Online Multilingual Topic Models with Multi-Level Hyperpriors

Kriste Krstovski[†,§], **David A. Smith**[‡] **and Michael J. Kurtz** [†]
[†]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA
[§]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA
[‡]College of Computer and Information Science, Northeastern University, Boston, MA
kkrstovski@cfa.harvard.edu, dasmith@ccs.neu.edu, kurtz@cfa.harvard.edu

Abstract

For topic models, such as LDA, that use a bag-of-words assumption, it becomes especially important to break the corpus into appropriately-sized "documents". Since the models are estimated solely from the term cooccurrences, extensive documents such as books or long journal articles lead to diffuse statistics, and short documents such as forum posts or product reviews can lead to sparsity. This paper describes practical inference procedures for hierarchical models that smooth topic estimates for smaller sections with hyperpriors over larger documents. Importantly for large collections, these online variational Bayes inference methods perform a single pass over a corpus and achieve better perplexity than "flat" topic models on monolingual and multilingual data. Furthermore, on the task of detecting document translation pairs in large multilingual collections, polylingual topic models (PLTM) with multi-level hyperpriors (mlhPLTM) achieve significantly better performance than existing online PLTM models while retaining computational efficiency.

1 Introduction

Bag of words models simplify the representation of documents by discarding grammatical information and simply relying on document-level word cooccurrence statistics. Topic models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003), use this representation. A major drawback of the bag of words representation, especially in collections of large documents, is that the word co-occurrence statistics are computed on a document level and as such they do not capture the effect of words co-occurring close to each other versus words co-occurring further apart.

One alternative approach to longer documents that has received attention in the past has been to directly model local—i.e., Markov—dependencies among tokens. For example, the topical n-gram model (TNG) introduced by Wang et al. (2007) models unigram and n-gram phrases as mixture of topics based on the nearby word context. More recently, Jameel & Lam (2013) proposed an LDA extension that uses word sequence information to generate topic distribution over n-grams and performs topic segmentation using segment and paragraph information. While these and many other approaches offer a better and more realistic modeling of word sequences, they don't model topical variations across document sections either in mono- or multilingual collections.

In this paper, we focus on hierarchical models for improving topic models of long documents. In the past, document-topic based hierarchical prior structures have been explored for LDA. For example, Wallach et al. (2009) showed that Gibbs sampling implementation of asymmetric Dirichlet priors provide better modeling of documents, across the whole collection, compared to the original LDA approach. More recently, Kim et al. (2013) introduced tiLDA, a topic model of monolingual document collections with nested hierarchies. In order to achieve reasonable performance over large document collections with deep hierarchies, tiLDA utilizes parallel variational Bayes (VB) inference. While VB is known to converge faster than Gibbs sampling, and paral-

lel implementations are even faster, they, as with Gibbs sampling, still require multiple iterations over the whole collection besides the overhead of parallelizing the model parameters. Furthermore these approaches focus on monolingual collections.

We propose an online VB inference approach for topic models that captures the document specific effect of local and long range word co-occurrence by modeling individual document sections using multi-level Dirichlet prior structure. The proposed models assign Dirichlet priors to individual document sections that are coupled by a document level hierarchical Dirichlet prior which facilitates explicit modeling of the variation in topics across documents in mono- and multilingual collections. This in turn streamlines the use of topic models in collections of large documents where there is a predetermined section structure. Our contribution is twofold: (1) we present an online VB inference approach for topic models with multi-level Dirchlet prior structure and more importantly (2) introduce a polylingual topic model (PLTM) with multi-level hyperpriors (mlhPLTM) which is capable of efficiently modeling topical variations across document sections in large multilingual collections.

