A Nonparametric N-Gram Topic Model with Interpretable Latent Topics*

Shoaib Jameel and Wai Lam

Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong
{msjameel,wlam}@se.cuhk.edu.hk

Abstract. Most nonparametric topic models such as Hierarchical Dirichlet Processes, when viewed as an infinite-dimensional extension to the Latent Dirichlet Allocation, rely on the bag-of-words assumption. They thus lose the semantic ordering of the words inherent in the text which can give an extra leverage to the computational model. We present a new nonparametric topic model that not only maintains the word order in the topic discovery process, but also generates topical n-gram words leading to more interpretable latent topics in the family of the nonparametric topic models. Our experimental results show an improved performance over the current state-of-the-art topic models in document modeling and generating n-gram words in topics.

Keywords: Bayesian Nonparametrics, N-gram words, Perplexity, N-gram topic model, Collocations.

1 Introduction

Nonparametric topic models such as Hierarchical Dirichlet Processes (HDP) [28], when viewed as an infinite-dimensional extension to the fixed-dimension Latent Dirichlet Allocation (LDA) model [4] and [3], have gained immense popularity in recent years because in that one does not need to explicitly provide the number of topics apriori. However, a limitation of the HDP model is that it loses important structural information present in the text leading to undesirable effects such as producing ambiguous terms in topics. For example, due to its bag-of-words assumption, HDP discovers unigrams such as "networks" in a topic which does not seem to be that insightful. Instead finding n-gram words can convey more interpretable meaning to readers [31] and [23], for example, "neural networks". Also, word order is important to many aspects of linguistic processing [26] and [21]. Related works in parametric topic modeling, such as the bigram topic model (BTM) [30], the LDA Collocation model [18] (LDACOL), the topical n-gram model [31] (TNG), which maintain the order of the words in the document have shown to perform better than the bag-of-words counterpart models.

* The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: CUHK413510) and the Direct Grant of the Faculty of Engineering, CUHK (Project Code: 2050522). This work is also affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies.

R.E. Banchs et al. (Eds.): AIRS 2013, LNCS 8281, pp. 74–85, 2013.

Processing documents by keeping the word ordering intact, such as the existing parametric topic models mentioned earlier, does incorporate additional computational burden, nonetheless it gives an upper-hand over traditional bag-of-words topic models [19]. One useful advantage is to discover more interpretable latent topics. However, one common limitation of these existing parametric n-gram topic models which consider word order is that they require the number of topics to be supplied by the user in advance. In reality a user is completely ignorant about the number of topics that may uncover the true latent structure of the corpus. It is therefore more reasonable to develop a nonparametric model which can automatically infer a desirable number of latent topics via the data characteristics inherent in a collection of text documents.

We develop a new nonparametric topic model, which we name as NHDP, by extending the HDP model so that word order is taken into consideration during the topic discovery process. Our proposed NHDP model not only maintains the document's word order information, but also discovers topical n-gram words based on context. By generating n-gram in topics helps in better topic interpretation because n-gram words are more insightful to the reader than unigram words [23].

2 Related Work

Considering the importance of the word order in nonparametric setting is becoming to attract attention. For example, Goldwater et al. [15] presented two nonparametric word segmentation models where one of the models, called the bigram HDP model, maintains the ordering in text. Related extensions are described in [5] and [14]. They are well catered to the word segmentation task. In their model, contextual dependencies are distributed according to a Dirichlet Process (DP) specific to the words in a document and it closely resembles the hierarchical Pitman-Yor processes model [27] and [16]. In [11], the author introduced a nonparametric model that can extract phrasal terms based on the mutual rank relation. This model first extracts phrases and subsequently ranks them. It employs a heuristic measure for the identification of phrasal terms. In [25], the authors introduced the notion of extension pattern, which is a formalization of the idea of extending lexical association measures defined for bigrams. In [33], the authors presented a Bayesian nonparametric model for symbolic chord sequences. Their model is designed to handle n-grams in chord sequences for music information retrieval. Our proposed model is significantly different than the ones mentioned above. First, our model is an n-gram nonparametric topic discovery model capturing word dependencies in text. Consequently, it can generate more interpretable topics.

Some nonparametric language models have been proposed recently which maintain the word order, for example, [27], [32], etc. But there are differences between language models and topic models [29]. For example, language models do not discover topics, which typically is a probability distribution over words. Also, language models focus on representing local linguistic structure, as expressed by word order [8] whereas topic models focus on finding topics.

