tends to find topic words appearing in only a few documents.

It is also interesting to see SOFTLINK having a relatively good balance between document and word levels, with consistently the most balanced transfer strengths across all models and languages.

8. Remarks and Conclusions

Multilingual topic models use corpora in multiple languages as input with additional language resources as supervision. The traits of these models inevitably lead to a wide variety of training scenarios, especially when a language's resources are scarce, whereas most previous studies on multilingual topic models have not analyzed in depth the appropriateness of different models for different training situations and resource availability. For example, experiments are most often done in European languages, with models that are typically trained on parallel or comparable corpora.

The contributions of our study are providing a unifying framework of these different models, and systematically analyzing their efficacy in different training situations. We conclude by summarizing our findings along two dimensions: training corpora characteristics and dictionary characteristics, since these are the necessary components to enable crosslingual knowledge transfer.

8.1 Model Selection

Document-level models are shown to work best when the corpus is parallel or at least comparable. In terms of learning high-quality topics, DOCLINK and C-BILDA yield very

similar results. However, since C-BILDA has a "language selector" mechanism in the generative process, it is slightly more efficient for training Wikipedia articles in low-resource languages, where the document lengths have large gaps compared to English. SOFTLINK, on the other hand, only needs a small dictionary to enable document-level transfer, and yields very competitive results. This is especially useful for low-resource languages when the dictionary size is small and only a small number of comparable document pairs are available for training.

Word-level models are harder to achieve full potential of transfer, due to limits in the dictionary size and training sets, and unrealistic assumptions of the generative process regarding dictionary coverage. The representative model, VOCLINK, has similarly good performance on document classification as other models, but the topic qualities according to coherence-based metrics are lower. Comparing to SOFTLINK, which also requires a dictionary as resource, directly modeling word translations in VOCLINK turns out to be a less efficient way of transferring dictionary knowledge. Therefore, when using dictionary information, we recommend SOFTLINK over VOCLINK.

8.2 Crosslingual Representations

As an alternative method to learning crosslingual representations, crosslingual word embeddings have been gaining attention (Ruder, Vulic, and Søgaard 2019; Upadhyay et al. 2016). Recent crosslingual embedding architectures have been applied to a wider range of applications in natural language processing, and achieve state-of-the-art performance. Similar to the topic space in multilingual topic models, crosslingual embeddings learn semantically consistent features in a shared embedding space for all languages.

Both approaches—topic modeling and embedding—have advantages and limitations. Multilingual topic models still rely on supervised data to learn crosslingual representations. The choice of such supervision and model is important, which leads to our main discussion of this work. Topic models have the advantage of being interpretable. Embedding methods are powerful in many natural language processing tasks, and the representations are more fine-grained. Recent advancements in crosslingual embedding training do not require crosslingual supervision resources such as dictionary or parallel data (Artetxe, Labaka, and Agirre 2018; Lample et al. 2018), which is a large step toward generalization of crosslingual modeling. Although it is an open problem on how to interpret the results and how to reduce the heavy computing resources required, embedding based methods are a promising research direction.

Relations to Topic Models. A very common strategy for learning crosslingual embeddings is to use a projection matrix as supervision or sub-objective to learn a projection matrix that projects independently trained monolingual embeddings into a shared crosslingual space (Dinu and Baroni 2014; Faruqui and Dyer 2014; Tsvetkov and Dyer 2016; Vulić and Korhonen 2016).

In multilingual topic models, the supervision matrix δ plays the role of a projection matrix between languages. In DOCLINK, for example, $\delta_{d_{\ell_2}, d_{\ell_1}}$ projects document d_{ℓ_2} to the document space of ℓ_1 (Equation (15)). SOFTLINK provides a simple extension by forming δ to a matrix of transfer distirbutions based on word-level document similarities. VOCLINK applies projections in the form of word translations.

Thus, we can see that the formation of projection matrices in multilingual topic models is still static and restricted to an identity matrix or a simple pre-calculated matrix. A generalization would be to add learning the projection matrix itself as an objective into multilingual topic models. This could be a way to improve VOCLINK

by extending word associations to polysemy across languages, and making it less dependent on context.

8.3 Future Directions

Our study inspires future work in two directions. The first direction is to increase the efficiency of word-level knowledge transfer. For example, it is possible to use co-location information of translated words to transfer knowledge, though cautiously, to untranslated words. It has been shown that word-level models can help find new word translations, for example, by using the existing dictionary as "seed," and gradually adding more internal nodes to the tree structure using trained topic-word distributions. Additionally, our analysis showed the benefits of using a "language selector" in C-BILDA to make the generative process of DOCLINK more realistic, and one could also implement a similar mechanism in VOCLINK to make the conditional distributions for tokens less dependent on specific context.

