Bilingual Segmented Topic Model

Akihiro Tamura and **Eiichiro Sumita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, JAPAN
{akihiro.tamura, eiichiro.sumita}@nict.go.jp

Abstract

This study proposes the bilingual segmented topic model (BiSTM), which hierarchically models documents by treating each document as a set of segments, e.g., sections. While previous bilingual topic models, such as bilingual latent Dirichlet allocation (BiLDA) (Mimno et al., 2009; Ni et al., 2009), consider only cross-lingual alignments between entire documents, the proposed model considers cross-lingual alignments between segments in addition to document-level alignments and assigns the same topic distribution to aligned segments. This study also presents a method for simultaneously inferring latent topics and segmentation boundaries, incorporating unsupervised topic segmentation (Du et al., 2013) into BiSTM. Experimental results show that the proposed model significantly outperforms BiLDA in terms of perplexity and demonstrates improved performance in translation pair extraction (up to +0.083 extraction accuracy).

1 Introduction

Probabilistic topic models, such as probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) and latent Dirichlet allocation (LDA) (Blei et al., 2003), are generative models for documents that have been used as unsupervised frameworks to discover latent topics in document collections without prior knowledge. These topic models were originally applied to monolingual data; however, various recent studies have proposed the use of probabilistic topic models in multilingual settings[1], where latent topics are shared across multiple languages. These models have improved several multilingual tasks, such as translation pair extraction and cross-lingual text classification (see the survey paper by Vulić et al. (2015) for details).

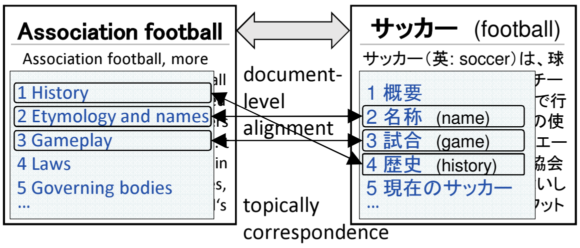

Figure 1: Wikipedia Article Example

Most multilingual topic models, including bilingual LDA (BiLDA) (Mimno et al., 2009; Ni et al., 2009), model a document-aligned comparable corpus, such as a collection of Wikipedia articles, where aligned documents are topically similar but are not direct translations[2]. In particular, these models assume that the documents in each tuple share the same topic distribution and that each cross-lingual topic has a language-specific word distribution.

Existing multilingual topic models consider only document-level alignments. However, most documents are hierarchically structured, i.e., a document comprises segments (e.g., sections and paragraphs) that can be aligned across languages. Figure 1 shows a Wikipedia article example, which contains a set of sections. Sections 1, 2, and 3 in the English article correspond topically to sections 4, 2, and 3 in the Japanese counterpart, re-

[1] In this work, we deal with a bilingual setting, but our approach can be extended straightforwardly to apply to more than two languages.

[2] In this study, we focus on models for a document-aligned comparable corpus. We describe other types of multilingual topic models and their limitations in Section 7.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1266–1276,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

spectively. To date, such segment-level alignments have been ignored; however, we consider that such corresponding segments must share the same topic distribution.

Du et al. (2010) have shown that segment-level topics and their dependencies can improve modeling accuracy in a monolingual setting. Based on that research, we expect that segment-level topics can also be useful for modeling multilingual data.

This study proposes a bilingual segmented topic model (BiSTM) that extends BiLDA to capture segment-level alignments through a hierarchical structure. In particular, BiSTM considers each document as a set of segments and models a document as a document-segment-word structure. The topic distribution of each segment (per-segment topic distribution) is generated using a Pitman–Yor process (PYP) (Pitman and Yor, 1997), in which the base measure is the topic distribution of the related document (per-document topic distribution). In addition, BiSTM introduces a binary variable that indicates whether two segments in different languages are aligned. If two segments are aligned, their per-segment topic distributions are shared; if they are not aligned, they are independently generated.

BiSTM leverages existing segments from a given segmentation. However, a segmentation is not always given, and a given segmentation might not be optimal for statistical modeling. Therefore, this study also presents a model, BiSTM+TS, that incorporates unsupervised topic segmentation into BiSTM. BiSTM+TS integrates point-wise boundary sampling into BiSTM in a manner similar to that proposed by Du et al. (2013) and infers segmentation boundaries and latent topics jointly.

Experiments using an English–Japanese and English–French Wikipedia corpus show that the proposed models (BiSTM and BiSTM+TS) significantly outperform the standard bilingual topic model (BiLDA) in terms of perplexity, and that they improve performance in translation extraction (up to +0.083 top 1 accuracy). The experiments also reveal that BiSTM+TS is comparable to BiSTM, which uses manually provided segmentation, i.e., section boundaries in Wikipedia articles.

2 Bilingual LDA

This section describes the BiLDA model (Mimno et al., 2009; Ni et al., 2009), which we take as our baseline. BiLDA is a bilingual extension of basic monolingual LDA (Blei et al., 2003) for a document-aligned comparable corpus. While monolingual LDA assumes that each document has its own topic distribution, BiLDA assumes that aligned documents share the same topic distribution and discovers latent cross-lingual topics.

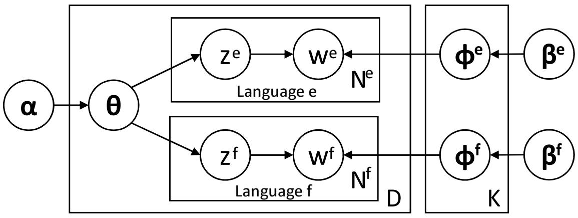

Figure 2: Graphical Model of BiLDA

Algorithm 1 Generative Process of BiLDA

```
 1: for each topic $k \in \{1, ..., K\}$ do
 2:   for each language $l \in \{e, f\}$ do
 3:     choose $\boldsymbol{\phi}_{\boldsymbol{k}}^{\boldsymbol{l}} \sim$ Dirichlet($\boldsymbol{\beta}^{\boldsymbol{l}}$)
 4:   end for
 5: end for
 6: for each document pair $d_i$ ($i \in \{1, ..., D\}$) do
 7:   choose $\boldsymbol{\theta}_{\boldsymbol{i}} \sim$ Dirichlet($\boldsymbol{\alpha}$)
 8:   for each language $l \in \{e, f\}$ do
 9:     for each word $w_{im}^l$ ($m \in \{1, ..., N_i^l\}$) do
10:       choose $z_{im}^l \sim$ Multinomial($\boldsymbol{\theta}_{\boldsymbol{i}}$)
11:       choose $w_{im}^l \sim p(w_{im}^l | z_{im}^l, \boldsymbol{\phi}^{\boldsymbol{l}})$
12:     end for
13:   end for
14: end for
```