Parametric and nonparametric syntax based models also capture word dependencies using an extra layer of Hidden Markov Model (HMM). But they are different from our model in that we do not incorporate a HMM model in our NHDP model to capture word

dependencies. For example, in [17] the authors introduced a parametric Bayesian topic model which can not only capture the semantic information inherent in the text, but also capture the syntax in the document by introducing an extra layer of HMM in the model. This model was later extended to a nonparametric setting [10] and [6] where the author introduced HDP model instead of its parametric counterpart, Latent Dirichlet Allocation (LDA) [4]. In [13], the authors presented the sticky HDP-HMM model for speaker diarization. Their model segments a piece of audio discourse using an augmented HDP-HMM that provides effective control over the switching rate in the audio data. The existing HMM based topic models are designed to capture the syntactic classes such as part-of-speech. In contrast, our model does not assume that syntax information is available.

Some existing parametric topic models discover n-gram words. But these models assume that the number of topics is known in advance. We believe that this is a major shortcoming because the desirable number of topics that describes the collection is typically not known in advance. One approach to solving this problem is to train several models with different numbers of topics and choose the one that performs reasonably well according to a performance measure [9]. But this is cumbersome and time consuming [10]. Note that selecting less number of topics than what the data can actually accommodate will result in under-fitting whereas selecting more number of topics will result in over-fitting. The LDA model [4], which is a basic parametric topic model, assumes "exchangeability" [1] among the words in the document. In [18], the authors proposed an extension to the LDA model, called the LDA Collocation (LDACOL) model. This model introduces a set of random variables which capture whether words in order form collocations. Each word has a topic assignment and a collocation assignment. The collocation variable can take on two values, namely, 0 and 1. If the collocation variable is 1, then the word is generated from the distribution based on just the previous word. Otherwise, the word is generated from a distribution associated with its topic. In this way, the model can generate both unigram and bigram words. Wang et al. [31] extended the LDA Collocation model and proposed the topical n-gram (TNG) model which makes it possible to decide whether to form a bigram for the same two consecutive words depending on their nearby context. However, this model suffers from some drawbacks such as words within a topical n-gram do not share the same topic. Moreover, the topic-specific bigram distributions share no probability mass between each other or with the unigram distributions. These shortcomings were addressed recently in another parametric topic model [23] based on the Hierarchical Pitman-Yor Processes [27]. However, their model becomes overly complex and it is inefficient for handling large datasets. Wallach in [30] proposed the bigram topic model which is an extension to the LDA and it maintains the word order in the document, but the model only generates bigrams in topics. In [20] Johnson described a connection between probabilistic context-free grammars PCFG and the LDA model. This paper shows how the LDA model can be expressed as a PCFG. The LDA model is employed to generate collocations of words apart from applying the model in other natural language processing task. The difference between Johnson's work in [20] and our paper is that we generate word collocations in a nonparametric setting whereas Johnson used the LDA model to generate word collocations. Recently, in [22] the authors presented a study where they

considered bigrams as a single token and used bigrams as features to be given to a topic model. The authors presented extensive experiments how collocations can help improve a topic model in empirical evaluations. Their method has a limitation in that one has to manually supply bigrams to the model rather than the bigrams automatically discovered by the model itself. In [2] the authors presented an application of topic models to recommender systems. The authors presented topic models where the models maintain the ordering of the words in sequence and in turn obtain better empirical results in their experimental analysis. However, their model does not generate n-gram words based on the co-occurrences in the data.

3 Background

In order to circumvent the limitation prevalent in parametric topic models, Teh et al. [28] proposed the Hierarchical Dirichlet Processes (HDP) model. This model can be regarded as a nonparametric version of the LDA model [10]. We will mainly describe the HDP model in the context of topic modeling.

HDP is a nonparametric Bayesian model which is a Bayesian model on an ∞-dimensional parameter space. For nonparametric models, the number of parameters grows with the sample size. Here we give a succinct description of the HDP model whose one of the applications is also topic modeling. Inquisitive readers are requested to consult [28] for more details.

Given a collection of text documents, HDP is characterized by a set of random probability measures G_d for each document d in the collection. In addition, a global random probability measure G_0 which itself is drawn from a Dirichlet Process (DP) with the base probability measure H. The global measure G_0 selects all the possible topics from the base measure H, and then each G_d draws the topics necessary for the document d from G_0. The model is defined as:

$$\begin{aligned} G_0|\gamma, H &\sim \mathbf{DP}(\gamma, H) \\ G_d|\alpha, G_0 &\sim \mathbf{DP}(\alpha, G_0) \\ z_{di}|G_d &\sim G_d \\ w_{di}|z_{di} &\sim \mathbf{Multinomial}(z_{di}) \end{aligned}$$

where γ and α are the concentration parameters that govern the variability around G_0 and G_d respectively. The base probability measure H provides the prior distribution for the factors or topics z_{di}. Each z_{di} is a factor corresponding to a single observation w_{di} which is the word at the position i in the document d.

