Polylingual Tree-Based Topic Models for Translation Domain Adaptation

Yuening Hu†
Computer Science
University of Maryland
ynhu@cs.umd.edu

Ke Zhai†
Computer Science
University of Maryland
zhaike@cs.umd.edu

Vladimir Eidelman
FiscalNote Inc.
Washington DC
vlad@fiscalnote.com

Jordan Boyd-Graber
iSchool and UMIACS
University of Maryland
jbg@umiacs.umd.edu

Abstract

Topic models, an unsupervised technique for inferring translation domains improve machine translation quality. However, previous work uses only the source language and completely ignores the target language, which can disambiguate domains. We propose new polylingual tree-based topic models to extract domain knowledge that considers both source and target languages and derive three different inference schemes. We evaluate our model on a Chinese to English translation task and obtain up to 1.2 BLEU improvement over strong baselines.

1 Introduction

Probabilistic topic models (Blei and Lafferty, 2009), exemplified by *latent Dirichlet allocation* (Blei et al., 2003, LDA), are one of the most popular statistical frameworks for navigating large unannotated document collections. Topic models discover—without any supervision—the primary themes presented in a dataset: the namesake topics.

Topic models have two primary applications: to aid human exploration of corpora (Chang et al., 2009) or serve as a low-dimensional representation for downstream applications. We focus on the second application, which has been fruitful for computer vision (Li Fei-Fei and Perona, 2005), computational biology (Perina et al., 2010), and information retrieval (Kataria et al., 2011).

In particular, we use topic models to aid *statistical machine translation* (Koehn, 2009, SMT). Modern machine translation systems use millions of examples of translations to learn translation rules. These systems work best when the training corpus has consistent genre, register, and topic. Systems that are robust to systematic variation in the training set are said to exhibit *domain adaptation*.

As we review in Section 2, topic models are a promising solution for automatically discovering domains in machine translation corpora. However, past work either relies solely on monolingual source-side models (Eidelman et al., 2012; Hasler et al., 2012; Su et al., 2012), or limited modeling of the target side (Xiao et al., 2012). In contrast, machine translation uses inherently multilingual data: an SMT system must translate a phrase or sentence from a *source* language to a different *target* language, so existing applications of topic models (Eidelman et al., 2012) are wilfully ignoring available information on the target side that could aid domain discovery.

This is not for a lack of multilingual topic models. Topic models bridge the chasm between languages using document connections (Mimno et al., 2009), dictionaries (Boyd-Graber and Resnik, 2010), and word alignments (Zhao and Xing, 2006). In Section 2, we review these models for discovering topics in multilingual datasets and discuss how they can improve SMT.

However, no models combine multiple bridges between languages. In Section 3, we create a model—the polylingual tree-based topic models (ptLDA)—that uses information from both external dictionaries and document alignments simultaneously. In Section 4, we derive both MCMC and variational inference for this new topic model.

In Section 5, we evaluate our model on the task of SMT using aligned datasets. We show that ptLDA offers better domain adaptation than other topic models for machine translation. Finally, in Section 6, we show how these topic models improve SMT with detailed examples.

2 Topic Models for Machine Translation

Before considering past approaches using topic models to improve SMT, we briefly review lexical weighting and domain adaptation for SMT.

---

† Yuening Hu and Ke Zhai made equal contributions.

2.1 Statistical Machine Translation

Statistical machine translation casts machine translation as a probabilistic process (Koehn, 2009). For a parallel corpus of aligned source and target sentences $(\mathcal{F}, \mathcal{E})$, a phrase $\bar{f} \in \mathcal{F}$ is translated to a phrase $\bar{e} \in \mathcal{E}$ according to a distribution $p_w(\bar{e}|\bar{f})$. One popular method to estimate the probability $p_w(\bar{e}|\bar{f})$ is via lexical weighting features.

Lexical Weighting In phrase-based SMT, lexical weighting features estimate the phrase pair quality by combining lexical translation probabilities of words in a phrase (Koehn et al., 2003). Lexical conditional probabilities $p(e|f)$ are maximum likelihood estimates from relative lexical frequencies $c(f,e)/\sum_e c(f,e)$, where $c(f,e)$ is the count of observing lexical pair (f,e) in the training dataset. The phrase pair probabilities $p_w(\bar{e}|\bar{f})$ are the normalized product of lexical probabilities of the aligned word pairs within that phrase pair (Koehn et al., 2003). In Section 2.2, we create topic-specific lexical weighting features.

Cross-Domain SMT A SMT system is usually trained on documents with the same genre (e.g., sports, business) from a similar style (e.g., newswire, blog-posts). These are called *domains*. Translations within one domain are better than translations across domains since they vary dramatically in their word choices and style. A correct translation in one domain may be inappropriate in another domain. For example, "潜水" in a newspaper usually means "underwater diving". On social media, it means a non-contributing "lurker".

Domain Adaptation for SMT Training a SMT system using diverse data requires *domain adaptation*. Early efforts focus on building separate models (Foster and Kuhn, 2007) and adding features (Matsoukas et al., 2009) to model domain information. Chiang et al. (2011) combine these approaches by directly optimizing genre and collection features by computing separate translation tables for each domain.

However, these approaches treat domains as hand-labeled, constant, and known *a priori*. This setup is at best expensive and at worst infeasible for large data. Topic models provide a solution where domains can be automatically induced from raw data: treat each topic as a domain.[1]

2.2 Inducing Domains with Topic Models

Topic models take the number of topics K and a collection of documents as input, where each document is a bag of words. They output two distributions: a distribution over topics for each document d; and a distribution over words for each topic. If each topic defines a SMT domain, the document's topic distribution is a soft domain assignment for that document.

