Nonparametric Topic Modeling Using Chinese Restaurant Franchise with Buddy Customers*

Shoaib Jameel, Wai Lam, and Lidong Bing

Key Lab of High Confidence Software Technologies,
Ministry of Education (CUHK Sub-Lab)
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong
{msjameel,wlam,ldbing}@se.cuhk.edu.hk

Abstract. Many popular latent topic models for text documents generally make two assumptions. The first assumption relates to a finite-dimensional parameter space. The second assumption is the bag-of-words assumption, restricting such models to capture the interdependence between the words. While existing nonparametric admixture models relax the first assumption, they still impose the second assumption mentioned above about bag-of-words representation. We investigate a nonparametric admixture model by relaxing both assumptions in one unified model. One challenge is that the state-of-the-art posterior inference cannot be applied directly. To tackle this problem, we propose a new metaphor in Bayesian nonparametrics known as the "Chinese Restaurant Franchise with Buddy Customers". We conduct experiments on different datasets, and show an improvement over existing comparative models.

1 Introduction

Assuming the bag-of-words representation in documents has been the holy-grail in probabilistic topic modeling such as Latent Dirichlet Allocation (LDA) [1]. The bag-of-words assumption simplifies the modeling [1], and has an advantage for computational efficiency [2]. However, this assumption has some disadvantages. One major disadvantage is that many unigram words discovered in the latent topics are not very insightful to a reader [3]. Another disadvantage is that the model is not able to consider semantic information that is conveyed by the order of the words in the document [2]. This results in an inferior performance in some text mining tasks as shown by different topic models [4,5,6,7]. These models may discover many general words in latent topics with high probability instead of relevant content words [8]. In order to tackle this problem, general words are commonly removed from the corpus during text pre-processing [9], but this

* 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 Microsoft Research Asia Urban Informatics Grant FY14-RES-Sponsor-057. This work is also affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies.

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 648–659, 2015.

leads to further problems especially when processing natural language or speech data [8]. Wallach [8] has described that incorporating word order removes many general words dominating the latent topics. McCallum et al., in [9] have shown that using asymmetric priors in the LDA model can also help reduce the problem, but still the topic interpretability problem remains [3].

In order to address the limitations inherent in the unigram based topic models, some parametric topic models have been proposed which maintain the order of the words in the document. Such models are able to not only discover phrasal terms in topics [3], but also demonstrate a superior performance on several text mining tasks such as document classification [2] and document modeling [8]. It is intuitive that generating a phrasal term such as “air conditioner” is more insightful than just discovering “air” and “conditioner” independently [3,10,11]. These models have a fixed parameter space and some parameters, such as the number of topics, need to be pre-defined by the user. This might be impractical because the user may not always know the true number of latent topics inherent in the data.

One way to address the model selection issue is to train several models with different number of topics, and choose the one that has the best performance [12]. But this is not a principled approach and it is very time consuming [12]. A desirable way to deal with the problem is to automatically infer the number of latent topics based on the characteristic of document collection. Such models are known as nonparametric probabilistic topic models which are characterized by an infinite-dimensional parameter space. Most importantly, these nonparametric latent topic models impose as few assumptions as possible [13] making them more powerful than parametric latent topic models. Parametric models might face over-fitting and under-fitting issues when there is a mis-match between the model complexity and the data. In contrast, nonparametric models are less prone to this problem [14]. Models such as Hierarchical Dirichlet Processes (HDP) [15] when used as a topic model [16,17] can automatically infer the number of latent topics based on the data characteristic, but it imposes the bag-of-words assumption in documents. The “Chinese Restaurant Franchise” (CRF) metaphor has been proposed to compute the posterior distribution of HDP, which generates data from an exchangeable distribution. It thus inherits some of the limitations of the unigram based topic models.

