Online Polylingual Topic Models for Fast Document Translation Detection

Kriste Krstovski
School of Computer Science
University of Massachusetts Amherst
Amherst, MA, 01003
kriste@cs.umass.edu

David A. Smith
School of Computer Science
University of Massachusetts Amherst
Amherst, MA, 01003
dasmith@cs.umass.edu

Abstract

Many tasks in NLP and IR require efficient document similarity computations. Beyond their common application to exploratory data analysis, latent variable topic models have been used to represent text in a low-dimensional space, independent of vocabulary, where documents may be compared. This paper focuses on the task of searching a large multilingual collection for pairs of documents that are translations of each other. We present (1) efficient, online inference for representing documents in several languages in a common topic space and (2) fast approximations for finding near neighbors in the probability simplex. Empirical evaluations show that these methods are as accurate as—and significantly faster than—Gibbs sampling and brute-force all-pairs search.

1 Introduction

Statistical topic models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003), have proven to be highly effective at discovering hidden structure in document collections (Hall et al., 2008, e.g.). Often, these models facilitate exploratory data analysis, by revealing which collocations of terms are favored in different kinds of documents or which terms and topics rise and fall over time (Blei and Lafferty, 2006; Wang and McCallum, 2006). One of the greatest advantages in using topic models to analyze and process large document collections is their ability to represent documents as probability distributions over a small number of topics, thereby mapping documents into a low-dimensional latent space—the T-dimensional probability simplex, where T is the number of topics. A document, represented by some point in this simplex, is said to have a particular "topic distribution".

Representing documents as points in a low-dimensional shared latent space abstracts away from the specific words used in each document, thereby facilitating the analysis of relationships between documents written using different vocabularies. For instance, topic models have been used to identify scientific communities working on related problems in different disciplines, e.g., work on cancer funded by multiple Institutes within the NIH (Talley et al., 2011). While vocabulary mismatch occurs within the realm of one language, naturally this mismatch occurs across different languages. Therefore, mapping documents in different languages into a common latent topic space can be of great benefit when detecting document translation pairs (Mimno et al., 2009; Platt et al., 2010). Aside from the benefits that it offers in the task of detecting document translation pairs, topic models offer potential benefits to the task of creating translation lexica, aligning passages, etc.

The process of discovering relationship between documents using topic models involves: (1) representing documents in the latent space by inferring their topic distributions and (2) comparing pairs of topic distributions to find close matches. Many widely used techniques do not scale efficiently, however, as the size of the document collection grows. Posterior inference by Gibbs sampling, for instance, may make thousands of passes through the data. For the task of comparing topic distributions, recent work has also resorted to comparing all pairs of documents (Talley et al., 2011).

This paper presents efficient methods for both

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 252–261,
Sofia, Bulgaria, August 8-9, 2013

of these steps and performs empirical evaluations on the task of detected translated document pairs embedded in a large multilingual corpus. Unlike some more exploratory applications of topic models, translation detection is easy to evaluate. The need for bilingual training data in many language pairs and domains also makes it attractive to mitigate the quadratic runtime of brute force translation detection. We begin in §2 by extending the online variational Bayes approach of Hoffman et al. (2010) to polylingual topic models (Mimno et al., 2009). Then, in §3, we build on prior work on efficient approximations to the nearest neighbor problem by presenting theoretical and empirical evidence for applicability to topic distributions in the probability simplex and in §4, we evaluate the combination of online variational Bayes and approximate nearest neighbor methods on the translation detection task.

2 Online Variational Bayes for Polylingual Topic Models

Hierarchical generative Bayesian models, such as topic models, have proven to be very effective for modeling document collections and discovering underlying latent semantic structures. Most current topic models are based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In some early work on the subject, Blei and Jordan (2003) showed the usefulness of LDA on the task of automatic annotation of images. Hall et al. (2008) used LDA to analyze historical trends in the scientific literature; Wei and Croft (2006) showed improvements on an information retrieval task. More recently Eisenstein et al. (2010) modeled geographic linguistic variation using Twitter data.