Algorithm 1 and Figure 2 show the generative process and graphical model, respectively, of BiLDA. BiLDA models a document-aligned comparable corpus, i.e., a set of D document pairs in two languages, e and f. Each document pair d_i ($i \in \{1, ..., D\}$) comprises aligned documents in the language e and f: d_i=(d_i^e, d_i^f). BiLDA assumes that each topic $k \in \{1, ..., K\}$ comprises the set of a discrete distribution over words for each language. Each language-specific per-topic word distribution $\boldsymbol{\phi}_{\boldsymbol{k}}^{\boldsymbol{l}}$ ($l \in \{e, f\}$) is drawn from a Dirichlet distribution with the prior $\boldsymbol{\beta}^{\boldsymbol{l}}$ (Steps 1-5). To generate a document pair d_i, the per-document topic distribution $\boldsymbol{\theta}_{\boldsymbol{i}}$ is first drawn from a Dirichlet distribution with the prior $\boldsymbol{\alpha}$ (Step 7). Thus, aligned documents d_i^e and d_i^f share the same topic distribution. Then, for each word at $m \in \{1, ..., N_i^l\}$ in document d_i^l in language l, a latent topic assignment z_{im}^l is drawn from a multinomial

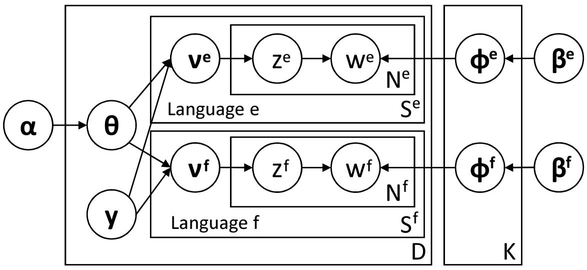

Figure 3: Graphical Model of BiSTM

Algorithm 2 Generative Process of BiSTM

```
 1: for each topic $k \in \{1, ..., K\}$ do
 2:   for each language $l \in \{e, f\}$ do
 3:     choose $\boldsymbol{\phi^l_k} \sim$ Dirichlet($\boldsymbol{\beta^l}$)
 4:   end for
 5: end for
 6: for each document pair $d_i$ $(i \in \{1, ..., D\})$ do
 7:   choose $\boldsymbol{\theta_i} \sim$ Dirichlet($\boldsymbol{\alpha}$)
 8:   if $\boldsymbol{y_i}$ are not given then
 9:     choose $\gamma_i \sim$ Beta($\eta_0, \eta_1$)
10:     choose $\boldsymbol{y_i} \sim$ Bernoulli($\gamma_i$)
11:   end if
12:   generate aligned segment sets $\boldsymbol{AS_i}$ = genAS($\boldsymbol{y_i}$)
13:   for each set $\boldsymbol{AS_{ig}}$ $(g \in \{1, ..., |\boldsymbol{AS_i}|\})$ do
14:     choose $\boldsymbol{\nu_{ig}} \sim$ PYP($a, b, \boldsymbol{\theta_i}$)
15:   end for
16:   for each language $l \in \{e, f\}$ do
17:     for each segment $s^l_{ij}$ $(j \in \{1, ..., S^l_i\})$ do
18:       get index of $s^l_{ij}$ in $\boldsymbol{AS_i}$: $g$ =get_idx($\boldsymbol{AS_i}$,$s^l_{ij}$)
19:       for each word $w^l_{ijm}$ $(m \in \{1, ..., N^l_{ij}\})$ do
20:         choose $z^l_{ijm} \sim$ Multinomial($\boldsymbol{\nu_{ig}}$)
21:         choose $w^l_{ijm} \sim p(w^l_{ijm}|z^l_{ijm}, \boldsymbol{\phi^l})$
22:       end for
23:     end for
24:   end for
25: end for
```

distribution with the prior $\boldsymbol{\theta_i}$ (Step 10). Later, a word w^l_{im} is drawn from a probability distribution $p(w^l_{im}|z^l_{im}, \boldsymbol{\phi^l})$ given the topic z^l_{im} (Step 11).

3 Bilingual Segmented Topic Model

Here, we describe BiSTM, which extends BiLDA to capture segment-level alignments. Algorithm 2 and Figure 3 show the generative process and graphical model, respectively, of BiSTM. As can be seen in Figure 3, BiSTM introduces a segment-level layer between the document- and word-level layers in both languages. In other words, per-segment topic distributions for each language, $\boldsymbol{\nu^e}$ and $\boldsymbol{\nu^f}$, are introduced between per-document topic distributions $\boldsymbol{\theta}$ and topic assignments for words, $\boldsymbol{z^e}$ and $\boldsymbol{z^f}$. In addition, BiSTM incorporates binary variables $\boldsymbol{y}$ to represent segment-level alignments.

Each document d^l_i in a pair of aligned documents d_i is divided into S^l_i segments: $d^l_i = \bigcup_{j=1}^{S^l_i} s^l_{ij}$. BiSTM makes the same assumption for per-topic word distributions as BiLDA, i.e., $\boldsymbol{\phi^l_k}$ are language-specific and drawn from Dirichlet distributions (Steps 1-5).

In the generative process for a document pair d_i, the per-document topic distribution $\boldsymbol{\theta_i}$ is first drawn in the same way as in BiLDA (Step 7). Thus, in BiSTM, each document pair shares the same topic distribution.