To tackle the above issues, we propose a new metaphor in Bayesian nonparametrics called “Chinese Restaurant Franchise with Buddy[1] Customers” (CRF-BC) that not only maintains the word order, but also infers automatically the number of latent topics based on the data characteristic. Our metaphor falls in the class of non-exchangeable distributions for Bayesian nonparametric models [18]. Using the buddy assignment scheme, our model can discover n-gram words in topics. By n-gram we mean that we can discover a unigram or a bigram or even a higher-order-gram depending upon the buddy assignments. One challenge is that the state-of-the-art posterior inference cannot be applied directly. We refine the traditional Gibbs sampling algorithm for nonparametric topic modeling for

[1] Buddy is an informal term meaning a close friend. -Source: Wikipedia.

our metaphor. We conduct experiments on document modeling and show that our framework can outperform state-of-the-art topic models.

2 Related Work

Much work has been done in the parametric topic modeling literature where the order of words in documents is maintained. There are some models which use the LDA model to discover n-gram words, for example, [4]. Wallach [8] proposed the Bigram Topic Model (BTM) for text data that maintains the order of the words. Griffiths et al., [19] extended the BTM model and proposed the LDA-Collocation Model (LDACOL). In the Topical N-gram model (TNG) [10] the topic assignments for the two words in a bigram may not be alike. Lindsey et al., [3] proposed a topic model that incorporates the Hierarchical Pitman-Yor Processes (HPYP) in the LDA model. But the main concern is that the model cannot scale to accommodate large text collections due to the HPYP model [20]. Lau et al., [21] presented a study investigating whether word collocations can help improve topic models. In Johri et al., [22], the authors introduced a multi-word enhanced author-topic model for text data. In [23], the authors proposed some improvements to the n-gram topic models. Their method uses Chinese Restaurant Process (CRP) for sampling, with a fixed dimensional parameter space.

The seminal nonparametric topic model is the Hierarchical Dirichlet Processes (HDP) model proposed by Teh et al., [15]. This model assumes that words in a document are exchangeable, and thus cannot capture short-range word dependencies. CRF metaphor is also used to describe this model [17]. Considering the order of words in Bayesian nonparametrics[2] has attracted some attention recently. Goldwater et al., [25] presented two nonparametric models for word segmentation. Observing that the ordering of words could play a dominant role, Goldwater et al., extended the unigram based model to a bigram based model called the "Bigram HDP" model. The model closely resembles the HPYP model and cannot generate latent topics. It is well suited for the word segmentation task. Johnson [26] incorporated nonparametric adaptor grammars to discover word collocations instead of just unigrams. However, one disadvantage is that it adopts a two-stage approach towards collocation discovery whereas our approach can tackle it in a single model. In [27], the author introduced a nonparametric model that can extract phrasal terms based on the mutual rank relation. It employs a heuristic measure for the identification of phrasal terms. In [28], the authors introduced the notion of an extension pattern, which is a formalization of the idea of extending lexical association measures defined for bigrams. In [29], 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. Recently, we have proposed a nonparametric topic model to discover more interpretable latent topics in [6]. One main weakness of the model is that only the first term in the bigram has a topic assignment

[2] Due to space limit, we do not present a detailed background of Bayesian nonparametrics. We request inquisitive readers to consult some excellent resources [24,13].

whereas the second term does not. The model uses existing posterior inference schemes to discover collocations. Our model proposed in this paper bears some theoretical resemblance with the Distance Dependent Chinese Restaurant Process (ddCRP) [30] in which customers are first assigned to each other and this customer-customer assignment can directly be related to a clustering property. In our model, customers are first assigned to each other using the buddy assignment scheme and then the customers are assigned to tables. A franchise based model based on the ddCRP has been proposed in [31], but this model does not consider the order of words in the document. Some interesting extensions have been proposed in the past with slight modifications to the basic CRF metaphor. For example, Fox et al., [32] proposed the "Chinese Restaurant Franchise with Loyal Customers". "Chinese Restaurant Franchise with Preferred Seating" has been proposed in [33].

Our proposed model is different from the above models. In contrast with [6,10], our framework gives the same topic assignment to all the words in an n-gram. We derive a posterior inference scheme which is different from the one employed in existing models.

3 Our Proposed Model

3.1 Chinese Restaurant Franchise (CRF) Background

One perspective associated with the HDP mechanism can be expressed by the Chinese Restaurant Franchise (CRF) [15] which is an extension of the Chinese Restaurant Process (CRP). The HDP model makes use of this metaphor to generate samples from the posterior distribution given the observations. In order to describe the 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 latent 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 which corresponds to discrete topic distributions. Multiple tables in multiple restaurants can serve the same dish. A table can be regarded as the topic assignment of the words in documents.

