# Multilingual Topic Models for Bilingual Dictionary Extraction

XIAODONG LIU, KEVIN DUH, and YUJI MATSUMOTO, Nara Institute of Science and Technology

A machine-readable bilingual dictionary plays a crucial role in many natural language processing tasks, such as statistical machine translation and cross-language information retrieval. In this article, we propose a framework for extracting a bilingual dictionary from comparable corpora by exploiting a novel combination of topic modeling and word aligners such as the IBM models. Using a multilingual topic model, we first convert a comparable *document*-aligned corpus into a parallel *topic*-aligned corpus. This novel topic-aligned corpus is similar in structure to the *sentence*-aligned corpus frequently employed in statistical machine translation and allows us to extract a bilingual dictionary using a word alignment model.

The main advantages of our framework is that (1) no seed dictionary is necessary for bootstrapping the process, and (2) multilingual comparable corpora in more than two languages can also be exploited. In our experiments on a large-scale Wikipedia dataset, we demonstrate that our approach can extract higher precision dictionaries compared to previous approaches and that our method improves further as we add more languages to the dataset.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Algorithms, Experimentation, Languages

Additional Key Words and Phrases: Bilingual dictionary, multilingual topic model, comparable corpus

**ACM Reference Format:**
Liu, X., Duh, K., and Matsumoto, Y. 2015. Multilingual topic models for bilingual dictionary extraction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 14, 3, Article 11 (June 2015), 22 pages.
DOI:http://dx.doi.org/10.1145/2699939

## 1. INTRODUCTION

A machine-readable bilingual dictionary plays a very important role in many natural language processing tasks. In statistical machine translation (SMT), dictionaries can help in the domain adaptation setting [Daume III and Jagarlamudi 2011]. In cross-lingual information retrieval (CLIR), dictionaries serve as efficient means for query translation [Resnik et al. 2001]. Many other multilingual applications also rely on bilingual dictionaries as integral components. For example, Volkova et al. [2013] use a bilingual dictionary to analyze multilingual sentiment in social media; Zhang et al. [2010] incorporate a Chinese-English dictionary into a probabilistic topic model to explore bilingual latent topics in Chinese and English texts.

One typical approach for building a bilingual dictionary resource uses parallel sentence-aligned corpora. This is often done in the context of SMT, using word alignment algorithms such as the IBM models [Brown et al. 1993; Och and Ney 2003].

This work is an extension of our conference paper [Liu et al. 2013].
This work is partially supported by the China Scholarship Council (CSC) and JSPS KAKENHI Grant Number 26730121.
Authors' addresses: X. Liu (corresponding author), K. Duh, and Y. Matsumoto, Computational Linguistics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan; email: Iiuxiaodongby@gmail.com; xiaodong-l@is.naist.jp.

2375-4699/2015/06-ART11 $15.00
DOI:http://dx.doi.org/10.1145/2699939

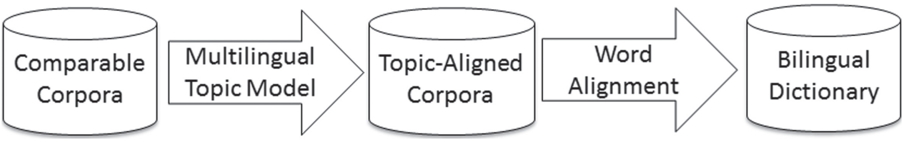

Fig. 1. The proposed framework for bilingual dictionary extraction. A multilingual topic model is used for converting a document-aligned comparable corpus to topic-aligned corpora. Given a topic, word alignment models are used to model co-occurrence across languages.

Unfortunately, parallel corpora may be scarce for certain language pairs or domains of interest (e.g., medical and microblog). Thus, the use of comparable corpora for bilingual dictionary extraction has become an active research topic [Haghighi et al. 2008; Vulić et al. 2011; Liu et al. 2013]. Here, a comparable corpus is defined as a collection of document pairs written in different languages, but talking about the same topic [Koehn 2010], such as interconnected Wikipedia articles. The challenge with bilingual dictionary extraction from comparable corpus is that existing word alignment methods developed for parallel corpus cannot be directly applied because of assumptions of sentence alignment. We solve this problem by converting comparable corpora, which are aligned at the *document* level, to *topic*-aligned corpora, then extracting the dictionary by conventional word alignment methods. This general framework is shown in Figure 1.

Intuitively, our framework works as follows: Let $w^e$ be an English word and $w^f$ be a French word, our goal is to obtain $(w^e, w^f)$ pairs in which the probability $p(w^e|w^f)$ or $p(w^f|w^e)$ is high. It is difficult to reliably estimate $p(w^f|w^e)$ from comparable corpora directly because the alignment is only at the document level, so the number of translation choices is high. Our approach differs by modeling $p(w^e|w^f, t)$ or $p(w^f|w^e, t)$ instead, where $t$ is a topic. The key idea is that it is easier to tease out the translation of a word $e$ given $p(w^f|w^e, t)$ rather than $p(w^f|w^e)$, since topic information can restrict the set of possible translation choices.

Unlike most previous works in bilingual dictionary extraction from comparable corpora, which are based on the context vector idea [Rapp 1995; Fung and Lo 1998], our topic modeling framework requires absolutely no seed dictionary to bootstrap the extraction process. Thus our framework is advantageous in low-resource scenarios and can be used as a way to obtain high-precision seed dictionaries for other bilingual dictionary extraction frameworks.

A second advantage of our topic modeling approach is that we can exploit comparable corpora in more than two languages, a situation which is becoming increasingly prevalent due to the spread of the multilingual Web. We show how we can improve the extraction of Japanese-English dictionaries using comparable data not only from Japanese and English, but also from other languages such as Chinese and French.

The article is organized as follows. In the next section, we review previous works on dictionary extraction using comparable corpora. We then describe our multilingual topic model in Section 3 and introduce our overall framework in Section 4. In Section 5, we show how our method extracts a high-precision Japanese-English dictionary from a large-scale comparable Wikipedia corpus consisting of Japanese, English, French, and Chinese.

## 2. RELATED WORK

The numerous works on bilingual lexicon from comparable corpora can be divided into two broad categories: context vector approaches (Section 2.1) and projection-based approaches (Section 2.2). We also briefly touch upon research on pivot languages

(Section 2.3) and multilingual word representation learning (Section 2.4); to the best of our knowledge, these are promising approaches but have not yet been employed for bilingual lexicon extraction.

### 2.1. Context Vector Approach

The context vector approach, starting with the seminal works of Rapp [1995] and Fung and Lo [1998], is built on the assumption that a word and its corresponding translation tend to appear in similar contexts across languages, also known as the *distributional hypothesis*. A typical context vector approach for the bilingual dictionary extraction consists of three steps.

(1) Represent contexts of a word using an existing seed dictionary. This ranges from simple representations based on bag-of-words [Fung and Lo 1998; Rapp 1995] or TF-IDF of words in a window context [Rapp 1995], to more elaborate representations such as dependency trees [Andrade et al. 2011].
(2) Measure similarity/distance between words in this common space, for example, using cosine similarity [Koehn and Knight 2002; Fung and Lo 1998] or Manhattan distance [Rapp 1995].
(3) Extract word pairs with high similarity.

Methods differ in how the seed dictionary is acquired [Déjean et al. 2002; Koehn and Knight 2002] and how similarity is defined [Fung and Cheung 2004; Tamura et al. 2012]. It is important to note that all these methods critically rely on a seed dictionary to ensure that word in different languages are represented in the same space. To alleviate the dependence on the size of the seed dictionary, Tamura et al. [2012] use an unsupervised label propagation method to improve robustness.

### 2.2. Projection-Based Approach

Projection-based approaches have also been proposed, though they can be shown to be related to the aforementioned distributional approaches [Gaussier et al. 2004]; for example, Haghighi et al. [2008] use canonical correlation analysis (CCA) to map vectors in different languages into the same latent space. Laroche and Langlais [2010] presents a good summary for the project-based approaches.

Vulić et al. [2011] pioneer a new approach to bilingual dictionary extraction. The main idea is to first map words in different languages into the same semantic space using multilingual topic models. Then, several statistical measures, such as Kullback-Leibler divergence, are used to compute similarity between words in cross-languages; finally, extract lexicons with high resulting probability. This method is a totally unsupervised learning style and does not require any seed dictionary.

Our approach is motivated by Vulić et al. [2011]. However, we exploit the topic model in a very different way (explained in Section 4.2). They do not use word alignments as we do, and as a result, their approach requires training topic models with a large number of topics, which may limit the scalability of the approach. Further, we explore extensions of multilingual topic models in more than two languages.