Then, if segment-level alignments are not given, $\boldsymbol{y_i}$ are generated (Steps 8-11). We assume that each document pair d_i has a probability γ_i that indicates comparability between segments across languages. γ_i is drawn from a Beta distribution with the priors η_0 and η_1 (Step 9). Then, each of $\boldsymbol{y_i}$ is drawn from a Bernoulli distribution with the prior γ_i (Step 10). Here, $y_{ijj'}$ = 1 if and only if s^e_{ij} and $s^f_{ij'}$ are aligned; otherwise, $y_{ijj'}$ = 0. Note that if segment-level alignments are observed, then Steps 8-11 are skipped. Later, a set of aligned segment sets $\boldsymbol{AS_i}$ is generated based on $\boldsymbol{y_i}$ (Step 12). For example, given $d^e_i = \{s^e_{i1}, s^e_{i2}\}, d^f_i = \{s^f_{i1}, s^f_{i2}, s^f_{i3}\}$, y_{i11} and y_{i12} are 1, and the other y's are 0, $\boldsymbol{AS_i} = \{\boldsymbol{AS_{i1}} = \{s^e_{i1}, s^f_{i1}, s^f_{i2}\}, \boldsymbol{AS_{i2}} = \{s^e_{i2}\}, \boldsymbol{AS_{i3}} = \{s^f_{i3}\}\}$ is generated in Step 12. Then, for each aligned segment set $\boldsymbol{AS_{ig}}$ $(g \in \{1, ..., |\boldsymbol{AS_i}|\})$, the per-segment topic distribution $\boldsymbol{\nu_{ig}}$ is obtained from a Pitman–Yor process with the base measure $\boldsymbol{\theta_i}$, the concentration parameter a, and the discount parameter b (Step 14). Through Steps 12-15, aligned segments indicated by $\boldsymbol{y}$ share the same per-segment topic distribution. For instance, s^e_{i1}, s^f_{i1}, and s^f_{i2} have the same topic distribution $\boldsymbol{\nu_{i1}} \sim$ PYP(a, b, $\boldsymbol{\theta_i}$) in the above example.

Then, for each word at $m \in \{1, ..., N^l_{ij}\}$ in segment s^l_{ij} in document d^l_i in language l, a latent topic assignment z^l_{ijm} is drawn from a multinomial distribution with the prior $\boldsymbol{\nu_{ig}}$ (Step 20), where g denotes the index of the element set of $\boldsymbol{AS_i}$ that includes the segment s^l_{ij}, e.g., g for s^f_{i2} is 1. Subsequently, a word w^l_{ijm} is drawn based on the assigned topic z^l_{ijm} and the language-specific per-topic word distribution $\boldsymbol{\phi^l}$ in the same manner as in BiLDA (Step 21).

t_{igk}	Table count of topic k in the CRP for aligned segment set g in document pair i.
$\boldsymbol{t_{ig}}$	K-dimensional vector, where k-th value is t_{igk}.
$t_{ig\cdot}$	Total table count in aligned segment set g in document pair i, i.e., $\sum_k t_{igk}$.
n_{igk}	Total number of words with topic k in aligned segment set g in document pair i.
$n_{ig\cdot}$	Total number of words in aligned segment set g in document pair i, i.e., $\sum_k n_{igk}$.
M^l_{kw}	Total number of word w with topic k in language l.
$\boldsymbol{M^l_k}$	$\lvert\boldsymbol{W^l}\rvert$-dimensional vector, where w-th value is M^l_{kw}.

Table 1: Statistics used in our Inference

3.1 Inference for BiSTM

In inference, we find the set of latent variables $\boldsymbol{\theta}$, $\boldsymbol{\nu}$, $\boldsymbol{z}$, and $\boldsymbol{\phi}$ that maximizes their posterior probability given the model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and observations $\boldsymbol{w}$, $\boldsymbol{y}$, i.e., $p(\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{z}, \boldsymbol{\phi}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{y})$. Here, a language-dependent variable without a superscript denotes both of the variable in language e and that in f, e.g., $\boldsymbol{z} = \{\boldsymbol{z^e}, \boldsymbol{z^f}\}$. Unfortunately, as in other probabilistic topic models, such as LDA and BiLDA, we cannot compute this posterior using an exact inference method. This section presents an approximation method for BiSTM based on blocked Gibbs sampling, inspired by Du et al. (2013).

In our inference, the hierarchy in BiSTM, i.e., the generation of $\boldsymbol{\nu}$ and $\boldsymbol{z}$, is explained by the Chinese restaurant process (CRP), through which the parameters $\boldsymbol{\theta}$, $\boldsymbol{\nu}$, and $\boldsymbol{\phi}$ are integrated out, and the statistics on table counts in the CRP, $\boldsymbol{t}$, are introduced. Table 1 lists all statistics used in our inference, where $\boldsymbol{W^l}$ denotes a vocabulary set in language l. Moreover, to accelerate convergence, we introduce an auxiliary binary variable δ^l_{ijm} for w^l_{ijm}, indicating whether w^l_{ijm} is the first customer on a table ($\delta^l_{ijm} = 1$) or not ($\delta^l_{ijm} = 0$), and t_{igk} is computed based on $\boldsymbol{\delta}$ in the same manner as in Chen et al. (2011):

$$t_{igk} = \sum_{s^l_{ij} \in \boldsymbol{AS_{ig}}} \sum_{m=1}^{N^l_{ij}} \delta^l_{ijm} I(z^l_{ijm} = k), \text{ where } I(x)$$

is a function that returns 1 if the condition x is true and 0 otherwise.

Our inference groups z^l_{ijm} and δ^l_{ijm} (each group is called a "block") and jointly samples them. Moreover, if $\boldsymbol{y}$ is not observed, our inference alternates two different kinds of blocks, $(z^l_{ijm}, \delta^l_{ijm})$ and $y_{ijj'}$. In each sampling, individual variables are resampled, conditioned on all other variables. In the following, we describe each sampling stage.