One perspective associated with the HDP mechanism can be expressed by the Chinese Restaurant Franchise (CRF) [28] which is an extension of the Chinese Restaurant Process (CRP). In order to describe sharing among the groups, the notion of "franchise" has been introduced that serves the same set of dishes globally. When applied to text data, each restaurant corresponds to a document. Each customer corresponds to a word. Each dish corresponds to a topic. A customer sits at a table, one dish is ordered for that table and all subsequent customers who sit at that table share that dish. The dishes are sampled from the base distribution H which corresponds to discrete topic distributions.

Multiple tables in multiple restaurants can serve the same dish. The factor values are shared both between and amongst documents. For a complete mathematical derivation of the CRF metaphor, we direct the reader to review [28].

4 Our N-Gram HDP Model (NHDP)

4.1 Model Description

We describe our n-gram nonparametric topic model, called NHDP, which is an extension to the basic HDP model described in Section 3. Unlike the basic HDP model, our proposed NHDP model is no longer invariant to the reshuffling of words in a document.

We introduce a set of binary random variables $\mathbf{x}$ which we term as the concatenation indicator variable that assume either of the two values which are 0 or 1. This variable indicates whether two words in consecutive order can be concatenated or not. Note that NHDP uses the first order Markov assumption on the words. There are two assignments per word w_{di} at position i in the document d, and $1 \leq i \leq N_d$ where N_d is the number of words (unigrams) in the document d. One assignment is the topic and the other assignment is the concatenation indicator variable x_{di} which relates to whether the word w_{di} can be concatenated with the previous word $w_{d,i-1}$. If $x_{di} = 1$, then w_{di} is part of a concatenation and the word is generated from a distribution that is dependent only on $w_{d,i-1}$. x_{di} is drawn from $P(x_{di}|w_{d,i-1})$. On the other hand, if $x_{di} = 0$, then w_{di} is generated from the distribution associated with its topic. We assume that the first indicator variable x_{d1} in a document is observed and set to 1, and only a unigram is allowed at the beginning of the document. In fact, we can also enforce other constraints in the model. Some examples are: no concatenation is allowed for sentence or paragraph boundary, only a unigram is allowed after a stopword is removed from that position, etc.

Note that NHDP can capture word dependencies in the document. The conditional probability $P(w_{di}|w_{d,i-1})$ can be written as:

$$P(w_{di}|w_{d,i-1}) = P(w_{di}|w_{d,i-1}, x_{di} = 1)P(x_{di} = 1|w_{d,i-1}) + P(w_{di}|w_{d,i-1}, x_{di} = 0) P(x_{di} = 0|w_{d,i-1}) \tag{1}$$

We can observe that $P(w_{di}|w_{d,i-1}, x_{di} = 0)$ can be computed using the basic HDP model. The full definition of our NHDP model is given as follows:

$G_0|\gamma, H \sim \mathbf{DP}(\gamma, H)$;
$G_d|\alpha, G_0 \sim \mathbf{DP}(\alpha, G_0)$;
$z_{di}|G_d \sim G_d$;
$x_{di}|w_{d,i-1} \sim \mathbf{Bernoulli}(\psi_{w_{d,i-1}})$;
if $x_{di} = 1$ **then**
  $w_{di}|w_{d,i-1} \sim \mathbf{Multinomial}(\sigma_{w_{d,i-1}})$
end
else
  $w_{di}|z_{di} \sim F(z_{di})$
end

Note that in the definition of our model the hyperprior of σ is δ. The hyperprior value of ψ is ϵ. Just as in the HDP model described earlier, the distribution $F(z_{di})$, is the Multinomial distribution in line 9 in the above generative process. We can obtain higher order n-grams by concatenating the current concatenated words with the next n-gram based on the value obtained by the next concatenation indicator variable. Although our model does not directly generate topic-wise n-grams, an n-gram can be associated with a topic via a simple post-processing strategy. One strategy is to take the topic of the first term in the n-gram as the topic for the whole n-gram. This technique has been used in [24] for the LDACOL model. Another strategy is to assume the topic of the n-gram as the most common topic occurring in the words involving in that n-gram [24].

