Inverted Bilingual Topic Models for Lexicon Extraction from Non-parallel Data

Tengfei Ma
IBM T. J. Watson Research Center
Tengfei.Ma1@ibm.com

Tetsuya Nasukawa
IBM Research-Tokyo
NASUKAWA@jp.ibm.com

Abstract

Topic models have been successfully applied in lexicon extraction. However, most previous methods are limited to document-aligned data. In this paper, we try to address two challenges of applying topic models to lexicon extraction in non-parallel data: 1) hard to model the word relationship and 2) noisy seed dictionary. To solve these two challenges, we propose two new bilingual topic models to better capture the semantic information of each word while discriminating the multiple translations in a noisy seed dictionary. We extend the scope of topic models by inverting the roles of "word" and "document". In addition, to solve the problem of noise in seed dictionary, we incorporate the probability of translation selection in our models. Moreover, we also propose an effective measure to evaluate the similarity of words in different languages and select the optimal translation pairs. Experimental results using real world data demonstrate the utility and efficacy of the proposed models.

1 Introduction

Bilingual lexicons play an important role in cross-lingual information retrieval and text mining tasks. However, there is often no existing dictionary for some technical data or low-resourced language pairs. For example, in some special domains, there are always novel words or new expressions emerging, and a generic dictionary can hardly keep up with them. Thus, automatically extracting translation pairs [Andrade *et al.*, 2010],[Bollegala *et al.*, 2015] has attracted a lot of attention.

As one of the most successful methods for latent semantic analysis in the past years, topic models (e.g. Latent Dirichlet Allocation [Blei *et al.*, 2003]) have demonstrated usefulness for lexicon extraction [Vulić *et al.*, 2011; Vulic and Moens, 2013; Mimno *et al.*, 2009; Ni *et al.*, 2009]. A classical bilingual topic model utilizes the alignment relationship of documents in different languages, and projects them into a shared latent semantic space. The basic idea is that an aligned document pair should have the same topic distribution, thus the topics in different languages can be connected. Then the shared topics could be used to calculate

the similarity of words in different languages, e.g. through Bayesian rules and K-L divergence [Vulić *et al.*, 2011]. Despite their success, there are two problems of the bilingual topic models in lexicon extraction.

Firstly, almost all previous cross-lingual topic models are developed for document-aligned data, however in practice corpora are often not aligned. Extracting lexicons from non-parallel corpora is more valuable yet far more challenging. Although topic models could easily integrate the document relationship [Dietz *et al.*, 2007], [Chang and Blei, 2009], they have difficulties representing the word relationship, which is needed for lexicon extraction from non-parallel data.

Secondly, to our best knowledge, few of existing bilingual topic models considered the probability of the multiple translations in their models, or handled the noise issue in their dictionary. Although some initial effort has been made, for example, Boyd-Graber and Blei [2009] integrated the prior of word matching to the bilingual topic models in non-parallel data. However, their results indicate the model had no effect of finding new word translations.

In this work, we propose two novel bilingual topic models to extend their applications to lexicon extraction from non-parallel data. We reverse the roles of documents and words to represent each word as a pseudo document and then model the words rather than original documents. To be specific, we use inverted indexing to represent a word as a list of documents where it occurs. After obtaining the pseudo documents, each word is then modeled as a topic distribution. Different from some previous work related to cross-lingual inverted indexing [Søgaard *et al.*, 2015], we do not consider connections between documents but only between words. Each translation pair is assumed to own the same topic distribution. In this way, topics in different languages can also be connected.

Next, in order to solve the problem of noisy translations in the seed dictionary, in the generative process of our models we select only one proper translation from all candidates in the seed dictionary probabilistically instead of using all of them equivalently. In addition, in the original bilingual LDA[Vulić *et al.*, 2011] only the aligned document pairs are modeled. In our case, however, we only have a subset of words included in the seed dictionary. The remaining words do not have any connection with words in other languages, but they are also included in our models, making the models essentially semi-supervised. We use Gibbs sampling for pos-

terior inference and get the similarity between words across languages on the basis of their topic distributions. In contrast to traditional cosine similarity and KL divergence, we define the similarity measure as the probability of a word generating another. Given a word in a source language, the word with the most similar topic distributions in the target language is then regarded as its translation.