2 Efficient Multi-level Hyperpriors

The original LDA model and its multilingual variant, PLTM, use symmetric Dirichlet priors over the document-topic distributions θ_d and topic-word distributions φ_k which means that the concentration parameter α of the Dirichlet distribution is fixed and that the base measure u across all topics is uniform. Symmetric Dirichlet priors assume that all documents in the collection are drawn from the same family of distributions. This assumption is not suitable for collections of documents that cover a diverse set of topics. In the past this issue has been addressed with asymmetric priors where the base measures are non-uniform. One way to assign asymmetric priors to individual documents is to treat the base measures vector u as a hidden variable and assign a symmetric Dirichlet prior to it which creates a hierarchical Dirichlet prior structure over all document-topic distributions in the collection. This approach was used by Wallach et al. (2009). Unlike Wallach et al. (2009), who use a single document-topic distribution θ_d, we introduce section-topic distributions θ_s. The existing symmetric Dirichlet prior over θ_d creates a hierarchical Dirichlet prior over θ_s ($\theta = \theta_d, \theta_{s_1}, \theta_{s_2}, ..., \theta_{s_S}$):

$$p(\theta|\alpha_d u, \alpha_s) \propto p(\theta_d|\alpha_d u) \prod_s p(\theta_s|\alpha_s \theta_d) \qquad (1)$$

In this setting the most widely used approach for estimating θ_d is Minka's (2000) fixed-point iteration approach which is also used in (Kim et al., 2013). Instead we use a more efficient approach for estimating the Dirichlet-multinomial hyperparameters by approximating the digamma differences in Minka's approach which was showcased in (Wallach, 2008) to be more efficient. Figure 1 shows the graphical model representation (left) of our model, which we refer to as multi-level hyperpriors LDA (mlhLDA), along with the free variational parameters for approximating the posteriors (right).

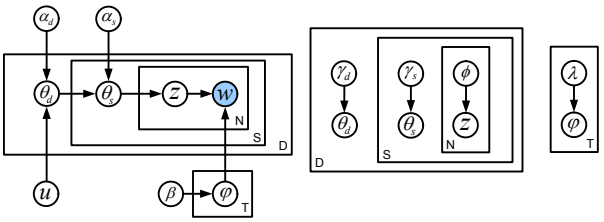


Figure 1: mlhLDA: Graphical representation (left); Free variational parameters for the online VB approximation (right).

2.1 Inference using Online VB

Due to its ease of implementation, the most widely used approach for inferring LDA posterior distributions is Gibbs sampling (Griffiths and Steyvers, 2004). For example, this approach was used by Wallach et al. (2009) and was originally used for PLTM. On the other hand the VB approach (Blei et al., 2003) offers more efficient computation but as in the case of Gibbs sampling requires iterating over the whole collection multiple times (e.g. Kim et al. (2013)). More recently Hoffman et al. (2010) introduced online LDA (oLDA) that relies on online stochastic optimization and requires a single pass over the whole collection. The same approach was also extended to PLTM (oPLTM) (Krstovski and Smith, 2013). In our work we also utilize online VB to implement multi-level hyperprior (mlh) structure in LDA and PLTM. Similar to batch VB, in online

VB locally optimal values of the free variational parameters γ and ϕ, which are used to approximate the posterior θ and z, are computed in the E step of the algorithm but on a batch b of documents d_i (rather than the whole collection D as in the case of batch VB) while holding the topic-word variational parameter λ fixed. In the M step, λ is updated using stochastic gradient algorithm by first computing the optimal values of $\tilde{\lambda}$ using the batch optimal values of ϕ^b: $\tilde{\lambda}_{kw} = \eta + \frac{D}{|b|} \sum_{i=1}^{|b|} n_{d_i w} \phi_{wk}^{d_i}$. This value is then combined with value of λ computed on the previous batch through weighted average:

$$\lambda_{kw}^{b} \leftarrow (1 - \rho_b)\, \lambda_{kw}^{b-1} + \rho_b \tilde{\lambda}_{kw} \tag{2}$$

When computing the section-topic variational parameters we follow the proof of the lower bound which was derived by Kim et al. (2013). This lower bound, which is looser than the original VB Evidence Lower Bound (ELBO), allows for the batch VB approach to be used with asymmetric priors. More specifically, given the document-topic variational parameter γ_{dk} in the E step of our online VB approach the update for the section-topic variational parameter γ_{sk} becomes:

$$\gamma_{sk} = \alpha_s(\frac{\gamma_{dk}}{\sum_k \gamma_{dk}}) + \sum_w n_w^s\, \phi_{wk}^s \tag{3}$$

3 Online PLTM with multi-level Dirichlet Priors

Given an aligned multilingual document tuple, PLTM assumes that: (1) there exists a single tuple-specific distribution across topics and (2) sets of language specific topic-word distributions. Each word is generated from a language- and topic-specific multinomial distribution φ_k^l as selected by the topic assignment variable z_n^l:

$$w_n^l \sim p\left(w_n^l \mid z_n^l, \varphi_k^l\right) \tag{4}$$

We extend this model by introducing sections specific topic distributions θ_s across the different languages in the tuple which are coupled by the tuple specific document-topic distribution θ_d.

