Topic Modeling: Beyond Bag-of-Words

Hanna M. Wallach HMW26@CAM.AC.UK
Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK

Abstract

Some models of textual corpora employ text generation methods involving n-gram statistics, while others use latent topic variables inferred using the "bag-of-words" assumption, in which word order is ignored. Previously, these methods have not been combined. In this work, I explore a hierarchical generative probabilistic model that incorporates both n-gram statistics and latent topic variables by extending a unigram topic model to include properties of a hierarchical Dirichlet bigram language model. The model hyperparameters are inferred using a Gibbs EM algorithm. On two data sets, each of 150 documents, the new model exhibits better predictive accuracy than either a hierarchical Dirichlet bigram language model or a unigram topic model. Additionally, the inferred topics are less dominated by function words than are topics discovered using unigram statistics, potentially making them more meaningful.

1. Introduction

Recently, much attention has been given to generative probabilistic models of textual corpora, designed to identify representations of the data that reduce description length and reveal inter- or intra-document statistical structure. Such models typically fall into one of two categories—those that generate each word on the basis of some number of preceding words or word classes and those that generate words based on latent topic variables inferred from word correlations independent of the order in which the words appear.

n-gram language models make predictions using observed marginal and conditional word frequencies. While such models may use conditioning contexts of arbitrary length, this paper deals only with bigram models—*i.e.*, models that predict each word based on the immediately preceding word.

To develop a bigram language model, marginal and conditional word counts are determined from a corpus $\boldsymbol{w}$. The marginal count N_i is defined as the number of times that word i has occurred in the corpus, while the conditional count $N_{i|j}$ is the number of times word i immediately follows word j. Given these counts, the aim of bigram language modeling is to develop predictions of word w_t given word w_{t-1}, in any document. Typically this is done by computing estimators of both the marginal probability of word i and the conditional probability of word i following word j, such as $f_i = N_i/N$ and $f_{i|j} = N_{i|j}/N_j$, where N is the number of words in the corpus. If there were sufficient data available, the observed conditional frequency $f_{i|j}$ could be used as an estimator for the predictive probability of i given j. In practice, this does not provide a good estimate: only a small fraction of possible i, j word pairs will have been observed in the corpus. Consequently, the conditional frequency estimator has too large a variance to be used by itself.

To alleviate this problem, the bigram estimator $f_{i|j}$ is smoothed by the marginal frequency estimator f_i to give the predictive probability of word i given word j:

$$P(w_t = i|w_{t-1} = j) = \lambda f_i + (1 - \lambda) f_{i|j}. \qquad (1)$$

The parameter λ may be fixed, or determined from the data using techniques such as cross-validation (Jelinek & Mercer, 1980). This procedure works well in practice, despite its somewhat *ad hoc* nature.

The hierarchical Dirichlet language model (MacKay & Peto, 1995) is a bigram model that is entirely driven by principles of Bayesian inference. This model has a similar predictive distribution to models based on equation (1), with one key difference: the bigram statistics $f_{i|j}$ in MacKay and Peto's model are not smoothed with marginal statistics f_i, but are smoothed with a quantity related to the number of different contexts in which each word has occurred.

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

Latent Dirichlet allocation (Blei et al., 2003) provides an alternative approach to modeling textual corpora. Documents are modeled as finite mixtures over an underlying set of latent topics inferred from correlations between words, independent of word order.

This bag-of-words assumption makes sense from a point of view of computational efficiency, but is unrealistic. In many language modeling applications, such as text compression, speech recognition, and predictive text entry, word order is extremely important. Furthermore, it is likely that word order can assist in topic inference. The phrases "the department chair couches offers" and "the chair department offers couches" have the same unigram statistics, but are about quite different topics. When deciding which topic generated the word "chair" in the first sentence, knowing that it was immediately preceded by the word "department" makes it much more likely to have been generated by a topic that assigns high probability to words related to university administration.

In practice, the topics inferred using latent Dirichlet allocation are heavily dominated by function words, such as "in", "that", "of" and "for", unless these words are removed from corpora prior to topic inference. While removing these may be appropriate for tasks where word order does not play a significant role, such as information retrieval, it is not appropriate for many language modeling applications, where both function and content words must be accurately predicted.

In this paper, I present a hierarchical Bayesian model that integrates bigram-based and topic-based approaches to document modeling. This model moves beyond the bag-of-words assumption found in latent Dirichlet allocation by introducing properties of MacKay and Peto's hierarchical Dirichlet language model. In addition to exhibiting better predictive performance than either MacKay and Peto's language model or latent Dirichlet allocation, the topics inferred using the new model are typically less dominated by function words than are topics inferred from the same corpora using latent Dirichlet allocation.

2. Background

I begin with brief descriptions of MacKay and Peto's hierarchical Dirichlet language model and Blei et al.'s latent Dirichlet allocation.