To summarize, it is worthwhile to highlight the following contributions of the proposed model:

- We adopt the inverted indexing technique to extend the scope of topic models to the task of lexicon extraction from non-parallel data.

- We extend the classical bilingual LDA by incorporating the probability of multiple translations in the model, thus solving the noise issue for the seed dictionary.

- We propose a new similarity measure from the conditional generating probability for two words across languages to handle the correlation of topics. Experimental results demonstrate its advantage over other traditional measures.

The following sections are organized as follows: we first review related work in Section 2. Then we propose our problem definition and new models in Section 3. Section 4 explains how we use our topic model to measure the word similarity. In Section 5 we describe our experiments. At last we conclude our work in Section 6.

2 Related Work

We have introduced the topic modeling methods for lexicon extraction. In this section, we will focus more on the history of **lexicon extraction from non-parallel data**.

The most well-established work on lexicon extraction is based on word alignment in parallel datasets. We can easily use a statistical machine translation system [Lopez, 2008] to induce translation pairs from parallel data. However, parallel data is not plentiful for all language pairs or all domains. This restricts the usefulness of these methods.

Lexicon extraction from non-parallel data was pioneered by [Rapp, 1995] and [Fung and Yee, 1998]. Instead of parallel/comparable documents, they use a seed dictionary as the pivots. Generally, this kind of approach can be factorized into two steps: 1, construct a context vector for each word, and 2, compute the context similarities on the basis of pivot words (i.e., seed dictionary entries). A common hypothesis is that a word and its translation tend to occur in similar contexts. Previous research has defined various correlation measures to construct a context vector representation for a word, including tf-idf [Fung and Yee, 1998] and pointwise mutual information (PMI) [Andrade et al., 2010]. As for the similarity computation, cosine similarity [Fung and Yee, 1998], non-aligned signatures (NAS) [Shezaf and Rappoport, 2010], and Johnson-Shannon divergence [Pekar et al., 2006], etc. can be used.

The context similarity-based models rely on the quality and the size of seed dictionaries. When a seed dictionary is small, the context vector will be too sparse and the similarity measure is not accurate enough. Recent work has used graph-based methods to propagate the seed dictionaries [Laws et al., 2010; Tamura et al., 2012]. There are also some methods that project the word vectors in different languages into the same low-dimensional space, such as linear transformation for cross-lingual word embedding [Mikolov et al., 2013]. Our motivation is similar to the graph-based and word embedding-based models in that we use a topic model to represent each word as a topic distribution in order to avoid the sparseness of context vectors. However, while the previous approaches generally just select the reliable translations as seeds [Mikolov et al., 2013; Haghighi et al., 2008], we assume our seed dictionary is noisy. We add the probability of existing translations as a new latent variable and make our model more robust and generalizable. Most recently, Duong et. al.[2016] and Zhang et. al. [2017] propose new bilingual word embedding methods to deal with the noise of the seed dictionary. But the word embedding methods cannot explicitly interpret the uncertainty of the multiple translations. As generative models, our topic models have obviously better interpretability for translation selection.

3 Proposed Model

In this section, we start by making a formal definition of our problem. Then we will describe the details of our new model, following a brief introduction to the background knowledge of cross-lingual topic models.

3.1 Problem Definition

Assume that we are given only two mono-lingual corpora in different languages, C^e and C^j. They are neither sentence-aligned nor document-aligned, but are in the same domain. The documents in C^e are noted as $\{d_i^e\}$ for $i = 1, ...N^e$ where N^e is the number of documents in C^e; while the documents in C^j are noted as $\{d_i^j\}$ for $i = 1, ...N^j$ where N^j is the number of documents in C^j. Other than the data corpora, we also have a set of seed dictionaries. We assume that the seed dictionary comes from the generic domain, and is noisy. It means one term in the seed dictionary may have several translations, within which some translations are not correct in this domain.

Now given a term in the source language t^j which appears in C^j, we want to find the most possible translation term in C^e.

3.2 Model Description

Inverted Indexing and Topic Models

Our approach to lexicon extraction is to first use topic models to model the cross-lingual data and obtain the topic distribution of each word. Then we can compare the topic distributions to compute the word similarities and get the translation.