Aside from their widespread use on monolingual text, topic models have also been used to model multilingual data (Boyd-Graber and Blei, 2009; Platt et al., 2010; Jagarlamudi and Daumé, 2010; Fukumasu et al., 2012), to name a few. In this paper, we focus on the Polylingual Topic Model, introduced by Mimno et al. (2009). Given a multilingual set of aligned documents, the PLTM assumes that across an aligned multilingual document tuple, there exists a single, tuple-specific, distribution across topics. In addition, PLTM assumes that for each language–topic pair, there exists a distribution over words in that language β_l. As such, PLTM assumes that the multilingual corpus is created through a generative process where

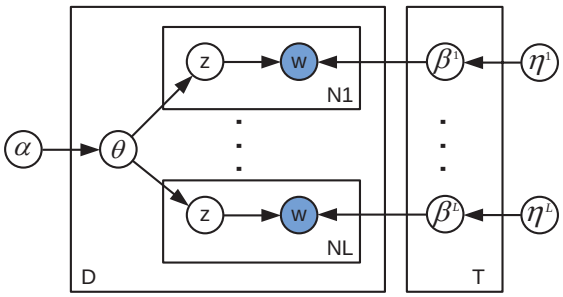


Figure 1: Polylingual topic model (PLTM)

first a document tuple is generated by drawing a tuple-specific distribution over topics θ[1] which, as it is the case with LDA, is drawn from a Dirichlet prior $\theta \sim Dir(\alpha)$. For each of the languages l in the tuple and for each of the N words w_n^l in the document the generative process: first chooses a topic assignment $z_n^l \sim Multinomial(\theta)$ which is then followed by choosing a word w_n^l from a multinomial distribution conditioned on the topic assignment and the language specific topics distribution over words $\beta_l \sim Dir(\eta_l)$. Both α and $\eta_{1,\ldots,L}$ are symmetric priors, i.e. the priors are exchangeable Dirichlet distributions. Finally, each word is generated from a language- and topic-specific multinomial distribution β_t^l as selected by the topic assignment variable z_n^l:

$$w_n^l \sim p\left(w_n^l \mid z_n^l, \beta_n^l\right) \tag{1}$$

Figure 1 shows a graphical representation of the PLTM using plate notation. In their original work Mimno et al. (2009) used the Gibbs sampling approach as a posterior inference algorithm to assign topics distributions over their test collection. While more straightforward to implement, this sampling approach is inherently slow when applied to large collections which makes the original PLTM work practically infeasible to be used on real-world data sets.

In general, performing posterior inference over the latent variables of a Bayesian model is usually done with two of the three approximate approaches, Gibbs sampling, variational Bayes (VB) and expectation-propagation. While Gibbs Sampling is a variation of Markov Chain Monte Carlo method (MCMC) which generates a sample from the true posterior after converging to a stationary

[1] In the traditional LDA model θ is used to specify the document specific distribution over topics.

distribution; in VB, a set of free variational parameters characterizes a simpler family of probability distributions. These variational parameters are then optimized by finding the minimum Kullback-Leibler (KL) divergence between the variational distribution $q(\theta, z, \beta|\gamma, \phi, \lambda)$ and the true posterior $P(\theta, z, \beta|w, \alpha, \eta)$. From an algorithmic perspective, the variational Bayes approach follows the Expectation-Maximization (EM) procedure where for a given document, the E-step updates the per document variational parameters γ_d and ϕ_d while holding the per words-topic distribution parameter λ fixed. It then updates the variational parameter λ using the sufficient statistics computed in the E step. In order to converge to a stationary point, both approaches require going over the whole collection multiple times which makes their time complexity to grown linearly with the size of the data collection. The mere fact that they require continuous access to the whole collection makes both inference approaches impracticable to use on very large or streaming collections. To alleviate this problem, several algorithms have been proposed that draws from belief propagation (Zeng et al., 2012), the Gibbs sampling approach such as (Canini et al., 2009), variational Bayes (Hoffman et al., 2010) as well as a combination of the latter two (Hoffman et al., 2012) to name a few. In this paper we use Hoffman et al. (2010) approach. Hoffman et al. (2010) proposed a new inference approach called Online LDA which relies on the stochastic gradient descent to optimize the variational parameters. This approach can produce good estimates of LDA posteriors in a single pass over the whole collection.