Given a collection of document tuples d where each tuple contains l documents that are translations of each other in different languages, mlhPLTM assumes the following generative process. For each language l in the collection the model first generates a set of $k \in \{1, 2, ..., K\}$ topic-word distributions, φ_k^l which are drawn from a Dirichlet prior with language specific hyperparameter β^l: $\varphi_k^l \sim Dirichlet(\beta^l)$. For each document d^l with s_d sections in tuple d, mlhPLTM then assumes the following generative process:

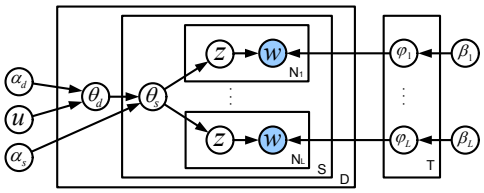


Figure 2: mlhPLTM: Graphical model representation.

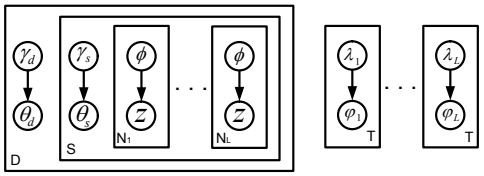


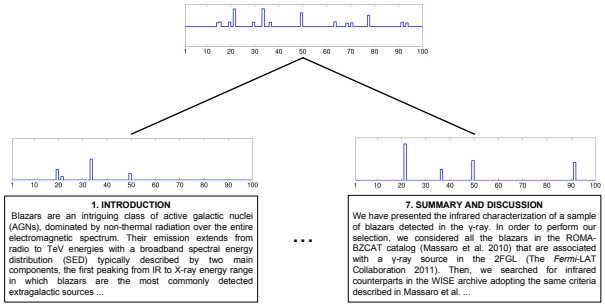
Figure 3: mlhPLTM: Graphical representation of the free variational parameters for the online VB approximation.

- Choose $\theta_d \sim Dir.(\alpha_d)$
- For each section s_d in document tuple d:
- Choose $\theta_s \sim Dir.(\alpha_s \theta_d)$
  - For each language l in section s:
    - For each word w in section s_d^l:
      - Choose a topic $z \sim Multi.(\theta_s^l)$
      - Choose a word $w \sim Multi.(\varphi_z^l)$

Figure 2 shows the graphical representation of mlhPLTM. The free variational parameters for the online VB approximation of the posteriors are shown in Figure 3.

4 Modeling Sections in Scientific Articles

We explore the ability of mlhLDA to model variations across document sections found in scientific articles using a collection of journal articles from the Astrophysics Data System (ADS) (Kurtz et al., 2000). Our collection consists of 130k training articles (888,346 sections) and a held-out set of 8,078 articles (54,502 sections). Figure 4 shows an example mlhLDA representation of an ApJ article with 100 topics. Shown on the top is the inferred topic representation of the whole document (θ_d) which, in the mlhLDA model, serves as a prior for the section-topic distributions (θ_s). Shown on the bottom are ex-



1. INTRODUCTION

Blazars are an intriguing class of active galactic nuclei (AGNs), dominated by non-thermal radiation over the entire electromagnetic spectrum. Their emission extends from radio to TeV energies with a broadband spectral energy distribution (SED) typically described by two main components, the first peaking from IR to X-ray energy range in which blazars are the most commonly detected extragalactic sources ...

...

7. SUMMARY AND DISCUSSION

We have presented the infrared characterization of a sample of blazars detected in the γ-ray. In order to perform our selection, we considered all the blazars in the ROMA-BZCAT catalog (Massaro et al. 2010) that are associated with a γ-ray source in the 2FGL (The *Fermi*-LAT Collaboration 2011). Then, we searched for infrared counterparts in the WISE archive adopting the same criteria described in Massaro et al. ...