2.1. Hierarchical Dirichlet Language Model

Bigram language models are specified by a conditional distribution $P(w_t = i|w_{t-1} = j)$, described by $W(W-1)$ free parameters, where W is the number of words in the vocabulary. These parameters are denoted by the matrix Φ, with $P(w_t = i|w_{t-1} = j) \equiv \phi_{i|j}$. Φ may be thought of as a transition probability matrix, in which the j^{th} row, the probability vector for transitions from word j, is denoted by the vector $\boldsymbol{\phi}_j$.

Given a corpus $\boldsymbol{w}$, the likelihood function is

$$P(\boldsymbol{w}|\Phi) = \prod_i \prod_j \phi_{i|j}^{N_{i|j}}, \tag{2}$$

where $N_{i|j}$ is the number of times that word i immediately follows word j in the corpus.

MacKay and Peto (1995) extend this basic framework by placing a Dirichlet prior over Φ:

$$P(\Phi|\beta\boldsymbol{m}) = \prod_j \text{Dirichlet}(\boldsymbol{\phi}_j|\beta\boldsymbol{m}), \tag{3}$$

where $\beta > 0$ and $\boldsymbol{m}$ is a measure satisfying $\sum_i m_i = 1$.

Combining equations (2) and (3), and integrating over Φ, yields the probability of the corpus given the hyperparameters $\beta\boldsymbol{m}$, also known as the "evidence":

$$P(\boldsymbol{w}|\beta\boldsymbol{m}) = \prod_j \frac{\prod_i \Gamma(N_{i|j} + \beta m_i)}{\Gamma(N_j + \beta)} \frac{\Gamma(\beta)}{\prod_i \Gamma(\beta m_i)}. \tag{4}$$

It is also easy to obtain a predictive distribution for each context j given the hyperparameters $\beta\boldsymbol{m}$:

$$P(i|j, \boldsymbol{w}, \beta\boldsymbol{m}) = \frac{N_{i|j} + \beta m_i}{N_j + \beta}. \tag{5}$$

To make the relationship to equation (1) explicit, $P(i|j, \boldsymbol{w}, \beta\boldsymbol{m})$ may be rewritten as

$$P(i|j, \boldsymbol{w}, \beta\boldsymbol{m}) = \lambda_j m_i + (1 - \lambda_j) f_{i|j}, \tag{6}$$

where $f_{i|j} = N_{i|j}/N_j$ and

$$\lambda_j = \frac{\beta}{N_j + \beta}. \tag{7}$$

The hyperparameter m_i is now taking the role of the marginal statistic f_i in equation (1).

Ideally, the measure $\beta\boldsymbol{m}$ should be given a proper prior and marginalized over when making predictions, yielding the true predictive distribution:

$$P(i|j, \boldsymbol{w}) = \int P(\beta\boldsymbol{m}|\boldsymbol{w}) P(i|j, \boldsymbol{w}, \beta\boldsymbol{m})\ \boldsymbol{d}(\beta\boldsymbol{m}). \tag{8}$$

However, if $P(\beta\boldsymbol{m}|\boldsymbol{w})$ is sharply peaked in $\beta\boldsymbol{m}$ so that that it is effectively a delta function, then the true predictive distribution may be approximated by $P(i|j, \boldsymbol{w}, [\beta\boldsymbol{m}]^{\text{MP}})$, where $[\beta\boldsymbol{m}]^{\text{MP}}$ is the maximum of

$P(\beta\boldsymbol{m}|\boldsymbol{w})$. Additionally, the prior over $\beta\boldsymbol{m}$ may be assumed to be uninformative, yielding a minimal data-driven Bayesian model in which the optimal $\beta\boldsymbol{m}$ may be determined from the data by maximizing the evidence. MacKay and Peto show that each element of the optimal $\boldsymbol{m}$, when estimated using this "empirical Bayes" procedure, is related to the number of contexts in which the corresponding word has appeared.

2.2. Latent Dirichlet Allocation

Latent Dirichlet allocation (Blei et al., 2003) represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words. Each word w_t in a corpus $\boldsymbol{w}$ is assumed to have been generated by a latent topic z_t, drawn from a document-specific distribution over T topics.

Word generation is defined by a conditional distribution $P(w_t = i|z_t = k)$, described by $T(W-1)$ free parameters, where T is the number of topics and W is the size of the vocabulary. These parameters are denoted by Φ, with $P(w_t = i|z_t = k) \equiv \phi_{i|k}$. Φ may be thought of as an emission probability matrix, in which the k^{th} row, the distribution over words for topic k, is denoted by $\boldsymbol{\phi}_k$. Similarly, topic generation is characterized by a conditional distribution $P(z_t = k|d_t = d)$, described by $D(T-1)$ free parameters, where D is the number of documents in the corpus. These parameters form a matrix Θ, with $P(z_t = k|d_t = d) \equiv \theta_{k|d}$. The d^{th} row of this matrix is the distribution over topics for document d, denoted by $\boldsymbol{\theta}_d$.