Recently, there has been much interest in multilingual topic models (MLTM) [Jagarlamudi and Daume 2010; Mimno et al. 2009; Ni et al. 2009; Boyd-Graber and Blei 2009]. Many of these models give $p(t|e)$ and $p(t|f)$ but stop short of extracting a bilingual lexicon. Although topic models can group related $e$ and $f$ in the same topic cluster, the extraction of a high-precision dictionary requires additional effort. One of our contributions here is an effective way to do this extraction using word alignment methods.

### 2.3. Pivot-Language in Lexicon Extraction and Machine Translation

Multilingual corpora in more than two languages have been exploited to various degrees. This is often done using a pivot language: for example, given a Japanese-English dictionary and an English-Chinese dictionary, one can exploit transitive properties with English serving as the pivot to find Japanese-Chinese translations. When available, such multilingual information has been shown to improve the quality of bilingual lexicon [Aker et al. 2014; Kwon et al. 2013; Sadat et al. 2002].

This pivot language idea has proven beneficial in applications such as cross-lingual information retrieval [Gollins and Sanderson 2001] and machine translation [Paul et al. 2009; Wu and Wang 2009]. It is also used for bootstrapping the construction of WordNet for low-resource languages (c.f. [Bond et al. 2008]) and for directly creating multilingual lexical resources [de Melo and Weikum 2009; Magnini et al. 1994; Mausam et al. 2009].

Our multilingual topic model approach differs in that there is no concept of pivot: data from all languages are treated equally. In this respect, extension to many languages is straightforward as long as computation efficiency issues can be solved.

### 2.4. Multilingual Word Representation Learning

Multilingual word representation learning, which is an extension of monolingual word representation learning, is a set of deep learning algorithms that enables new ways to do cross-lingual processing. It works by mapping words in different languages into the same low-dimensional space in order to capture syntactic and semantic similarities across languages. For instance, Klementiev et al. [2012] propose training bilingual word representations by jointly training monolingual neural language models together with a regularizer that enforces seed translations to have similar representations. Chandar et al. [2014] propose a novel autoencoder algorithm for learning bilingual word representations; importantly, their algorithm only depends on bag-of-words representations of aligned sentences and does not rely on word alignments.

These works focus on cross-lingual classification tasks, but conceivably, their results could be adapted to our comparable lexicon extraction task. For example, the vectors might be used as seed within context-based models like with Rapp [1995]. Alternatively, these vectors could be used within our framework to generate something like the topic-aligned corpora; that is, words with similar vectors are grouped together and given to our word aligner, analogously to how we group words with the same topic together. We believe the use of vector representations is an interesting area of future work.

## 3. MULTILINGUAL TOPIC MODEL

We adopt the Multilingual Topic Model (MLTM) proposed by Ni et al. [2009] and Mimno et al. [2009], which extends the monolingual Latent Dirichlet Allocation model [Blei et al. 2003]. MLTM learns word-topic distributions and topic-document distributions from comparable corpora. In the original works, it is assumed that each document tuple $t_m$ in the comparable document is *fully-connected*; for example, if we have a quad-lingual comparable corpus consisting of Chinese (c), English (e), French (f), and Japanese (j), it is assumed that all document tuples in the collection contains documents in all four languages (i.e., $t_m = [d_m^c, d_m^e, d_m^f, d_m^j]\ \forall m$), where $m$ indexes tuples in the collection and $d_m^c$ represents a Chinese document, $d_m^e$ represents an English document, etc.

However, such fully-connected comparable corpora are rare in practice. Taking the entire Wikipedia as an example, Arai et al. [2008] show that among all Japanese documents, only 64.4% have links to English entries and only 22.5% have links to Chinese

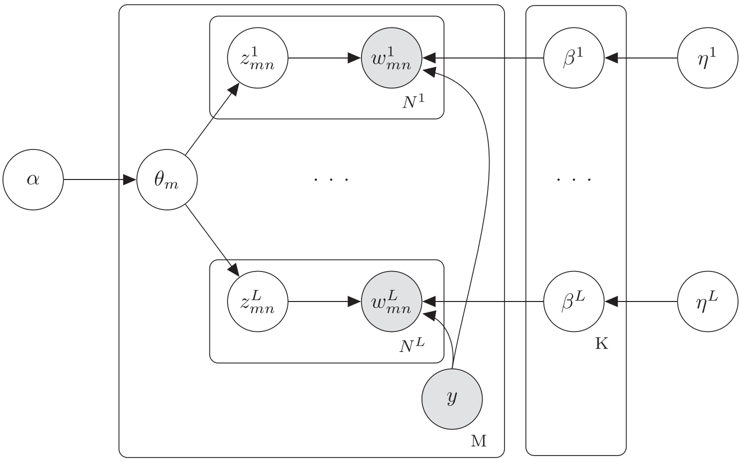

Fig. 2. Graphical representation of partially-connected multilingual topic model.

entries. In our own Wikipedia crawl (described in Section 5.1), we find that within our target set of 14K Japanese documents, the proportion of linked English, Chinese, and French documents is only around 20–30%; if we restrict to tuples that contain documents in all four languages, this number drops to 12%. In some cases, this is because the *interlanguage-link* information is missing, but in most cases, this disparity is largely the result of different human editors contributing independently in different languages [Duh et al. 2013].

We assume that a tuple does not necessarily contain documents in all languages and call such comparable corpora *partially-connected*. In the following, we extend MLTM [Mimno et al. 2009; Ni et al. 2009] to handle partially-connected corpora with a maximum of $L$ languages per tuple.

### 3.1. Generative Process

The generative process of our proposed partially-connected multilingual topic model is this: First, we define our comparable corpus as a collection of $M$ tuples in different languages (i.e., $t_m = [d^1_m, ..., d^l_m, ..., d^L_m]$ with $m \in \{1, ..., M\}$). Given a tuple of documents $t_m = [d^1_m, ..., d^l_m, ..., d^L_m]$, there is a corresponding auxiliary variable $y_m = [y^l_m, ..., y^l_m, ..., y^L_m]$, which is an $L$-dimensional binary vector that indicates the presence or absence of documents in the language $l$ in tuple $m$. The value of the $l$th of vector $m$, $y^l_m \in \{0, 1\}$, which 1 indicates presence and 0 indicates absence. For example, a tuple of documents which may contain Chinese, English, and Japanese can be represented by a 3-dimensional binary vector: $[1, 1, 1]$ denotes that it contains all of the three languages; while $[0, 1, 0]$ denotes that it only contains English document. It is not difficult to see that it is very flexible to encode the relationship of a tuple of a document for the multilingual topic model.

The generative story is shown in Algorithm 1, and a graphical representation is shown in Figure 2.

Here, language-specific topic word distributions $\beta^l$ are drawn from symmetric Dirichlet distributions with prior $\eta^l$; $\theta_m$ is the topic proportion of a tuple of documents $t_m$ drawn from symmetric Dirichlet distribution with prior $\alpha$; $z^l$ are topic indices in language $l$; words $w^l$ are drawn from language-specific distributions $p(w^l|z^l,\beta^l,y^l_m)$, where $l \in \{1,...,L\}$.

**Algorithm 1:** Generative story for partially-connected multilingual topic model

```
for each topic $k$ do
    for $l \in \{1,...,L\}$ do
        sample $\varphi^l_k \sim Dirichlet(\beta^l)$
    end
end
for each tuple $t_m$ in corpus do
    sample $y_m$
    for each document pair $t_m$ do
        sample $\theta_m \sim Dirichlet(\alpha)$
        for $l \in \{1,...,L\}$ do
            if $y^l_m == 1$ then
                sample $z^l \sim Multinomial(\theta_m)$
                for each word $w^l$ in $d^l_i$ do
                    sample $w^l \sim p(w^l|z^l,\varphi^l,y^l_m)$
                end
            end
        end
    end
end
```

### 3.2. Inference

The central computational problem for a partial multilingual topic model is approximating the posterior given a tuple of documents. In general, it is hard to estimate the posterior of a Bayesian model using exact inference methods. Therefore, approximate inference algorithms are always selected to deal with these kind of models. One of the approximate inference methods is based on Markov Chain Monte Carlo (MCMC) [Blei and Jordan 2006; Mimno et al. 2009], a sampling approach. The basic idea of MCMC is that first a Markov chain is constructed, and then its stationary distribution, which is the posterior of interest, is computed.

In this article, we develop a collapsed Gibbs sampling [Heinrich 2004; Mimno et al. 2009; Ni et al. 2009], a type of MCMC, to estimate the posterior given a tuple of documents. Concretely, given a tuple of documents $m$, the possibility of topic $k$ of the $i$ word in the language $l$ yields:

$$p(z^l_i = k|\vec{w}^l, \vec{z}^l{}_{\neg i}, \beta^1, ..., \beta^L, \alpha) \propto \frac{n^{(v)}_{l,k,\neg i} + \eta^l}{\sum_{v'=1}^{V^l} n^{(v')}_{l,k,\neg i} + \eta^l \cdot V^l} \cdot (\sum_{l'=1}^{L} y^{l'}_m \cdot (n^{(k)}_{l',m})_{\neg l,i} + \alpha). \tag{1}$$

Here, the document in language $l$ denotes $\vec{w}^l = \{w^l_i = v, w^l_{\neg i}\}$ with the corresponding topic states $\vec{z}^l = \{z^l_i = k, \vec{z}^l{}_{\neg i}\}$; the counts $n^{(v)}_{l,k,\neg i}$ indicate that token $i$ is excluded from the corresponding document $l$ in the tuple; the counts $(n^{(k)}_{l',m})_{\neg l,i}$ denote that token $i$ is excluded from the corresponding topic $k$ when $l = l'$ is held in the tuple; $V^l$ denotes vocabulary in language $l$.