3.2 Our Proposed CRF-BC Model

We propose a new class of non-exchangeable metaphor which considers the order of words in the document. In this metaphor, customers are first assigned to each other outside the restaurant, and subsequently, individual customers enter the restaurant and sit at tables just as in the CRF metaphor. However in order to capture n-grams words, we need to refine the existing HDP model and its inference framework which uses CRF because the existing framework does not consider word order. Our new metaphor known as "Chinese Restaurant Franchise with Buddy Customers" (CRF-BC) can capture friendship associations between customers in the entire customer-franchise setup. Our model follows a Markovian assumption on the order of words and also imposes a transitive property on that order in sequence to discover n-grams. It means that if w_i^d (w_i^d is a word at position i in the document d) is a buddy of w_{i-1}^d, and w_{i-1}^d is a buddy of w_{i-2}^d, then w_i^d is also a buddy of w_{i-2}^d. Similarly, if w_{i-1}^d is a buddy of w_{i-2}^d, and w_{i-2}^d is a buddy of w_{i-3}^d, then w_{i-3}^d and w_i^d are also buddies. Following this rule, we can obtain higher order n-grams. One can certainly impose higher order Markovian assumptions, but it would impose problems with data sparsity and high computational complexity. The idea of employing first order Markovian assumption on word order has also been used in other parametric topic models such as [2].

The general idea behind this metaphor can be described in this way. Consider a Chinese franchise with a shared menu which is shared across the restaurants. Each restaurant has an infinite set of tables as in the original CRF scheme and each restaurant corresponds to a document. Consider a set of customers, which are mainly words in the document. Some of the customers have pre-planned their visit so that they can spend time together with their "good old buddies" and eat the same food in the table. These buddies have already reserved their tables beforehand. In this scheme, we assume that the customers are waiting in the queue outside the restaurant in the same order as that of the words in a document. This assumption is different from the CRF metaphor. There might be "loners" too in the same queue who may have no buddies. They too can sit and eat in the same restaurant in any of the other unreserved tables or share the table with other lonely customers. Just as in the CRF metaphor, we assume that the loners share the same dish with other customers in that table. Note that inside the restaurant, exchangeability is still valid i.e. tables are exchangeable and so are customers who are sitting at those tables as buddies can sit in any seat at the reserved table. As every customer carries with herself a table, a buddy and word order assignments, we can easily get n-gram words in topics from these three information. We present a detailed generative mechanism of our probabilistic CRF-BC in the "restaurant-franchise representation" below.

1. Draw ϕ from **Dirichlet**$(\beta\tau)$, where β is the concentration parameter, and τ is the corpus-wide distribution over vocabulary. ϕ is the word-topic distribution matrix. We place a **Dirichlet**(κ_τ) prior over τ. We also place a **Gamma**$(\kappa_\beta^1, \kappa_\beta^2)$ over β. $\kappa_\beta^1, \kappa_\beta^2$ are the shape and scale parameters respectively. One can notice that we infer the priors by placing priors over those priors to find their posteriors. Thus the resulting inferences are less influenced by these "hyper-hyperparameters" than they are by fixing the original hyperparameters to specific values [13].
2. Draw μ from **GEM**(η).
   We place a **Gamma**$(\kappa_\eta^1, \kappa_\eta^2)$ prior over η to compute its posterior. $(\kappa_\eta^1, \kappa_\eta^2)$ are the shape and scale parameters of the Gamma distribution respectively. Readers can consult [13] for description about GEM distribution. μ actually supplies the corpus-wide distribution over topics information which follows the stick-breaking representation.
3. Draw **Discrete**(σ) from **Dirichlet**(δ).
   σ is the distribution over "buddies", and δ is its conjugate prior. We place a **Gamma**$(\kappa_\delta^1, \kappa_\delta^2)$ prior over this prior to compute the posterior of this prior.