2.1 Algorithmic Implementation

We now derive an online variational Bayes algorithm for PLTM to infer topic distributions over multilingual collections. Figure 2 shows the variational model and free parameters used in our approach. As in the case of Hoffman et al. (2010), our algorithm updates the variational parameters γ_d^l and ϕ_d^l on each batch of documents while the variational parameter λ is computed as a weighted average of the value on the previous batch and its approximate version $\tilde{\lambda}$. Averaging is performed using a decay function whose parameters control the rate at which old values of λ^l are forgotten. Within the E step of the VB approach, we compute the updates over the variational parameter ϕ_l

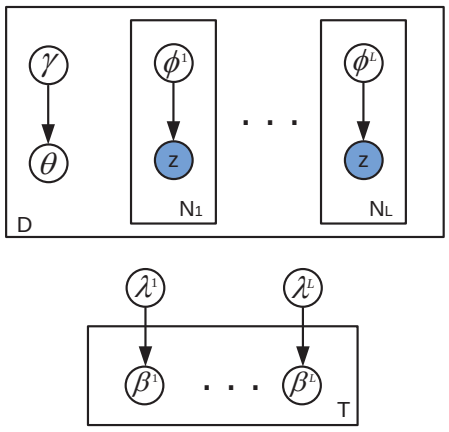


Figure 2: Graphical model representation of the free variational parameters for the online variational Bayes approximation of the PLTM posterior

for each language L present in our document tuple while the update on the γ parameter accumulates the language specific sufficient statistics:

$$\gamma_k^m = \alpha + \sum_l \sum_w \phi_{wk}^{ml} \, n_w^{ml} \qquad (2)$$

We detail these steps in Algorithm 1.

2.2 Performance Analysis

To demonstrate the efficacy of online PLTM, we ran topic inference on a subset of the English-Spanish Europarl collection consisting of ∼64k parallel speeches and compared the accuracy results vs. the training and inference speed against the original PLTM model using topic sets of T=50,100, 200 and 500. We explain in details the evaluation task and the performance metric used in §4. Shown in Figure 3 are the results of these comparisons. Our speed measurements were performed on Xeon quad processors with a clock speed of 2.66GHz and a total of 16GB of memory.

As we increase the number of topics we gain in accuracy over the evaluation task across both inference approaches. When we increase the number of topics from 50 to 500 the speed improvement obtained by Online VB PLTM drops by a factor of 2.9 within the training step and by a factor of 4.45 in the test step. Our total running time for the Online VB PLTM with T=500 approaches the running time of the Gibbs sampling approach with T=50. The gradual drop in speed improvement with the increase of the number topics is mostly attributed to the commutation of the

Algorithm 1 Online variational Bayes for PLTM

initialize λ_l randomly
obtain the tth mini-batch of tuples M_t
for $t = 1$ to ∞ **do**
  $\rho_t \leftarrow \left(\frac{1}{t_0+t}\right)^{\kappa}$
  E step:
  initialize γ_t randomly
  for each document tuple in mini-batch t
  for m in M_t **do**
    repeat
      for $l \in 1, \ldots,$L **do**
        $\phi_{wk}^{ml} \propto$
        $\exp\{E_q[\log \theta_k^m]\} *$
        $\exp\{E_q[\log \beta_{kw}^{ml}]\}$
      end for
      $\gamma_k^m = \alpha + \sum_l \sum_w \phi_{wk}^{ml}\, n_w^{ml}$
    until convergence
  end for
  M step:
  for $l \in 1, \ldots,$L **do**
    $\tilde{\lambda}_{kw}^{l} = \eta + D \sum_m \phi_{wk}^{ml} n_w^{ml}$
    $\lambda_{kw}^{lt} \leftarrow (1-\rho_t)\, \lambda_{kw}^{l(t-1)} + \rho_t \tilde{\lambda}_{kw}^{l}$
  end for
end for

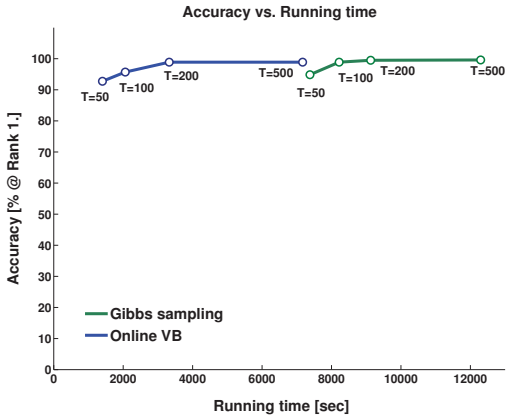


Figure 3: Speed vs. accuracy comparison between Online VB PLTM and Gibbs Sampling PLTM at T=50,100, 200 and 500. We used a Python implementation of Online VB and Mallet's Java implementation of PLTM with in-memory Gibbs Sampling using 1000 iterations.