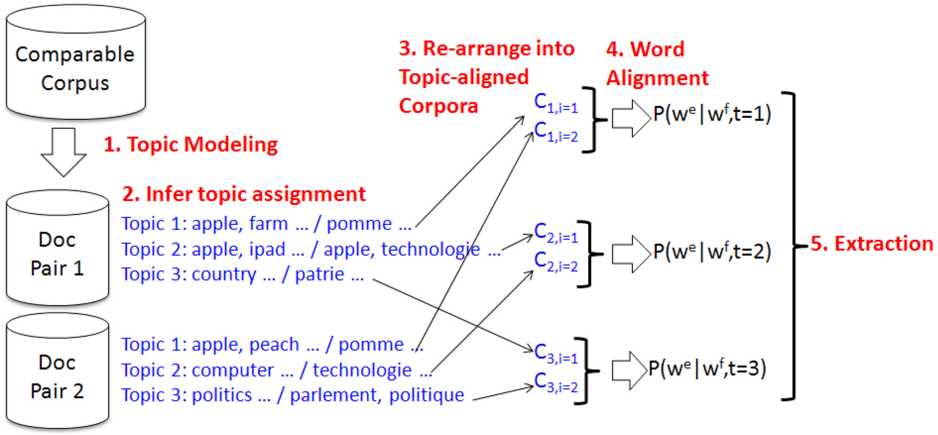
Comparable Corpus
1. Topic Modeling
Doc Pair 1
Doc Pair 2
2. Infer topic assignment
Topic 1: apple, farm … / pomme …
Topic 2: apple, ipad … / apple, technologie …
Topic 3: country … / patrie …
Topic 1: apple, peach … / pomme …
Topic 2: computer … / technologie …
Topic 3: politics … / parlement, politique
3. Re-arrange into Topic-aligned Corpora
C1,i=1
C1,i=2
C2,i=1
C2,i=2
C3,i=1
C3,i=2
4. Word Alignment
P(we|wf,t=1)
P(we|wf,t=2)
P(we|wf,t=3)
5. Extraction

Fig. 3. Construction of topic-aligned corpora.

Finally, we compute the multinomial parameter sets of $\Theta$ and $B$:

$$\beta^{l}_{k,v} = \frac{n^{(v)}_{l,k,} + \eta^{l}}{\sum_{v'=1}^{V^{l}} n^{(v')}_{l,k} + \eta^{l} \cdot V^{l}}, \tag{2}$$

$$\theta_{m,k} = \frac{\sum_{l'=1}^{L} y^{l'}_{m} \cdot n^{(k)}_{l',m} + \alpha}{\sum_{k'} \sum_{l'=1}^{L} y^{l'}_{m} \cdot n^{(k')}_{l',m}}. \tag{3}$$

Direchlet hyperparameters $\alpha$ and $\eta$ can be optimized by a simple and stable fixed-point iteration for a maximum likelihood estimator [Minka 2000].

## 4. PROPOSED FRAMEWORK FOR BILINGUAL DICTIONARY EXTRACTION

The general idea of our proposed framework is sketched in Figure 1: First, we run a multilingual topic model to convert the comparable corpora to topic-aligned corpora; second, we run a word alignment algorithm on the topic-aligned corpora in order to extract translation pairs. The innovation is in how this topic-aligned corpora is defined and constructed—the link between the two stages. We describe how this is done in Section 4.1 and show how existing approaches are subsumed in our general framework in Section 4.2.

### 4.1. Topic-Aligned Corpora

Suppose the original comparable corpus has $D$ document pairs $[d^{e}_{i}, d^{f}_{i}]_{i=1,\ldots,D}$. We run a multilingual topic model with $K$ topics, where $K$ is user-defined (Section 3). The topic-aligned corpora is defined hierarchically as a *set of sets*: on the first level, we have a set of $K$ topics, $\{t_1, \ldots, t_k, \ldots, t_K\}$; on the second level, for each topic $t_k$, we have a set of $D$ "word collections“ $\{C_{k,1}, \ldots, C_{k,i}, \ldots, C_{k,D}\}$. Each word collection $C_{k,i}$ represents the English and foreign words that occur simultaneously in topic $t_k$ and document $d_i$.

For clarity, let us describe the topic-aligned corpora construction process with a flow chart in Figure 3.

(1) Train a multilingual topic model (Section 3).
(2) Infer a topic assignment for each token in the comparable corpora and generate a list of word collections $C_{k,i}$ occurring under a given topic.

(3) Rearrange the word collections such that $C_{k,i}$ belonging to the same topic are grouped together. This resulting set of sets is called topic-aligned corpora, since it represents word collections linked by the same topics.

(4) For each topic $t_k$, we run IBM Model 1 [Brown et al. 1993; Och and Ney 2003] on $\{C_{k,1}, \ldots, C_{k,i}, \ldots, C_{k,D}\}$. In analogy to statistical machine translation, we can think of this dataset as a parallel corpus of $D$ "sentence pairs," where each "sentence pair" contains the English and foreign word tokens that co-occur under the same topic and the same document. Note that word alignment is run independently for each topic, resulting in $K$ topic-dependent lexicons $p(w^e|w^f, t_k)$.[1]

(5) To extract a bilingual dictionary, we find pairs $(w^e, w^f)$ with high probability under the following model.

$$p(w^e|w^f) = \sum_k p(w^e|w^f, t_k)p(t_k|w^f). \tag{4}$$

The first term is the topic-dependent bilingual lexicon from Step 4; the second term is estimated as follows using topic model parameters.

$$p(t_k|w^f) = \frac{p(t_k, w^f)}{\sum_k p(t_k, w^f)} \propto p(w^f|t_k)p(t_k). \tag{5}$$

If we assume that $p(t_k)$ is the uniform distribution over topics, $p(t_k|w^f)$ is defined as

$$p(t_k|w^f) \propto p(w^f|t_k). \tag{6}$$

Here, $p(w^f|t_k)$ is the topic posterior from the topic model in Step 1.

In practice, we compute the probabilities of Equation (4) in both directions: $p(w^e|w^f)$ as in Eq. (4) and $p(w^f|w^e) = \sum_k p(w^f|w^e, t_k)p(t_k|w^e)$. Subsequently, several options are conceivable for extracting a bilingual lexicon: option (a) is to set a threshold $\delta$ and extract all pairs $(\tilde{e}, \tilde{f})$ with $p(w^f = \tilde{f}|w^e = \tilde{e}) + p(w^e = \tilde{e}|w^f = \tilde{f}) > \delta$; option (b) is to set thresholds $\delta_1, \delta_2$ and extract lexicons based on the following bidirectional constraint that a pair $(\tilde{e}, \tilde{f})$ is extracted only if

$$p(w^e = \tilde{e}|w^f = \tilde{f}) > \delta_1; p(w^f = \tilde{f}|w^e = \tilde{e}) > \delta_2. \tag{7}$$

We show results from both options in our experiments. Option (a) is useful for generating a ranked list and computing precision-recall curves, since $\delta$ can be adjusted to allow for different numbers of extracted pairs. Option (b) gives very high precision extractions, since it takes the intersection from both $p(w^f|w^e)$ and $p(w^e|w^f)$; however, it is not easy to tune $\delta_1, \delta_2$ to extract a given number of pairs, since the intersection is not known beforehand. In our experiments, we "set" $\delta_1, \delta_2$ to retrieve only one candidate translation per model, extracting a pair $(\tilde{e}, \tilde{f})$ if the following holds:

$$\tilde{e} = \arg\max_{w^e} p(w^e|w^f = \tilde{f}); \tilde{f} = \arg\max_{w^f} p(w^f|w^e = \tilde{e}). \tag{8}$$

[1] Specifically, let $\boldsymbol{e} = [w^e_1, w^e_2, \ldots w^e_{|\boldsymbol{e}|}]$ be a bag of $|\boldsymbol{e}|$ English words in $C_{k,i}$ and $\boldsymbol{f} = [w^f_1, w^f_2, \ldots w^f_{|\boldsymbol{f}|}]$ be a bag of $|\boldsymbol{f}|$ foreign words in $C_{k,i}$. We model the alignment as $p(\boldsymbol{e}, a, |\boldsymbol{f}, t_k) \approx \prod_{i=1}^{|\boldsymbol{e}|} p(w^e_i|w^f_{a(i)}, t_k)$, where $p(w^e_i|w^f_{a(i)}, t_k)$ captures the translation probability of the $i$th English word from the aligned foreign word at position $j = a(i)$. The actual alignment $a$ is a hidden variable learned via EM and $t_k$ is held constant.