4. Draw **Bernoulli**(ω) from **Beta**(γ_0, γ_1), where γ_0 and γ_1 are the shape parameters of the Beta distribution.
   ω is the distribution over "buddy assignment variables".
5. For each document d,
   (a) Draw **Multinomial**$(\widetilde{\theta}^d)$ from **Dirichlet**(α).
   The variable $\widetilde{\theta}^d$ will contain the per-document topic distribution, α is the prior or concentration parameter, and we determine the value of this prior by placing another prior, for example, **Gamma**$(\kappa_\alpha^1, \kappa_\alpha^2)$, where $\kappa_\alpha^1, \kappa_\alpha^2$ are the shape and the scale parameters of the Gamma distribution respectively.
   (b) Draw k_t^d from μ, where k_t^d is the topic index variable for each table t in d. μ comes from the stick breaking process.
   (c) For each word w_i^d at the position i in the document d (we are considering the word order here),
      i. Draw b_i^d from **Bernoulli**$(\omega_{t_{i-1}^d w_{i-1}^d})$.
      This is where we conduct buddy assignments. The underlying meaning is that, if $b_i^d = 0$, where b_i^d is a buddy assignment variable, then the customer (word) is a "loner" and is not a buddy with the previous customer standing in that queue, and if $b_i^d = 1$, then customer who is waiting outside the restaurant is a "buddy" with the previous customer (word) standing in the same queue. Previous customer means a customer standing in front of the current customer in the queue. This partitioning of customers or buddy assignments outside the restaurant is done based on corpus wide statistics. The first customer in the queue assumes $b_i^d = 0$. Buddy assignments not only consider the co-occurrence information, but also consider the latent topic of the previous word. In the initial run of the algorithm, this assignment is done randomly which may change by the sampler during future iterations.
      ii. Draw t_i^d from $\widetilde{\theta}^d$ if $b_i^d = 0$, otherwise $t_i^d = t_{i-1}^d$.
      This process says that if the current customer is not a buddy with the previous customer then the current customer draws a new table assignment for herself. Otherwise, if the new customer is a buddy and sits at the same table as its previous buddy and shares the same dish. t is a table or an indication of a cluster for the word i in the document d.
      iii. Draw w_i^d from $\phi_{k_{t_i^d}^d}$ if $b_i^d = 0$ else draw $\sigma_{w_{i-1}^d}$. $\phi_{k_{t_i^d}^d}$ refers to a specific value in the matrix ϕ by following the path of the table and dish assignments if the customers are not buddies. Otherwise, buddies are drawn from a distribution of the previous buddy (word). Another way to describe the process is that the customer w_i^d in the restaurant d, sat at table t_i^d while the table t in the restaurant d serves the dish k_t^d.

3.3 Posterior Inference in CRF-BC

To find the latent variables that best explain the observed data, we use Gibbs sampling. One of the main advantages of using this sampling is that it samples from a true posterior. It requires some resources on book-keeping leading to a more effective algorithm [15]. Note that in our model, we have to make significant changes at the restaurant level, and little at the franchise level of the CRF metaphor as the buddy allocation happens outside the restaurant. Due to space constraint, we present an outline of our algorithm.

We will sample t_i^d which is the table index for each word w at the position i in the document d. Let K be the total number of topics, which can either increase or decrease as the number of iterations of the sampler increases. Let $\hat{k}$ denote the new topic being sampled. We will then sample k_t^d which is the topic (dish) index variable for each table t in d. Let n_{tk}^d be the number of customers at restaurant d, sitting at table t eating dish k. We define $\mathbf{w}$ as $(w_i^d : \forall d, i)$ and $\mathbf{w}_t^d$ as $(w_i^d : \forall i \text{ with } t_i^d = t)$, $\mathbf{t}$ as $(t_i^d : \forall d, i)$ and $\mathbf{k}$ as $(k_t^d : \forall d, t)$. Let $m_{.k}$ denote the number of tables belonging to the topic k in the corpus. Let $m_{..}$ denote the total number of tables in the corpus. $f_{\hat{k}}^{\neg w_i^d}(w_i^d)$ is the prior density of w_i^d. When a sign $\neg$ in the 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 is removed from the set or from the calculation of the count. Let $f_k^{\neg w_i^d}(.)$ denote the conditional likelihood density for some previously used table, which can be derived based on the type of the problem we are solving. In [15], the authors only presented HDP in general and not for topic modeling in particular. In case of topic modeling, we can follow a widely used Dirichlet-Multinomial paradigm, where the base measure is a Dirichlet, and the density F (same F as used in [15]) as Multinomial. We also introduce a notion of reserved tables using r. We use v to denote an unreserved table. We use the symbol $\hat{t}$ or $\hat{k}$ to denote a new table and dish, respectively. Also, note that buddies will be in their own buddy circles (commonly known as friendship circle) waiting outside the restaurant in queue, so different buddy groups take their own reserved tables. The likelihood of w_i^d who is a loner for $t_i^d = \hat{t}$, where $\hat{t}$ is the new table being sampled, is written as:

$$P(w_i^d = \texttt{Loner} | t_i^d = \hat{t} = v, \mathbf{t}^{\neg di}, \mathbf{k}, b_i^d = 0, w_{i-1}^d, t_{i-1}^d) = \sum_{k=1}^{K} \frac{m_{.k}}{m_{..} + \eta} f_k^{\neg w_i^d}(w_i^d) + \frac{\eta}{m_{..} + \eta} f_{\hat{k}}^{\neg w_i^d}(w_i^d) \qquad (1)$$

The above equation lays a restriction on the "loner" not to occupy the reserved table. This is because $b_i^d = 0$ associated with the loner will disallow this loner to occupy any of the reserved tables. But the loner can request a new table of the same topic (by ordering the same dish k as those of the reserved tables) as that of the reserved table or a different dish $\hat{k}$, with probability value proportional to α. The loner can also share an unreserved table with other loners with a value proportional to n_{tk}^d. The mechanism for buddies choosing a table is different. b_i^d indicates whether a customer is a buddy with the previous customer. The first buddy, w_i^d, who enters the restaurant carries with herself $b_i^d = 0$ because this customer is not a buddy with the previous customer who has just entered the restaurant. This customer is certainly not a loner, but will follow Equation 1 due to the buddy assignment variable. Therefore, this customer can either share an unreserved table with other loners, or requests a new table and sits alone. But when the second customer w_{i+1}^d in that buddy group enters the restaurant, this customer knows that the previous customer is her buddy. So this customer requests new table serving the same dish if the previous customer sat at an unreserved shared table, or shares the table with the previous buddy in case that buddy had requested a new table for herself and happens to be the first customer to sit there. The table is then set to reserved. The changes made by w_i^d using Equation 1 (if used) have to be reset to the previous state. This is where we make slight changes at the franchise level where we decrement the count from the existing unreserved table where w_i^d sat. The previous buddy then joins the buddy in that table. The scheme at the restaurant level can be expressed as:

$$P(w_i^d = \texttt{First}|t_i^d, \mathbf{t}^{\neg di}, \mathbf{k}, b_i^d = 1, w_{i-1}^d, t_{i-1}^d) = \begin{cases} \frac{\eta}{m_{..}+\eta} f_{\hat{k}}^{\neg w_i^d}(w_i^d) \text{ \& } k_{i-1}^d = k_i^d \text{ if } t_i^d = \hat{t}, b_{i-1}^d = 0, b_i^d = 1 \\ \sum_{k=1}^{K} \frac{m_{.k}}{m_{..}+\eta} f_k^{\neg w_i^d}(w_i^d) \text{ \& } t_i^d = t_{i-1}^d, \hat{t} = r \text{ if } b_{i-1}^d = 0, t_{i-1}^d = \hat{t}, b_i^d = 1 \end{cases} \tag{2}$$

Others, in the buddy group sit in the same table one by one requested by the "First Buddy" (denoted by `First` in Equation 2) i.e. ($t_i^d = t_{i-1}^d$) and share the same dish k.

$$P(w_i^d = \texttt{Other}|t_i^d = r, \mathbf{t}^{\neg di}, \mathbf{k}, b_i^d = 1, w_{i-1}^d, t_{i-1}^d) = \sum_{k=1}^{K} \frac{m_{.k}}{m_{..}+\eta} f_k^{\neg w_i^d}(w_i^d) \text{ \& } t_i^d = t_{i-1}^d, k_i^d = k_{i-1}^d \tag{3}$$