Rank	Topic=33	Topic=19	Topic=49
1	spectral	aperture	measured
2	amplification	measured	uncertainties
3	isotropic	total	catalog
4	dropout	exposure	matching
5	competition	position	estimated
6	caustic	ratio	respectively
7	detected	selected	final
8	antenna	color	cathode
9	function	spread	total
10	color	objects	limit

Rank	Topic=21	Topic=49	Topic=91
1	entanglement	measured	ferroelectric
2	color	uncertainties	population
3	distance	catalog	rational
4	magnitude	matching	fraction
5	accretion	estimated	starburst
6	similar	respectively	shielding
7	modulus	final	similar
8	objects	cathode	emitting
9	right	total	reputation
10	parameters	limit	respectively

Figure 4: mlhLDA representation of the ApJ article "Infrared Colors of the Gamma-Ray Detected Blazers".

amples of 2 article sections (out of 7), their inferred topic distributions along with the top 10 words for each of the top 3 section topics.

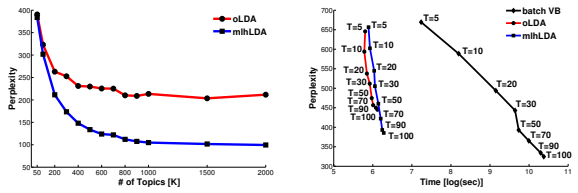


Figure 5: oLDA vs. mlhLDA: perplexity comparison (left); speed vs. perplexity comparisons with batch VB (right).

The left side of Figure 5 shows the held-out perplexity comparison between oLDA and mlhLDA across 13 different topic configurations. For this set of experiments we used the above training set of 130k articles and the set of 8,078 held-out articles. From these comparisons we clearly see the advantage of using the multi-level Dirchlet prior structure. Another way of evaluating topic models is through an extrinsic evaluation task which was not available for this collection. In the case of oLDA, article sections were treated as individual documents. In the original oLDA[1] implementation the per document concentration parameter α_d was set to $\frac{1}{K}$ which we also use in our case for both the symmetric θ_d and asymmetric θ_s (same goes for PLTM and mlhPLTM). Since in our case we perform relative comparison between oLDA and mlhLDA we weren't concerned with experimenting with different concentration parameters but we rather used the default one implemented in oLDA.

With a random subset of 10k training and 1k held-out articles we compared the performance of oLDA and mlhLDA with the original batch VB[2] implementation of Blei et al. (2003). Unlike the implementations of oLDA and mlhLDA which are written in Python the original VB algorithm is written in C and requires multiple iterations over the whole collection. The right side of Figure 5 shows the speed (in natural log scale) vs. perplexity comparison across the three models.

5 Modeling and Retrieving Speeches in Europarl Sessions

We compared the modeling performance of oPLTM and mlhPLTM on a subset of the English-Spanish Europarl collection (Koehn, 2005). The subset consists of ∼64k training pairs of English-Spanish speeches that are translations of each other which originate from 374 sessions of the European Parliament (Europarl) and a test set of ∼14k speech translation pairs from 112 sessions. With oPLTM we modeled individual speech pairs while with mlhPLTM we utilized the session hierarchy and modeled pairs of speeches as document sections. Comparisons were performed intrinsically (using perplexity) and extrinsically on a cross-language information retrieval (CLIR) task. This task, along with the Europarl subset, have been previously defined by Mimno et al. (2009) and used across other publications (Platt et al., 2010; Krstovski and Smith, 2013). Given a query English speech, the CLIR task is to retrieve its Spanish translation equivalent. It involves performing comparison across topic representations of all Spanish speeches using Jensen-Shannon divergence and sorting the results. Models are evaluated using precision at rank one (P@1). Figure 6 shows the CLIR task performance comparisons results using 13 different topic configurations. We performed comparisons across three different settings of the concentration parameters α_d and α_s (α_d=α_s=$\frac{1}{K}$, 0.4 and 1.0).