The joint probability of a corpus $\boldsymbol{w}$ and a set of corresponding latent topics $\boldsymbol{z}$ is

$$P(\boldsymbol{w}, \boldsymbol{z}|\Phi, \Theta) = \prod_i \prod_k \prod_d \phi_{i|k}^{N_{i|k}} \theta_{k|d}^{N_{k|d}}, \tag{9}$$

where $N_{i|k}$ is the number of times that word i has been generated by topic k, and $N_{k|d}$ is the number of times topic k has been used in document d.

Blei et al. place a Dirichlet prior over Φ,

$$P(\Phi|\beta\boldsymbol{m}) = \prod_k \text{Dirichlet}(\boldsymbol{\phi}_k|\beta\boldsymbol{m}), \tag{10}$$

and another over Θ,

$$P(\Theta|\alpha\boldsymbol{n}) = \prod_d \text{Dirichlet}(\boldsymbol{\theta}_d|\alpha\boldsymbol{n}). \tag{11}$$

Combining these priors with equation (9) and integrating over Φ and Θ gives the evidence for hyperparameters $\alpha\boldsymbol{n}$ and $\beta\boldsymbol{m}$, which is also the probability of the corpus given the hyperparameters:

$$P(\boldsymbol{w}|\alpha\boldsymbol{n}, \beta\boldsymbol{m}) = \sum_{\boldsymbol{z}} \left(\prod_k \frac{\prod_i \Gamma(N_{i|k} + \beta m_i)}{\Gamma(N_k + \beta)} \frac{\Gamma(\beta)}{\prod_i \Gamma(\beta m_i)} \prod_d \frac{\prod_k \Gamma(N_{k|d} + \alpha n_k)}{\Gamma(N_d + \alpha)} \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha n_k)} \right). \tag{12}$$

N_k is the total number of times topic k occurs in $\boldsymbol{z}$, while N_d is the number of words in document d.

The sum over $\boldsymbol{z}$ cannot be computed directly because it does not factorize and involves T^N terms, where N is the total number of words in the corpus. However, it may be approximated using Markov chain Monte Carlo (Griffiths & Steyvers, 2004).

Given a corpus $\boldsymbol{w}$, a set of latent topics $\boldsymbol{z}$, and optimal hyperparameters $[\alpha\boldsymbol{n}]^{\text{MP}}$ and $[\beta\boldsymbol{m}]^{\text{MP}}$, approximate predictive distributions for each topic k and document d are given by the following pair of equations:

$$P(i|k, \boldsymbol{w}, \boldsymbol{z}, [\beta\boldsymbol{m}]^{\text{MP}}) = \frac{N_{i|k} + [\beta m_i]^{\text{MP}}}{N_k + \beta^{\text{MP}}} \tag{13}$$

$$P(k|d, \boldsymbol{w}, \boldsymbol{z}, [\alpha\boldsymbol{n}]^{\text{MP}}) = \frac{N_{k|d} + [\alpha n_k]^{\text{MP}}}{N_d + \alpha^{\text{MP}}}. \tag{14}$$

These may be rewritten as

$$P(i|k, \boldsymbol{z}, \boldsymbol{w}, \beta\boldsymbol{m}) = \lambda_k f_{i|k} + (1 - \lambda_k) m_i \tag{15}$$

$$P(k|d, \boldsymbol{z}, \boldsymbol{w}, \alpha\boldsymbol{n}) = \gamma_d f_{k|d} + (1 - \gamma_d) n_d, \tag{16}$$

where $f_{i|k} = N_{i|k}/N_k$, $f_{k|d} = N_{k|d}/N_d$ and

$$\lambda_k = \frac{\beta}{N_k + \beta} \tag{17}$$

$$\gamma_d = \frac{\alpha}{N_d + \alpha}. \tag{18}$$

$f_{i|k}$ is therefore being smoothed by the hyperparameter m_i, while $f_{k|d}$ is smoothed by n_k. Note the similarity of equations (15) and (16) to equation (6).

3. Bigram Topic Model

This section introduces a model that extends latent Dirichlet allocation by incorporating a notion of word order, similar to that employed by MacKay and Peto's hierarchical Dirichlet language model. Each topic is now represented by a set of W distributions.

Word generation is defined by a conditional distribution $P(w_t = i|w_{t-1} = j, z_t = k)$, described by $WT(W-1)$ free parameters. As before, these parameters form a matrix Φ, this time with WT rows.

Each row is a distribution over words for a particular context j, k, denoted by $\boldsymbol{\phi}_{j,k}$. Each topic k is now characterized by the W distributions specific to that topic. Topic generation is the same as in latent Dirichlet allocation: topics are drawn from the conditional distribution $P(z_t = k|d_t = d)$, described by $D(T-1)$ free parameters, which form a matrix Θ.