4.2 Posterior Inference

Our inference scheme is based on the Chinese Restaurant Franchise scheme [28] with some modifications. In our scheme, we have to handle two different conditions. The first condition is concerned with $x_{di} = 0$ whereas the second condition is concerned with $x_{di} = 1$. Note that for some observed x_{di}, only z_{di} needs to be drawn.

In the document modeling setting, each document is referred to as a restaurant and words in the document are referred to as customers. The set of documents share a global menu of topics. The words in the document are divided into groups, each of which shares a table. Each table is associated with a topic and words around each table are associated with the table's topic.

The First Condition: The first condition refers to $x_{di} = 0$. In this setting, most of the modeling will resemble the HDP model as presented in [28], but in our case we need to derive updates for the HDP model for text data.

We will sample t_{di} which is the table index for each word w_{di} at the position i in the document d. We will then sample k_{dt} which is the topic index variable for each table t in d. $k_{d\hat{t}}$ is the new topic index variable created for a new table. Note that we will only sample the index variables here rather than the distributions themselves [10]. We define $\mathbf{w}$ as $(w_{di} : \forall d, i)$ and $\mathbf{w}_{dt}$ as $(w_{di} : \forall i \text{ with } t_{di} = t)$, $\mathbf{t}$ as $(t_{di} : \forall d, i)$ and $\mathbf{k}$ as $(k_{dt} : \forall d, t)$. In addition, we also define $\mathbf{x}$ as $(x_{di} : \forall d, i)$. When a superscript is attached to a set of variables or count, for example, $(\mathbf{k}^{\neg dt}, \mathbf{t}^{\neg di})$, it means that the variables corresponding to the superscripted index are removed from the set or from the calculation of the count. Each word whose $x_{di} = 0$ is assumed to be drawn from $F(z)$ whose density is written as $f(.|\phi)$ (f is just one part obtained from F). This density is the multinomial distribution with the parameter ϕ. The likelihood of w_{di} for $t_{di} = t$ where t is an existing table, denoted as $f_k^{\neg w_{di}}(w_{di})$, is the conditional density of w_{di} given all words in topic k except w_{di}:

$$f_k^{\neg w_{di}}(w_{di}) = \frac{\int f(w_{di}|\phi_k) \prod_{d'i' \neq di, z_{d'i'}=k} f(w_{d'i'}|\phi_k) h(\phi_k) \mathrm{d}\phi_k}{\int \prod_{d'i' \neq di, z_{d'i'}=k} f(w_{d'i'}|\phi_k) h(\phi_k) \mathrm{d}\phi_k} \tag{2}$$

where h is a probability density function of H and H is a Dirichlet distribution over a fixed vocabulary of size V. $h(.)$ is the Dirichlet distribution with the parameter η. ϕ_k is one of the global topics with which each table is associated which is indicated with a table-specific topic index k_{dt}. Furthermore, Equation 2 can be simplified as:

$$f_k^{\neg w_{di}}(w_{di} = \vartheta) = \frac{n_{..k}^{\neg w_{di},\vartheta} + \eta}{n_{..k}^{\neg w_{di}} + V\eta} \tag{3}$$

where $n_{..k}^{\neg w_{di}}$ is the number of words belonging to the topic k in the corpus whose $x_{di} = 0$ excluding w_{di}. $n_{..k}^{\neg w_{di},\vartheta}$ is the number of times the word ϑ is assigned with the topic k excluding w_{di} and whose x_{di} is 0. Furthermore, V is the number of words in the vocabulary which is typically fixed and is known. The likelihood of w_{di} for $t_{di} = \hat{t}$, where $\hat{t}$ is the new table being sampled, is written as:

$$P(w_{di}|t_{di} = \hat{t}, \mathbf{t}^{\neg di}, \mathbf{k}) = \sum_{k=1}^{K} \frac{m_{.k}}{m_{..} + \gamma} f_k^{\neg w_{di}}(w_{di}) + \frac{\gamma}{m_{..} + \gamma} f_{\hat{k}}^{\neg w_{di}}(w_{di}) \tag{4}$$

where $\hat{k}$ is the new topic being sampled. $m_{.k}$ is the number of tables belonging to the topic k in the corpus. $m_{..}$ is the total number of tables in the corpus. $f_{\hat{k}}^{\neg w_{di}}(w_{di}) = \int f(w_{di}|\phi)h(\phi)\mathrm{d}\phi$ is the prior density of w_{di}. γ is the concentration parameter as described in Section 3. Since we follow the standard Chinese Restaurant Franchise sampling procedure, the conditional density for t_{di} for Gibbs sampling, the conditional densities for $k_{d\hat{t}}$ and k_{dt} can be found in [10].