Sampling $(\boldsymbol{z}, \boldsymbol{\delta})$:

The joint posterior distribution of $\boldsymbol{z}$, $\boldsymbol{w}$, and $\boldsymbol{\delta}$ is induced in a manner similar to that in Du et al. (2010; 2013): $p(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\delta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y})$

$$= \prod_{i=1}^{D} \Bigg(\frac{\text{Beta}_K(\boldsymbol{\alpha} + \sum_{\boldsymbol{AS_i}} \boldsymbol{t_{ig}})}{\text{Beta}_K(\boldsymbol{\alpha})}$$

$$\prod_{\boldsymbol{AS_i}} \Bigg(\frac{(b|a)_{t_{ig\cdot}}}{(b)_{n_{ig\cdot}}} \prod_{k=1}^{K} S\Big(n_{igk}, t_{igk}, a\Big) \binom{n_{igk}}{t_{igk}}^{-1} \Bigg) \Bigg)$$

$$\prod_{k=1}^{K} \Bigg(\frac{\text{Beta}_{W^e}(\boldsymbol{\beta^e} + \boldsymbol{M^e_k})}{\text{Beta}_{W^e}(\boldsymbol{\beta^e})} \frac{\text{Beta}_{W^f}(\boldsymbol{\beta^f} + \boldsymbol{M^f_k})}{\text{Beta}_{W^f}(\boldsymbol{\beta^f})} \Bigg),$$

where $\text{Beta}_K(\cdot)$ and $\text{Beta}_{W^l}(\cdot)$ are K- and $|\boldsymbol{W^l}|$-dimensional beta functions, respectively, $(b|a)_n$ is the Pochhammer symbol[3], and $(b)_n$ is given by $(b|1)_n$. $S(n, m, a)$ is a generalized Stirling number of the second kind (Hsu and Shiue, 1998), which is given by the linear recursion $S(n + 1, m, a) = S(n, m - 1, a) + (n - ma)S(n, m, a)$. To reduce computational cost, the Stirling numbers are preliminarily calculated in a logarithm format (Buntine and Hutter, 2012). Then, the cached values are used in our sampling.

The joint conditional distributions of z^l_{ijm} and δ^l_{ijm} are obtained from the above joint distribution using Bayes' rule: $p(z^l_{ijm} = k, \delta^l_{ijm} = 1|\boldsymbol{z}^{-z^l_{ijm}}, \boldsymbol{w}, \boldsymbol{\delta}^{-\delta^l_{ijm}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y})$

$$= \frac{\beta^l_{w^l_{ijm}} + M^l_{kw^l_{ijm}}}{\sum_{w \in \boldsymbol{W^l}} (\beta^l_w + M^l_{kw})} \frac{\alpha_k + \sum_{\boldsymbol{AS_i}} t_{igk}}{\sum_{k=1}^{K} (\alpha_k + \sum_{\boldsymbol{AS_i}} t_{igk})}$$

$$\frac{b + at_{ig'\cdot}}{b + n_{ig'\cdot}} \frac{S(n_{ig'k} + 1, t_{ig'k} + 1, a)}{S(n_{ig'k}, t_{ig'k}, a)} \frac{t_{ig'k} + 1}{n_{ig'k} + 1},$$

$p(z^l_{ijm} = k, \delta^l_{ijm} = 0|\boldsymbol{z}^{-z^l_{ijm}}, \boldsymbol{w}, \boldsymbol{\delta}^{-\delta^l_{ijm}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y})$

$$= \frac{\beta^l_{w^l_{ijm}} + M^l_{kw^l_{ijm}}}{\sum_{w \in \boldsymbol{W^l}} (\beta^l_w + M^l_{kw})} \frac{1}{b + n_{ig'\cdot}}$$

$$\frac{S(n_{ig'k} + 1, t_{ig'k}, a)}{S(n_{ig'k}, t_{ig'k}, a)} \frac{n_{ig'k} + 1 - t_{ig'k}}{n_{ig'k} + 1},$$

where s^l_{ij} is included in $\boldsymbol{AS_{ig'}}$.

Sampling $\boldsymbol{y}$:

In our inference, each aligned segment set corresponds to a restaurant in the CRP. We regard the sampling of $y_{ijj'}$ as the choice of splitting or merging restaurant(s) in a manner similar to that

[3] $(b|a)_n = \prod_{t=0}^{n-1}(b + ta)$.

in the sampling of segmentation boundaries in Du et al. (2013). In particular, if $y_{ijj'} = 0$, then one aligned segment set $\boldsymbol{AS_m}$ is split into two aligned segment sets $\boldsymbol{AS_l}$ and $\boldsymbol{AS_r}$, where $\boldsymbol{AS_l}$, $\boldsymbol{AS_r}$, and $\boldsymbol{AS_m}$ include s^e_{ij}, $s^f_{ij'}$, and both, respectively. If $y_{ijj'} = 1$, then $\boldsymbol{AS_l}$ and $\boldsymbol{AS_r}$ are merged to $\boldsymbol{AS_m}$. For simplicity, our inference specifies $\boldsymbol{AS_l}$ and $\boldsymbol{AS_r}$ based on the current $\boldsymbol{y}$ as follows: if $AS_i(s^e_{ij}) = AS_i(s^f_{ij'})$, then $\boldsymbol{AS_l} = \{s^e_{ij}\} \cup AS^f_i(s^e_{ij}) \setminus \{s^f_{ij'}\}$ and $\boldsymbol{AS_r} = \{s^f_{ij'}\} \cup AS^e_i(s^f_{ij'}) \setminus \{s^e_{ij}\}$; otherwise, $\boldsymbol{AS_l} = AS_i(s^e_{ij})$ and $\boldsymbol{AS_r} = AS_i(s^f_{ij'})$. Here, $AS_i(j)$ is the element set of $\boldsymbol{AS_i}$ that includes the segment j, and $AS^l_i(j)$ is the set of segments in language l included in $AS_i(j)$. For example, in the example in Section 3, $AS_i(s^f_{i1}) = \boldsymbol{AS_{i1}} = \{s^e_{i1}, s^f_{i1}, s^f_{i2}\}$, $AS^e_i(s^f_{i1}) = \{s^e_{i1}\}$, and $AS^f_i(s^f_{i1}) = \{s^f_{i1}, s^f_{i2}\}$. In addition, if $y_{i11} = 0$, then $\boldsymbol{AS_m} = \{s^e_{i1}, s^f_{i1}, s^f_{i2}\}$ is split into $\boldsymbol{AS_l} = \{s^e_{i1}\} \cup AS^f_i(s^e_{i1}) \setminus \{s^f_{i1}\} = \{s^e_{i1}, s^f_{i2}\}$ and $\boldsymbol{AS_r} = \{s^f_{i1}\} \cup AS^e_i(s^f_{i1}) \setminus \{s^e_{i1}\} = \{s^f_{i1}\}$. If $y_{i23} = 1$, then $\boldsymbol{AS_l} = AS_i(s^e_{i2}) = \{s^e_{i2}\}$ and $\boldsymbol{AS_r} = AS_i(s^f_{i3}) = \{s^f_{i3}\}$ are merged to $\boldsymbol{AS_m} = \{s^e_{i2}, s^f_{i3}\}$.

