Incorporating Lexical Priors into Topic Models

Jagadeesh Jagarlamudi
University of Maryland
College Park, USA
jags@umiacs.umd.edu

Hal Daumé III
University of Maryland
College Park, USA
hal@umiacs.umd.edu

Raghavendra Udupa
Microsoft Research
Bangalore, India
raghavu@microsoft.com

Abstract

Topic models have great potential for helping users understand document corpora. This potential is stymied by their purely unsupervised nature, which often leads to topics that are neither entirely meaningful nor effective in extrinsic tasks (Chang et al., 2009). We propose a simple and effective way to guide topic models to learn topics of specific interest to a user. We achieve this by providing *sets of seed words* that a user believes are representative of the underlying topics in a corpus. Our model uses these seeds to improve *both* topic-word distributions (by biasing topics to produce appropriate seed words) and to improve document-topic distributions (by biasing documents to select topics related to the seed words they contain). Extrinsic evaluation on a document clustering task reveals a significant improvement when using seed information, even over other models that use seed information naïvely.

1 Introduction

Topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have emerged as a powerful tool to analyze document collections in an unsupervised fashion. When fit to a document collection, topic models implicitly use document level co-occurrence information to group semantically related words into a single topic. Since the objective of these models is to maximize the probability of the observed data, they have a tendency to explain only the most obvious and superficial aspects of a corpus. They effectively sacrifice performance on rare topics to do a better job in modeling frequently occurring words. The user is then left with a skewed impression of the corpus, and perhaps one that does not perform well in extrinsic tasks.

To illustrate this problem, we ran LDA on the most frequent five categories of the Reuters-21578 (Lewis et al., 2004) text corpus. This document distribution is very skewed: more than half of the collection belongs to the most frequent category (“Earn”). The five topics identified by the LDA are shown in Table 1. A brief observation of the topics reveals that LDA has roughly allocated topics 1 & 2 for the most frequent class (“Earn”) and one topic for the subsequent two frequent classes (“Acquisition” and “Forex”) and merged the least two frequent classes (“Crude” and “Grain”) into a single topic. The red colored words in topic 5 correspond to the “Crude” class and blue words are from the “Grain” class.

This leads to the situation where the topics identified by LDA are not in accordance with the underlying topical structure of the corpus. This is a problem not just with LDA: it is potentially a problem with any extension thereof that have focused on improving the semantic coherence of the words in each topic (Griffiths et al., 2005; Wallach, 2005; Griffiths et al., 2007), the document topic distributions (Blei and McAuliffe, 2008; Lacoste-Julien et al., 2008) or other aspects (Blei. and Lafferty., 2009).

We address this problem by providing some additional information to the model. Initially, along with the document collection, a user may provide higher level view of the document collection. For instance, as discussed in Section 4.4, when run on historical NIPS papers, LDA fails to find topics related to Brain Imaging, Cognitive Science or Hardware, even though we know from the call for

mln, dlrs, billion, year, pct, company, share, april, record, cts, quarter, march, earnings, stg, first, pay
mln, NUM, cts, loss, net, dlrs, shr, profit, revs, year, note, oper, avg, shrs, sales, includes
lt, company, shares, corp, dlrs, stock, offer, group, share, common, board, acquisition, shareholders
bank, market, dollar, pct, exchange, foreign, trade, rate, banks, japan, yen, government, rates, today
oil, tonnes, prices, mln, wheat, production, pct, gas, year, grain, crude, price, corn, dlrs, bpd, opec

Table 1: Topics identified by LDA on the frequent-5 categories of the Reuters corpus. The categories are Earn, Acquisition, Forex, Grain and Crude (in the order document frequency).

1	company, billion, quarter, shrs, earnings
2	acquisition, procurement, merge
3	exchange, currency, trading, rate, euro
4	grain, wheat, corn, oilseed, oil
5	natural, gas, oil, fuel, products, petrol

Table 2: An example for sets of seed words (*seed topics*) for the frequent-5 categories of the Reuters-21578 categorization corpus. We use them as running example in the rest of the paper.

papers that such topics should exist in the corpus. By allowing the user to provide some *seed words* related to these underrepresented topics, we encourage the model to find evidence of these topics in the data. Importantly, we *only* encourage the model to follow the seed sets and do *not* force it. So if it has compelling evidence in the data to overcome the seed information then it still has the freedom to do so. Our seeding approach in combination with the interactive topic modeling (Hu et al., 2011) will allow a user to both *explore* a corpus, and also guide the exploration towards the distinctions that he/she finds more interesting.