The Second Condition: The second condition refers to $x_{di} = 1$. We only need to sample the probability of a topic in a document as the current word w_{di} is generated by the previous word $w_{d,i-1}$. In order to do this, we proceed as follows:

$$P(k_{dt} = k|\mathbf{t}, \mathbf{k}^{\neg dt}) \propto \begin{cases} m_{.k}^{\neg dt} f_k^{\neg \mathbf{w}_{dt}}(\mathbf{w}_{dt}) & \text{if } k \text{ is already used} \\ \gamma f_{\hat{k}}^{\neg \mathbf{w}_{dt}}(\mathbf{w}_{dt}) & \text{if } k = \hat{k} \end{cases} \tag{5}$$

where $f_k^{\neg \mathbf{w}_{dt}}(\mathbf{w}_{dt})$, which is the conditional density of $\mathbf{w}_{dt}$ given all words associated with the topic k leaving out $\mathbf{w}_{dt}$ is defined as:

$$f_k^{\neg \mathbf{w}_{dt}}(\mathbf{w}_{dt}) = \frac{\Gamma(n_{..k}^{\neg \mathbf{w}_{dt}} + V\eta)}{\Gamma(n_{..k}^{\neg \mathbf{w}_{dt}} + n^{\mathbf{w}_{dt}} + V\eta)} \times \frac{\prod_{\vartheta} \Gamma(n_{..k}^{\neg \mathbf{w}_{dt},\vartheta} + n^{\mathbf{w}_{dt},\vartheta} + \eta)}{\prod_{\vartheta} \Gamma(n_{..k}^{\neg \mathbf{w}_{dt},\vartheta} + \eta)} \tag{6}$$

where $n^{\mathbf{w}_{dt}}$ is the total number of words at the table t whose $x_{di} = 0$. $n^{\mathbf{w}_{dt},\vartheta}$ is the number of times the word ϑ appears at the table t with the assignment $x_{di} = 0$. $n_{..k}^{\neg \mathbf{w}_{dt}}$ is the number of words belonging to topic k in the corpus except $\mathbf{w}_{dt}$.

Sampling the Concatenation Indicator Variables: We present how to sample the values of the indicator variables. The idea is to compute the probabilities of how often two words consecutively occur in sequence. Then based on the probability value, the indicator variable is set to either 0 or 1. Let $n_0^{w_{d,i-1}}$ and $n_1^{w_{d,i-1}}$ be the number of times word $w_{d,i-1}$ has been drawn from a topic or formed a part of a concatenation respectively and all counts exclude the current case. ϵ_0 and ϵ_1 are the priors of the binomial distribution. $n_{w_{di}}^{w_{d,i-1}}$ is the number of times the word w_{di} comes after the word $w_{d,i-1}$. $n_{..k}^{\neg w_{di},\vartheta}$ and $n_{..k}^{\neg w_{di}}$ have been defined in Equation 3.

$$P(x_{di}=0|\mathbf{x}_{\neg di},\mathbf{w},\mathbf{k}) \propto \frac{n_0^{w_{d,i-1}}+\epsilon_0}{\sum_{c=0}^{1} n_c^{w_{d,i-1}}+\epsilon_0+\epsilon_1} \times \frac{n_{..k}^{\neg w_{di},\vartheta}+\eta}{n_{..k}^{\neg w_{di}}+V\eta} \tag{7}$$

$$P(x_{di}=1|\mathbf{x}_{\neg di},\mathbf{w},\mathbf{k}) \propto \frac{n_1^{w_{d,i-1}}+\epsilon_1}{\sum_{c=0}^{1} n_c^{w_{d,i-1}}+\epsilon_0+\epsilon_1} \times \frac{n_{w_{di}}^{w_{d,i-1}}+\delta}{\sum_{v=1}^{V} n_v^{w_{d,i-1}}+V\delta} \tag{8}$$

where δ is same as described in Section 4.1.