The joint probability of a corpus $\boldsymbol{w}$ and a single set of latent topic assignments $\boldsymbol{z}$ is

$$P(\boldsymbol{w}, \boldsymbol{z}|\Phi, \Theta) = \prod_i \prod_j \prod_k \prod_d \phi_{i|j,k}^{N_{i|j,k}} \theta_{k|d}^{N_{k|d}}, \quad (19)$$

where $N_{i|j,k}$ is the number of times word i has been generated by topic k when preceded by word j. As in latent Dirichlet allocation, $N_{k|d}$ is the number of times topic k has been used in document d.

The prior over Θ is chosen to be the same as that used in latent Dirichlet allocation:

$$P(\Theta|\alpha\boldsymbol{n}) = \prod_d \text{Dirichlet}(\boldsymbol{\theta}_d|\alpha\boldsymbol{n}). \quad (20)$$

However, the additional conditioning context j in the distribution that defines word generation affords greater flexibility in choosing a hierarchical prior for Φ than in either latent Dirichlet allocation or the hierarchical Dirichlet language model. The priors over Φ used in both MacKay and Peto's language model and Blei et al.'s latent Dirichlet allocation are "coupled" priors: learning the probability vector for a single context, $\boldsymbol{\phi}_j$ the case of MacKay and Peto's model and $\boldsymbol{\phi}_k$ in Blei et al.'s, gives information about the probability vectors in other contexts, j' and k' respectively. This dependence comes from the hyperparameter vector $\beta\boldsymbol{m}$, shared, in the case of the hierarchical Dirichlet language model, between all possible previous word contexts j and, in the case of latent Dirichlet allocation, between all possible topics k. Since word generation is conditioned upon both j and k in the new model presented in this paper, there is more than one way in which hyperparameters for the prior over Φ might be shared in this model.

Prior 1: Most simply, a single hyperparameter vector $\beta\boldsymbol{m}$ may be shared between all j, k contexts:

$$P(\Phi|\beta\boldsymbol{m}) = \prod_j \prod_k \text{Dirichlet}(\boldsymbol{\phi}_{j,k}|\beta\boldsymbol{m}). \quad (21)$$

Here, knowledge about the probability vector for one $\boldsymbol{\phi}_{j,k}$ will give information about the probability vectors $\boldsymbol{\phi}_{j',k'}$ for all other j', k' contexts.

Prior 2: Alternatively there may be T hyperparameter vectors—one for each topic k:

$$P(\Phi|\{\beta_k\boldsymbol{m}_k\}) = \prod_j \prod_k \text{Dirichlet}(\boldsymbol{\phi}_{j,k}|\beta_k\boldsymbol{m}_k). \quad (22)$$

Information is now shared between only those probability vectors with topic context k. Intuitively, this is appealing. Learning about the distribution over words for a single context j, k yields information about the distributions over words for other contexts j', k that share this topic, but not about distributions with other topic contexts. In other words, this prior encapsulates the notion of similarity between distributions over words for a given topic context.

Having defined the distributions that characterize word and topic generation in the new model and assigned priors over the parameters, the generative process for a corpus $\boldsymbol{w}$ is:

1. For each topic k and word j:
   (a) Draw $\boldsymbol{\phi}_{j,k}$ from the prior over Φ: either Dirichlet($\boldsymbol{\phi}_{j,k}|\beta\boldsymbol{m}$) (prior 1) or Dirichlet($\boldsymbol{\phi}_{j,k}|\beta_k\boldsymbol{m}_k$) (prior 2).
2. For each document d in the corpus:
   (a) Draw the topic mixture $\boldsymbol{\theta}_d$ for document d from Dirichlet($\boldsymbol{\theta}_d|\alpha\boldsymbol{n}$).
   (b) For each position t in document d:
      i. Draw a topic $z_t \sim \text{Discrete}(\boldsymbol{\theta}_d)$.
      ii. Draw a word w_t from the distribution over words for the context defined by the topic z_t and previous word w_{t-1}, Discrete($\boldsymbol{\phi}_{w_{t-1},z_t}$).