Given the soft domain assignments, Eidelman et al. (2012) extract lexical weighting features conditioned on the topics, optimizing feature weights using the *Margin Infused Relaxed Algorithm* (Crammer et al., 2006, MIRA). The topics come from source documents *only* and create topic-specific lexical weights from the per-document topic distribution $p(k \mid d)$. The lexical probability conditioned on the topic is expected count $e_k(e,f)$ of a word translation pair under topic k,

$$\hat{c}_k(e,f) = \textstyle\sum_d p(k|d) c_d(e,f), \tag{1}$$

where $c_d(\bullet)$ is the number of occurrences of the word pair in document d. The lexical probability conditioned on topic k is the unsmoothed probability estimate of those expected counts

$$p_w(e|f;k) = \frac{\hat{c}_k(e,f)}{\sum_e \hat{c}_k(e,f)}, \tag{2}$$

from which we can compute the phrase pair probabilities $p_w(\bar{e}|\bar{f};k)$ by multiplying the lexical probabilities and normalizing as in Koehn et al. (2003).

For a test document d, the document topic distribution $p(k \mid d)$ is inferred based on the topics learned from training data. The feature value of a phrase pair $(\bar{e}, \bar{f})$ is

$$f_k(\bar{e}|\bar{f}) = -\log\left\{p_w(\bar{e}|\bar{f};k) \cdot p(k|d)\right\}, \tag{3}$$

a combination of the topic dependent lexical weight and the topic distribution of the document, from which we extract the phrase. Eidelman et al. (2012) compute the resulting model score by combining these features in a linear model with other standard SMT features and optimizing the weights.

Conceptually, this approach is just reweighting examples. The probability of a topic given a document is never zero. Every translation observed in the training set will contribute to $p_k(e|f)$; many of the expected counts, however, will be less than one. This obviates the explicit smoothing used in other domain adaptation systems (Chiang et al., 2011).

---

[1] Henceforth we will use the term "topic" and "domain" interchangeably: "topic" to refer to the concept in topic models and "domain" to refer to SMT corpora.

We adopt this framework in its entirety. Our contribution are topics that capture *multilingual* information and thus better capture the domains in the parallel corpus.

2.3 Beyond Vanilla Topic Models

Eidelman et al. (2012) ignore a wealth of information that could improve topic models and help machine translation. Namely, they only use monolingual data from the source language, ignoring all target-language data and available lexical semantic resources between source and target languages.

Different complement each other to reduce ambiguity. For example, "木马" in a Chinese document can be either "hobbyhorse" in a children's topic, or "Trojan virus" in a technology topic. A short Chinese context obscures the true topic. However, these terms are unambiguous in English, revealing the true topic.

While vanilla topic models (LDA) can only be applied to monolingual data, there are a number of topic models for parallel corpora: Zhao and Xing (2006) assume aligned word pairs share same topics; Mimno et al. (2009) connect different languages through comparable documents. These models take advantage of word or document *alignment information* and infer more robust topics from the aligned dataset.

On the other hand, *lexical information* can induce topics from multilingual corpora. For instance, orthographic similarity connects words with the same meaning in related languages (Boyd-Graber and Blei, 2009), and dictionaries are a more general source of information on which words share meaning (Boyd-Graber and Resnik, 2010).

These two approaches are not mutually exclusive, however; they reveal different connections across languages. In the next section, we combine these two approaches into a polylingual tree-based topic model.

3 Polylingual Tree-based Topic Models

In this section, we bring existing tree-based topic models (Boyd-Graber et al., 2007, tLDA) and polylingual topic models (Mimno et al., 2009, pLDA) together and create the polylingual tree-based topic model (ptLDA) that incorporates both word-level correlations and document-level alignment information.

Word-level Correlations Tree-based topic models incorporate the correlations between words by encouraging words that appear together in a **concept** to have similar probabilities given a topic. These concepts can come from WordNet (Boyd-Graber and Resnik, 2010), domain experts (Andrzejewski et al., 2009), or user constrains (Hu et al., 2013). When we gather concepts from bilingual resources, these concepts can connect different languages. For example, if a bilingual dictionary defines "电脑" as "computer", we combine these words in a concept.

We organize the vocabulary in a tree structure based on these concepts (Figure 1): words in the same concept share a common parent node, and then that concept becomes one of many children of the root node. Words that are not in any concept—**uncorrelated words**—are directly connected to the root node. We call this structure the **tree prior**.

When this tree serves as a prior for topic models, words in the same concept are correlated in topics. For example, if "电脑" has high probability in a topic, so will "computer", since they share the same parent node. With the tree priors, each topic is no longer a distribution over word types, instead, it is a distribution over paths, and each path is associated with a word type. The same word could appear in multiple paths, and each path represents a unique sense of this word.

Document-level Alignments Lexical resources connect languages and help guide the topics. However, these resources are sometimes brittle and may not cover the whole vocabulary. Aligned document pairs provide a more corpus-specific, flexible association across languages.

Polylingual topic models (Mimno et al., 2009) assume that the aligned documents in different languages share the same topic distribution and each language has a unique topic distribution over its word types. This level of connection between languages is flexible: instead of requiring the exact matching on words and sentences, only a coarse document alignment is necessary, as long as the documents discuss the same topics.

Combine Words and Documents We propose polylingual tree-based topic models (ptLDA), which connect information across different languages by incorporating both word correlation (as in tLDA) and document alignment information (as in pLDA). We initially assume a given tree structure, deferring the tree's provenance to the end of this section.