In a conventional topic model, only the documents are represented by topic distributions, while the topic distribution for a word is not explicit. In addition, it is relatively easy to model document pairs or other document relationships by various topic models, as discussed earlier. However, in our problem setting, we only have a seed dictionary and non-parallel data corpora, so it is difficult to find document rela-

tionships but easy to get word translation pairs. The motivation is that if we can transfer a word into a pseudo document, we can utilize the word relationship in seed dictionaries.

In order to implement this idea, we invert the document-word index so that a word is constructed by a list of document IDs. If we assume a word w that appears in d_1 twice, d_2 once, and d_3 once, it is represented as (d_1, d_1, d_2, d_3). We also keep the word frequency in this representation.

As far as we know, this is the first work to integrate inverted indexing and topic models. We can also use other ways to construct the pseudo documents, such as using neighbor words. However, there are far fewer documents than context words, so we can reduce the computational cost. In addition, using inverted indexing-based representation enables us to easily calculate $p(d|w) = \sum_z p(d|z)p(z|w)$ from the topic distributions. So we can easily achieve the conditional probability of all documents when given a search term in another language. This might be useful for cross-lingual information retrieval tasks (although this is not our focus in this paper).

To avoid confusion, in the following sections we use "word" to indicate the pseudo document in topic models and use "document" to indicate the basic element in a pseudo document. Thus, a topic is a distribution of documents, and a word is a mixture of topics. That is to say, we have reversed the roles of words and documents in conventional topic models.

Once we obtain the pseudo documents, we can use them to train a Bilingual LDA model[Vulić *et al.*, 2011]. If two words are translations to each other, they are assumed to have similar topic distributions. The problem is that we only have a subset of words that are translated, and a word in a seed dictionary may have several translations. Therefore, first we need to construct one-to-one word pairs, the same as what Bilingual LDA does for documents.

Intuitively, it is not a good choice to make all translations modeled because the different translations will own the same topic distribution if a word has polysemy. Instead, we just select the most frequent term in the translation list to form a translation pair. Then, for all translation pairs, we use the same model as the Bilingual LDA. Words that do not have translations are modeled together using the original LDA.

- For each translation pair t^j, t^e,
    - Sample a topic distribution $\theta \sim \text{Dirichlet}(\alpha)$
- For each word t_l ($l \in \{j, e\}$) without translation,
    - Sample a topic distribution $\theta_l \sim \text{Dirichlet}(\alpha)$

Following this process, we sample the topics for each token d^e and d^j from θ and then draw documents from the topic. We also performed experiments to try out another way to obtain translation pairs. Instead of just selecting one translation, we randomly select a translation in each sample iteration, which means we finally use all the translations over all iterations. We call this model **BiLDA_all**, while the previous one is called **BiLDA**. A comparison of the two models is given in Section 5. They are both used as our baseline systems.

If we select just one translation, there is a risk of losing a lot of information. This is especially problematic when the

seed dictionary is not large, as the lost information will cause a serious performance decrease. On the other hand, using all translations without discrimination is not ideal either, as we discussed previously. We therefore came up with a solution to properly select the correct translation for each word.

ProbBiLDA

We developed two approaches to model the probability of translation selection. The first approach is to add a selection variable for each token (i.e. each document) d^j in word t^j, such that the topic distribution of each t^j is a mixture of its translations. This is similar to the idea of citation models [Dietz *et al.*, 2007], which model the probability of citation as the influence rate. The difference is that we have two sets of topics for the two respective languages. We do not directly share the topics of the "cited" pseudo document, opting instead to use the "cited" topic distribution to sample a new topic in its own language. We call this model **ProbBiLDA** (probabilistically linked bilingual LDA). The generative process of the target language e is same as the original LDA. To save space, we only list the generative process for source language j as follows.