The evidence, or probability of a corpus $\boldsymbol{w}$ given the hyperparameters, is either (prior 1)

$$P(\boldsymbol{w}|\alpha\boldsymbol{n}, \beta\boldsymbol{m}) = \sum_{\boldsymbol{z}} \left(\prod_j \prod_k \frac{\prod_i \Gamma(N_{i|j,k} + \beta m_i)}{\Gamma(N_{j,k} + \beta)} \frac{\Gamma(\beta)}{\prod_i \Gamma(\beta m_i)} \prod_d \frac{\prod_k \Gamma(N_{k|d} + \alpha n_k)}{\Gamma(N_d + \alpha)} \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha n_k)} \right) \quad (23)$$

or (prior 2)

$$P(\boldsymbol{w}|\alpha\boldsymbol{n}, \{\beta_k\boldsymbol{m}_k\}) = \sum_{\boldsymbol{z}} \left(\prod_j \prod_k \frac{\prod_i \Gamma(N_{i|j,k} + \beta_k m_{i|k})}{\Gamma(N_{j,k} + \beta_k)} \frac{\Gamma(\beta)}{\prod_i \Gamma(\beta_k m_{i|k})} \prod_d \frac{\prod_k \Gamma(N_{k|d} + \alpha n_k)}{\Gamma(N_d + \alpha)} \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha n_k)} \right). \quad (24)$$

As in latent Dirichlet allocation, the sum over $\boldsymbol{z}$ is intractable, but may be approximated using MCMC.

For a single set of latent topics $\boldsymbol{z}$, and optimal hyperparameters $[\beta\boldsymbol{m}]^{\text{MP}}$ or $\{[\beta_k\boldsymbol{m}_k]^{\text{MP}}\}$, the approximate predictive distribution over words given previous word j and current topic k is either (prior 1)

$$P(i|j,k,\boldsymbol{w},\boldsymbol{z},[\beta\boldsymbol{m}]^{\text{MP}}) = \frac{N_{i|j,k} + [\beta m_i]^{\text{MP}}}{N_{j,k} + \beta^{\text{MP}}} \tag{25}$$

or (prior 2)

$$P(i|j,k,\boldsymbol{w},\boldsymbol{z},\{[\beta_k\boldsymbol{m}_k]^{\text{MP}}\}) = \frac{N_{i|j,k} + [\beta_k m_{i|k}]^{\text{MP}}}{N_{j,k} + \beta_k^{\text{MP}}}. \tag{26}$$

In equation (25), the statistic $N_{i|j,k}/N_{j,k}$ is always smoothed by the quantity m_i, regardless of the conditioning context, j,k. Meanwhile, in equation (26), $N_{i|j,k}/N_{j,k}$ is smoothed by $m_{i|k}$, which is will vary depending on the conditioning topic k.

Given $[\alpha\boldsymbol{n}]^{\text{MP}}$, the approximate predictive distribution over topics for document d is

$$P(k|d,\boldsymbol{w},\boldsymbol{z},[\alpha\boldsymbol{n}]^{\text{MP}}) = \frac{N_{k|d} + [\alpha n_k]^{\text{MP}}}{N_d + \alpha^{\text{MP}}}. \tag{27}$$

4. Inference of Hyperparameters

Previous sampling-based treatments of latent Dirichlet allocation have not included any method for optimizing hyperparameters. However, the method described in this section may be applied to both latent Dirichlet allocation and the model presented in this paper.

Given an uninformative prior over $\alpha\boldsymbol{n}$ and $\beta\boldsymbol{m}$ or $\{\beta_k\boldsymbol{m}_k\}$, the optimal hyperparameters, $[\alpha\boldsymbol{n}]^{\text{MP}}$ and $[\beta\boldsymbol{m}]^{\text{MP}}$ or $\{[\beta_k\boldsymbol{m}_k]^{\text{MP}}\}$, may be found by maximizing the evidence, given in equation (23) or (24).

The evidence contains latent variables $\boldsymbol{z}$ and must therefore be maximized with respect to the hyperparameters using an expectation-maximization (EM) algorithm. Unfortunately, the expectation with respect to the distribution over the latent variables involves a sum over T^N terms, where N is the number of words in the entire corpus. However, this sum may be approximated using a Markov chain Monte Carlo algorithm, such as Gibbs sampling, resulting in a Gibbs EM algorithm (Andrieu et al., 2003). Given a corpus $\boldsymbol{w}$, and denoting the set of hyperparameters as $U = \{\alpha\boldsymbol{n}, \beta\boldsymbol{m}\}$ or $U = \{\alpha\boldsymbol{n}, \{\beta_k\boldsymbol{m}_k\}\}$, the optimal hyperparameters may be found by using the following steps:

1. Initialize $\boldsymbol{z}^{(0)}$ and $U^{(0)}$ and set $i = 1$.
2. Iteration i:
   (a) **E-step:** Draw S samples $\{\boldsymbol{z}^{(s)}\}_{s=1}^{S}$ from $P(\boldsymbol{z}|\boldsymbol{w}, U^{(i-1)})$ using a Gibbs sampler.
   (b) **M-step:** Maximize
   $$U^{(i)} = \arg\max_U \frac{1}{S}\sum_{s=1}^{S} \log P(\boldsymbol{w}, \boldsymbol{z}^{(s)}|U)$$
3. $i \leftarrow i + 1$ and go to 2.