Generative Process As in LDA, each word token is associated with a topic. However, tree-based topic models introduce an additional step of selecting a concept in a topic responsible for generating each word token. This is represented by a path $y_{d,n}$ through the topic's tree.

The probability of a path in a topic depends on the transition probabilities in a topic. Each concept i in topic k has a distribution over its children nodes is governed by a Dirichlet prior: $\pi_{k,i} \sim \text{Dir}(\beta_i)$. Each path ends in a word (i.e., a leaf node) and the probability of a path is the product of all of the transitions between topics it traverses. Topics have correlations over words because the Dirichlet parameters can encode positive or negative correlations (Andrzejewski et al., 2009).

With these correlated in topics in hand, the generation of documents are very similar to LDA. For every document d, we first sample a distribution over topics θ_d from a Dirichlet prior $\text{Dir}(\alpha)$. For every token in the documents, we first sample a topic z_{dn} from the multinomial distribution θ_d, and then sample a path y_{dn} along the tree according to the transition distributions specified by topic z_{dn}. Because every path y_{dn} leads to a word w_{dn} in language l_{dn}, we append the sampled word w_{dn} to document $d_{l_{dn}}$. Aligned documents have words in both languages; monolingual documents only have words in a single language.

The full generative process is:

```
for topic $k \in 1, \cdots, K$ do
    for each internal node $n_i$ do
        draw a distribution $\pi_{ki} \sim \text{Dir}(\beta_i)$
for document set $d \in 1, \cdots, D$ do
    draw a distribution $\theta_d \sim \text{Dir}(\alpha)$
    for each word in documents $d$ do
        choose a topic $z_{dn} \sim \text{Mult}(\theta_d)$
        sample a path $y_{dn}$ with probability $\prod_{(i,j) \in y_{dn}} \pi_{z_{dn},i,j}$
        $y_{dn}$ leads to word $w_{dn}$ in language $l_{dn}$
        append token $w_{dn}$ to document $d_{l_{dn}}$
```

If we use a flat symmetric Dirichlet prior instead of the tree prior, we recover pLDA; and if all documents are monolingual (i.e., with distinct distributions over topics θ), we recover tLDA. ptLDA connects different languages on both the word level (using the word correlations) and the document level (using the document alignments). We compare these models' machine translation performance in Section 5.

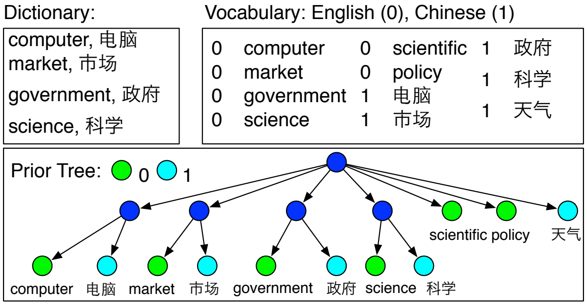


Figure 1: An example of constructing a prior tree from a bilingual dictionary: word pairs with the same meaning but in different languages are concepts; we create a common parent node to group words in a concept, and then connect to the root; uncorrelated words are connected to the root directly. Each topic uses this tree structure as a prior.

Build Prior Tree Structures One remaining question is the source of the word-level connections across languages for the tree prior. We consider two resources to build trees that correlate words across languages. The first are a multilingual dictionaries (*dict*), which match words with the same meaning in different languages together. These relations between words are used as the concepts in the prior tree (Figure 1).

In addition, we extract the word alignments from aligned sentences in a parallel corpus. The word pairs define concepts for the prior tree (*align*). We use both resources for our models (denoted as ptLDA-*dict* and ptLDA-*align*) in our experiments (Section 5) and show that they yield comparable performance in SMT.

4 Inference

Inference of probabilistic models discovers the posterior distribution over latent variables. For a collection of D documents, each of which contains N_d number of words, the latent variables of ptLDA are: transition distributions $\boldsymbol{\pi}_{ki}$ for every topic k and internal node i in the prior tree structure; multinomial distributions over topics $\boldsymbol{\theta}_d$ for every document d; topic assignments z_{dn} and path y_{dn} for the n^{th} word w_{dn} in document d. The joint distribution of polylingual tree-based topic models is

$$\begin{aligned} p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi}; \alpha, \beta) = & \textstyle\prod_k \prod_i p(\boldsymbol{\pi}_{ki} | \beta_i) \\ & \cdot \textstyle\prod_d p(\boldsymbol{\theta}_d | \alpha) \cdot \prod_d \prod_n p(z_{dn} | \boldsymbol{\theta}_d) \\ & \cdot \textstyle\prod_d \prod_n \left(p(y_{dn} | z_{dn}, \boldsymbol{\pi}) p(w_{dn} | y_{dn}) \right). \end{aligned} \tag{4}$$

Exact inference is intractable, so we turn to approximate posterior inference to discover the latent variables that best explain our data. Two widely used approximation approaches are *Markov chain Monte Carlo* (Neal, 2000, MCMC) and *variational Bayesian inference* (Blei et al., 2003, VB). Both frameworks produce good approximations of the posterior mode (Asuncion et al., 2009). In addition, Mimno et al. (2012) propose hybrid inference that takes advantage of parallelizable variational inference for global variables (Wolfe et al., 2008) while enjoying the sparse, efficient updates for local variables (Neal, 1993). In the rest of this section, we discuss all three methods in turn.

We explore multiple inference schemes because while all of these methods optimize likelihood because they might give different results on the translation task.