The conditional distributions of $y_{ijj'}$ are as follows:

$$
\begin{aligned}
&p(y_{ijj'} = 0 | \boldsymbol{y}^{-y_{ijj'}}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\delta}, \boldsymbol{\alpha}, a, b, \eta_0, \eta_1) \\
&\propto \frac{\eta_0 + c_{i0}}{\eta_0 + \eta_1 + c_{i0} + c_{i1}} \mathrm{Beta}_K\Bigg(\boldsymbol{\alpha} + \sum_{\boldsymbol{AS_i}} \boldsymbol{t_{ig}}\Bigg) \\
&\qquad \prod_{g \in \{\boldsymbol{AS_l}, \boldsymbol{AS_r}\}} \frac{(b|a)_{t_{ig\cdot}}}{(b)_{n_{ig\cdot}}} \prod_{k=1}^{K} S(n_{igk}, t_{igk}, a), \\
&p(y_{ijj'} = 1 | \boldsymbol{y}^{-y_{ijj'}}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{t} \setminus \mathbb{T}, \boldsymbol{\alpha}, a, b, \eta_0, \eta_1) \\
&\propto \sum_{\mathbb{T}} \Bigg(\frac{\eta_1 + c_{i1}}{\eta_0 + \eta_1 + c_{i0} + c_{i1}} \mathrm{Beta}_K\Bigg(\boldsymbol{\alpha} + \sum_{\boldsymbol{AS_i}} \boldsymbol{t_{ig}}\Bigg) \\
&\qquad \frac{(b|a)_{t_{i,\boldsymbol{AS_m},\cdot}}}{(b)_{n_{i,\boldsymbol{AS_m},\cdot}}} \prod_{k=1}^{K} S(n_{i,\boldsymbol{AS_m},k}, t_{i,\boldsymbol{AS_m},k}, a) \Bigg),
\end{aligned}
$$

where $\mathbb{T}$ is the set of t_{igk} such that for either or both of $\boldsymbol{AS_l}$ and $\boldsymbol{AS_r}$, $t_{igk} = 1$. c_{i0} and c_{i1} are the total number of y_i's whose values are 0 and that of y_i's whose values are 1, respectively. Note that we change y_i's that relate to the selected action (merging or splitting), in addition to $y_{ijj'}$ to maintain consistency between $\boldsymbol{y}$ and the aligned segment sets.

Inference of $\boldsymbol{\theta}$, $\boldsymbol{\nu}$, $\boldsymbol{\phi}$:

Although our inference does not directly estimate $\boldsymbol{\theta}$, $\boldsymbol{\nu}$, and $\boldsymbol{\phi}$, these variables can be inferred from the following posterior expected values via sampling:

$$
\begin{aligned}
\hat{\theta}_{ik} &= \mathbb{E}_{\boldsymbol{z_i}, \boldsymbol{t_i} | \boldsymbol{w_i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y}} \left[\frac{\alpha_k + \sum_{\boldsymbol{AS_i}} t_{igk}}{\sum_{k=1}^{K} (\alpha_k + \sum_{\boldsymbol{AS_i}} t_{igk})} \right], \\
\hat{\nu}_{igk} &= \mathbb{E}_{\boldsymbol{z_i}, \boldsymbol{t_i} | \boldsymbol{w_i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y}} \left[\frac{n_{igk} - a t_{igk}}{b + n_{ig\cdot}} + \theta_{ik} \frac{a t_{ig\cdot} + b}{b + n_{ig\cdot}} \right], \\
\hat{\phi}^l_{kw} &= \mathbb{E}_{\boldsymbol{z}, \boldsymbol{t} | \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b, \boldsymbol{y}} \left[\frac{\beta^l_w + M^l_{kw}}{\sum_{w' \in \boldsymbol{W^l}} (\beta^l_{w'} + M^l_{kw'})} \right].
\end{aligned}
$$

Algorithm 3 Generative Process for Segments

```
1: for each document $d^l_i$ ($i \in \{1, ..., D\}$) do
2:   choose $\pi^l_i \sim$ Beta($\lambda_0$,$\lambda_1$)
3:   for each passage $u^l_{ih}$ ($h \in \{1, ..., U^l_i\}$) do
4:     choose $\rho^l_{ih} \sim$ Bernoulli($\pi^l_i$)
5:   end for
6:   $\boldsymbol{s^l_i}$ = concatenate($\boldsymbol{u^l_i}$, $\boldsymbol{\rho^l_i}$)
7: end for
```

4 Integration of Topic Segmentation into BiSTM (BiSTM+TS)

To infer segmentation boundaries simultaneously with cross-lingual topics, we integrate the unsupervised Bayesian topic segmentation method proposed by Du et al. (2013) into the proposed BiSTM (BiSTM+TS).

We assume that each segment is a sequence of topically-related passages. In particular, we consider a sentence as a passage. Our segmentation model defines a segment in document d^l_i by a boundary indicator variable ρ^l_{ih} for each passage u^l_{ih} ($h \in \{1, ..., U^l_i\}$); ρ^l_{ih} is 1 if there is a boundary after passage u^l_{ih} (otherwise 0). For example, $\boldsymbol{\rho^l_i} = (0, 1, 0, 0, 1)$ indicates that the document d^l_i comprises the two segments $\{u^l_{i1}, u^l_{i2}\}$ and $\{u^l_{i3}, u^l_{i4}, u^l_{i5}\}$.

Algorithm 3 shows the generative process for segments. The generative process of BiSTM+TS inserts Algorithm 3 between Steps 7 and 8 of Algorithm 2. Note that two documents $(d^e_i, d^f_i) \in d_i$ are segmented independently. BiSTM+TS assumes that each document d^l_i has its own topic shift probability π^l_i. For each document d^l_i, π^l_i is first drawn from a Beta distribution with the priors λ_0 and λ_1 (Step 2). Then, for each passage u^l_{ih} ($h \in \{1, ..., U^l_i\}$), ρ^l_{ih} is drawn from a Bernoulli distribution with the prior π^l_i (Step 4). Finally, segments $\boldsymbol{s^l_i}$ are generated by concatenating passages based on $\boldsymbol{\rho^l_i}$ (Step 6).