We present the buddy assignment scheme below which is based on global statistics. The idea is to compute the probabilities of how often two customers (words) consecutively come in sequence. Then based on the probability value, the buddy indicator variable is set to either 0 or 1. Let $p_{t_{i-1}^d w_{i-1}^d b_i^d}$ be the number of times the buddy indicator variable b_i^d has been set to 0 or 1 given the previous word and the table of the previous word. $n_{w_i^d}^{w_{i-1}^d}$ is the number of times the word w_i^d comes after the word w_{i-1}^d in the entire corpus. Let V be the total number of words in the vocabulary. n_{kw} is the number of times a word has appeared in topic k.

$$P(b_i^d = 0|\mathbf{b}^{\neg di}, \mathbf{w}, \mathbf{t}) = \frac{p_{t_{i-1}^d w_{i-1}^d 0} + \omega_0}{\sum_{c=0}^{1} p_{t_{i-1}^d w_{i-1}^d c} + \omega_0 + \omega_1} \times \frac{(\beta\tau_{w_i^d} + n_{k w_i^d} - 1)}{\sum_{v=1}^{V}(\beta\tau_v + n_{kv}) - 1} \tag{4}$$

$$P(b_i^d = 1|\mathbf{b}^{\neg di}, \mathbf{w}, \mathbf{t}) = \frac{p_{t_{i-1}^d w_{i-1}^d 1} + \omega_1}{\sum_{c=0}^{1} p_{t_{i-1}^d w_{i-1}^d c} + \omega_0 + \omega_1} \times \frac{n_{w_i^d}^{w_{i-1}^d} + \delta_{w_i^d} - 1}{\sum_{v=1}^{V}(n_v^{w_{i-1}^d} + \delta_v) - 1} \text{ and } t_i^d = t_{i-1}^d \tag{5}$$

Using the above equations at the restaurant level and the franchise level of the CRF, one can compute the posterior estimates to get the topic distributions for a corpus.

4 Experiments and Results

In our experiments, we evaluate different aspects of our model in terms of its generalization ability on unseen data and the words generated in the topics. In all experiments, the Gibbs sampler was run for 1000 iterations. We found that this number of iterations is sufficient because the joint likelihood of the sampled hidden variables and the words indicated convergence in the Markov chain. The topic models were run for five times, and the average of those five runs was taken.

4.1 Document Modeling

Document modeling using perplexity has been widely used in topic modeling. We use the same formula for perplexity as used in [15]. We use both small and large scale datasets for this experiment. The datasets that we use are: 1) AQUAINT-1 that comes with TREC HARD track (1,033,461 documents), 2) NIPS dataset (1,830 documents) commonly used for topic models 3) OHSUMED, a popular dataset used in the information retrieval community (233,448 documents), 4) Reuters collection (806,791 documents). We used the same text pre-processing strategy as used in [2], which also maintains order of words. We create five folds for each of these datasets and conduct five-fold cross validation. Each fold is created by randomly sampling 75% of the entire documents into the training set, and the rest into the test set.

The comparative methods that we use in experiments consist of both parametric and nonparametric topic models. The parametric topic models are: `LDA` [1], `BTM` [8], `LDACOL` [19], `TNG` [10], and a recently proposed method `NTSeg` [2]. The nonparametric topic models are `HDP` [15], and a recently proposed model `NHDP` [6]. We use the best experimental settings including hyperparameter sampling for these models as described in their respective works. `HDP` and `NHDP` both use CRF to sample from the posterior.

We use a tuning method to determine the number of latent topics in the parametric models. In the tuning process, in each fold, we first divide the training set into the development set which is 75% of the total number of documents in the training set, and the rest goes into the tuning set. We train the model using the development set and vary the number of topics. Then we compute the perplexity for each number of topics using the tuning set in each fold. Note that we also run the Gibbs sampler with 1000 iterations in each fold. Then we choose the best performing model through this procedure i.e. the model with the lowest average perplexity. We repeat five times and take the average. The number of topics with the lowest average perplexity is chosen as the output of the tuning process. We then merge the development and the tuning sets together to get the