4.1 Markov Chain Monte Carlo Inference

We use a collapsed Gibbs sampler for tree-based topic models to sample the path y_{dn} and topic assignment z_{dn} for word w_{dn},

$$\begin{aligned} & p(z_{dn} = k, y_{dn} = s | \neg \boldsymbol{z}_{dn}, \neg \boldsymbol{y}_{dn}, \boldsymbol{w}; \alpha, \boldsymbol{\beta}) \\ & \quad \propto \mathbb{I}\left[\Omega(s) = w_{dn}\right] \cdot \frac{N_{k|d} + \alpha}{\sum_{k'} (N_{k'|d} + \alpha)} \\ & \quad \cdot \prod_{i \to j \in s} \frac{N_{i \to j|k} + \beta_{i \to j}}{\sum_{j'} (N_{i \to j'|k} + \beta_{i \to j'})}, \end{aligned}$$

where $\Omega(s)$ represents the word that path s leads to, $N_{k|d}$ is the number of tokens assigned to topic k in document d and $N_{i \to j|k}$ is the number of times edge $i \to j$ in the tree assigned to topic k, excluding the topic assignment z_{dn} and its path y_{dn} of current token w_{dn}. In practice, we sample the latent variables using efficient sparse updates (Yao et al., 2009; Hu and Boyd-Graber, 2012).

4.2 Variational Bayesian Inference

Variational Bayesian inference approximates the posterior distribution with a simplified *variational distribution* q over the latent variables: document topic proportions θ, transition probabilities π, topic assignments z, and path assignments y.

Variational distributions typically assume a mean-field distribution over these latent variables, removing all dependencies between the latent variables. We follow this assumption for the transition probabilities $q(\boldsymbol{\pi} \,|\, \boldsymbol{\lambda})$ and the document topic proportions $q(\boldsymbol{\theta} \,|\, \boldsymbol{\gamma})$; both are variational Dirichlet distributions. However, due to the tight coupling between the path and topic variables, we must model this joint distribution as one multinomial, $q(\boldsymbol{z}, \boldsymbol{y} \,|\, \boldsymbol{\phi})$. If word token w_{dn} has K topics and S paths, it has a $K * S$ length variational multinomial ϕ_{dnks}, which represents the probability that the word takes path s in topic k. The complete variational distribution is

$$\begin{aligned} q(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{y} | \gamma, \boldsymbol{\lambda}, \phi) = \textstyle\prod_d q(\boldsymbol{\theta}_d | \boldsymbol{\gamma}_d) \cdot \\ \textstyle\prod_k \prod_i q(\boldsymbol{\pi}_{ki} | \lambda_{ki}) \cdot \prod_d \prod_n q(z_{dn}, y_{dn} | \boldsymbol{\phi}_{dn}). \end{aligned} \tag{5}$$

Our goal is to find the variational distribution q that is closest to the true posterior, as measured by the *Kullback-Leibler* (KL) divergence between the true posterior p and variational distribution q. This induces an "evidence lower bound" (ELBO, $\mathcal{L}$) as a function of a variational distribution q: $\mathcal{L} =$

$$\begin{aligned} & \mathbb{E}_q[\log p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi})] - \mathbb{E}_q[\log q(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{y})] \\ & = \textstyle\sum_k \sum_i \mathbb{E}_q[\log p(\boldsymbol{\pi}_{ki} | \beta_i)] \\ & + \textstyle\sum_d \mathbb{E}_q[\log p(\boldsymbol{\theta}_d | \alpha)] \\ & + \textstyle\sum_d \sum_n \mathbb{E}_q[\log p(z_{dn}, y_{dn} | \boldsymbol{\theta}_d, \boldsymbol{\pi}) p(w_{dn} | y_{dn})] \\ & + \mathbb{H}[q(\boldsymbol{\theta})] + \mathbb{H}[q(\boldsymbol{\pi})] + \mathbb{H}[q(\boldsymbol{z}, \boldsymbol{y})], \end{aligned} \tag{6}$$

where $\mathbb{H}[\bullet]$ represents the entropy of a distribution. Optimizing $\mathcal{L}$ using coordinate descent provides the following updates:

$$\begin{aligned} \phi_{dnkt} \propto \exp\{\Psi(\gamma_{dk}) - \Psi(\textstyle\sum_k \gamma_{dk}) \\ + \textstyle\sum_{i \to j \in s} \left(\Psi(\lambda_{k,i \to j}) - \Psi(\sum_{j'} \lambda_{k,i \to j'})\right)\}; \end{aligned} \tag{7}$$

$$\gamma_{dk} = \alpha_k + \textstyle\sum_n \sum_{s \in \Omega^{-1}(w_{dn})} \phi_{dnkt}; \tag{8}$$

$$\begin{aligned} \lambda_{k,i \to j} = \beta_{i \to j} \\ + \textstyle\sum_d \sum_n \sum_{s \in \Omega'(w_{dn})} \phi_{dnkt} \mathbb{I}\left[i \to j \in s\right]; \end{aligned} \tag{9}$$

where $\Omega'(w_{dn})$ is the set of all paths that lead to word w_{dn} in the tree, and t represents one particular path in this set. $\mathbb{I}\left[i \to j \in s\right]$ is the indicator of whether path s contains an edge from node i to j.

4.3 Hybrid Stochastic Inference

Given the complementary strengths of MCMC and VB, and following hybrid inference proposed by Mimno et al. (2012), we also derive hybrid inference for ptLDA.

The transition distributions $\boldsymbol{\pi}$ are treated identically as in variational inference. We posit a variational Dirichlet distribution $\boldsymbol{\lambda}$ and choose the one that minimizes the KL divergence between the true posterior and the variational distribution.