2 Incorporating Seeds

Our approach to allowing a user to guide the topic discovery process is to let him provide *seed information* at the level of word type. Namely, the user provides sets of seed words that are representative of the corpus. Table 2 shows an example of seed sets one might use for the Reuters corpus. This kind of supervision is similar to the seeding in bootstrapping literature (Thelen and Riloff, 2002) or prototype-based learning (Haghighi and Klein, 2006). Our reliance on seed sets is orthogonal to existing approaches that use external knowledge, which operate at the level of documents (Blei and McAuliffe, 2008), tokens (Andrzejewski and Zhu, 2009) or pair-wise constraints (Andrzejewski et al., 2009).

We build a model that uses the seed words in two ways: to improve both topic-word and document-topic probability distributions. For ease of exposition, we present these ideas separately and then in combination (Section 2.3). To improve topic-word distributions, we set up a model in which each topic prefers to generate words that are related to the words in a seed set (Section 2.1). To improve document-topic distributions, we encourage the model to select document-level topics based on the existence of input seed words in that document (Section 2.2).

Before moving on to the details of our models, we briefly recall the generative story of the LDA model and the reader is encouraged to refer to (Blei et al., 2003) for further details.

1. For each topic $k = 1 \cdots$ T,
   - choose $\phi_k \sim \text{Dir}(\beta)$.
2. For each document d, choose $\theta_d \sim \text{Dir}(\alpha)$.
   - For each token $i = 1 \cdots N_d$:
     (a) Select a topic $z_i \sim \text{Mult}(\theta_d)$.
     (b) Select a word $w_i \sim \text{Mult}(\phi_{z_i})$.

where T is the number of topics, α, β are hyperparameters of the model and ϕ_k and θ_d are topic-word and document-topic Multinomial probability distributions respectively.

2.1 Word-Topic Distributions (Model 1)

In regular topic models, each topic k is defined by a Multinomial distribution ϕ_k over words. We extend this notion and instead define a topic as a mixture of *two* Multinomial distributions: a “seed topic” distribution and a “regular topic” distribution. The seed topic distribution is constrained to *only* generate words from a corresponding seed set. The regular topic distribution may generate any word (*including* seed words). For example, seed topic 4 (in Table 2) can only generate the five words in its set. The word “oil” can be generated by seed topics 4 and 5, as well as any regular

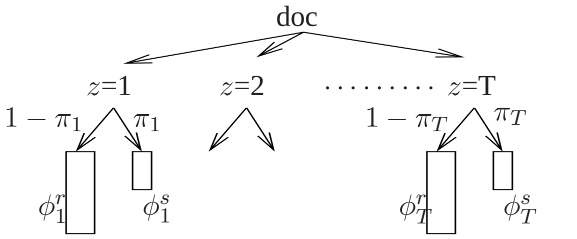


Figure 1: Tree representation of a document in Model 1.

topic. We want to emphasize that, like any regular topic, each seed topic is a *non*-uniform probability distribution over the words in its set. The user only inputs the sets of seed words and the model will infer their probability distributions.

For the sake of simplicity, we describe our model by assuming a one-to-one correspondence between seed and regular topics. This assumption can be easily relaxed by duplicating the seed topics when there are more regular topics. As shown in Fig. 1, each document is a mixture over T topics, where each of those topics is a mixture of a regular topic ($\phi^r_\cdot$) and its associated seed topic ($\phi^s_\cdot$) distributions. The parameter π_k controls the probability of drawing a word from the seed topic distribution versus the regular topic distribution. For our first model, we assume that the corpus is generated based on the following generative process (its graphical notation is shown in Fig. 2(a)):