Table 1. Document modeling results

Model	Perplexity			
	AQUAINT-1	NIPS	OHSUMED	Reuters
LDA	4599.48	834.45	2305.32	3490.12
BTM	4578.57	833.75	2229.96	3411.98
LDACOL	4501.44	831.45	2398.22	3298.76
TNG	4423.76	828.32	2315.72	3108.43
NTSeg	4400.76	811.32	2295.72	3112.43
HDP	4322.32	825.43	2240.23	3192.54
NHDP	4495.32	820.56	2299.45	3102.53
Our	4107.75	766.90	2192.44	3089.44

training set where we train the model using the number of topics obtained from the tuning process. We test the model using the same number of topics on the test set in each fold, by running five times and compute the average.

Table 1 depicts the result of document modeling. In all the four datasets, we see that our model, labeled as "Our" is the best performing one. The improvements are statistically significant based on two-tailed test with $p < 0.05$ against each of the comparative methods. The reason why our model performs better in generalizability is mainly due to its ability to determine the number of topics based on the data characteristic. In addition, considering word order is another advantage. Our model also performs better than the n-gram parametric models. For parametric models, despite using the tuning step, the data fitting might be an issue in the test set. Unigram models cannot capture word order information.

4.2 Qualitative Results

We present some high probability words in decreasing order obtained from the nonparametric topic models in Table 2. Following the result illustration technique from [10], we present unigrams and n-grams separately as we are comparing with the HDP model. We show the results obtained from AQUAINT-1 (presented left) and Reuters (presented right). The topics shown in the tables have been selected randomly from these two collections. Although qualitative comparison in topic models is not a strong predictor for measuring the robustness of a model, we can see from the results that our model has discovered better topical words than the comparative models. Bigrams such as "january february" do not convey much meaning in a topic in the NHDP model in Reuters. Similarly, in the same collection, the word "report" discovered by the HDP model is not very insightful. In AQUAINT-1, n-gram such as "talk real person" by the NHDP model is also not very insightful, and same goes for word "new" discovered by the HDP model.

Table 2. High probability words in descending order obtained from a topic in two different collections. The table on the left shows results from AQUAINT-1 collection, and on the right, we depict results from Reuters collections.

HDP	NDHP		Our	
	Unigrams	N-grams	Unigrams	N-grams
year	test	internet sale	phone	web site
game	computer	search engine	digit	cell phone
music	year	create search engine	computers	high technology
computer	project	internet user	technology	microsoft windows
train	modern	index html	information	computer technology
new	service	state department	web	computer device
team	software	computer software	mail	laptop equipment
church	internet	computer bulletin	user	recognition software
transit	editor	latin america	online	large comfortable keyboard
time	technology	talk real person	network	speech technology

HDP	NDHP		Our	
	Unigrams	N-grams	Unigrams	N-grams
report	year	oil product	oil	oil price
bank	japan	crude oil	trade	gulf war
win	iraq	new oil product	cargo	oil stock
pakistan	oil	january february	high	crude oil
oil	crude	saudi arabia	market	domestic crude
rate	demand	total product	price	iraq ambassador
net	gasoline	crude export	fuel	oil product
french	saudi	gasoline distillation	tonne	indian oil
launch	arabia	thousand barrel	crude	run oil company
qatar	uae	oil import	week	world price