### 4.2. Alternative Approaches

To the best of our knowledge, Vulić et al. [2011] is the only work that focuses on using topic models for bilingual lexicon extraction like ours, but they exploit the topic model results in a different way. Their *Cue Method* computes

$$p(w^e|w^f) = \sum_k p(w^e|t_k)p(t_k|w^f). \tag{9}$$

This can be seen as a simplification of our Eq. (4), where Eq. (9) replaces $p(w^e|w^f, t_k)$ with the simpler $p(w^e|t_k)$. This is a strong assumption which essentially claims that the topic distribution $t_k$ summarizes all information about $w^f$ for predicting $w^e$. Our formulation can be considered more realistic because we do not have the assumption that $w^e$ is independent of $w^f$ given $t_k$; we model $p(w^e|w^f, t_k)$ directly and estimate its parameters with word alignment methods.

Another variant proposed by Vulić et al. [2011] is the so-called Kullback-Leibler (KL) method. It scores translation pairs by

$$KL(w^e, w^f) = Divergence(p(t_k|w^e)||p(t_k|w^f)) = -\sum_k p(t_k|w^e) \log p(t_k|w^e)/p(t_k|w^f). \tag{10}$$

The information content is the same as the Cue Method (Eq. (9)); it is simply a different scoring equation. In our experiment, we find that a symmetric version of KL, known as Jensen–Shannon Divergence, gave better results:

$$JS(w^e, w^f) = \frac{1}{2}KL(w^e, w^{ef}) + \frac{1}{2}KL(w^f, w^{ef}), \tag{11}$$

where $w^{ef}$ denotes the average of the *word-topic* distributions of both $e$ and $f$, that is, $w^{ef} = [p(t|w^e) + p(t|w^f)]/2$.[2]

## 5. EXPERIMENTS

First, we describe our experiment setup in Section 5.1. Section 5.2 compares our method with previous works, and Section 5.3 shows how our method improves given additional languages in the comparable data. Section 5.4 discusses practical issues, such as hyper-parameter selection and runtime, while Section 5.5 provides detailed analyses of the results. Finally, Section 5.6 demonstrates how our approach can be used to provide high-precision dictionaries to bootstrap existing context vector methods.

### 5.1. Experiment Setting

We perform experiments based on the Kyoto Wiki Corpus.[3] We choose the Kyoto Wiki Corpus because it is a *parallel* corpus, where the Japanese edition of Wikipedia is translated manually into English sentence by sentence (14K document pairs, 472K sentences). This enables us to use standard word alignment methods to create a "gold-standard" lexicon for large-scale automatic evaluation. First, we run IBM Model 4 on this parallel corpus. Then we extract 166K $(\tilde{e}, \tilde{f})$ pairs based on the strict bidirectional requirement of Eq. (7), with threshod $\delta_1 = \delta_2 = 0.3$. We refer to these pairs as a "gold standard" bilingual lexicon. Due to the large data size and the strict bidirectional requirement, these "gold standard" bilingual dictionary items are of high quality (92%

[2] The third and final variant by Vulić et al. [2011], TF-ITF, performs poorly and is not reported.
[3] `http://alaginrc.nict.go.jp/WikiCorpus/index_E.html`.

Table I. Statistics of our Multilingual Wiki Crawl Dataset

| Wiki | #document | #vocabulary | $\#(j \cap e)$ | $\#(j \cap e \cap c)$ | $\#\ (j \cap e \cap c \cap f)$ |
|---|---|---|---|---|---|
| **Japanese** (j) | 14,033 | 40k | — | — | — |
| **English** (e) | 4,087 | 20k | 4,087 | — | — |
| **Chinese** (c) | 3,494 | 23k | — | 2,338 | — |
| **French** (f) | 2,871 | 12k | — | — | 1,719 |

*Note:* $\#(l \cap l')$ denotes the number of fully-connected document tuples by intersecting languages $l$ and $l'$.

Table II. Statistics of Stop-Words in Different Languages

| Language | Japanese | English | Chinese | French |
|---|---|---|---|---|
| Number of stop-words | 44 | 571 | 125 | 463 |

precision by a manual check on 500 random items). Note that sentence alignments are used only for creating this gold standard and are not used in subsequent experiments.

To evaluate the proposed framework, we use a real comparable corpus crawled from Wikipedia (denoted as Wiki). We keep the Japanese side of the original Kyoto Wiki Corpus but crawl the online English, Chinese, and French editions by following the inter-language links from the Japanese page. The statistics of the crawl are shown in Table I. Observe that this is a partially-connected comparable corpus: the number of corresponding articles in English, Chinese, and French is much smaller, consisting of only 20–30% of the original Japanese. The number of fully-connected tuples in all four languages is only 12%, as seen in the intersection $(j \cap e \cap c \cap f)$.

For preprocessing, we did word segmentation on Japanese and Chinese using Kytea [Neubig et al. 2011] and Porter stemming on English and French using NLTK Version 3.0.[4] Finally, we remove the 2,000 rarest words and stop-words from each language as shown in Table II. To facilitate future work in this area, our stop-word lists for these four languages are available at `https://bitbucket.org/allenLao/stopwords`.

### 5.2. Lexical Extraction Results: Comparison with Baselines

We begin by comparing with previous topic-modeling approaches to bilingual lexicon extraction, namely, Vulić et al. [2011]. Using the automatically-created "gold-standard" lexicon, we evaluate methods by precision, defined as $\frac{|\{Gold(e,f)\} \bigcap \{Extracted(e,f)\}|}{\#Extracted}$.

Table III shows the precision of our proposed method compared with the baseline Cue and JS methods [Vulić et al. 2011]. All these methods first run our MLTM with $K = 400$ topics[5] on the partially-connected Japanese-English Wiki dataset, which consists of $14,033 + 4,087 = 18,120$ documents.

Our method extracts a total of 1,457 pairs using the bidirectional constraint in Eq. (8). This achieves a precision of 0.742. For comparison, we adjust the threshold $\delta$ (Option (a) discussed in Section 4.1), such that the Cue (Eq. (9)) and JS (Eq. (11)) methods give roughly the same number of extracted pairs as our proposed method. The resulting precision of Cue and JS are very poor, at 0.073 and 0.091, respectively. Vulić et al. [2011] reports that a large number of topics is necessary for good results, so we re-ran the baselines with $K = 2,000$, the suggested value in Vulić et al. [2011]. Despite the long runtime of MLTM for large $K$, the precision only increased to 0.104 and 0.123 for Cue and JS, respectively.

[4] `www.nltk.org/api/nltk.stem.html`

[5] MLTM hyperparameters are $\alpha = 50/K$ and $\beta = 0.01$ following Vulić et al. [2011].

Table III. Comparison with Baselines using the Japanese-English Part of Wiki Dataset

| System | Precision | #Extracted |
|---|---|---|
| Proposed (K=400) | 0.742 | 1,457 |
| Cue (K=400) | 0.073 | 1,400 |
| JS (K=400) | 0.091 | 1,400 |
| Cue (K=2,000) | 0.104 | 1,400 |
| JS (K=2,000) | 0.123 | 1,400 |
| IBM-1 | 0.521 | 1,400 |

It can be seen that our proposed method is much more effective at extracting bilingual lexicon, in particular in large-vocabulary datasets (the vocabulary size in [Vulić et al. 2011] is 7K and 9K in Italian and English, respectively). We have a hypothesis as to why Cue and JS depend on large $K$. Eq. (4) is a valid expression for $p(w^e|w^f)$ that makes little assumptions. We can view Eq. (9) as simplifying the first term of Eq. (4) from $p(w^e|t_k, w^f)$ to $p(w^e|t_k)$. Both probability tables have the same output space ($w^e$), so the same number of parameters is needed in reality to describe this distribution. By throwing out $w^f$, which has large cardinality, $t_k$ needs to grow in cardinality to compensate for the loss of expressiveness.

As an additional baseline, we directly run IBM Model 1 on the fully-connected Japanese-English comparable corpora, treating each document pair as a sentence pair. This IBM-1 baseline does not employ MLTM, and the score of a pair $(e,f)$ is defined as the average lexical probabilities obtained from IBM Model 1 in both directions. Interestingly, this baseline achieves a precision of 0.52, better than Cue and JS. But our proposed method still performs better, implying that the combination of existing word alignment models and MLTM attains good synergy.