- For each topic $z^l \in \{1, ...K\}$ (language $l \in \{j, e\}$):
    - Sample a document distribution $\phi_{z^l}^l \sim \text{Dir}(\beta)$
- For each word t^j:
    - If t^j not in seed dictionary::
        * Sample a topic distribution $\theta_{t^j} \sim \text{Dir}(\alpha)$
        * For each position i in the word:
            · Sample a topic assignment $z_i^j \sim \text{Multi}(\theta_{t^j})$
    - If t^j has S translations:
        * Draw a probability distribution $\psi_{t^j} \sim \text{Dir}(\alpha_\psi)$ over all translations
        * For each position i in the word:
            · Sample a translation $s_i \sim \text{Multi}(\psi_{t^j})$ from the S translations in language e
            · Draw a topic $z_i^j \sim \text{Multi}(\theta_{s_i})$
            · Draw a document $d_i^j \sim \text{Multi}(\phi_{z_i^j}^j)$

BlockProbBiLDA

Another way to model the probability of translations is to add the probability variable to the word itself instead of to each document in that word. That is to say, we select a translation for the whole word, and all the documents in that word follow the same topic distribution. For example, a word $t = (d1, d2)$ has three translations $t1, t2, t3$. If we use the ProbBiLDA, the topic of each document in word t is sampled from different translations, e.g., $z_{d1} \sim \theta_{t2}$ and $z_{d2} \sim \theta_{t3}$. However, in the new model, we require that all documents in t can only select one same translation in each iteration. If $t2$ is selected as the translation of t, we must have $z_{d1} \sim \theta_{t2}$ and $z_{d2} \sim \theta_{t2}$.

As all the documents select translations together like a block, we call this model **BlockProbBiLDA**. This model is essentially more similar to the original Bilingual LDA. Compared to Bilingual LDA, it does not fix the translation pairs but rather assigns a prior to each translation. Compared to
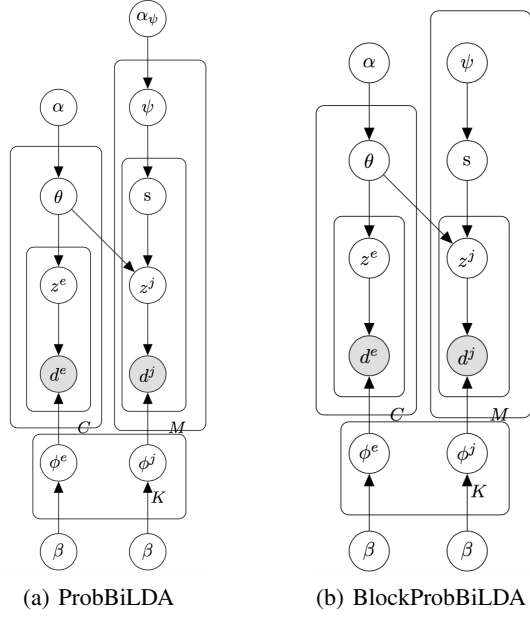
(a) ProbBiLDA   (b) BlockProbBiLDA

Figure 1: Graphic Representation of the two new Bilingual topic models. Note that in the figure we only represent the generative process for words in seed dictionaries; while for other words, they are modeled as same as an original LDA.

the generative process of ProbBiLDA, it only changes the position of s and uses a uniform prior distribution ψ instead of Dirichlet prior. The graphical representation of BlockProb-BiLDA is shown in Figure 1; and its generative process for words in source language is as follows:

- For each word t^j:
  - If t^j not in seed dictionary:
    * Sample a topic distribution $\theta_{t^j} \sim Dir(\alpha)$
    * For each position i in the word:
      · Sample a topic assignment $z_i^j \sim \text{Multi}(\theta_{t^j})$
  - If t^j has S translations.
    * Sample a translation $s \sim Multi(\psi_{t^j})$ from the S translations, where ψ_{t^j} is a uniform distribution over all translations
    * For each position i in the word:
      · Draw a topic $z_i^j \sim Multi(\theta_s)$,
      · Draw a document $d_i^j \sim Multi(\phi_{z_i^j}^j)$

3.3 Posterior Inference

For both of the two new models, we use collapsed Gibbs sampling to approximate the posterior. We iteratively update latent variables (including topic assignment z) given other variables.