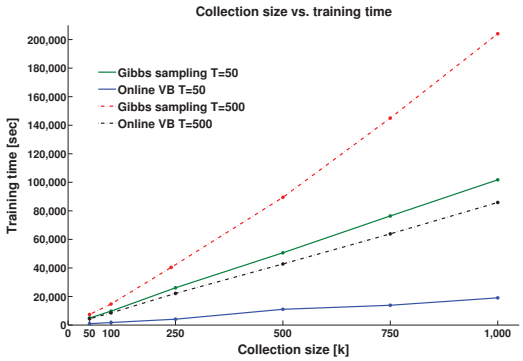


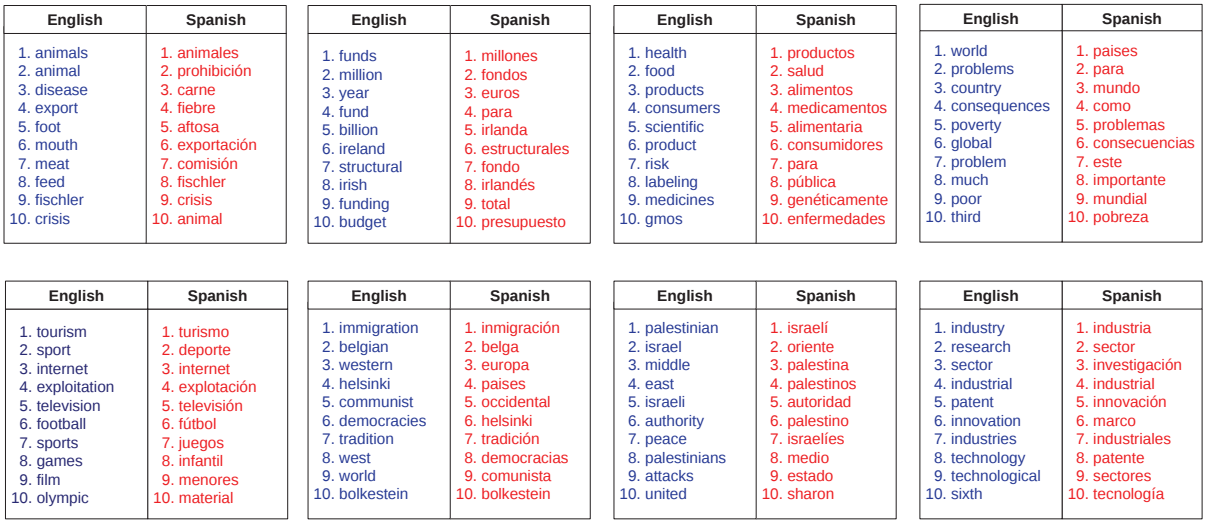
Figure 4: Collection size vs. training time comparison between Online VB PLTM and Gibbs Sampling PLTM using multilingual collections of 50k, 100k, 250k, 500k, 750k and 1M speech pairs.

digamma function (Asuncion et al., 2009) whose time complexity increases linearly with the number of topics.

While a multilingual collection of ∼64k document pairs is considered relatively big, our goal of deriving the Online VB PLTM approach was to be able to utilize PLTM on very large multilingual collections. To analyze the potential of using Online VB PLTM on such collections we ran speed comparisons within the training step by creating multilingual collections of different lengths multiplying the original English-Spanish Europarl collection. Speed comparisons using collections of length 50K, 100K, 250K, 500K, 750K and 1M are shown in Figure 4. Training was performed with the number of topics T set to T=50 and T=500.

As we increase the collection size we observe the real benefit of using Online VB compared to Gibbs sampling. This is mostly attributed to the fact that the Gibbs sampling approach requires multiple iterations over the whole collection in order to achieve a convergence point. For collection sizes of 50k and 100k the training time for the Online VB PLTM with T=500 approaches the training time of Gibbs sampling with T=50 and as we increase the collection size this proximity dissipates.

In Figure 5 we show a sample set of the aligned topics extracted using Online VB PLTM with T=400 on the English-Spanish Europarl collection. For a given topic tuple words are ordered based on probability of occurrence within the given topic.