5 Empirical Evaluation

5.1 Test Collections

We used several corpora in our experiments. One corpus is the NIPS[1] collection often used in the topic modeling literature. Note that the original raw NIPS corpus consists of 17 years of conference papers. But we supplemented this corpus by including some new raw NIPS documents[2] and it has 19 years of papers in total. Our NIPS collection consists of 2,741 documents comprising of 4,536,069 non-unique words and 94,961 words in the vocabulary. The second corpus is the Associated Press (AP) corpus. We have obtained this corpus from the LDA-C[3] package. This corpus consists of 2,243 documents with 38,631 words in the vocabulary. We also use one of the datasets from the 20 Newsgroups corpus for showing qualitative results. We have chosen the computer (indexed as "comp" available in the corpus) dataset from the 20 Newsgroups corpus. We removed stopwords[4] from the collections, but did not perform word stemming.

5.2 Comparative Methods

One of the comparative methods is the basic `HDP` model proposed in [28]. We also chose the `LDACOL` model proposed in [18] for our comparative study. This model can be regarded as a parametric version close to our model. In addition to our proposed full `NHDP`, we also investigated a variant of our model, where we set all $x_{di}=1$. We call this model as `Bi-NHDP` in the experiments. Note that we do not expect the `Bi-NHDP` model to generate interpretable latent topics because it always generates bigrams just like the `BTM` [30] model. We do not compare with the `TNG` model [31] because the `TNG` model performs topic sampling for every word in a bigram. Neither our model nor `LDACOL` employ this sampling. Also we only chose strong closely related comparative methods here. It has already been demonstrated through quantitative analysis in [31] that the `LDACOL` model is more powerful than the `BTM` model. Also, in [30] it has been has shown that the `BTM` outperforms the `LDA` model in several quantitative experiments. Hence we do not compare our model with the `BTM` and the `LDA` models.

[1] `http://www.cs.nyu.edu/~roweis/data.html`
[2] `http://ai.stanford.edu/~gal/Data/NIPS/`
[3] `http://www.cs.princeton.edu/~blei/topicmodeling.html`
[4] `http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop`

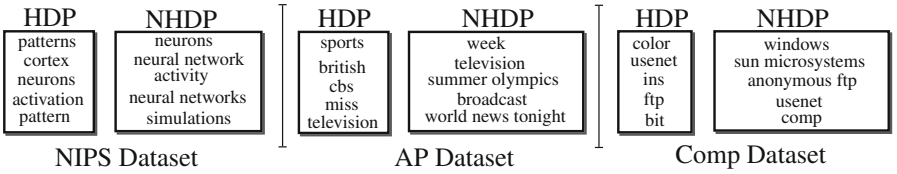

Fig. 1. Qualitative comparison of our model NHDP with the HDP model

5.3 Experimental Setup

The number of iterations for the Gibbs sampler for all models is 1,000. We have set $\delta = 0.1$, $\epsilon_0 = 0.1$, and $\epsilon_1 = 0.1$ which are the new parameters introduced in our model NHDP. We used a symmetric Dirichlet distribution with parameter of 0.5 for the prior H over topic distributions and the concentration parameters were set as $\gamma \sim \text{Gamma}(1, 0.1)$ and $\alpha \sim \text{Gamma}(1, 1)$ for our and the HDP models. η was set to 0.01. For the LDACOL model, following the notations described in [31], we set $\beta = 0.01$, $\alpha = 50/T$, $\gamma_0 = 0.1$, $\gamma_1 = 0.1$ and $\delta = 0.1$. Note that T is the number of topics which is pre-defined for the LDACOL model. The same hyperparameter values are also used in the LDACOL implementation available publicly[5]. Since the LDACOL model requires the number of topics to be supplied by the user, we conducted several runs by varying the number of topics and measured the performance at different number of topics. The best result value was chosen based on all values obtained.

5.4 Qualitative Results

We first show how our model generates more interpretable topics with topical n-grams. Here we employed a strategy that using the topic of the first word as the topic of the entire n-gram (refer Section 4.1 for more details). Our main comparative method for qualitative analysis is the HDP model which belongs to a family of nonparametric topic models. We manually chose top five words occurring with high probability in some topics. From Figure 1, we can see that compared to the HDP model, our NHDP model has generated words which are more coherent and has provided an extremely salient summary about "neural networks" in the NIPS collection, media related information obtained from the Associated Press (AP), and computer technology related words from the 20 Newsgroups computer dataset (denoted as "Comp Dataset" in Figure 1). The results show that our model has produced interpretable topics.

Both LDACOL and NHDP can generate bigrams such as "neural networks", etc. But the merit of NHDP lies in the fact that it does not require the number of topics to be specified explicitly by the user and it is automatically inferred from the data characteristics. Moreover, NHDP can produce topical n-grams, not only restricted to bigrams. We thus investigate how well they perform in some typical text analysis tasks, such as document modeling, as described next.