Posterior Inference for ProbBiLDA

In the model of ProbBiLDA, for each document $d_{t^j,i}^j$ in a word t^j, we assume that it selects a translation word c in target language e, i.e. it is drawn from the topic distribution of this word. Given the translation selection, and other topic assignments, we sample the topic for document $d_{t^j,i}^j$ according to:

$$p(z_i^j = k | z_{-i,t^j}^j, s_i = c, d_{t^j,i}^j = n, \theta) \tag{1}$$
$$\propto \frac{nmk(c,k)+cmk(c,k)+\alpha-1}{nm(c)+cm(c)+K*\alpha-1} * \frac{nkvj(k,n)+\beta-1}{nkj(k)+V_j*\beta-1}$$

where $nmk(c,k)$ denotes the number of documents in word c that are assigned to topic k; $nm(c)$ denotes the total number of documents in word c; $cmk(c,k)$ is the number of documents with topic k in language e that select c as the translation of its associated word; and $cm(c)$ is the total number of documents in language e with translation selection c. $nkvj(k,n)$ is the number of times when document n is assigned to topic k in language j; and accordingly $nkj(k)$ is the sum of $nkvj(k,n)$ over all documents in language j; V_j is the total number of documents in language j.

Given these topic assignments, we can sample the translation selection:

$$p(s_i = c | s_{-i}, z_i^j = k, d_{t^j,i}^j = n, \theta) \tag{2}$$
$$\propto \prod_i \frac{nmk(c,k)+cmk(c,k)+\alpha-1}{nm(c)+cmk(c)+K*\alpha-1} * \frac{nms(t^j,c)+\alpha_\psi-1}{nm(t^j)+S(t^j)*\alpha_\psi-1}$$

where $nms(t^j,c)$ denotes the number of documents in word t^j which selects translation c; and $S(t^j)$ is the number of translation candidates for word t^j.

The above sampling scheme is for the source language. While for target language, we only need to care about the topic assignments.

$$p(z_i^e = k | z_{-i,t^e}^e, d_{t^e,i}^e = n, \theta) \tag{3}$$
$$\propto \frac{nmk(t^e,k)+cmk(t^e,k)+\alpha-1}{nm(t^e)+cm(t^e)+K*\alpha-1} * \frac{nkve(k,n)+\beta-1}{nke(k)+V_e*\beta-1}$$

where the denotations of the variables are similar to the ones defined in (1).

Given all the topic assignments, we can then derive the topic distribution $\theta_m = (\theta_{m,1}, \theta_{m,2}, ..., \theta_{m,K})$ for word m.

$$\theta_{m,k} = \frac{nmk(m,k)+\alpha}{nm(m)+K*\alpha} \tag{4}$$

The topic variables are derived from:

$$\phi_k^e = \frac{nkve(k,n)+\beta}{nke(k)+V_e*\beta} \tag{5}$$

$$\phi_k^j = \frac{nkvj(k,n)+\beta}{nkj(k)+V_j*\beta} \tag{6}$$

We run 1500 iterations for inference while the first 1000 iterations are discarded as burn-in steps. After the sampling chain converges, we average the value of θ_m to get the final per-word topic distribution.

Posterior Inference for BlockProbBiLDA

For each word t^e, we sample its topic according to:

$$p(z_i^e = k | z_{-i,t^e}^e, d_{t^e,i}^e = n, \theta) \tag{7}$$
$$\propto \frac{nmk(t^e,k)+cmk(t^e,k)+\alpha-1}{nm(t^e)+cm(t^e)+K*\alpha-1} * \frac{nkve(k,n)+\beta-1}{nke(k)+V_e*\beta-1}$$

For each word t^j, if it is in the dictionary, and it selects c as its translation in the previous iteration, then

$$p(z_i^j = k | z_{-i,t^j}^j, d_{t^j,i}^j = d, \theta) \qquad (8)$$

$$\propto \frac{nmk(t^j, k) + nmk(c, k) + \alpha - 1}{nm(t^j) + nm(c) + K * \alpha - 1} * \frac{nkvj(k, d) + \beta - 1}{nkj(k) + V_j * \beta - 1}$$

The selection of translations is sampled by

$$p(s^j = t^e | z^j, d_{t^j,i}^j = d, \theta) \qquad (9)$$

$$\propto \prod_i \frac{nmk(t^e, z_i^j) + \alpha + \sum_{m \in C(t^e)/\{t^j\}} nmk(m, z_i^j)}{nm(t^e) + K * \alpha + \sum_{m \in C(t^e)/\{t^j\}} nm(m)}$$