English	Spanish
1. animals	1. animales
2. animal	2. prohibición
3. disease	3. carne
4. export	4. fiebre
5. foot	5. aftosa
6. mouth	6. exportación
7. meat	7. comisión
8. feed	8. fischler
9. fischler	9. crisis
10. crisis	10. animal

English	Spanish
1. funds	1. millones
2. million	2. fondos
3. year	3. euros
4. fund	4. para
5. billion	5. irlanda
6. ireland	6. estructurales
7. structural	7. fondo
8. irish	8. irlandés
9. funding	9. total
10. budget	10. presupuesto

English	Spanish
1. health	1. productos
2. food	2. salud
3. products	3. alimentos
4. consumers	4. medicamentos
5. scientific	5. alimentaria
6. product	6. consumidores
7. risk	7. para
8. labeling	8. pública
9. medicines	9. genéticamente
10. gmos	10. enfermedades

English	Spanish
1. world	1. paises
2. problems	2. para
3. country	3. mundo
4. consequences	4. como
5. poverty	5. problemas
6. global	6. consecuencias
7. problem	7. este
8. much	8. importante
9. poor	9. mundial
10. third	10. pobreza

English	Spanish
1. tourism	1. turismo
2. sport	2. deporte
3. internet	3. internet
4. exploitation	4. explotación
5. television	5. televisión
6. football	6. fútbol
7. sports	7. juegos
8. games	8. infantil
9. film	9. menores
10. olympic	10. material

English	Spanish
1. immigration	1. inmigración
2. belgian	2. belga
3. western	3. europa
4. helsinki	4. paises
5. communist	5. occidental
6. democracies	6. helsinki
7. tradition	7. tradición
8. west	8. democracias
9. world	9. comunista
10. bolkestein	10. bolkestein

English	Spanish
1. palestinian	1. israelí
2. israel	2. oriente
3. middle	3. palestina
4. east	4. palestinos
5. israeli	5. autoridad
6. authority	6. palestino
7. peace	7. israelíes
8. palestinians	8. medio
9. attacks	9. estado
10. united	10. sharon

English	Spanish
1. industry	1. industria
2. research	2. sector
3. sector	3. investigación
4. industrial	4. industrial
5. patent	5. innovación
6. innovation	6. marco
7. industries	7. industriales
8. technology	8. patente
9. technological	9. sectores
10. sixth	10. tecnología

Figure 5: Sample set of topics extracted from Europarl English-Spanish collection of 64k speeches using Online PLTM with T=400 ordered based on their probability of occurrence within the topic.

3 Approximate NN Search in the Probability Simplex

One of the most attractive applications for topic models has involved using the latent variables as a low-dimensional representation for document similarity computations (Hall et al., 2008; Boyd-Graber and Resnik, 2010; Talley et al., 2011). After computing topic distributions for documents, however, researchers in this line of work have almost always resorted to brute-force all-pairs similarity comparisons between topic distributions.

In this section, we present efficient methods for approximate near neighbor search in the probability simplex in which topic distributions live. Measurements for similarity between two probability distributions are information-theoretic, and distance metrics, typical for the metric space, are not appropriate (measurements such as Euclidean, cosine, Jaccard, etc.). Divergence metrics, such as Kullback-Leibler (KL), Jensen-Shannon (JS), and Hellinger distance are used instead. Shown in Figure 6 are the formulas of the divergence metrics along with the Euclidean distance. When dealing with a large data set of N documents, the $O(N^2)$ time complexity of all-pairs comparison makes the task practically infeasible. With some distance measures, however, the time complexity on near neighbor tasks has been alleviated using approximate methods that reduce the time complexity of each query to a sub-linear number of comparisons. For example, Euclidean distance (3) has been efficiently used on all-pairs comparison tasks in large data sets thanks to its approximate based versions developed using locality sensitive hashing (LSH) (Andoni et al., 2005) and k-d search trees (Friedman et al., 1977). In order to alleviate the all-pairs computational complexity in the probability simplex, we will use a reduction of the Hellinger divergence measure (4) to Euclidean distance and therefore utilize preexisting approximation techniques for the Euclidean distance in the probability simplex.

This reduction comes from the fact that both measurements have similar algebraic expressions. If we discard the square root used in the Euclidean distance, Hellinger distance (4) becomes equivalent to the Euclidean distance metric (3) between $\sqrt{p_i}$ and $\sqrt{q_i}$. The task of finding nearest neighbors for a given point (whether in the metric space or the probability simplex) involves ranking all nearest points discovered and as such not computing the square root function does not affect the overall ranking and the nearest neighbor discovery. Moreover, depending on its functional form, the Hellinger distance is often defined as square root over the whole summation. Aside from the Hellinger distance, we also approximate Jensen-Shannon divergence which is a symmetric version of the Kullback-Liebler divergence. For the JS approximation, we will use a constant factor relationship between the Jensen-Shannon divergence an Hellinger distance previously explored by (Topsøe, 2000). More specifically, we will be using its more concise form (7) also presented by