[1] http://www.cs.princeton.edu/~mdhoffma

[2] http://www.cs.princeton.edu/~blei/lda-c

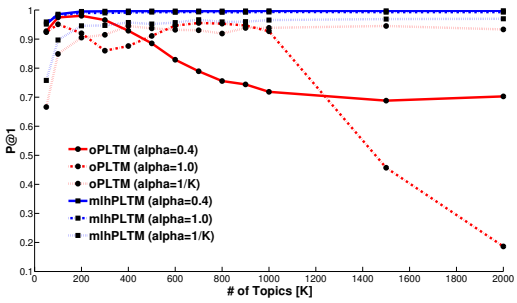


Figure 6: oPLTM vs. mlhPLTM: Performance comparison on the CLIR task using chronological ordering of sessions across different hyperparameter settings, α_d=α_s=$\frac{1}{K}$, 0.4 and 1.0 .

Across the different concentration parameter values and across the 13 different topic configurations we observe that the performance of oPLTM fluctuates as we increase the numbers of topics. On the other hand, across the three different concentration parameter settings, mlhPLTM performance is very steady and tends to increase with the number of topics. Across the different topic configurations both models provide the best performance with $\alpha_d = \alpha_s = 0.4$. Setting the concentration parameters to $\frac{1}{K}$ gives the overall worst performance.

In our initial experiments we unintentionally reordered our set of training Europarl sessions based on two digit years which was different from the experimental setup in (Mimno et al., 2009) and (Krstovski and Smith, 2013) where the order of the presentation data (Europarl speeches) was chronological. This emphasized the fact that in online VB, order of presentation of documents plays an important role especially in the training step where the model learns the per topic-word distributions. Figure 7 shows the performance comparison results between oPLTM and mlhPLTM when documents in the training and test steps are ordered numerically. In our initial experimental setup concentration parameters where set to $\alpha_d = \alpha_s = \frac{1}{K}$. To the left is the perplexity comparison between the two models. The CLIR task performance comparisons results are shown on the right. Unordered mlhPLTM achieves high P@1 after 2,000 topics. While it takes much longer in terms of the number of topics unordered mlhPLTM ultimately achieves similar performance results as ordered mlhPLTM.

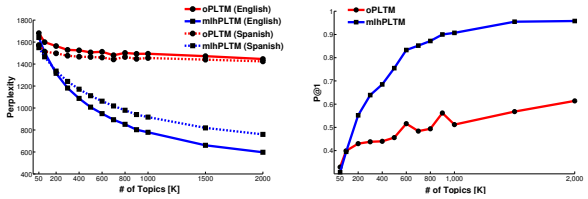


Figure 7: oPLTM vs. mlhPLTM: perplexity comparison (left); performance comparison on the CLIR task (right). Documents were presented out of chronological order and thus performance is lower, especially for oPLTM.

6 Conclusion

We presented online topic models with multi-level Dirichlet prior structure that provide better modeling of topical variations across document sections in mono- and multilingual collections. We showed that documents with rich sub-document level structure could be modeled with higher likelihood compared to regular online LDA and PLTM models while offering the same efficiency. Furthermore on the task of retrieving document translations we showed that mlhPLTM achieves significantly better retrieval results compared to online PLTM.

Acknowledgments

This work was supported in part by the Harvard-Smithsonian CfA predoctoral fellowship, 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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3:993–1022.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online learning for latent Dirichlet allocation. In *NIPS*, pages 856–864.

Shoaib Jameel and Wai Lam. 2013. An unsupervised topic segmentation model incorporating word order. In *SIGIR*, pages 203–212.

Do-Kyum Kim, Geoffrey Voelker, and Lawrence K. Saul. 2013. A variational approximation for topic modeling of hierarchical corpora. In *ICML*, pages 55–63.

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

Kriste Krstovski and David A. Smith. 2013. Online polylingual topic models for fast document translation detection. In *WMT*, pages 252–261.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and Joyce M. Watson. 2000. The nasa astrophysics data system: Overview. *Astronomy and Astrophysics Supplement Series*, 143:41–59.

David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*, pages 880–889.

Thomas P. Minka. 2000. Estimating a dirichlet distribution. Technical report, MIT.

John Platt, Kristina Toutanova, and Wen tau Yih. 2010. Translingual document representations from discriminative projections. In *EMNLP*, pages 251–261.

Hanna M. Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *NIPS*, pages 1973–1981.

Hanna M. Wallach. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*, pages 697–702.