[5] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

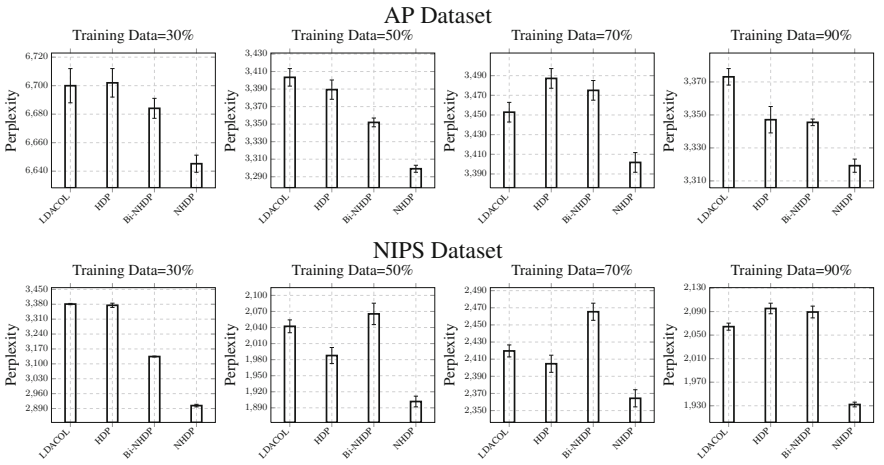

Fig. 2. Perplexity results for the AP and NIPS datasets

5.5 Document Modeling

Several works in the topic modeling literature have used perplexity analysis to compare the generalization ability of the model on unseen data as exemplified in [12]. Evaluation using perplexity is important for our model because it is well suited for models where word order in the document is maintained [7]. Generally the documents in the collection are treated as unlabeled. Thus, our goal is density estimation. We wish to achieve a high likelihood on the held-out test set. We first train the parameters of the model using a training set and subsequently the unseen data is fed to the learnt model in order to measure its generalization ability. A commonly used metric is the perplexity score. A lower perplexity score indicates better generalization performance. More formally, for a test set $\mathbf{D}$ consisting of D documents, the perplexity score is written as:

$$\text{Perplexity}(\mathbf{D}) = \exp\left\{ -\frac{\sum_{d=1}^{D} \log \mathrm{P}(\mathbf{w}_d)}{\sum_{d=1}^{D} N_d} \right\} \tag{9}$$

where $\mathbf{w}_d$ are the words in the document d and N_d is the number of words in the document d.

We conducted perplexity analysis by experimenting on two datasets described earlier (AP and NIPS datasets). Perplexity analysis was conducted by running the Gibbs sampler 3 times each with 1,000 iterations and the average of the three perplexity values was taken. For all the datasets, we split the datasets randomly into two subsets each. One subset is the training set and the other subset is the testing set. We conducted several runs by varying the split proportion obtaining different amounts of the training set from 30% to 90% in steps of 20%. The purpose is to study how the models perform on different sizes of the training set.

From the results depicted in Figure 2, our model outperforms all comparative models in terms of generalizing on the unseen data. The computed average perplexity values for

our NHDP model are statistically significant compared to the HDP and LDACOL models according to a two-tailed statistical significance test with $p < 0.05$ in all corpora. The variant of our model, namely, Bi-NHDP does not perform at par with our NHDP model. The LDACOL model also does not generalize well on the unseen data. In Figure 2, we see the effect of varying the training size on different models. For example, in the AP and NIPS datasets, our model generally performs extremely well in different training portions. One can note that even when the training data is less, our NHDP model generalizes well on the unseen data. In contrast, HDP loses important structural information in the document.

6 Conclusions and Future Work

We have presented a new nonparametric n-gram topic model that maintains the order of the words in the document. Word ordering plays a vital role in many linguistic tasks. An important innovation that we introduce in our work is generating n-gram words in topics where the number of topics need not be specified by the user. We have shown better quantitative performance in generalizing on an unseen data in two document collections. Our model generates more interpretable latent topics with n-gram words, whereas the existing nonparametric topic model HDP fails to generate such n-grams which are more insightful to a reader.

In the future, we intend to extend our model in generating n-gram words in topics over time as test collections in general are dynamic and topics change over time. Another direction which we wish to investigate is to incorporate text segmentation in a nonparametric setting and capturing n-gram words in each segment.