### 5.3. Lexicon Extraction Results: Additional Languages

We now examine the effects of adding additional languages to the Japanese-English lexicon extraction. Table IV shows how precision improves as we add Chinese (3,494 comparable documents in addition to the original 18,120 Japanese-English corpora), as well as both Chinese and French (3,494+2,871=6,365 comparable documents). Since the number of extractions changes (because probability value changes affect the bidirectional constraint of Eq. (8)), we also manually evaluated precision (ManualPrec) on a fixed random set of 100 pairs.

From Table IV, we see that adding Chinese documents improves the (automatic) precision from 0.742 to 0.761. Adding both Chinese and French documents further improves results, with (automatic) precision gaining 3% (0.742 → 0.774) and manual precision gaining 9% (0.62 → 0.71). We observe these improvements because adding more languages and data improves the estimation of the MLTM. Specifically, in our bilingual extraction equation (Eq. (4)), more data can directly improve the estimation of the topic distribution $p(t_k|w^f)$; further, more data may also indirectly improve the estimation of the the topic-dependent bilingual lexicon $p(w^e|w^f, t_k)$ via better posterior inference results for input into the word alignment step. Note that the word alignment part is the same for the various systems in Table IV, so improvements come from better MLTM.

We also show results using only the fully-connected Japanese-English comparable corpus (Full-Japanese-English). This system only runs MLTM on 4,087 document pairs, and as a result, the precision is lower than the partially-connected case (0.612 vs. 0.742). This demonstrates our MLTM is effective in exploiting monolingual documents in estimating its parameters.

Table IV. Comparison of Proposed Method using Additional Languages in the **Wiki** Dataset

| System | Precision | ManualPrec | #Extracted |
|---|---|---|---|
| Full-Japanese-English | 0.612 | 0.51 | 1,745 |
| Japanese-English (= **Proposed** in Table III) | 0.742 | 0.62 | 1,457 |
| Japanese-English-Chinese | 0.761 | 0.64 | 1,365 |
| Japanese-English-Chinese-French | 0.774 | 0.71 | 1,372 |

*Note:* $K = 400$ and MLTM hyperparameters are same as described in Section 5.2.

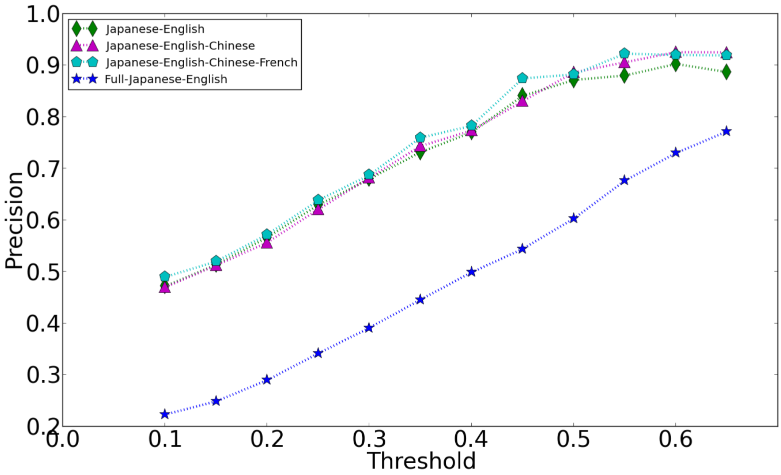

Fig. 4. Effect of using additional languages: precision vs. threshold curve.

Finally, we also compare the systems not by using the bidirectional constraint, but by varying the threshold $\delta$ (as discussed in Option (a) at the end of Section 4.1. Figure 4 shows how precision varies as we lower the threshold. Figure 5 plots the number of extracted pairs versus threshold on the same data. We observe a large overall gain in precision regardless of threshold as we move from fully-connected to partially-connected data, which corroborates the results in Table IV. The number of extractions are roughly similar for the various partially-connected systems, while fully-connected has slightly larger numbers (but lower precision). The differences between the various systems using partially-connected corpora does not seem very large, but this is not surprising given the large amount of monolingual Japanese documents (140,033) in our dataset compared to additional Chinese and French documents (around 3,000). Nevertheless, we do observe that the Japanese-English-Chinese-French system does indeed have the best precision curve.

### 5.4. Practical Issues: Model Selection and Runtime

The most important parameter in our approach is the number of topics $K$ in MLTM. As $K$ goes to one, the proposed approach becomes equivalent to running word alignment directly on comparable documents, treating each document pair as a sentence pair. As $K$ increases, the topic-aligned corpora become more fine grained, and the lexicon extraction precision improves. However, if $K$ is too large, then each word collection $C_{k,i}$ in the topic-aligned corpora becomes too small, and if the topic model incorrectly assigns translation pairs to different topics, it becomes impossible to extract it in the subsequent word alignment step.

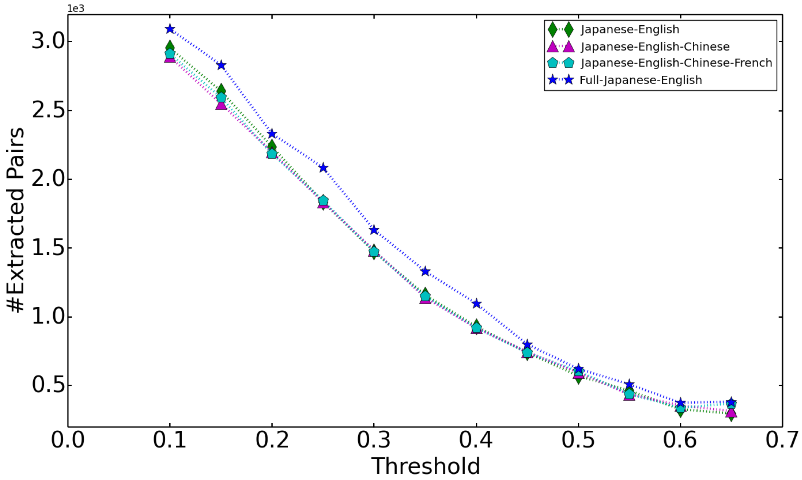

Fig. 5. Effect of using additional languages: number of extraction vs. threshold curve.

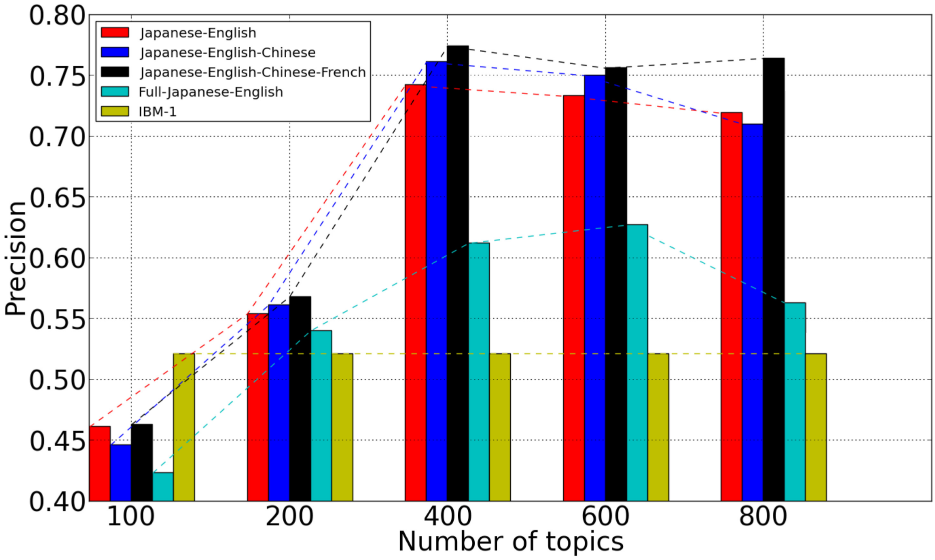

Fig. 6. Precision by number of topics ($K$).

First, we show how precision varies with different values of $K$ in Figure 6. We observe that for low values of $K$ (e.g., 100, 200), the precision is relatively low, around 0.4–0.6. The best precision is achieved with $K = 400$, followed closely by $K = 600$ and $K = 800$, all in the 0.7–0.8 range.