For topic $\boldsymbol{z}$ and path $\boldsymbol{y}$, instead of variational updates, we use a Gibbs sampler within a document. We sample z_{dn} and y_{dn} conditioned on the topic and path assignments of all other document tokens, based on the variational expectation of $\boldsymbol{\pi}$,

$$\begin{aligned} q(z_{dn} = k, y_{dn} = s|\neg \boldsymbol{z}_{dn}, \neg \boldsymbol{y}_{dn}; \boldsymbol{w}) \propto \\ (\alpha + \textstyle\sum_{m \neq n} \mathbb{I}\left[z_{dm} = k\right]) \\ \cdot \exp\{\mathbb{E}_q[\log p(y_{dn}|z_{dn}, \boldsymbol{\pi}) p(w_{dn}|y_{dn})]\}. \end{aligned} \tag{10}$$

This equation embodies how this is a hybrid algorithm: the first term resembles the Gibbs sampling term encoding how much a document prefers a topic, while the second term encodes the expectation under the variational distribution of how much a path is preferred by this topic,

$$\begin{aligned} \mathbb{E}_q[\log p(y_{dn}|z_{dn}, \boldsymbol{\pi}) p(w_{dn}|y_{dn})] = \mathbb{I}_{[\Omega(y_{dn}) = w_{dn}]} \\ \cdot \textstyle\sum_{i \to j \in y_{dn}} \mathbb{E}_q[\log \lambda_{z_{dn}, i \to j}]. \end{aligned}$$

For every document, we sweep over all its tokens and resample their topic z_{dn} and path y_{dn} conditioned on all the other tokens' topic and path assignments $\neg \boldsymbol{z}_{dn}$ and $\neg \boldsymbol{y}_{dn}$. To avoid bias, we discard the first B burn-in sweeps and take the following M samples. We then use the empirical average of these samples update the global variational parameter $q(\boldsymbol{\pi}|\boldsymbol{\lambda})$ based on how many times we sampled these paths

$$\begin{aligned} \lambda_{k, i \to j} = \tfrac{1}{M} \textstyle\sum_d \sum_n \sum_{s \in \Omega^{-1}(w_{dn})} \big(\mathbb{I}\left[i \to j \in s\right] \\ \cdot \mathbb{I}\left[z_{dn} = k, y_{dn} = s\right]\big) + \beta_{i \to j}. \end{aligned} \tag{11}$$

For our experiments, we use the recommended settings $B = 5$ and $M = 5$ from Mimno et al. (2012).

5 Experiments

We evaluate our new topic model, ptLDA, and existing topic models—LDA, pLDA, and tLDA—on their ability to induce domains for machine translation and the resulting performance of the translations on standard machine translation metrics.

Dataset and SMT Pipeline We use the NIST MT Chinese-English parallel corpus (NIST), excluding non-UN and non-HK Hansards portions as our training dataset. It contains 1.6M sentence pairs, with 40.4M Chinese tokens and 44.4M English tokens. We replicate the SMT pipeline of Eidelman et al. (2012): word segmentation (Tseng et al., 2005), align (Och and Ney, 2003), and symmetrize (Koehn et al., 2003) the data. We train a modified Kneser-Ney trigram language model on English (Chen and Goodman, 1996). We use CDEC (Dyer et al., 2010) for decoding, and MIRA (Crammer et al., 2006) for parameter training. To optimize SMT system, we tune the parameters on NIST MT06, and report results on three test sets: MT02, MT03 and MT05.[2]

Topic Models Configuration We compare our polylingual tree-based topic model (ptLDA) against tree-based topic models (tLDA), polylingual topic models (pLDA) and vanilla topic models (LDA).[3] We also examine different inference algorithms—Gibbs sampling (**gibbs**), variational inference (**variational**) and hybrid approach (**variational-hybrid**)—on the effects of SMT performance. In all experiments, we set the per-document Dirichlet parameter $\alpha = 0.01$ and the number of topics to 10, as used in Eidelman et al. (2012).

Resources for Prior Tree To build the tree for tLDA and ptLDA, we extract the word correlations from a Chinese-English bilingual dictionary (Denisowski, 1997).[4] We filter the dictionary using the NIST vocabulary, and keep entries mapping single Chinese and single English words. The prior tree has about 1000 word pairs (*dict*).

We also extract the bidirectional word alignments between Chinese and English using GIZA++ (Och and Ney, 2003). We then remove the word pairs appearing more than 50K times or fewer than 500 times and construct a second prior tree with about 2500 word pairs (*align*).

We apply both trees to tLDA and ptLDA, denoted as tLDA-*dict*, tLDA-*align*, ptLDA-*dict*, and ptLDA-*align*. However, tLDA-*align* and ptLDA-*align* do worse than tLDA-*dict* and ptLDA-*dict*, so we omit tLDA-*align* in the results.

Domain Adaptation using Topic Models We examine the effectiveness of using topic models for domain adaptation on standard SMT evaluation metrics—BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). We report the results on three different test sets (Figure 2), and all SMT results are averaged over five runs.

We refer to the SMT model without domain adaptation as **baseline**.[5] LDA marginally improves machine translation (less than half a BLEU point).

---

[2] The NIST datasets contain 878, 919, 1082 and 1664 sentences for MT02, MT03, MT05 and MT06 respectively.

[3] For Gibbs sampling, we use implementations available in Hu and Boyd-Graber (2012) for tLDA; and Mallet (McCallum, 2002) for LDA and pLDA.

[4] This is a two-level tree structure. However, one could build a more sophisticated tree prior with a hierarchical dictionary such as multilingual WordNet.

[5] Our replication of Eidelman et al. (2012) yields slightly higher baseline performance, but the trend is consistent.