, where $C(t^e)$ is the set of all words which cite t^e as their translations in last iteration; $C(t^e)/\{t^j\}$ means to exclude t^j in this set. As the product of the probabilities is usually very small, $p(s^j = t^e)$ has different orders of magnitude for each t^e, so the sampling of s^e can be approximated by selecting the one with largest probability. We use the following equation instead:

$$s^j \approx \qquad (10)$$

$$\arg\max_{t^e} \sum_i \log \frac{nmk(t^e, z_i^j) + \alpha + \sum_{m \in C(t^e)/\{t^j\}} nmk(m, z_i^j)}{nm(t^e) + K * \alpha + \sum_{m \in C(t^e)/\{t^j\}} nm(m)}$$

After sampling the translation selection $s^j = t^e$ for t^j, we update the $C(t^e)$ as well as $C(t^{e\prime})$, where $C(t^{e\prime})$ is the previous selection of s^j. Then we use a scheme similar to the one used in 3.3 to get topic distribution θ,

4 Computing Word Similarities to Obtain Translations

Once we get the topic distribution of each word, we can use them to calculate the similarity between words. The simplest way to do this is to regard each topic distribution as a vector representation of a word. We can then calculate the cosine similarity between these vectors.

$$Cosine(\theta_m, \theta_c) = \frac{\sum_{k=1}^{K} \theta_{mk} \theta_{ck}}{\sqrt{\sum_{k=1}^{K} \theta_{mk}^2} \sqrt{\sum_{k=1}^{K} \theta_{ck}^2}} \qquad (11)$$

Another measure is to use the Kullback-Leibler (KL) divergence. KL divergence is a measure of difference between two probability distributions that is widely used in previous topic model-based approaches.

$$D_{KL}(\theta_m || \theta_c) = \sum_{k=1}^{K} \theta_{mk} \log \frac{\theta_{mk}}{\theta_{ck}} \qquad (12)$$

Neither cosine similarity nor KL divergence considers the correlation between topics. For a topic model, as we know the topic distribution of each word in addition to knowing the topic itself, we can take advantage of the topic structures by directly modeling the probability of $p(w^e | w^j)$ as the similarity between words w^e and w^j. This tells us how likely it is we can generate w^e from w^j. We call this similarity measure selProb (selection probability).

$$selProb = p(w^e | w^j) \propto p(w^j | \theta_{w^e})$$

$$= \prod_{i=1}^{n} \sum_{z^j=1}^{K} p(d_i^j | z^j, \phi^j) p(z^j | \theta_{w^e})$$

Then, we can select the most similar word in the target language as the translation.

$$\arg\max_{w^e} \log p(w^j | \theta_{w^e})$$

$$= \arg\max_{w^e} \sum_{i=1}^{n} \log \sum_{z^j=1}^{K} p(d_i^j | z^j, \phi^j) p(z^j | \theta_{w^e})$$

5 Experimental Results

5.1 Experiment Data

We use two Japanese-English domain-specific corpora for our experiments. The first corpus comes from a bilingual law dataset [1]. We selected it for our experiments because it has an associated law dictionary that can be directly used as test data. Although this corpus was originally parallel, we do not use it as a parallel data source. We randomly select 150,000 paragraphs for each language (each paragraph is seen as one document) and the paragraphs are not kept aligned. The other dataset is a collection of car complaints from MLIT[2] and NHTSA[3]. The data, which is unbalanced, includes 351,811 short English documents and 32,059 short Japanese documents.

The Japanese texts are processed by our own NLP tool to obtain the segmentation and the English texts are tokenized and lemmatized by NLTK[4]. We removed stop words for both Japanese and English texts. We also removed documents containing less than five words. For both corpora, we use the generic Japanese-English dictionary Edict[5] and exclude any words that do not appear in our corpora. For the law data, we use the associated law dictionary as the test data. After erasing any words not in the corpora, we are left with 840 words for test data, within which 133 words are not in the generic dictionary and 254 do not have correct translations in the generic dictionary. In terms of the car data, we do not have any technical dictionary for cars, so we manually annotate 150 technical words not covered by the seed dictionary. In contrast to [Andrade *et al.*, 2010] and [Tamura *et al.*, 2012], we test not only nouns but also words of other parts of speech.