4.1. E-Step

Gibbs sampling involves sequentially sampling each variable of interest, z_t here, from the distribution over that variable given the current values of all other variables and the data. Letting the subscript $-t$ denote a quantity that excludes data from the t^{th} position, the conditional posterior for z_t is either (prior 1)

$$P(z_t = k|\boldsymbol{z}_{-t},\boldsymbol{w},\alpha\boldsymbol{n},\beta\boldsymbol{m}) \propto \frac{\{N_{w_t|w_{t-1},k}\}_{-t} + \beta m_{w_t}}{\{N_k\}_{-t} + \beta} \frac{\{N_{k|d_t}\}_{-t} + \alpha n_k}{\{N_{d_t}\}_{-t} + \alpha} \tag{28}$$

or (prior 2)

$$P(z_t = k|\boldsymbol{z}_{-t},\boldsymbol{w},\alpha\boldsymbol{n},\{\beta_k\boldsymbol{m}_k\}) \propto \frac{\{N_{w_t|w_{t-1},k}\}_{-t} + \beta_k m_{w_t|k}}{\{N_{w_{t-1},k}\}_{-t} + \beta_k} \frac{\{N_{k|d_t}\}_{-t} + \alpha n_k}{\{N_{d_t}\}_{-t} + \alpha}. \tag{29}$$

Drawing a single set of topics $\boldsymbol{z}$ takes time proportional to the size of the corpus N and the number of topics T. The E-step therefore takes time proportional to N, T and the number of iterations for which the Markov chain is run in order to obtain the S samples.

Note that the samples used to approximate the E-step must come from a single Markov chain. The model is unaffected by permutations of topic indices. Consequently, there is no correspondence between topic indices across samples from different Markov chains: topics that have index k in two different Markov chains need not have similar distributions over words.

4.2. M-Step

Given $\{\boldsymbol{z}^{(s)}\}_{s=1}^{S}$, the optimal $\alpha\boldsymbol{n}$ can be computed using the fixed-point iteration

$$[\alpha n_k]^{\text{new}} = \alpha n_k \frac{\sum_s \sum_d \left(\Psi(N_{k|d}^{(s)} + \alpha n_k) - \Psi(\alpha n_k)\right)}{\sum_s \sum_d \left(\Psi(N_d + \alpha) - \Psi(\alpha)\right)}, \tag{30}$$

where $N_{k|d}^{(s)}$ is the number of times topic k has been used in document d in the s^{th} sample. Similar fixed-point iterations can be used to determine $[\beta m_i]^{\text{MP}}$ and $\{[\beta_k\boldsymbol{m}_k]^{\text{MP}}\}$ (Minka, 2003).

In my implementation, each fixed-point iteration takes time that is proportional to S and (at worst) N. For latent Dirichlet allocation and the new model with prior 1, the time taken to perform the M-step is therefore at worst proportional to S, N and the number of iterations taken to reach convergence. For the new model with prior 2, the time taken is also proportional to T.

5. Experiments

To evaluate the new model, both variants were compared with latent Dirichlet allocation and MacKay and Peto's hierarchical Dirichlet language model. The topic models were trained identically: the Gibbs EM algorithm described in the previous section was used for both the new model (with either prior) and latent Dirichlet allocation. The hyperparameters of the hierarchical Dirichlet language model were inferred using the same fixed-point iteration used in the M-step. The results presented in this section are therefore a direct reflection of differences between the models.

Language models are typically evaluated by computing the information rate of unseen test data, measured in bits per word: the better the predictive performance, the fewer the bits per word. Information rate is a direct measure of text compressibility. Given corpora $\boldsymbol{w}$ and $\boldsymbol{w}_{\text{test}}$, information rate is defined as

$$R = -\frac{\log_2 P(\boldsymbol{w}_{\text{test}}|\boldsymbol{w})}{N_{\text{test}}}, \tag{31}$$

where N_{test} is the number of words in the test corpus. The information rate may be computed directly for the hierarchical Dirichlet language model. For the topic models, computing $P(\boldsymbol{w}_{\text{test}}|\boldsymbol{w})$ requires summing over $\boldsymbol{z}$ and $\boldsymbol{z}^{\text{test}}$. As mentioned before, this is intractable. Instead, the information rate may be computed using a single set of topics $\boldsymbol{z}$ for the training data, in this case obtained by running a Gibbs sampler for 20000 iterations after the hyperparameters have been inferred. Given $\boldsymbol{z}$, multiple sets of topics for the test data $\{\boldsymbol{z}_{\text{test}}^{(s)}\}_{s=1}^{S}$ may be obtained using the predictive distributions. Given hyperparameters U, $P(\boldsymbol{w}_{\text{test}}|\boldsymbol{w})$ may be approximated by taking the harmonic mean of $\{P(\boldsymbol{w}_{\text{test}}|\boldsymbol{z}_{\text{test}}^{(s)}, \boldsymbol{w}, \boldsymbol{z}, U)\}_{s=1}^{S}$ (Kass & Raftery, 1995).