While it is expected that results vary somewhat by $K$, the important question is whether the best $K$ can be selected a priori in an unsupervised manner. Now we show that the per-word log-likelihood on the held-out dataset is effective for model selection. *Per-word log-likelihood*, which is widely used in the machine learning and statistics

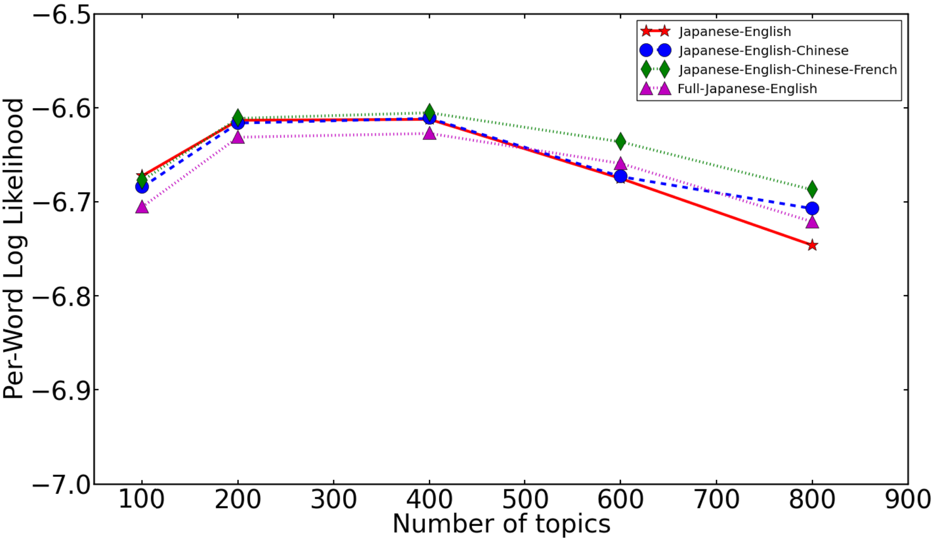

Fig. 7. Per-word likelihood by number of topics. Note that this figure correlates with Figure 6, suggesting per-word likelihood is a reasonable unsupervised metric for model selection.

community, is defined as the geometric mean of the inverse marginal probability of each word in the held-out (dev) set of documents $D_{dev}$:

$$likelihood_{pw} = \frac{\sum_{t \in D_{dev}} \log p(t|D_{train})}{\sum_{t \in D_{dev}} n_t}. \tag{12}$$

Here, $n_t$ denotes the number of words for the $t$th tuple of documents in dev corpus. Following Teh et al. [2006], we estimate $p(t|D_{train}) = \prod_l p(w_t^l|D_{train}) = \prod_l \prod_{w \in w^l} \sum_k \theta_{t,k} \beta_{k,w}^l$. The hidden variables $\theta_{t,k}$ and $\beta_{k,w}^l$ can be computed as Eq. (2) and Eq. (3). A higher per-word log-likelihood score indicates better performance [Blei et al. 2003; Hoffman et al. 2010; Teh et al. 2006].

Figure 7 summarizes our results for the model selection, plotting the per-word likelihoods of a 100-tuple held-out dev set. We observe that per-word likelihood successively picks out $K = 400$ as the best model for various setups, which generally corresponds to the best precision results in Figure 6.

Finally, we show the runtime of MLTM on a 2.4GHz Opteron CPU for varying $K$ in Figure 8. As expected, runtime increases with $K$: on datasets as large as ours, training with $K = 400$ takes approximately 4 days, and $K = 800$ takes 8 days. Time complexity of MLTM is $O(NK \sum_{m=1}^{M} \sum_{l=1}^{L} w_m^l)$, where N indicates the number of iterations; K, M, L denotes the number of the topics, size of corpus, and numbers of the languages; $w_m^l$ denotes the number of words in tuple $m$ written in language $l$.

The overall time for various systems is shown in Table V. First, note that MLTM time dominates the overall time for all systems, so the training time does not differ much among methods if we use the same number of topics in MLTM, but in practice, Proposed requires fewer numbers of topics, so it is much faster to train. Second, assuming the same number of topics, the breakdown of training time shows that Proposed is still relatively fast because both GIZA++ and Eq. (4) are fast. Comparing Eq. (9) of Cue to Eq. (4) of Proposed, we see that both need to compute $p(t_k|w^f)$, but the $\sum_k$ in Eq. (4)

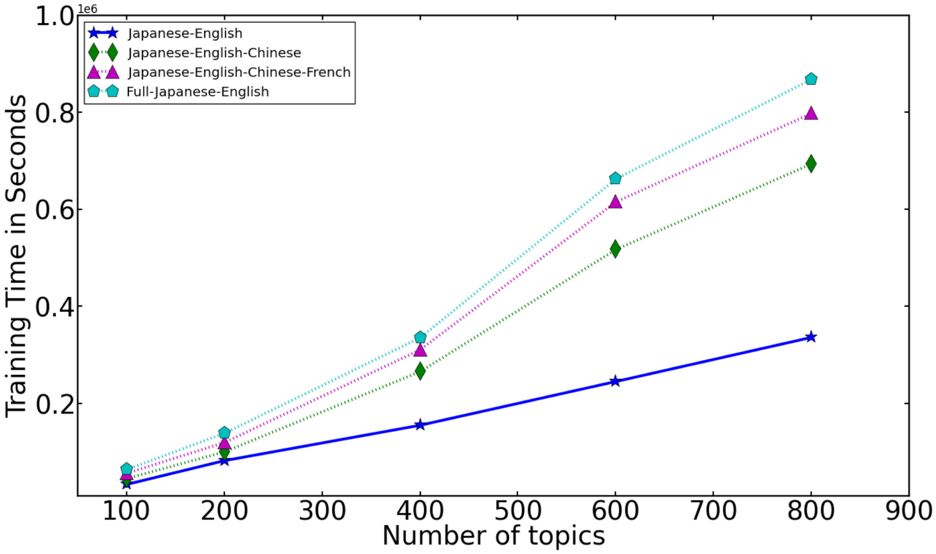

Fig. 8. Training time of MLTM by dataset and number of topics.

Table V. Wall-Clock Times in Seconds for Word Alignment (GIZA) and $p(w^e|w^f)$ calculation

| #Topic | MLTM (En-Ja) | Proposed(Giza++/Eq. (4)) | JS(Eq. (11)) | Cue (Eq. (9)) |
|---|---|---|---|---|
| 100 | 3.4e4 | 472 + 50 | 631 | 361 |
| 200 | 8.2e4 | 491 + 99 | 1,312 | 720 |
| 400 | 1.6e5 | 608 + 202 | 2,345 | 1,031 |
| 600 | 2.4e5 | 815 + 279 | 3,729 | 1,424 |
| 800 | 3.3e5 | 830 + 371 | 4,216 | 2,043 |

*Note:* Overall time for Proposed is the training time of MLTM, word alignment (Giza++), plus Eq. (4); for Cue it is the training time of MLTM plus Eq. (9); for JS it training time of MLTM plus Eq. (11). Here, Eq. (9) and Eq. (11) are computed in parallel with 100 threads. MLTM (En-Ja) denotes the training time of multilingual topic model on English and Japanese corpus. The training time of multilingual topics for different settings is shown in Figure 8.

tends to be faster because $p(w^e|w^f, t_k)$ in Eq. (4) tends to be sparse, while $p(w^e|t_k)$ in Eq. (9) is dense.

### 5.5. Detailed Analysis of Results

First, we present some interesting examples of bilingual lexicon found by our method. In particular, due to the topic-dependent translation probabilities $p(w^f|w^e, t_k)$, we are able to tease out the translation of polysemous words. In Table VI, we look at potentially polysemous English words $w^e$ and list the highest-probability Japanese translations $w^f$ conditioned on different $t_k$. We found many promising cases where the topic identification helps divide the different senses of the English word, leading to the correct Japanese translation achieving the highest probability.

Second, we perform an error analysis on the results of the Full-Japanese-English system. We find that the proposed method makes several types of incorrect lexicon extractions. First, *word segmentation* "errors" on Japanese could make it impossible to find a proper English translation (e.g., 高市皇子 should translate to "Prince-Takechi," but the system proposes "Takechi"). Second, an unrelated word pair $(w^e, w^f)$ may be

Table VI. Examples of Topic-Dependent Translations Given by $p(w^f|w^e, t_k)$

| English | Japanese1(gloss), Japanese2(gloss) |
|---|---|
| interest | 関心 (a sense of concern), 利息 (a charge of money borrowing) |
| count | 数え(act of reciting numbers), 伯爵 (nobleman) |
| free | 自由 (as in “free” speech), 無料 (as in “free” beer) |
| blood | 血縁 (line of descent), 血 (the red fluid) |
| demand | 需要 (as noun), 要求 (as verb) |
| draft | 提案 (as verb), 草稿 (as noun) |
| page | ページ (one leaf of e.g. a book), 侍童 (youthful attendant) |
| staff | スタッフ (general personel), 参謀 (as in political “chief of staff”) |
| director | 長官 (someone who controls), 理事 (board of directors) 監督 (movie director) |
| beach | 浜 (area of sand near water), 海水浴 (leisure spot at beach) |
| actor | 役者 (theatrical performer), 俳優 (movie actor) |

*Note:* The top portion shows examples of polysemous English words. The bottom shows examples where English is not decisively polysemous, but indeed has distinct translations in Japanese based on topic.