$$\mathrm{Eu}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (3)$$

$$\mathrm{He}(p,q) = \sum_{i=1}^{n}\left(\sqrt{p(x_i)} - \sqrt{q(x_i)}\right)^2 \quad (4)$$

$$\mathrm{KL}(p,q) = \sum_{i=1}^{n} p(x_i)\log\frac{p(x_i)}{q(x_i)} \quad (5)$$

$$\mathrm{JS}(p,q) = \frac{1}{2}\mathrm{KL}\left(p, \frac{p+q}{2}\right) + \frac{1}{2}\mathrm{KL}\left(q, \frac{p+q}{2}\right) \quad (6)$$

$$\frac{1}{2}\mathrm{He}(p,q) \leq \mathrm{JS}(p,q) \leq 2\ln(2)\mathrm{He}(p,q) \quad (7)$$

Figure 6: Distance measures and bounds

(Guha et al., 2006). The constant factor relationship provides us with the theoretical guarantees necessary for this approximation.

In practice, we can often do much better than this theoretical bound. Figure 7 shows the empirical relation of JS and Hellinger on a translation-detection task. As will be described in §4, we computed the JS and Hellinger divergences between topic distributions of English and Spanish Europarl speeches for a total of 1 million document pairs. Each point in the figure represents one Spanish-English document pair that might or might not be translations of each other. In this figure we emphasize the lower left section of the plot where the nearest neighbors (i.e., likely translations) reside, and the relationship between JS and Hellinger is much tighter than the theoretical bounds and from pratical perspective as we will show in the next section. As a summary for the reader, using the above approaches, we will approximate JS divergence by using the Euclidean based representation of the Hellinger distance. As stated earlier, the Euclidean based representation is computed using well established approximation approaches and in our case we will use two such approaches: the Exact Euclidean LSH (E2LSH) (Andoni et al., 2005) and the k-d trees implementation within the Approximate Nearest Neighbor (ANN) library (Mount and Arya, 2010).

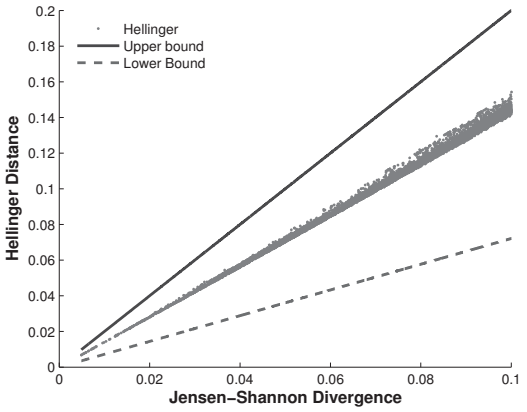


Figure 7: Empirical evidence of the bounds presented in Eq. 7 on 1 million document pairs—zoomed section where nearest neighbors reside. The lower bound is $He(p,q) = \frac{1}{2\ln(2)}JS(p,q)$ while the upper bound is $He(p,q) = 2JS(p,q)$.

4 Efficient Approximate Translation Detection

Mapping multilingual documents into a common, language-independent vector space for the purpose of improving machine translation (MT) and performing cross-language information retrieval (CLIR) tasks has been explored through various techniques. Mimno et al. (2009) introduced polylingual topic models (PLTM), an extension of latent Dirichlet allocation (LDA), and, more recently, Platt et al. (2010) proposed extensions of principal component analysis (PCA) and probabilistic latent semantic indexing (PLSI). Both the PLTM and PLSI represent bilingual documents in the probability simplex, and thus the task of finding document translation pairs is formulated as finding similar probability distributions. While the nature of both works was exploratory, results shown on fairly large collections of bilingual documents (less than 20k documents) offer convincing argument of their potential. Expanding these approaches to much large collections of multilingual documents would require utilizing fast NN search for computing similarity in the probability simplex. While there are many other proposed approaches to the task of finding document translation pairs that represent documents in metric space, such as Krstovski and Smith (2011) which utilizes LSH for cosine distance, there is no evidence that they yield good results on documents of small lengths such as paragraphs and even sen-

tences.

In this section, we empirically show how to utilize approaches that deal with representing documents in the probability simplex without a significant loss in accuracy while significantly improving the processing time. We use PLTM representations of bilingual documents. In addition, we show how the results as reported by Platt et al. (2010) can be obtained using the PLTM representation with a significant speed improvement.