5.2 Comparisons with Other Models

We compare our approaches with several previous approaches.

- **TFIDF** [Fung and Yee, 1998] is a classical lexicon extraction method. It uses tf-idf weights for contextual seed words to obtain the context vectors and then uses cosines similarity to rank candidate translations.

- **Label propagation (LP)** propagates seed distributions on a graph representing relations among words, and translation pairs are extracted by identifying word pairs with a high similarity in the seed distributions. It is claimed that it could resolve the sparseness problem of context vectors and better utilize the information of seed

---

[1] http://www.phontron.com/jaen-law/index-ja.html
[2] http://www.mlit.go.jp/jidosha/carinf/rcl/defects.html
[3] http://www-odi.nhtsa.dot.gov/downloads/index.cfm
[4] http://www.nltk.org
[5] http://www.edrdg.org/jmdict/edict.html

Table 1: Accuracies of translations on law data and car data.

	Law				Car	
	Acc1	Acc10	Acc1_new	Acc10_new	Acc1	Acc10
TFIDF	57.6%	72.1%	5.2%	20.3%	1.3%	3.3%
LP	44.2%	71.7%	3.8%	15.0%	2.7%	7.3%
Mixed word embedding	43.8%	71.9%	1.5%	15.0%	4.7%	8.7%
BiLDA + cosine	56.3%	73.4%	4.5%	9.8%	4%	12%
BiLDA + KLD	58.2%	74.8%	6.8%	18.0%	4.7%	9.3%
BiLDA + selProb	53.3%	74.9%	6.0%	17.3%	5.3%	12.7%
BiLDA_all + cosine	56.3%	72.6%	1.5%	9.0%	4.7%	9.3%
BiLDA_all + KLD	56.4%	72.7%	3.0%	10.5%	4%	6.7%
BiLDA_all + selProb	52.1%	73.8%	2.25%	12.0%	6%	13.3%
ProbBiLDA + cosine	55.7%	74.7%	9.8%	17.3%	2.7%	7.3%
ProbBiLDA + KLD	54.8%	75.2%	5.3%	18.0%	2.7	4.0%
ProbBiLDA + selProb	57.4%	77.4%	11.3%	24.1%	6%	11.3%
BlockProbBiLDA + cosine	58.1%	76.1%	9.0%	23.3%	6.7%	10%
BlockProbBiLDA + KLD	59.3%	77.6%	**14.3%**	29.3%	4.0%	8.0%
BlockProbBiLDA + selProb	**60.5%**	**78.1%**	**14.3%**	**31.6%**	**8.0%**	**14.6%**

dictionary. In our experiments, we use the same parameter setting as [Tamura *et al.*, 2012] (window size = 4).

- **Mixed word embedding** [Gouws and Søgaard, 2015] is a simple but efficient cross-lingual word embedding that can handle the problem of multiple translation (unlike the one-to-one seed dictionary used in [Mikolov *et al.*, 2013]). After obtaining the word embeddings, we use cosine similarity to rank the candidates.

We implemented two bilingual LDA 3.2 models as our baselines and compare them with the ProbBiLDA and Block-ProbBiLDA. For all the models, we set the hyperparameters as follows: $\alpha = \alpha_\phi = 0.5$, $\beta = 0.01$, and the topic number is 50. We use three similarity measures to rank the candidate translations: cosine similarity, KL divergence, and sel-Prob. We use Accuracy@K (K=1, 10) as the evaluation metric. Note that for law data, we test not only the overall accuracy of all law dictionary words (Acc1_full, Acc10_full) but also the accuracy w.r.t only new words (not in the generic dictionary) (Acc1_new, Acc10_new). As for car data, all test words are new, so we do not discriminate them. The results are listed in Table 1.

First let us compare the three similarity measures. It is obvious that selProb performs better than the other two in most cases. This supports our assumption that topics are correlated and selProb can better catch the word similarity. In the following comparison, we use only this measure for all topic models.