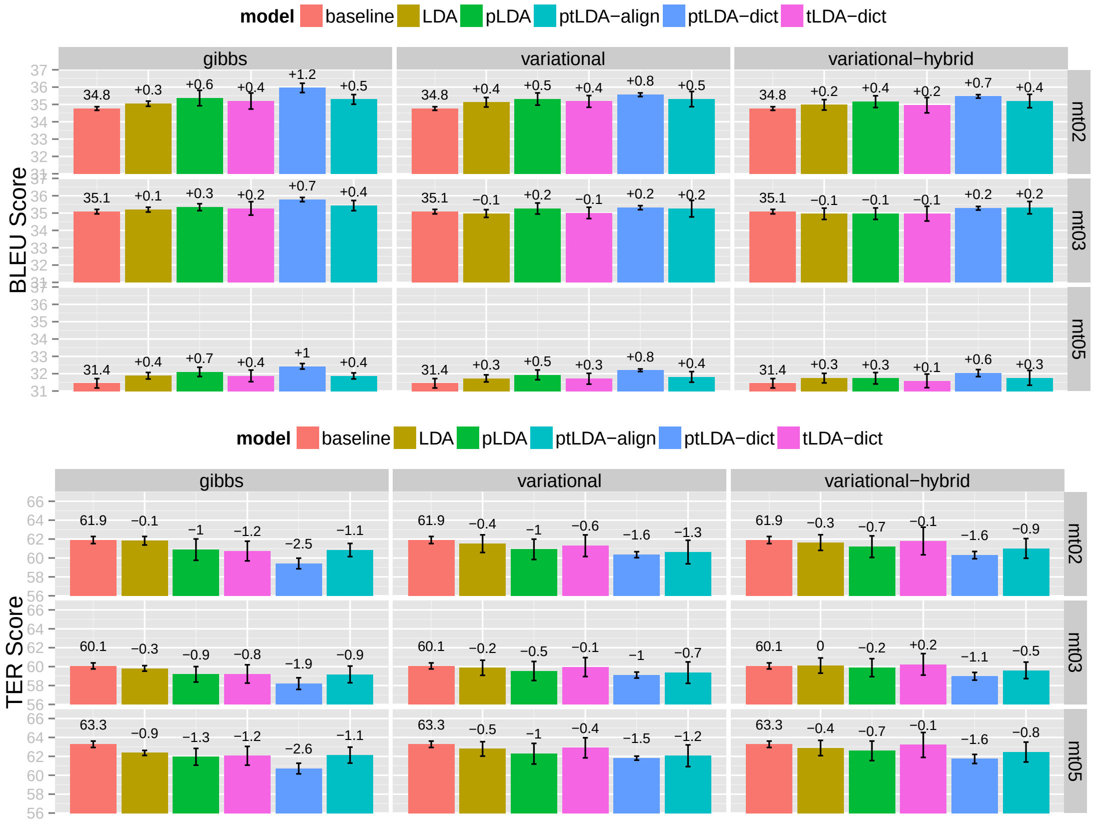


Figure 2: Machine translation performance for different models and inference algorithms against the baseline, on BLEU (top, higher the better) and TER (bottom, lower the better) scores. Our proposed ptLDA performs best. Results are averaged over 5 random runs. For model ptLDA-*dict* with different inference schemes, the BLEU improvement on three test sets is mostly significant with $p = 0.01$, except the results on MT03 using variational and variational-hybrid inferences.

Polylingual topic models pLDA and tree-based topic models tLDA-*dict* are consistently better than LDA, suggesting that incorporating additional bilingual knowledge improves topic models. These improvements are not redundant: our new ptLDA-*dict* model, which has aspects of both models yields the best performance among these approaches—up to a 1.2 BLEU point gain (higher is better), and -2.6 TER improvement (lower is better). The BLEU improvement is significant (Koehn, 2004) at $p = 0.01$,[6] except on MT03 with variational and variational-hybrid inference.

While ptLDA-*align* performs better than **baseline** SMT and LDA, it is worse than ptLDA-*dict*, possibly because of errors in the word alignments, making the tree priors less effective.

Scalability While **gibbs** has better translation scores than **variational** and **variational-hybrid**, it is less scalable to larger datasets. With 1.6M NIST training sentences, **gibbs** takes nearly a week to run 1000 iterations. In contrast, the parallelized **variational** and **variational-hybrid** approaches, which we implement in MapReduce (Dean and Ghemawat, 2004; Wolfe et al., 2008; Zhai et al., 2012), take less than a day to converge.

6 Discussion

In this section, we qualitatively analyze the translation results and investigate how ptLDA and its cousins improve SMT. We also discuss other approaches to improve unsupervised domain adaptation for SMT.

6.1 How do Topic Models Help SMT?

We present two examples of how topic models can improve SMT. The first example shows both LDA and ptLDA improve the **baseline**. The second example shows how LDA introduce biases that mislead SMT and how ptLDA's bilingual constraints correct these mistakes.

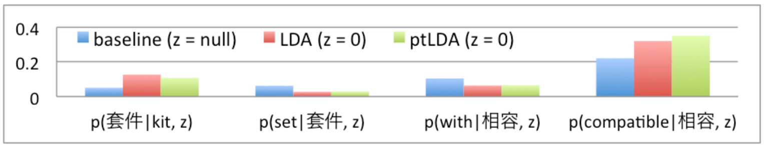
Figure 3 shows a sentence about a company

---

[6] Because we have multiple runs of each topic model (and thus different translation models), we select the run closest to the average BLEU for the translation significance test.

source	新力已在北美地区售出大约五十七万套窄频网路连结套件，每套售价约三十九美元, 相容游戏约有二十种。
reference	sony has already sold about 570,000 units of narrowband connection **kits** in north america at the price of about 39 us dollars and some 20 **compatible** games .

source	… 网路连结套件 ...	… 相容游戏约有二十种 。
reference	… connection **kits** ...	… some 20 **compatible** games .
baseline	… internet links **set** ...	… **with** about 20 of the game .
LDA	… internet links **kit** …	… , there are about 20 **compatible** games .
ptLDA	… internet links **kit** …	… , there are about 20 **compatible** games .