As in (Platt et al., 2010) and (Mimno et al., 2009) the task is to find document translation pairs in a multilingual collection of documents by representing documents in the probability simplex and computing similarity between their probability distribution representation across all document pairs. For this experimental setup, accuracy is defined as the number of times (in percentage) that the target language document was discovered at rank 1 (i.e. % @Rank 1.) across the whole test collection.

4.1 Experimental Setup

We use Mallet's (McCallum, 2002) implementation of the PLTM to train and infer topics on the same data set used in Platt et al. (2010). That paper used the Europarl (Koehn, 2005) multilingual collection of English and Spanish sessions. Their training collection consists of speeches extracted from all Europarl sessions from the years 1996 through 1999 and the year 2002 and a development set which consists of speeches from sessions in 2001. The test collection consists of Europarl speeches from the year 2000 and the first nine months of 2003. While Platt et al. (2010) do offer absolute performance comparison between their JPLSA approach and previous results published by (Mimno et al., 2009), these performance comparisons are not done on the same training and test sets—a gap that we fill below.

We train PLTM models with number of topics T set to 50, 100, 200, and 500. In order to compare exactly the same topic distributions when computing speed vs. accuracy of various approximate and exhaustive all-pairs comparisons we focus only on one inference approach - the Gibbs sampling and ignore the online VB approach as it yields similar performance. For all four topic models, we use the same settings for PLTM (hyperparameter values and number of Gibbs sampling iterations) as in (Mimno et al., 2009)[2]. Topic distributions were then inferred on the test collection using the trained topics. We then performed all-pairs comparison using JS divergence, Hellinger distance, and approximate, LSH and kd-trees based, Hellinger distance. We measured the total time that it takes to perform exhaustive all-pairs comparison using JS divergence, the LSH and kd-trees version on a single machine consisting of a core 2 duo quad processors with a clock speed of 2.66GHz on each core and a total of 8GB of memory. Since the time performance of the E2LSH depends on the radius R of data set points considered for each query point (Indyk and Motwani, 1998), we performed measurements with different values of R. For this task, the all-pairs JS code implementation first reads both source and target sets of documents and stores them in hash tables. We then go over each entry in the source table and compute divergence against all target table entries.We refer to this code implementation as hash map implementation.

4.2 Evaluation Task and Results

Performance of the four PLTM models and the performance across the four different similarity measurements was evaluated based on the percentage of document translation pairs (out of the whole test set) that were discovered at rank one. This same approach was used by (Platt et al., 2010) to show the absolute performance comparison. As in the case of the previous two tasks, in order to evaluate the approximate, LSH based, Hellinger distance we used values of R=0.4, R=0.6 and R=0.8. Since in (Platt et al., 2010) numbers were reported on the test speeches whose word length is greater or equal to 100, we used the same subset (total of 14150 speeches) of the original test collection. Shown in Table 1 are results across the four different measurements for all four PLTM models. When using regular JS divergence, our PLTM model with 200 topics performs the best with 99.42% of the top one ranked candidate translation documents being true translations. When using approximate, kd-trees based, Hellinger distance, we outperform regular JS and Hellinger divergence across all topics and for T=500 we achieve the best overall accuracy of 99.61%. We believe that this is due to the small amount of error

[2]We start off by first replicating the results as in (Mimno et al., 2009) and thus verifying the functionality of our experimental setup.

Divergence	T=50	100	200	500
JS	94.27	98.48	99.42	99.33
He	94.30	98.45	99.40	99.31
He LSH R=0.4	93.95	97.46	98.27	98.01
He LSH R=0.6	94.30	98.46	99.40	99.31
He LSH R=0.8	94.30	98.45	99.34	99.31
He kd-trees	94.86	98.90	99.50	99.61

Table 1: Percentage of document pairs with the correct translation discovered at rank 1: comparison of different divergence measurements and different numbers T of PLTM topics.

Divergence	T=50	100	200	500
JS	7.8	4.6	2.4	1.0
He LSH R=0.4	511.5	383.6	196.7	69.7
He LSH R=0.6	142.1	105.0	59.0	18.6
He LSH R=0.8	73.8	44.7	29.5	16.3
He kd-trees	196.7	123.7	76.7	38.5

Table 2: Relative speed improvement between all-pairs JS divergence and approximate He divergence via kd-trees and LSH across different values of radius R. The baseline is brute-force all-pairs comparison with Jensen-Shannon and 500 topics.

in the search introduced by ANN, due to its approximate nature, which for this task yields positive results. On the same data set, (Platt et al., 2010) report accuracy of 98.9% using 50 topics, a slightly different prior distribution, and MAP instead of posterior inference.