5 Conclusions

We have proposed a new metaphor in Bayesian nonparametrics called the Chinese Restaurant Franchise with Buddy Customers that takes into account the order of words in documents. Our model is able to discover n-gram words in latent topics. We have introduced a notion of buddy assignments in the basic CRF metaphor where we find out whether customers standing in order are friends with each other. All buddies occupy their reserved table in the restaurant which is not shared by other customers who do not belong to their friendship circle. We have tested our model on some text collections, and have shown that improvements are achieved in both quantitative performance and quality of topical words.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR 3, 993–1022 (2003)
2. Jameel, S., Lam, W.: An unsupervised topic segmentation model incorporating word order. In: SIGIR, pp. 472–479 (2013)
3. Lindsey, R.V., Headden III, W.P., Stipicevic, M.J.: A phrase-discovering topic model using hierarchical Pitman-Yor processes. In: EMNLP, pp. 214–222 (2012)
4. Kim, H.D., Park, D.H., Lu, Y., Zhai, C.: Enriching text representation with frequent pattern mining for probabilistic topic modeling. ASIST 49, 1–10 (2012)
5. Barbieri, N., Manco, G., Ritacco, E., Carnuccio, M., Bevacqua, A.: Probabilistic topic models for sequence data. Machine Learning 93, 5–29 (2013)
6. Jameel, S., Lam, W.: A nonparametric N-gram topic model with interpretable latent topics. In: Banchs, R.E., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., Lang, J. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 74–85. Springer, Heidelberg (2013)
7. Kawamae, N.: Supervised N-gram topic model. In: WSDM, pp. 473–482 (2014)
8. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: ICML, pp. 977–984 (2006)
9. McCallum, A., Mimno, D.M., Wallach, H.M.: Rethinking LDA: Why priors matter. In: NIPS, pp. 1973–1981 (2009)
10. Wang, X., McCallum, A., Wei, X.: Topical N-grams: Phrase and topic discovery, with an application to Information Retrieval. In: ICDM, pp. 697–702 (2007)
11. Fei, G., Chen, Z., Liu, B.: Review topic discovery with phrases using the Pólya urn model. In: COLING, pp. 667–676 (2014)

12. Darling, W.: Generalized Probabilistic Topic and Syntax Models for Natural Language Processing. PhD thesis (2012)
13. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. JACM 57, 7 (2010)
14. Teh, Y.W.: Dirichlet process. In: Encyclopedia of Machine Learning, pp. 280–287 (2010)
15. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. JASA 101, 1566–1581 (2006)
16. Teh, Y.W., Kurihara, K., Welling, M.: Collapsed variational inference for HDP. In: NIPS, pp. 1481–1488 (2007)
17. Sudderth, E.B.: Graphical models for visual object recognition and tracking. PhD thesis, Massachusetts Institute of Technology (2006)
18. Foti, N., Williamson, S.: A survey of non-exchangeable priors for Bayesian nonparametric models (2013)
19. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological Review 114, 211 (2007)
20. Bartlett, N., Pfau, D., Wood, F.: Forgetting counts: Constant memory inference for a dependent hierarchical Pitman-Yor process. In: ICML, pp. 63–70 (2010)
21. Lau, J.H., Baldwin, T., Newman, D.: On collocations and topic models. TSLP 10, 10:1–10:14 (2013)
22. Johri, N., Roth, D., Tu, Y.: Experts' retrieval with multiword-enhanced author topic model. In: NAACL. SS 2010, pp. 10–18 (2010)
23. Noji, H., Mochihashi, D., Miyao, Y.: Improvements to the Bayesian topic n-gram models. In: EMNLP, pp. 1180–1190 (2013)
24. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. Bayesian Nonparametrics, 158–207 (2010)
25. Goldwater, S., Griffiths, T., Johnson, M.: A Bayesian framework for word segmentation: Exploring the effects of context. Cognition 112, 21–54 (2009)
26. Johnson, M.: PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In: ACL, pp. 1148–1157 (2010)
27. Deane, P.: A nonparametric method for extraction of candidate phrasal terms. In: ACL, pp. 605–613 (2005)
28. Petrovic, S., Snajder, J., Basic, B.D.: Extending lexical association measures for collocation extraction. Computer Speech and Language 24, 383–394 (2010)
29. Yoshii, K., Goto, M.: A vocabulary-free infinity-gram model for nonparametric Bayesian chord progression analysis. In: ICMIR, pp. 645–650 (2011)
30. Blei, D.M., Frazier, P.I.: Distance dependent Chinese restaurant processes. JMLR 12, 2461–2488 (2011)
31. Kim, D., Oh, A.: Accounting for data dependencies within a hierarchical Dirichlet process mixture model. In: CIKM, pp. 873–878 (2011)
32. Fox, E., Sudderth, E., Jordan, M., Willsky, A.: A sticky HDP-HMM with application to speaker diarization. APS 5, 1020–1056 (2011)
33. Tayal, A., Poupart, P., Li, Y.: Hierarchical double Dirichlet process mixture of Gaussian processes. In: AAAI (2012)