4.1 Inference for BiSTM+TS

Our inference for BiSTM+TS alternates three different kinds of blocks, sampling of $\boldsymbol{\rho}$ and samplings for BiSTM (($\boldsymbol{z}$, $\boldsymbol{\delta}$) and $\boldsymbol{y}$). The conditional distribution of $\boldsymbol{\rho}$ comprises the Gibbs probability for splitting one segment s_m into two segments s_r and s_l by placing the boundary after u^l_{ih} ($\rho^l_{ih} = 1$) and that for merging s_r and s_l to s_m by removing the boundary after u^l_{ih} ($\rho^l_{ih} = 0$).

These probabilities are estimated in the same manner as the conditional probabilities of $y_{ijj'}$, where $\boldsymbol{y}$ ($y_{ijj'} = 0, 1$), $\boldsymbol{AS_l}$, $\boldsymbol{AS_r}$, $\boldsymbol{AS_m}$, η_0, and η_1 are replaced with $\boldsymbol{\rho}$ ($\rho^l_{ih} = 1, 0$), s_l, s_r, s_m, λ_1, and λ_0, respectively, and the statistics t and n are summed for every segment rather than for every aligned segment set (see Equation (6) and (9) in Du et al. (2013)).

Our inference assumes that sampling $\boldsymbol{\rho}$ does not depend on aligned segments in the other language, i.e., $\boldsymbol{y}$[4]. After splitting or merging, we set the y's of s_m, s_l, and s_r as follows: if s_m is split into s_l and s_r, then $AS(s_l) = AS(s_m)$ and $AS(s_r) = AS(s_m)$; if s_l and s_r are merged to s_m, then $AS(s_m) = AS(s_l) \cup AS(s_r)$.

5 Experiment

We evaluated the proposed models in terms of perplexity and performance in translation pair extraction, which is a well-known application that uses a bilingual topic model. We used a document-aligned comparable corpus comprising 3,995 document pairs, each of which is a Japanese Wikipedia article in the Kyoto Wiki Corpus[5] and its corresponding English Wikipedia article[6]. Note that the English articles were collected from the English Wikipedia database dump (2 June 2015)[7] based on inter-language links, even though the original Kyoto Wiki corpus is a parallel corpus, in which each sentence in the Japanese articles is manually translated into English. Thus, our experimental data is not a parallel corpus. We extracted texts from the collected English articles using an open-source script[8]. All Japanese and English texts were segmented using MeCab[9] and TreeTagger[10] (Schmid, 1994), respectively. Then, function words were removed, and the remaining words were lemmatized to reduce data sparsity.

For translation extraction experiments, we automatically created a gold-standard translation set according to Liu et al. (2013). We first computed $p(w^e|w^f)$ and $p(w^f|w^e)$ by running IBM Model 4 on the original Kyoto Wiki corpus, which is a parallel corpus, using GIZA++ (Och and Ney, 2003), and then extracted word pairs $(\hat{w^e}, \hat{w^f})$ that satisfy both of the following conditions: $\hat{w^e} = \mathrm{argmax}_{w^e} p(w^e|w^f = \hat{w^f})$ and $\hat{w^f} = \mathrm{argmax}_{w^f} p(w^f|w^e = \hat{w^e})$. Finally, we eliminated word pairs that do not appear in the document pairs in the document-aligned comparable corpus. We used all 7,930 Japanese words in the resulting gold-standard set as the evaluation input.

5.1 Competing Methods

We compared the proposed models (BiSTM and BiSTM+TS) with a standard bilingual topic model (BiLDA). BiSTM considers each section in Wikipedia articles as a segment. Note that alignments between sections are not given in our experimental data. Thus, $\boldsymbol{y}$ is inferred in both BiSTM and BiSTM+TS.

As in the proposed models, BiLDA was trained using Gibbs sampling (Mimno et al., 2009; Ni et al., 2009; Vulić et al., 2015). In the training of each model, each variable was first initialized. Here, z^l_{ijm} is randomly initialized to an integer between 1 and K, and each of δ^l_{ijm}, $y_{ijj'}$, and ρ^l_{ih} is randomly initialized to 0 or 1. We then performed 10,000 Gibbs iterations. We used the symmetric prior $\alpha_k = 50/K$ and $\beta^l_w = 0.01$ over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}^l$, respectively, in accordance with Vulić et al. (2011). The hyperparameters a, b, λ_0, and λ_1 were set to 0.2, 10, 0.1, and 0.1, respectively, in accordance with Du et al. (2010; 2013). Both η_0 and η_1 were set to 0.2 as a result of preliminary experiments. We used several values of K to measure the impact of topic size: we used $K = 100$ and $K = 400$ in accordance with Liu et al. (2013) in addition to the suggested value $K = 2,000$ in Vulić et al. (2011).

In the translation extraction experiments,

[4] We leave a bilingual extension of the topic segmentation, i.e., incorporation of $\boldsymbol{y}$, for future work.

[5] http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

[6] We filtered out the Japanese articles that do not have corresponding English articles.

[7] http://dumps.wikimedia.org/enwiki/

[8] https://github.com/attardi/wikiextractor/

[9] http://taku910.github.io/mecab/

[10] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Model	K=100	K=400	K=2,000
BiLDA	693.6	530.7	479.9
BiSTM	520.1	429.3	394.6
BiSTM+TS	537.5	445.3	411.8

Table 2: Test Set Perplexity

we used two translation extraction methods, i.e., *Cue* (Vulić et al., 2011) and *Liu* (Liu et al., 2013). Both methods first infer cross-lingual topics for words using a bilingual topic model (BiLDA/BiSTM/BiSTM+TS) and then extract word pairs (w^e, w^f) with a high value of the probability $p(w^e|w^f)$ defined by the inferred topics. *Cue* calculates $p(w^e|w^f) = \sum_{k=1}^{K} p(w^e|k)p(k|w^f)$, where $p(k|w) \propto \frac{p(w|k)}{\sum_{k=1}^{K} p(w|k)}$ and $p(w|k) = \phi_{kw}$. Liu first converts a document-aligned comparable corpus into a topic-aligned parallel corpus according to the topics of words and computes $p(w^e|w^f, k)$ by running IBM Model 1 on the parallel corpus. Liu then calculates $p(w^e|w^f) = \sum_{k=1}^{K} p(w^e|w^f, k)p(k|w^f)$. Hereafter, a bilingual topic model used in an extraction method is shown in parentheses, e.g., *Cue*(BiLDA) denotes *Cue* with BiLDA.