LDA-Topic 0 (business)

公司(company), 中国(China), 服务(service), 市场(market), 技术(technology), 企业(industry), 提供(provide), 发展(develop), 年(year), 产品(product), 上, 合作(coorporate), 中, 管理(manage), 投资(invest), 经济(economy), 国际(international), 系统(system), 银行(bank)

ptLDA-Topic 0 (business)

公司(company), 服务(service), 市场(market), 技术(technology), china, 企业(industry), 产品(product), market, company, technology, services, 系统(system), year, industry, products, business, 经济(economy), information, 管理(manage), 投资(invest), percent, 网络(internet), companies, world, system, 信息(information), 增长(increase), 设备(device), service, 业务(service)

Figure 3: Better SMT result using topic models for domain adaptation. Top row: the source sentence and its reference translation. Middle row: the highlighted translations from different approaches. Bottom row: the change of relevant translation probabilities after incorporating the domain knowledge from LDA and ptLDA. Right: most-probable words of the topic the source sentence is assigned to under LDA (top) and ptLDA (bottom). The Chinese translations are in parenthesis.

introducing new technology gadgets where both LDA and ptLDA improve translations. The **baseline** translates “套件” to “set” (red), and “相容” to “with” (blue), which do not capture the reference meaning of a *add-on device* that works with *compatible* games. Both LDA and ptLDA assign this sentence to a business domain, which makes the translations probabilities shift toward correct translations: the probability of translating “相容” to “compatible” and the probability of translating “套件” to “kit” in the business domain are both significantly larger than without the domain knowledge; and the probabilities of translating “相容” to “with” and the probability of translating “set” to “套件” in the business domain decrease.

The second example (Figure 4) illustrates how ptLDA offers further improvements over LDA. The source sentence discusses foreign affairs. The **baseline** correctly translates the word “影响” to “affect”. However, LDA—which only takes monolingual information from the source language—assigns this sentence to economic development. This misleads SMT to lower the probability for the correct translation “affect”; it chooses “impact” instead. In contrast, ptLDA—which incorporates bilingual constraints—successfully labels this sentence as foreign affairs and produces a softer, more nuanced translation that better matches the reference. The translation of “承诺” is very similar, except in this case, both the **baseline** and LDA produce the incorrect translation “the commitment of”. This is possible because the probabilities of translating “承诺” to “promised to” and translating “promised to” to “承诺” (the correct translation, in both directions) increase when conditioned on ptLDA’s correct topic but decrease when conditioned on LDA’s incorrect topic.

6.2 Other Approaches

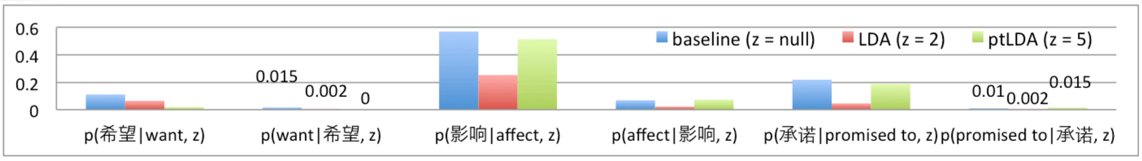
Other approaches have used topic models for machine translation. Xiao et al. (2012) present a topic similarity model based on LDA that produces a feature that weights grammar rules based on topic compatibility. They also model the source and target side of rules and compare the target similarity during decoding by projecting the target distribution into the source space. Hasler et al. (2012) use the source-side topic assignments from *hidden topic Markov models* (Gruber et al., 2007, HTMM) which models documents as a Markov chain and assign one topic to the whole sentence, instead of a mixture of topics. Su et al. (2012) also apply HTMM to monolingual data and apply the results to machine translation. To our knowledge, however, this is the first work to use *multilingual* topic models for domain adaptation in machine translation.

6.3 Improving Language Models

Topic models capture document-level properties of language, but a critical component of machine translation systems is the language model, which provides local constraints and preferences. Domain adaptation for language models (Bellegarda, 2004; Wood and Teh, 2009) is an important avenue for improving machine translation. Models that simultaneously discover global document themes as well as local, contextual domain-specific informa-

source	消息指出，韩国使馆人员向中方官员表示，韩国方面并没有支持朝鲜人以这种方法前往韩国，韩国并不希望这类事件再次发生，以免对中国和朝鲜半岛双方间的关系带来影响，韩国方面并向中国方面承诺，愿意协助中国管理好在京的韩国居民
reference	sources said rok embassy personnel told chinese officials that rok has not backed any dpr koreans to get to rok in such a manner and rok would not **like** **such** things happen again to **affect** relationship between china and the two sides of the korean peninsula . rok also **promised to** assist china in the administration of koreans in beijing .

source	… 不希望 ...	… 以免对…关系带来影响…	… 韩国方面并向中国方面承诺…
reference	… would not **like** ...	… to **affect** the relationship …	… rok also **promised to** the chinese side ...
baseline	… does not **want** ...	… so as not to **affect** the relations…	… south korea and **the commitment of** the chinese side ...
LDA	… does not **hope that** ...	… so as to avoid **impact** the relations…	… the rok side , and **the commitment of** the chinese side ...
ptLDA	… does not **hope that** ...	… so as not to **affect** the relations…	… south korea has **promised to** the chinese side ...