References

1. Aldous, D.: Exchangeability and related topics. Ecole d'Ete de Probabilites de Saint-Flour XIII-1983, pp. 1–198 (1985)
2. Barbieri, N., Manco, G., Ritacco, E., Carnuccio, M., Bevacqua, A.: Probabilistic topic models for sequence data. Machine Learning 93(1), 5–29 (2013)
3. Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (2012)
4. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. JMLR 3, 993–1022 (2003)
5. Blunsom, P., Cohn, T., Goldwater, S., Johnson, M.: A note on the implementation of hierarchical Dirichlet processes. In: Proc. of ACL-IJCNLP, pp. 337–340 (2009)
6. Boyd-Graber, J., Blei, D.M.: Syntactic topic models. In: Proc. of NIPS (2008)
7. Caballero, K.L., Barajas, J., Akella, R.: The generalized Dirichlet distribution in enhanced topic detection. In: Proc. of CIKM, pp. 773–782 (2012)
8. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proc. of ACL, pp. 310–318 (1996)
9. Claeskens, G., Hjort, N.: Model selection and model averaging. Cambridge Books (1993)
10. Darling, W.: Generalized Probabilistic Topic and Syntax Models for Natural Language Processing. Ph.D. thesis (2012)
11. Deane, P.: A nonparametric method for extraction of candidate phrasal terms. In: Proc. of ACL, pp. 605–613 (2005)
12. Fang, Y., Si, L., Somasundaram, N., Yu, Z.: Mining contrastive opinions on political texts using cross-perspective topic model. In: Proc. of WSDM, pp. 63–72 (2012)

13. Fox, E., Sudderth, E., Jordan, M., Willsky, A.: A sticky HDP-HMM with application to speaker diarization. The Annals of Applied Statistics 5(2A), 1020–1056 (2011)
14. Goldwater, S., Griffiths, T., Johnson, M.: A Bayesian framework for word segmentation: Exploring the effects of context. Cognition 112(1), 21–54 (2009)
15. Goldwater, S., Griffiths, T.L., Johnson, M.: Contextual dependencies in unsupervised word segmentation. In: Proc. of ACL, pp. 673–680 (2006)
16. Goldwater, S., Griffiths, T., Johnson, M.: Interpolating between types and tokens by estimating power-law generators. In: Proc. of NIPS, vol. 18, p. 459 (2006)
17. Griffiths, T.L., Steyvers, M., Blei, D., Tenenbaum, J.: Integrating topics and syntax. In: Proc. of NIPS, vol. 17, pp. 537–544 (2005)
18. Griffiths, T., Steyvers, M., Tenenbaum, J.: Topics in semantic representation. Psychological Review 114(2), 211–244 (2007)
19. Gruber, A., Rosen-Zvi, M., Weiss, Y.: Hidden topic Markov models. In: Proc. of AISTATS (2007)
20. Johnson, M.: PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In: Proc. of ACL, pp. 1148–1157 (2010)
21. Kim, H.D., Park, D.H., Lu, Y., Zhai, C.: Enriching text representation with frequent pattern mining for probabilistic topic modeling. JASIST 49(1), 1–10 (2012)
22. Lau, J.H., Baldwin, T., Newman, D.: On collocations and topic models. ACM Trans. Speech Lang. Process. 10(3), 10:1–10:14 (2013)
23. Lindsey, R.V., Headden, W.P., Stipicevic, M.J.: A phrase-discovering topic model using hierarchical Pitman-Yor processes. In: Proc. of EMNLP, pp. 214–222 (2012)
24. McCallum, A., Wang, X.: A note on topical N-grams. Department of Computer Science, University of Massachusetts, Amherst (2005)
25. Petrović, S., Šnajder, J., Bašić, B.: Extending lexical association measures for collocation extraction. Computer Speech & Language 24(2), 383–394 (2010)
26. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of Latent Semantic Analysis 427(7), 424–440 (2007)
27. Teh, Y.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proc. of ACL, pp. 985–992 (2006)
28. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. JASA 101(476), 1566–1581 (2006)
29. Wallach, H.M.: Structured topic models for language. Ph.D. thesis (2008)
30. Wallach, H.: Topic modeling: beyond bag-of-words. In: Proc. of ICML, pp. 977–984 (2006)
31. Wang, X., McCallum, A., Wei, X.: Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In: Proc. of ICDM, pp. 697–702 (2007)
32. Wood, F., Teh, Y.W.: A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. Journal of Machine Learning 5, 607–614 (2009)
33. Yoshii, K., Goto, M.: A vocabulary-free infinity-gram model for nonparametric Bayesian chord progression analysis. In: Proc. of ISMIR (2011)