5.2 Experimental Results

We evaluated the predictive performance of each model by computing the test set perplexity based on 5-fold cross validation. A lower perplexity indicates better generalization performance. Table 2 shows the perplexity of each model. As can be seen, BiSTM and BiSTM+TS are better than BiLDA in terms of perplexity.

We measured the performance of translation extraction with top N accuracy (ACC_N), the number of test words whose top N translation candidates contain a correct translation over the total number of test words (7,930). Table 3 summarizes ACC_1 and ACC_{10} for each model. As can be seen, *Cue*/*Liu*(BiSTM) and *Cue*/*Liu*(BiSTM+TS) significantly outperform *Cue*/*Liu*(BiLDA) ($p < 0.01$ in the sign test). This indicates that BiSTM and BiSTM+TS improve the performance of translation extraction for both the *Cue* and *Liu* methods by assigning more suitable topics.

Both experiments prove that capturing segment-level alignments is effective for modeling bilingual data. In addition, these experiments show that BiSTM+TS is comparable with BiSTM, indicating that the proposed model could yield a significant benefit even if the boundaries of segments are unknown.

ACC_1			
Method	K=100	K=400	K=2,000
Cue(BiLDA)	0.024	0.056	0.101
Cue(BiSTM)	0.055	0.112	0.184
Cue(BiSTM+TS)	0.052	0.107	0.176
Liu(BiLDA)	0.206	0.345	0.426
Liu(BiSTM)	0.287	0.414	0.479
Liu(BiSTM+TS)	0.283	0.406	0.467
ACC_{10}			
Method	K=100	K=400	K=2,000
Cue(BiLDA)	0.093	0.170	0.281
Cue(BiSTM)	0.218	0.286	0.410
Cue(BiSTM+TS)	0.196	0.274	0.398
Liu(BiLDA)	0.463	0.550	0.603
Liu(BiSTM)	0.531	0.625	0.671
Liu(BiSTM+TS)	0.536	0.612	0.667

Table 3: Performance of Translation Extraction

	Reference $y = 1$	Reference $y = 0$
Inference $y = 1$	195	174
Inference $y = 0$	43	1132

Table 4: Distribution of Segment-level Alignments

Tables 2 and 3 show that a larger topic size yields better performance for each model. Furthermore, *Liu* outperforms *Cue* regardless of the choice of bilingual topic models, which is consistent with previously reported results (Liu et al., 2013). The results of our experiments demonstrate that the proposed models have the same tendencies as BiLDA.

6 Discussion

6.1 Inferred Segment-level Alignments

We created a reference set to evaluate segment-level alignments $\boldsymbol{y}$ inferred by BiSTM (K=2,000). We randomly selected 100 document pairs from the comparable corpus and then manually identified cross-lingual alignments between sections. Table 4 shows the distribution of inferred $\boldsymbol{y}$ values and that of $\boldsymbol{y}$ values in the reference set. As can be seen, the accuracy of $\boldsymbol{y}$ is 0.859 (1,327/1,544).

The majority of false negatives (121/174) are sections that are not parallel but correspond partially. An example is the alignment between the

Model	Japanese article	English article
BiSTM	4.8	2.9
BiSTM+TS	10.6	4.1

Table 5: Average Number of Segments

Japanese section "history" and the English section "Bujutsu (old type of Budo)" in the "Budo (a Japanese martial art)" article pair, where a part of the English section "Bujutsu" is described in the Japanese section "history." Such errors might not necessarily have a negative effect, because partial alignments can be useful.

6.2 Inferred Segmentation Boundaries

This section compares segment boundaries inferred by BiSTM+TS (K=2,000) with section boundaries in the original articles, which have been referred to by BiSTM. The recall of BiSTM+TS for the original section boundaries is 0.727. This indicates that the unsupervised segmentation in BiSTM+TS finds drastic topical changes, i.e., section boundaries, with high recall.

Table 5 shows the average number of segments per article for each model. As can be seen, BiSTM+TS divides an article into segments smaller than the original sections. This seems to be reasonable, because some original sections include multiple topics. However, Tables 2 and 3 show that inferred boundaries do not work better than section boundaries. One reason for that is that some errors are caused by a sparseness problem, when BiSTM+TS separates an article into extremely fine-grained segments. In addition, Table 5 reveals that BiSTM+TS increases the gap between languages. Thus, segmentation with a comparable granularity between languages might be favorable for the proposed models.

6.3 Effectiveness for an English–French Wikipedia Corpus

We evaluated BiLDA, BiSTM, and BiSTM+TS in terms of perplexity and performance in translation extraction on an English–French Wikipedia corpus to verify the effectiveness of the proposed models for language pairs other than English–Japanese. The settings, e.g., parameters, for each model are the same as in Section 5. Note that we report only the performances of each model with $K = 2,000$, because all models achieved the best performances when $K = 2,000$.

Model	Test Set Perplexity
BiLDA	439.1
BiSTM	379.4
BiSTM+TS	396.6

Model	ACC_1	ACC_{10}
Cue(BiLDA)	0.219	0.556
Cue(BiSTM)	0.275	0.580
Cue(BiSTM+TS)	0.257	0.582
Liu(BiLDA)	0.715	0.838
Liu(BiSTM)	0.742	0.859
Liu(BiSTM+TS)	0.732	0.852

Table 6: Performance on an English–French Wikipedia Corpus ($K = 2,000$)

We collected French articles that correspond to the English articles used in the experiments in Section 5, from the French Wikipedia database dump (2 June 2015) based on inter-language links. As a result, our English–French corpus comprises 3,159 document pairs. The French articles were preprocessed in the same manner as the English articles: text extraction using the open-source script, segmentation using TreeTagger, removal of function words, and lemmatization.

We created a gold-standard translation set for translation extraction experiments using Google Translate service[11] in a manner similar to that in Gouws et al. (2015) and Coulmance et al. (2015), translating the French words in our corpus using Google Translate, and then eliminating word pairs that do not appear in the document pairs in our corpus. We used the top 1,000 most frequent French words in the resulting gold-standard set as the evaluation input.

Table 6 summarizes ACC_1, ACC_{10}, and perplexity. It shows that the proposed models are effective also for the English–French Wikipedia corpus. BiSTM and BiSTM+TS outperform BiLDA in terms of perplexity and performance of translation extraction, and BiSTM+TS works well even if the boundaries of segments are unknown.