Table VII. Counts of Various Error Types

| Word Segmentation Error | Incorrect Topic | Correct Topic, Incorrect Alignment | Reason Unknown |
|---|---|---|---|
| 14 | 29 | 40 | 7 |

incorrectly placed in the same topic, leading to an *incorrect topic* error. Third, even if $(w^e, w^f)$ intuitively belong to the same topic, they may not be direct translations; an extraction in this case would be a *correct topic, incorrect alignment* error (e.g., もんじゃ焼き, a particular panfried snack, is incorrectly translated as “panfry”).

Table VII shows the distribution of error types by manual classification. *Incorrect alignment* errors are the most frequent, implying the topic models are doing a reasonable job of generating the *topic*-aligned corpus. The amount of *incorrect topic* is not trivial, though, so we would still imagine more advanced topic models to help. *segmentation* errors are in general hard to solve, even with a better word segmenter, since one-to-one cross-lingual word correspondence is not consistent. We believe the solution is a system that naturally handles multiword expressions [Baldwin 2011].

Since word alignment errors were frequent, we conduct an additional experiment to compare several popular word alignment methods in statistical machine translation: (a) Giza-vb, a modification of IBM Models training using Variational Bayes EM learning [Riley and Gildea 2010]; (b) Giza-L0, a modification of IBM Models to generate sparser alignments using approximate L0-norm optimization [Vaswani et al. 2012]; and (c) Berkeley Aligner, which enforces agreement in bidirection word alignment [Liang et al. 2006]. Figure 9 shows the lexicon extraction results using different alignment tools. While the differences are in general not very large, we observe that the Berkeley Aligner appears slightly better than all the other IBM model variants, implying that bidirectional constraints may be helpful in this kind of topic-aligned data.

Finally, we attempt to quantitatively explain why our approach of topic-aligned corpus is more effective than directly extracting bilingual lexicon from topic model parameters, as suggested by Mimno et al. [2009] and executed by Vulić et al. [2011]. To do so, we first take the word-topic distribution from MLTM ($K = 400$ setup) and plot the number of topics each word type $w$ may appear in it (i.e., nonzero probabilities according to $p(w|t)$). Figure 10 shows that the plot of the number of word types that have $x$

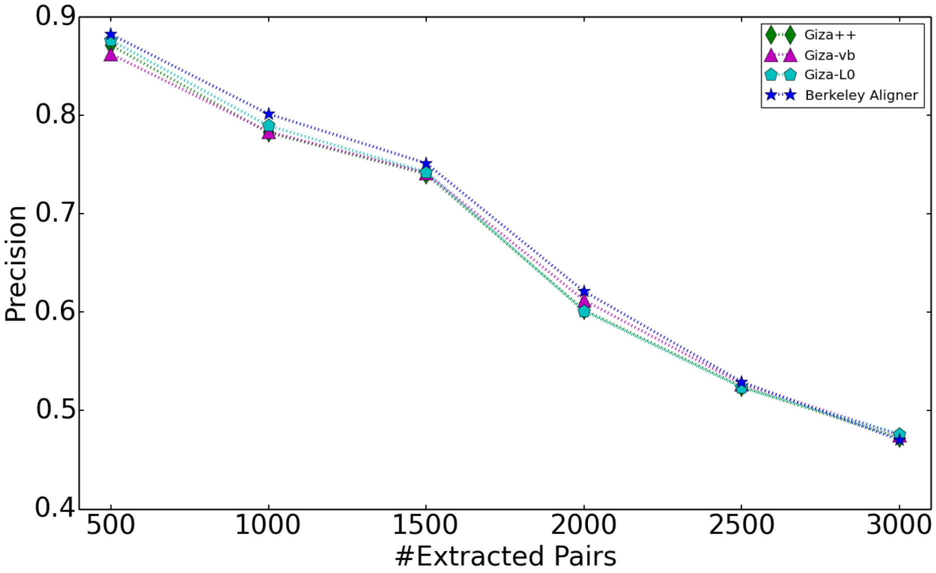

Fig. 9. Comparison of different word alignment tools: precision vs. #extracted pairs curve. The Dirichlet parameter $\alpha$ is set to 0.01 in Giza-vb; parameters of Giva-L0 are set to default, and bidirectional joint IBM1 settings are used in Berkeley Aligner.

number of topics behaves like a power-law. This suggests that it is not easy to directly extract a lexicon by taking the cross-product $(w^f, w^e)$ of the top-$n$ words in $p(w^f|t_k)$ and $p(w^e|t_k)$ for the same topic $t_k$, as suggested by Mimno et al. [2009]. The majority of words are grouped in the same few topics, making it difficult to discern the actual translations. When we attempt to do this using top-2 words per $p(w^f|t_k)$ and $p(w^e|t_k)$, we could only obtain precision of 0.37 for 1,600 extractions. The methods of Vulić et al. [2011] are essentially based on similar information, albeit with probability weighting. This skewed distribution similarly explains the poor performance of the **Cue** and **JS** baselines.

On the other hand, after constructing the topic-aligned corpora (Step 3 of Figure 3), we compute the ratio of distinct English word types versus distinct Japanese word types for each topic. If the ratio is close to one, this means the partition into topic-aligned corpora effectively separates the skewed word-topic distribution of Figure 10 into more uniformly-distributed word collections. We found that the mean ratio averaged across topics is low at 1.721 (variance is 1.316), implying that within each topic, word alignment is relatively easy.

Some examples of how the proposed multilingual model reduces translation errors are shown in Table VIII. Taking "music" as an example, if we only use the English-Japanese corpus, we erroneously find that 歌唱 (to sing) has high translation probability; this is understandable, though, because the words are roughly in the same topic. However, with additional language data (Chinese, French), the topic distributions become more precise, so the error disappears and the correct translation 音楽 (music) is left with higher probability.

### 5.6. Hybrid Systems: Bootstrapping for Context Vector Approaches

Context vector approaches are one of the most established ways to extract lexicon from comparable corpora. These approaches require a seed dictionary, which could be provided by our approach. We evaluate this hybrid approach as follows.

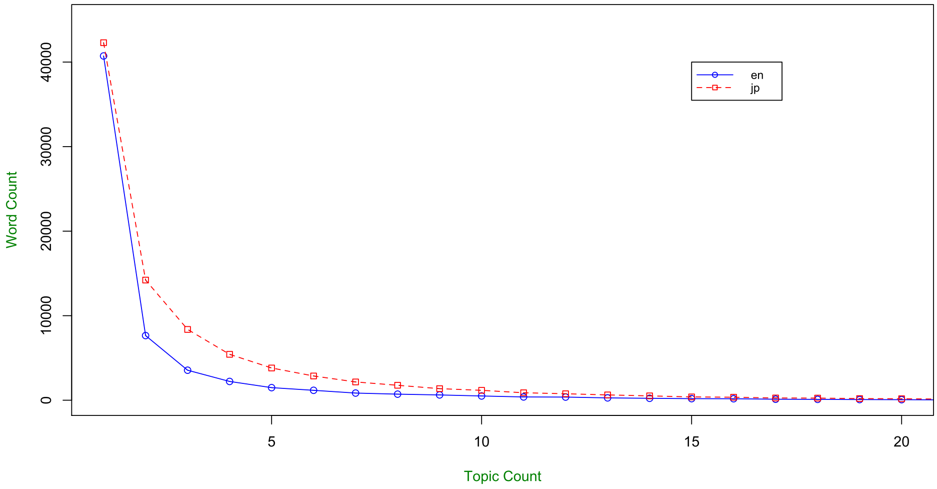

Fig. 10. Power-law distribution of number of word types with X number of topics.

Table VIII. Error Analysis on Multilingual Case

| English Words | English-Japanese | +Chinese | +Chinese+French |
|---|---|---|---|
| music | 音楽 [music] (0.323)<br>歌唱 [sing] (0.203) | 音楽 [music] (0.442) | 音楽 [music] (0.445) |
| ikoma | 生駒 [ikoma] (0.497)<br>石切 [ishikiri] (0.205)<br>近鉄 [train] (0.272) | 生駒 [ikoma] (0.574)<br>近鉄 [train] (0.252) | 生駒 [ikoma] (0.619)<br>近鉄 [train] (0.254) |
| yoshino | 吉野 [yoshino] (0.389) | 吉野 [yoshino] (0.603) | 吉野 [yoshino] (0.373)<br>吉野山 (0.204) |

*Note:* The words colored in grey indicate translation errors; the words in [] are the corresponding translation; the numbers in () are the translation probabilities.

- First, the 1,457-pair high-precision dictionary extracted by our proposed method in Table III is used as seed for the context vector approach of Rapp [1995]. This hybrid system is called *WikiSeeds*, and the resulting precision is reported in Figure 11.
- As comparison, we run the same [Rapp 1995] method with different amounts of "gold seeds," the purpose being to observe how many manual gold seed translations are necessary to attain the extraction result of our purely unsupervised WikiSeeds system. In particular, for a fair comparison, for cases under 1,457 seeds (GoldSeeds500 and GoldSeeds1000), we randomly sample 500 and 1,000 unique Japanese vocabularies in WikiSeeds and look up their corresponding English translation in the gold standard lexicon described in Section 5.1. For cases above 1,457 seeds (GoldSeeds1500 and GoldSeeds3000), we use the entire gold standard lexicon associated with the 1,457 vocabulary, with additional translation pairs randomly sampled from the gold standard lexicon.