5.1. Corpora

The models were compared using two data sets. The first was constructed by drawing 150 abstracts (documents) at random from the Psychological Review Abstracts data provided by Griffiths and Steyvers (2005). A subset of 100 documents were used to infer the hyperparameters, while the remaining 50 were used for evaluating the models. The second data set consisted of 150 newsgroup postings, drawn at random from the 20 Newsgroups data (Rennie, 2005). Again, 100 documents were used for inference, while 50 were retained for evaluating predictive accuracy.

Punctuation characters, including hyphens and apostrophes, were treated as word separators, and each number was replaced with a special "number" token to reduce the size of the vocabulary. To enable evaluation using documents containing tokens not present in the training corpus, all words that occurred only once in the training corpus were replaced with an "unseen" token u. Preprocessing the Psychological Review Abstracts data in this manner resulted in a vocabulary of 1374 words, which occurred 13414 times in the training corpus and 6521 times in the documents used for testing. The 20 Newsgroups data ended up with a vocabulary of 2281 words, which occurred 27478 times in the training data and 13579 times in the test data. Despite consisting of the same number of documents, the 20 Newsgroups corpora are roughly twice the size of the Psychological Review Abstracts corpora.

5.2. Results

The experiments involving latent Dirichlet allocation and the new model were run with 1 to 120 topics, on an Opteron 254 (2.8GHz). These models all required at most 200 iterations of the Gibbs EM algorithm described in section 4. In the E-step, a Markov chain was run for 400 iterations. The first 200 iterations were discarded and 5 samples were taken from the remaining iterations. The mean time taken for each iteration is shown for both variants of the new model as a function of the number of topics in figure 2. As expected, the time taken is proportional to both the number of topics and the size of the corpus.

The information rates of the test data are shown in figure 1. On both corpora, latent Dirichlet allocation and the hierarchical Dirichlet language model achieve similar performance. With prior 1, the new model improves upon this by between 0.5 and 1 bits per word. However, with prior 2, it achieves an information rate reduction of between 1 and 2 bits per word. For latent Dirichlet allocation, the information rate is reduced most by the first 20 topics. The new model uses a larger number of topics and exhibits a greater information rate reduction as more topics are added. In latent Dirichlet allocation, the latent topic for a given word is inferred using the identity of the word, the number of times the word has previously been assumed to be generated by each topic, and the number of times each topic has been used in the current document. In the new model, the previous word is also taken into ac-

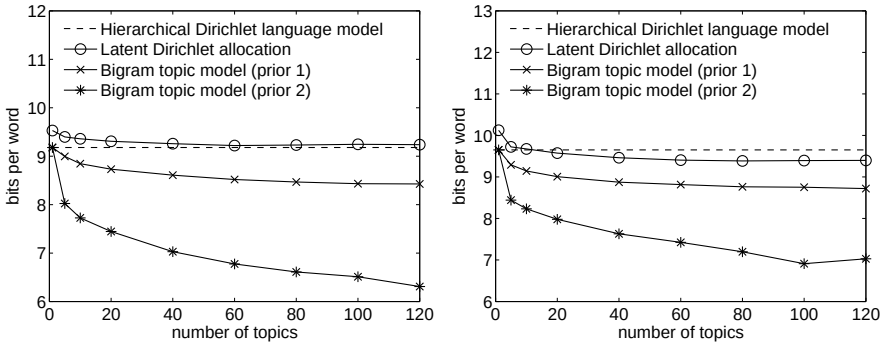

Figure 1. Information rates of the test data, measured in bits per word, under the different models versus number of topics. Left: Psychological Review Abstracts data. Right: 20 Newsgroups data.

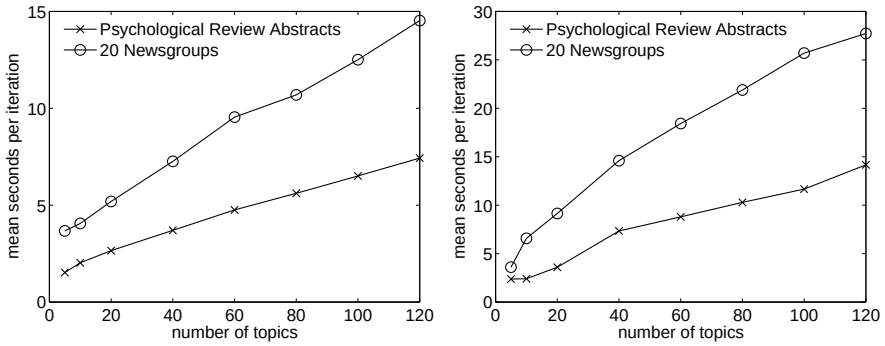

Figure 2. Mean time taken to perform a single iteration of the Gibbs EM algorithm described in section 4 as a function of the number of topics for both variants of the new model. Left: prior 1. Right: prior 2.

count. This additional information means words that were considered to be generated by the same topic in latent Dirichlet allocation, may now be assumed to have been generated by different topics, depending on the contexts in which they are seen. Consequently, the new model tends to use a greater number of topics.