Shown in Table 2 are the relative differences in time between all pairs JS divergence, approximate kd-trees and LSH based Hellinger distance with different value of R. Rather than showing absolute speed numbers, which are often influenced by the processor configuration and available memory, we show relative speed improvements where we take the slowest running configuration as a referent value. In our case we assign the referent speed value of 1 to the configuration with T=500 and all-pairs JS computation. Results shown are based on comparing running time of E2LSH and ANN against the all-pairs similarity comparison implementation that uses hash tables to store all documents in the bilingual collection which is significantly faster than the other code implementation.

For the approximate, LSH based, Hellinger distance with T=100 we obtain a speed improvement of 24.2 times compared to regular all-pairs JS divergence while maintaining the same performance compared to Hellinger distance metric and insignificant loss over all-pairs JS divergence. From Table 2 it is evident that as we increase the radius R we reduce the relative speed of performance since the range of points that LSH considers for a given query point increases. Also, as the number of topics increases, the speed benefit is reduced for both the LSH and k-d tree techniques.

5 Conclusion

Hierarchical Bayesian models, such as Polylingual Topic Models, have been shown to offer great potential in analyzing multilingual collections, extracting aligned topics and finding document translation pairs when trained on sufficiently large aligned collections. Online stochastic optimization inference allows us to generate good parameter estimates. By combining these two approaches we are able to infer topic distributions across documents in large multilingual document collections in an efficient manner. Utilizing approximate NN search techniques in the probability simplex, we showed that fast document translation detection could be achieved with insignificant loss in accuracy.

6 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Alexandr Andoni, Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab Mirrokni. 2005. Locality-sensitive hashing using stable distributions. In G. Shakhnarovich, T. Darrell, and P. Indyk, editors, *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*, pages 61–72. MIT Press.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 27–34, Arlington, Virginia, United States. AUAI Press.

David M. Blei and Michael I. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research*

and development in informaion retrieval, SIGIR '03, pages 127–134, New York, NY, USA. ACM.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

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

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 75–82, Arlington, Virginia, United States. AUAI Press.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 45–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin R. Canini, Lei Shi, and Thomas L. Griffiths. 2009. Online inference of topics with latent dirichlet allocation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. H. Friedman, J. L. Bentley, and R. A. Finkel. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226.

Kosuke Fukumasu, Koji Eguchi, and Eric Xing. 2012. Symmetric correspondence topic models for multilingual text analysis. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1295–1303.

Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. 2006. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864.

Matt Hoffman, David M. Blei, and David M. Mimno. 2012. Sparse stochastic inference for latent dirichlet allocation. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1599–1606, New York, NY, USA. ACM.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, STOC '98, pages 604–613, New York, NY, USA. ACM.

Jagadeesh Jagarlamudi and Hal Daumé. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32nd European conference on Advances in Information Retrieval*, ECIR'2010, pages 444–456, Berlin, Heidelberg. Springer-Verlag.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, pages 79–86.

Kriste Krstovski and David A. Smith. 2011. A minimally supervised approach for detecting and ranking document translation pairs. In *Proc. Workshop on Statistical MT*, pages 207–216.

Andrew Kachites McCallum, 2002. *MALLET: A Machine Learning for Language Toolkit*. `http://mallet.cs.umass.edu`.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M. Mount and Sunil Arya, 2010. *ANN: A Library for Approximate Nearest Neighbor Searching*. `http://www.cs.umd.edu/~mount/ANN/`.

John C. Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual document representations from discriminative projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 251–261, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edmund Talley, David Newman, David Mimno, Bruce Herr, Hanna Wallach, Gully Burns, Miriam Leenders, and Andrew McCallum. 2011. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8:443–444.

Flemming Topsøe. 2000. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Information Theory*, 44(4):1602–1609.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 424–433, New York, NY, USA. ACM.

Xing Wei and W. Bruce Croft. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185, New York, NY, USA. ACM.

Jia Zeng, Xiao-Qin Cao, and Zhi-Qiang Liu. 2012. Residual belief propagation for topic modeling. *CoRR*, abs/1204.6610.