From Figure 11, we find that WikiSeeds outperforms both 500 and 1,000 gold seeds in precision across all numbers of extracted pairs. As expected, a roughly equal number of gold seeds (1,500) should outperform WikiSeeds, but the differences are not large. Such observations imply that our extracted seeds can be used in the context vector approach when there are no large seeds existing in certain language pairs. The

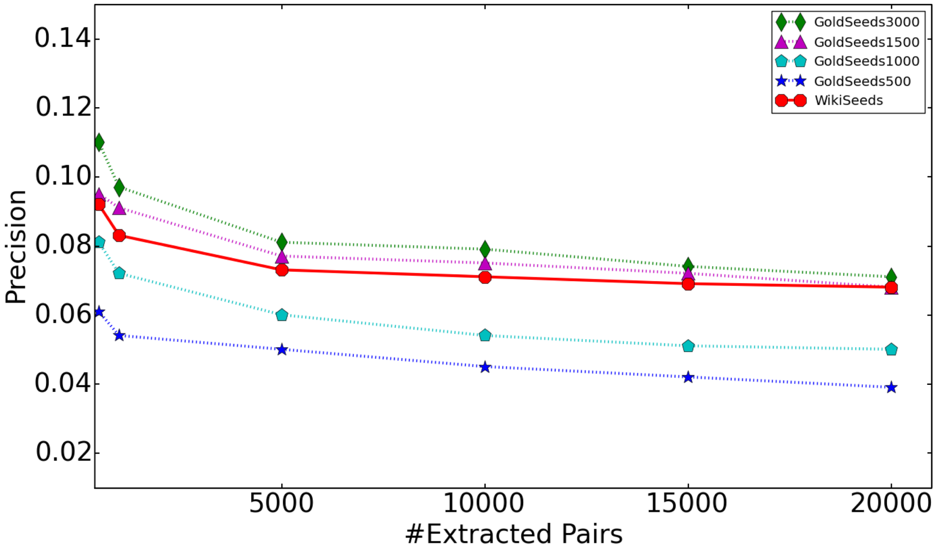

Fig. 11. Comparison of different seeds for the context vector approach: precision vs. #Extracted pairs curve. Note that GoldSeeds# denotes the size of "gold seeds".

code of the context vector approach is available at `https://bitbucket.org/allenLao/context_based_model_for_dic/src`.

## 6. CONCLUSION

We propose an effective way to extract bilingual dictionaries using a novel combination of topic modeling and word alignment techniques. The key innovation is the conversion of a comparable *document*-aligned corpus into a parallel *topic*-aligned corpus, which allows word alignment techniques to learn topic-dependent translation models of the form $p(w^e|w^f, t_k)$. The main advantages of our approach are that (1) it does not require any bilingual seed dictionary, and (2) it can effectively exploit comparable corpora consisting of documents in more than two languages.

Our large-scale experiments demonstrate that the proposed framework outperforms existing baselines under both automatic metrics and manual evaluation. Further, we show improvements in the precision of our Japanese-English lexicon as we include more languages (i.e., Chinese and French) to the comparable corpora. To facilitate further work in this area, all preprocessed data and topic modeling code is available at `https://bitbucket.org/allenLao/topic-modeling-gibbs`.

While our framework is purely unsupervised in the sense that it requires no seed dictionary, we can imagine several interesting extensions if such a seed dictionary were available. First, the seeds could be used as a prior for the multilingual topic model, for instance by employing the Dirichlet tree prior [Andrzejewski et al. 2009; Hu et al. 2014]. Second, the seed translation could also be incorporated into the word alignment step (as supervised alignments) to improve performance of the topic-dependent translations, $p(w^f|w^e, t_k)$. In general, the modularity of our method makes it relatively flexible to incorporate additional resources and knowledge into the lexicon extraction process.

## ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their comments.

## References

Ahmet Aker, Monica Lestari Paramita, Marcis Pinnis, and Robert Gaizauskas. 2014. Bilingual dictionaries for all EU languages. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*.

Daniel Andrade, Takuya Matsuzaki, and Jun'ichi Tsujii. 2011. Effective use of dependency structure for bilingual lexicon creation. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*. Lecture Notes in Computer Science, Vol. 6639, Springer, 80–92.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*. ACM, 25–32.

Yoshiaki Arai, Tomohiro Fukuhara, Hidetaka Masuda, and Hiroshi Nakagawa. 2008. Analyzing interlanguage links of Wikipedias. In *Proceedings of the Wikimania Conference*.

Timothy Baldwin. 2011. MWEs and topic modelling: Enhancing machine learning with linguistics. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World (MWE'11)*. Association for Computational Linguistics, Stroudsburg, PA, 1–1. http://dl.acm.org/citation.cfm?id=2021121.2021123.

David M. Blei and Michael I. Jordan. 2006. Variational inference for Dirichlet process mixtures. *Bayesian Anal. 1*, 1, 121–143.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Machine Learn. Res. 3*, 993–1022.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*.

Peter F. Brown, Vincent J. Della Pietra, Stephen Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics 19*, 2, 263–311.

Sarath Chandar A. P., Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'14)*.

Hal Daume III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'11)*. 407–412.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal Wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. ACM, 513–522.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*. 1–7.

Kevin Duh, Ching-Man Au Yeung, Tomoharu Iwata, and Masaaki Nagata. 2013. Managing information disparity in multilingual document collections. *ACM Trans. Speech Lang. Process. 10*, 1, (March 2013), Article 1. http://doi.acm.org/10.1145/2442076.2442077.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*.

Pascale Fung and Yuen Yee Lo. 1998. Translating unknown words using nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING'98)*.

Eric Gaussier, J. M. Renders, I. Matveeva, C. Goutte, and H. Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*. 526–533.

Tim Gollins and Mark Sanderson. 2001. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th ACM Conference of the Special Interest Group in Information Retrieval (SIGIR'01)*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'08)*. 771–779.

Gregor Heinrich. 2004. Parameter estimation for text analysis. Technical Report. rsonix GmbH and University of Leipzig.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent dirichlet allocation. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'10)*. 856–864.

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*. Volume 1: Long Papers, 1166–1176.

Jagadeesh Jagarlamudi and Hal Daume. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR'10)*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING'12)*. 1459–1474.

Philipp Koehn. 2010. *Statistical Machine Translation* (1st Ed.). Cambridge University Press, New York, NY.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*.

Hong-seok Kwon, Hyeong-won Seo, and Jae-hoon Kim. 2013. Bilingual lexicon extraction via pivot language and word alignment tool. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*. 11–15.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*.

Percy Liang, Taskar Ben, and Klein Dan. 2006. Alignment by agreement. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT'06)*. 104–111.

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 17th Conference on Computational Natural Language Learning (CoNLL'13)*. 212.

Bernardo Magnini, Carlo Strapparava, Fabio Ciravegna, and Emanuele Pianta. 1994. Multilingual lexical knowledge bases: Applied WordNet prospects. In *Proceedings of the International Workshop on the Future of the Dictionary*.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the 47th Meeting of the Association for Computational Linguistics (ACL'09)*.

David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*.

Thomas Minka. 2000. Estimating a Dirichlet distribution. Microsoft Research.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'11)*. 529–533.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International Conference on the World Wide Web (WWW'09)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist. 29*, 1, 19–51. DOI:http://dx.doi.org/10.1162/089120103321337421.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics—Human Language Technologies (NAACL/HLT'09)*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*.

Philip Resnik, Douglas Oard, and Gina Levow. 2001. Improved cross-language retrieval using backoff translation. In *Proceedings of the 1st International Conference on Human Language Technology Research*. 1–3.

Darcey Riley and Daniel Gildea. 2010. Improving the performance of GIZA++ using variational bayes. Technical Report. The University of Rochester, Computer Science Department.

Fatiha Sadat, Herve Dejean, and Eric Gaussier. 2002. A combination of models for bilingual lexicon extraction from comparable corpora. In *Proceedings of the Seminaire Papillon*.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*. 24–36.

Yee W. Teh, David Newman, and Max Welling. 2006. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'06)*. 1353–1360.

Ashish Vaswani, Liang Huang, and David Chiang. 2012. Smaller alignment models for better translations: Unsupervised word alignment with the l 0-norm. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*. 311–319.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*. 18–21.

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 Association for Computational Linguistics: Human Language Technologies (ACL'11)*. 479–484.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-AFNLP'09)*. 154–162.

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 (ACL'10)*. 1128–1137.

Received May 2014; revised September 2014; accepted November 2014