The second direction is more general. By systematically synthesizing various models and abstracting the knowledge transfer mechanism through an explicit transfer operation, we can construct models that shape the probabilistic distributions of a target language using that of a source language. By defining different transfer operations, more complex and robust models can be developed, and this transfer formulation may provide new ways of constructing models than with a traditional joint formulation (Hao and Paul 2019). For example, SOFTLINK is generalization DOCLINK based on transfer operations that does not have an equivalent joint formulation. This framework for thinking about multilingual topic models may lead to new ideas for other models.

References

Andrzejewski, David, 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*, pages 25–32, Montreal.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798, Melbourne.

Besag, Julian. 1975. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24:179–195.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, David M. 2018. Technical perspective: expressive probabilistic models and scalable method of moments. *Communications of the ACM*, 61(4):84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation.

Journal of Machine Learning Research, 3:993–1022.

Boyd-Graber, Jordan L. 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*, pages 75–82, Montreal.

Chang, Jonathan and David M. Blei. 2009. Relational topic models for document networks. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009*, pages 81–88, Clearwater Beach, FL.

Chang, Jonathan, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296, Vancouver.

Chen, Ning, Jun Zhu, Fei Xia, and Bo Zhang. 2013. Generalized relational topic models with data augmentation. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1273–1279, Beijing, China.

Christodoulopoulos, Christos and Mark Steedman. 2015. A massively parallel

corpus: The bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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

Dinu, Georgiana and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6568.

Faruqui, Manaal and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg.

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

Gutiérrez, E. Dario, 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.

Hao, Shudong, 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*, pages 1090–1100, New Orleans, LA.

Hao, Shudong and Michael J. Paul. 2018. Learning multilingual topics from incomparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 2595–2609, Santa Fe, NM.

Hao, Shudong and Michael J. Paul. 2019. Analyzing Bayesian crosslingual transfer in topic models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 1551–1565, Minneapolis, MN.

Heyman, Geert, Ivan Vulic, and Marie-Francine Moens. 2016. C-BiLDA extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content. *Data Mining and Knowledge Discovery*, 30(5):1299–1323.

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, CA.

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

Hu, Yuening, Ke Zhai, Vladimir Eidelman, and Jordan L. Boyd-Graber. 2014b. Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, pages 1166–1176, Baltimore, MD.

Jagarlamudi, Jagadeesh 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*, pages 444–456, Milton Keynes.

Kim, Do-kyum, Geoffrey M. Voelker, and Lawrence K. Saul. 2013. A variational approximation for topic modeling of hierarchical corpora. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, pages 55–63, Atlanta, GA.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit*, 5:79–86.

Koller, Daphne and Nir Friedman. 2009. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press.

Krstovski, Kriste and David A. Smith. 2011. A minimally supervised approach for detecting and ranking document translation pairs. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011*, pages 207–216, Edinburgh.

Krstovski, Kriste 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*, pages 1127–1132, San Diego, CA.

Krstovski, Kriste, 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*, pages 454–459, San Diego, CA.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver.

Lau, Jey Han, 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*, pages 530–539, Gothenburg.

Leppä-aho, Janne, 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.

Littman, Michael L., Susan T. Dumais, and Thomas K. Landauer. 1998. In Automatic cross-language information retrieval using latent semantic indexing, In G. Grefenstette, ed., *Cross-Language Information Retrieval*, Springer, pages 51–62.

Liu, Xiaodong, Kevin Duh, and Yuji Matsumoto. 2015. Multilingual topic models for bilingual dictionary extraction. *ACM Transactions on Asian & Low-Resource Language Information Processing*, 14(3):11:1–11:22.

Ma, Tengfei and Tetsuya Nasukawa. 2017. Inverted bilingual topic models for lexicon extraction from non-parallel data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 4075–4081, Melbourne.

Maaten, Laurens van der and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

McCallum, Andrew Kachites. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Mimno, David M., 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*, pages 880–889, Singapore.

Moens, Marie-Francine and Ivan Vulic. 2013. Monolingual and cross-lingual

probabilistic topic models and their applications in information retrieval. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013*, pages 874–877, Moscow.

Ni, Xiaochuan, 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*, pages 1155–1156, Madrid.

Ruder, Sebastian, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Seroussi, Yanir, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Smet, Wim De, 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*, pages 549–560, Shenzhen.

Søgaard, Anders, Zeljko Agic, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. 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*, pages 1713–1722, Beijing.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218, Istanbul.

Tsvetkov, Yulia and Chris Dyer. 2016. Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research*, 55:63–93.

Upadhyay, Shyam, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 1661–1670, Berlin.

Vulić, Ivan and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 247–257, Berlin.

Vulić, Ivan and Marie-Francine Moens. 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014*, pages 349–362, Doha.

Vulić, Ivan, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147.

Wallach, Hanna M., David M. Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems 22*, pages 1973–1981, Vancouver.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David M. Mimno.

2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, pages 1105–1112, Montreal.

Xu, Wei, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–273, Toronto.

Yuan, Jinhui, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. LightLDA: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015*, pages 1351–1361, Florence.