LDA-Topic 5 (economic development)	ptLDA-Topic 2 (foreign affairs)
发展(develop), 国(country), 两(two), 中国(China), 关系(relation), 中, 合作(cooperate), 经济(economy), 人民(people), 友好(friendly), 国家(country), 新(new), 问题(problem), 上, 加强(emphasize), 重要(important), 和平(peace), 共同(together), 建设(build), 世界(world)	china, 问题(issue), military, united, president, 国家(country), 地区(area), minister, 伊拉克(Iraq), 和平(peace), nuclear, people, 总统(president), peace, security, 联合国(UN), 军事(military), 以色列(Israel), iraq, foreign, international, 部队(army), beijing, world, defense, south, 安 全(security), war, 协议(agreement), 会议(conference)

Figure 4: Better SMT result using ptLDA compared to LDA and the baseline. Top row: the source sentence and a reference translation. Second row: the highlighted translations from different models. Third row: the change of relevant translation probabilities after incorporating domain knowledge from LDA and ptLDA. Bottom row: most-probable words for the topics the source sentence is assigned to under LDA (left) and ptLDA (right). The meanings of Chinese words are in parenthesis.

tion (Wallach, 2006; Boyd-Graber and Blei, 2008) may offer further improvements.

6.4 External Data

The topic models presented here only require weak alignment between documents at the document level. Extending to larger datasets for learning topics is straightforward in principle. For example, ptLDA could learn domains from a much larger corpus like Wikipedia and then apply the extracted domains to machine translation data. However, this presents further challenges, as Wikipedia's domains are not representative of newswire machine translation datasets; a flexible hierarchical topic model (Teh et al., 2006) would better distinguish useful domains from extraneous ones.

7 Conclusion

Topic models generate great interest, but their use in "real world" applications still lags; this is particularly true for multilingual topic models. As topic models become more integrated in commonplace applications, their adoption, understanding, and robustness will improve.

This paper contributes to the deeper integration of topic models into critical applications by presenting a new multilingual topic model, ptLDA, comparing it with other multilingual topic models on a machine translation task, and showing that these topic models improve machine translation. ptLDA models both source and target data to induce domains from both dictionaries and alignments. Further improvement is possible by incorporating topic models deeper in the decoding process and adding domain knowledge to the language model.

Acknowledgments

We would like to thank the anonymous reviewers, Doug Oard, and John Morgan for their helpful comments, and thank Junhui Li and Ke Wu for insightful discussions. This work was supported by NSF Grant IIS-1320538. Boyd-Graber is also supported by NSF Grant CCF-1018625. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the International Conference of Machine Learning*.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of Uncertainty in Artificial Intelligence*.

Jerome R. Bellegarda. 2004. Statistical language model adaptation: review and perspectives. volume 42, pages 93–108.

David M. Blei and John D. Lafferty. 2009. Visualizing topics with Multi-Word expressions. *arXiv*.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3.

Jordan Boyd-Graber and David M. Blei. 2008. Syntactic topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of Uncertainty in Artificial Intelligence*.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of Emperical Methods in Natural Language Processing*.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of Emperical Methods in Natural Language Processing*.

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

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Association for Computational Linguistics*.

David Chiang, Steve DeNeefe, and Michael Pust. 2011. Two easy improvements to lexical weighting. In *Proceedings of the Human Language Technology Conference*.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *Symposium on Operating System Design and Implementation*.

Paul Denisowski. 1997. CEDICT. http://www.mdbg.net/chindict/.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL System Demonstrations*.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the Association for Computational Linguistics*.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Amit Gruber, Michael Rosen-Zvi, and Yair Weiss. 2007. Hidden topic Markov models. In *Artificial Intelligence and Statistics*.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2012. Sparse lexicalised features and topic adaptation for SMT. In *Proceedings of IWSLT*.

Yuening Hu and Jordan Boyd-Graber. 2012. Efficient tree-based topic modeling. In *Proceedings of the Association for Computational Linguistics*.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2013. Interactive topic modeling. *Machine Learning Journal*.

Saurabh S. Kataria, Krishnan S. Kumar, Rajeev R. Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. 2011. Entity disambiguation with hierarchical topic models. In *Knowledge Discovery and Data Mining*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of Emperical Methods in Natural Language Processing*.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Li Fei-Fei and Pietro Perona. 2005. A Bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition*.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of Emperical Methods in Natural Language Processing*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of Emperical Methods in Natural Language Processing*.

David Mimno, Matthew Hoffman, and David Blei. 2012. Sparse stochastic inference for latent Dirichlet allocation. In *Proceedings of the International Conference of Machine Learning*.

Radford M. Neal. 1993. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto.

Radford M. Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29(21), pages 19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318.

Alessandro Perina, Pietro Lovato, Vittorio Murino, and Manuele Bicego. 2010. Biologically-aware latent Dirichlet allocation (balda) for the classification of expression microarray. In *Proceedings of the 5th IAPR international conference on Pattern recognition in bioinformatics*, PRIB'10.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of the Association for Computational Linguistics*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *SIGHAN Workshop on Chinese Language Processing*.

Hanna M. Wallach. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the International Conference of Machine Learning*.

Jason Wolfe, Aria Haghighi, and Dan Klein. 2008. Fully distributed EM for very large datasets. In *Proceedings of the International Conference of Machine Learning*, pages 1184–1191.

Frank Wood and Yee Whye Teh. 2009. A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 12.

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. 2012. A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the Association for Computational Linguistics*.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Knowledge Discovery and Data Mining*.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of World Wide Web Conference*.

Bing Zhao and Eric P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the Association for Computational Linguistics*.