In addition to comparing predictive accuracy, it is instructive to look at the inferred topics. Table 1 shows the words most frequently assigned to a selection of topics extracted from the 20 Newsgroups training data by each of the models. The "unseen" token was omitted. The topics inferred using latent Dirichlet allocation contain many function words, such as "the", "in" and "to". In contrast, all but one of the topics inferred by the new model, especially with prior 2, typically contain fewer function words. Instead, these are largely collected into the single remaining topic, shown in the last column of rows 2 and 3 in table 1. This effect is similar, though less pronounced, to that achieved by Griffiths et al.'s composite model (2004), in which function words are handled by a hidden Markov model, while content words are handled by latent Dirichlet allocation.

6. Future Work

There is a another possible prior over Φ, in addition to the two priors discussed in this paper. This prior has a hyperparameter vector for each previous word context j, resulting in W hyperparameter vectors:

$$P(\Phi|\{\beta_j \boldsymbol{m}_j\}) = \prod_j \prod_k \text{Dirichlet}(\boldsymbol{\phi}_{j,k}|\beta_j \boldsymbol{m}_j). \quad (32)$$

Here, information is shared between all distributions with previous word context j. This prior captures the notion of common bigrams—word pairs that always occur together. However, the number of hyperparameter vectors is extremely large—much larger than the number of hyperparameters in prior 2—with comparatively little data from which to infer them. To make effective use of this prior, each normalized measure $\boldsymbol{m}_j$ should itself be assigned a Dirichlet prior. This variant of the model could be compared with those presented

in this paper. To enable a direct comparison, Dirichlet hyperpriors could also be placed on the hyperparameters of the priors described in section 3.

7. Conclusions

Creating a single model that integrates bigram-based and topic-based approaches to document modeling has several benefits. Firstly, the predictive accuracy of the new model, especially when using prior 2, is significantly better than that of either latent Dirichlet allocation or the hierarchical Dirichlet language model. Secondly, the model automatically infers a separate topic for function words, meaning that the other topics are less dominated by these words.

Acknowledgments

Thanks to Phil Cowans, David MacKay and Fernando Pereira for useful discussions. Thanks to Andrew Suffield for providing sparse matrix code.

References

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

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

Griffiths, T. L., & Steyvers, M. (2005). Topic modeling toolbox. `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. *Advances in Neural Information Processing Systems*.

Jelinek, F., & Mercer, R. (1980). Interpolated estimation of Markov source parameters from sparse data. In E. Gelsema and L. Kanal (Eds.), *Pattern recognition in practice*, 381–402. North-Holland publishing company.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

MacKay, D. J. C., & Peto, L. C. B. (1995). A hierarchical Dirichlet language model. *Natural Language Engineering*, *1*, 289–307.

Minka, T. P. (2003). Estimating a Dirichlet distribution. `http://research.microsoft.com/~minka/papers/dirichlet/`.

Rennie, J. (2005). 20 newsgroups data set. `http://people.csail.mit.edu/jrennie/20Newsgroups/`.

Table 1. Top: The most commonly occurring words in some of the topics inferred from the 20 Newsgroups training data by latent Dirichlet allocation. Middle: Some of the topics inferred from the same data by the new model with prior 1. Bottom: Some of the topics inferred by the new model with prior 2. Each column represents a single topic, and words appear in order of frequency of occurrence. Content words are in bold. Function words, which are not in bold, were identified by their presence on a standard list of stop words: `http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words`. All three sets of topics were taken from models with 90 topics.

Latent Dirichlet allocation			
the	i	that	**easter**
"number"	is	**proteins**	**ishtar**
in	**satan**	the	a
to	the	of	the
espn	which	to	have
hockey	and	i	with
a	of	if	but
this	**metaphorical**	"number"	**english**
as	**evil**	you	and
run	there	**fact**	is

Bigram topic model (prior 1)			
to	the	the	the
party	**god**	and	a
arab	is	between	to
not	**belief**	**warrior**	i
power	**believe**	**enemy**	of
any	**use**	**battlefield**	"number"
i	there	a	is
is	**strong**	of	in
this	**make**	there	and
things	i	**way**	it

Bigram topic model (prior 2)			
party	**god**	"number"	the
arab	**believe**	the	to
power	about	**tower**	a
as	**atheism**	**clock**	and
arabs	**gods**	a	of
political	before	**power**	i
are	see	**motherboard**	is
rolling	**atheist**	**mhz**	"number"
london	most	**socket**	it
security	**shafts**	**plastic**	that