7 Related Work

Multilingual topic models other than BiLDA (Section 2) have been proposed for document-aligned comparable corpora. Fukumasu et al. (2012) applied SwitchLDA (Newman et al., 2006) and Correspondence LDA (Blei and Jordan, 2003), which

[11] `http://translate.google.com/`

were originally intended to work with multimodal data, such as annotated image data, to modeling multilingual text data. They also proposed a symmetric version of Correspondence LDA. Platt et al. (2010) projected monolingual models based on PLSA or Principal Component Analysis into a shared multilingual space with the constraint that document pairs must map to similar locations. Hu et al. (2014) proposed a multilingual tree-based topic model that uses a hierarchical bilingual dictionary in addition to document alignments. Note that these models do not consider segment-level alignments.

There are several multilingual topic models tailored for data other than a document-aligned comparable corpus, including bilingual topic models for word alignment and machine translation on parallel sentence pairs (Zhao and Xing, 2006; Zhao and Xing, 2008). Some models have mined multilingual topics from unaligned text data by bridging the gap between different languages using a bilingual dictionary (Jagarlamudi and Daumé III, 2010; Zhang et al., 2010; Negi, 2011). Boyd-Graber and Blei (2009) used parallel sentences in combination with a bilingual dictionary. However, these models have the drawback that they require a parallel corpus or a bilingual dictionary in advance, which cannot be obtained for some language pairs or domains.

In a monolingual setting, some topic models that consider segment-level topics have been proposed. Du et al. (2010) considered a document as a set of segments and generated each per-segment topic distribution from the topic distribution of the related document through a Pitman–Yor process. Others have considered a document as a sequence of segments. Cheng et al. (2009) reflected the underlying sequences of segments' topics by positing a permutation distribution over a document. Wang et al. (2011) modeled topical sequences in documents with a latent first-order Markov chain, and Du et al. (2012) generated each per-segment topic distribution from the topic distribution of its document and that of its previous segment. Note that none of these models have been extended to a multilingual setting.

8 Conclusions

In this paper, we proposed BiSTM, which models a document hierarchically and deals with segment-level alignments. BiSTM assigns the same topic distribution to both aligned documents and aligned segments. We also presented an extended model, BiSTM+TS, that infers segmentation boundaries in addition to latent topics by incorporating unsupervised topic segmentation (Du et al., 2013). Our experimental results show that capturing segment-level alignments improves perplexity and translation extraction performance, and that BiSTM+TS yields a significant benefit even if the boundaries of segments are not given.

This paper presented an extension to BiLDA, but hierarchical structures can also be incorporated into other bilingual topic models (Section 7). As future work, we would like to verify the effectiveness of the proposed models for other datasets or other cross-lingual tasks, such as cross-lingual document classification (Ni et al., 2009; Platt et al., 2010; Ni et al., 2011; Smet et al., 2011) and cross-lingual information retrieval (Vulić et al., 2013).

Acknowledgments

We thank Atsushi Fujita for valuable comments on earlier versions of this manuscript.

References

David M. Blei and Michael I. Jordan. 2003. Modeling Annotated Data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134.

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

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

Wray Buntine and Marcus Hutter. 2012. A Bayesian View of the Poisson-Dirichlet Process. `http://arxiv.org/pdf/1007.0296.pdf`.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Global Models of Document Structure using Latent Permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 371–379.

Changyou Chen, Lan Du, and Wray Buntine. 2011. Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process. In *Proceedings of the European Conference on Machine Learning and*

Principles and Practice of Knowledge Discovery in Databases 2011, pages 296–311.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Transgram, Fast Cross-lingual Word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113.

Lan Du, Wray Buntine, and Huidong Jin. 2010. A Segmented Topic Model Based on the Two-parameter Poisson-Dirichlet Process. *Machine Learning*, 81(1):5–19.

Lan Du, Wray Buntine, and Huidong Jin. 2012. Modelling Sequential Text with an Adaptive Topic Model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 535–545.

Lan Du, Wray Buntine, and Mark Johnson. 2013. Topic Segmentation with a Structured Topic Model. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200.

Kosuke Fukumasu, Koji Eguchi, and Eric P. Xing. 2012. Symmetric Correspondence Topic Models for Multilingual Text Analysis. In *Advances in Neural Information Processing Systems 25*, pages 1286–1294.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 748–756.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57.

Leetsch C. Hsu and Peter Jau-Shyong Shiue. 1998. A Unified Approach to Generalized Stirling Numbers. *Advances in Applied Mathematics*, 20(3):366–384.

Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan Boyd-Graber. 2014. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1166–1176.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting Multilingual Topics from Unaligned Comparable Corpora. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, pages 444–456.

Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2013. Topic Models + Word Alignment = A Flexible Framework for Extracting Bilingual Dictionary from Comparable Corpus. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 212–221.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889.

Sumit Negi. 2011. Mining Bilingual Topic Hierarchies from Unaligned Text. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 992–1000.

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Statistical Entity-topic Models. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 680–686.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining Multilingual Topics from Wikipedia. In *Proceedings of the 18th International World Wide Web Conference*, pages 1155–1156.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 375–384.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

Jim Pitman and Marc Yor. 1997. The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *The Annals of Probability*, 25(2):855–900.

John Platt, Kristina Toutanova, and Wen tau Yih. 2010. Translingual Document Representations from Discriminative Projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 251–261.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Wim De Smet, Jie Tang, and Marie-Francine Moens. 2011. Knowledge Transfer Across Multilingual Corpora via Latent Topics. In *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 549–560.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying Word Translations from Comparable Corpora Using Latent Topic Models. In *Proceedings of the 49th Annual Meeting of the Associ-*

ation for Computational Linguistics: Human Language Technologies, pages 479–484.

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

Ivan Vulić, 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.

Hongning Wang, Duo Zhang, and ChengXiang Zhai. 2011. Structural Topic Model for Latent Topical Structure Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1526–1535.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-Lingual Latent Topic Extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137.

Bing Zhao and Eric P. Xing. 2006. BiTAM: Bilingual Topic AdMixture Models for Word Alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 969–976.

Bing Zhao and Eric P. Xing. 2008. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. In *Advances in Neural Information Processing Systems 20*, pages 1689–1696.