By comparing LP and TFIDF, we find, somewhat surprisingly, that LP performs worse than TFIDF on law data. This is probably because the seed dictionary is very noisy for the law domain. Because of the noise, LP does not benefit from propagating the seed dictionary on the contrary, it is badly impacted because error translations are propagated. We may achieve a similar conclusion by comparing BiLDA and BiLDA_all. However, on car data, the comparison result is reversed, so we can only assume that the seed dictionary for the car domain is cleaner.

Now let us turn to our two probabilistically linked bilingual topic models, the ProbBiLDA and BlockProbBiLDA. Regardless of the datasets, BlockProbBiLDA is consistently the best model and the ProbBiLDA is the runner-up or at least comparable to other best models. This demonstrates the effectiveness of modeling translation probability. Among the two models, BlockProbBiLDA is better on both datasets. A possible explanation is that BlockProbBiLDA tends to exclude the noise when a word has several translations, while ProbBiLDA is a mixed average of all translations including noise; and BlockProbBiLDA enables words better correlated to their translations than ProbBiLDA.

6 Conclusion

In this paper, we proposed a new framework for extracting translations from non-parallel corpora. First we constructed pseudo documents using inverted indexing; and then we introduced two new bilingual topic models, ProbBiLDA and BlockProbBiLDA, to obtain topic distributions for each word. These models are extensions of the classical Bilingual LDA featuring a new hierarchy to integrate the translation probability for multiple translations in the seed dictionary. We advanced the generation of probability to measure the similarity between one candidate word and a given target word. Experimental results show that Bilingual Topic models as well as inverted representation can be effectively utilized for lexicon extraction and that modeling the probabilities of translations in the seed dictionary is helpful as well.

References

[Andrade *et al.*, 2010] Daniel Andrade, Tetsuya Nasukawa, and Jun'ichi Tsujii. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 19–27. Association for Computational Linguistics, 2010.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[Bollegala *et al.*, 2015] Danushka Bollegala, Georgios Kontonatsios, and Sophia Ananiadou. A cross-lingual similarity measure for detecting biomedical term translations. *PloS one*, 10(6):e0126196, 2015.

[Boyd-Graber and Blei, 2009] Jordan Boyd-Graber and David M Blei. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 75–82. AUAI Press, 2009.

[Chang and Blei, 2009] Jonathan Chang and Daokvid M Blei. Relational topic models for document networks. In *International conference on artificial intelligence and statistics*, pages 81–88, 2009.

[Dietz *et al.*, 2007] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM, 2007.

[Duong *et al.*, 2016] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 12851295, 2016.

[Fung and Yee, 1998] Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics, 1998.

[Gouws and Søgaard, 2015] Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of NAACL-HLT*, pages 1386–1390, 2015.

[Haghighi *et al.*, 2008] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. *ACL-08: HLT*, page 771, 2008.

[Laws *et al.*, 2010] Florian Laws, Lukas Michelbacher, Beate Dorow, Christian Scheible, Ulrich Heid, and Hinrich Schütze. A linguistically grounded graph model for bilingual lexicon extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 614–622. Association for Computational Linguistics, 2010.

[Lopez, 2008] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.

[Mikolov *et al.*, 2013] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[Mimno *et al.*, 2009] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum.

Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics, 2009.

[Ni *et al.*, 2009] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 1155–1156. ACM, 2009.

[Pekar *et al.*, 2006] Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266, 2006.

[Rapp, 1995] Reinhard Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics, 1995.

[Shezaf and Rappoport, 2010] Daphna Shezaf and Ari Rappoport. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 98–107. Association for Computational Linguistics, 2010.

[Søgaard *et al.*, 2015] Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. Inverted indexing for cross-lingual nlp. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, 2015.

[Tamura *et al.*, 2012] Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36. Association for Computational Linguistics, 2012.

[Vulic and Moens, 2013] Ivan Vulic and Marie-Francine Moens. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 106–116. ACL, 2013.

[Vulić *et al.*, 2011] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 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: short papers-Volume 2*, pages 479–484. Association for Computational Linguistics, 2011.

[Zhang *et al.*, 2017] Meng Zhang, Haoruo Peng3 Yang Liu, and Huanbo Luan1 Maosong Sun. Bilingual lexicon induction from non-parallel data with minimal supervision. 2017.