1. For each topic k=1$\cdots$T,
    - (a) Choose regular topic $\phi^r_k \sim$ Dir(β_r).
    - (b) Choose *seed* topic $\phi^s_k \sim$ Dir(β_s).
    - (c) Choose $\pi_k \sim$ Beta$(1, 1)$.
2. For each document d, choose $\theta_d \sim$ Dir(α).
    - For each token $i = 1 \cdots N_d$:
        - (a) Select a topic $z_i \sim$ Mult(θ_d).
        - (b) Select an indicator $x_i \sim$ Bern(π_{z_i})
        - (c) if x_i is 0
            - – Select a word $w_i \sim$ Mult$(\phi^r_{z_i})$. // choose from regular topic
        - (d) if x_i is 1
            - – Select a word $w_i \sim$ Mult$(\phi^s_{z_i})$. // choose from seed topic

The first step is to generate Multinomial distributions for both seed topics and regular topics. The seed topics are drawn in a way that *constrains* their distribution to only generate words in the corresponding seed set. Then, for each token in a document, we first generate a topic. After choosing a topic, we flip a (biased) coin to pick either the seed or the regular topic distribution. Once this distribution is selected we generate a word from it. It is important to note that although there are 2×T topic-word distributions in total, each document is still a mixture of *only* T topics (as shown in Fig. 1). This is crucial in relating seed and regular topics and is similar to the way topics and aspects are tied in TAM model (Paul and Girju, 2010).

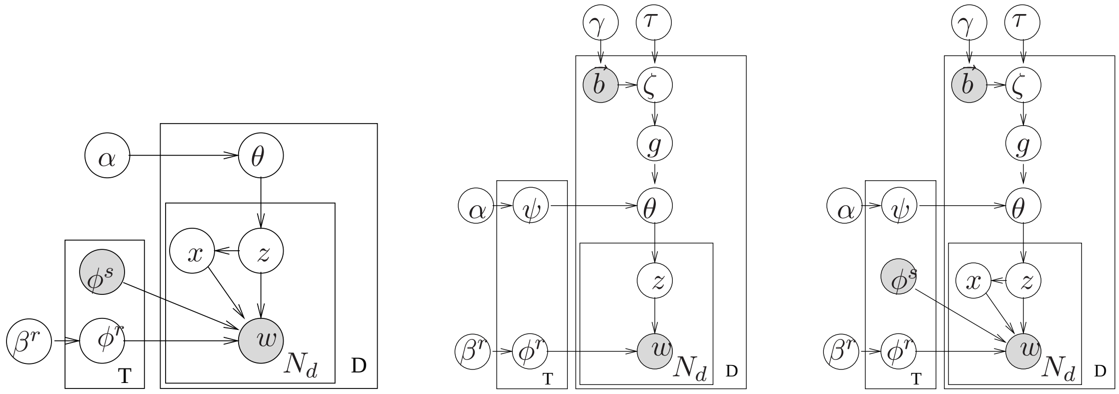
To understand how this model gathers words related to seed words, consider a seed topic (say the fourth row in Table 2) with seed words $\{$grain, wheat, corn, *etc.* $\}$. Now by assigning all the related words such as “tonnes”, “agriculture”, “production” *etc.* to its corresponding *regular topic*, the model can potentially put high probability mass on topic $z = 4$ for agriculture related documents. Instead, if it places these words in another regular topic, say $z = 3$, then the document probability mass has to be distributed among topics 3 and 4 and as a result the model will pay a steeper penalty. Thus the model uses seed topic to gather related words into its associated regular topic and as a consequence the document-topic distributions also become focussed.

We have experimented with two ways of choosing the binary variable x_i (step 2b) of the generative story. In the first method, we fix this sampling probability to a constant value which is independent of the chosen topic (*i.e.* $\pi_i = \hat{\pi}$, $\forall i = 1 \cdots$ T). And in the second method we learn the probability as well (Sec. 4).

2.2 Document-Topic distributions (Model 2)

In the previous model we used seed words to improve topic-word probability distributions. Here we propose a model to explore the use of seed words to improve document-topic probability distributions. Unlike the previous model, we will present this model in the general case where the number of seed topics is not equal to the number of regular topics. Hence, we associate each seed set (we refer seed set as group for conciseness) with a Multinomial distribution over the regular topics which we call group-topic distribution.

To give an overview of our model, first, we transfer the seed information from words onto



(a) Model 1 (b) Model 2 (c) SeededLDA

Figure 2: The graphical notation of all the three models. In Model 1 we use seed topics to improve the topic-word probability distributions. In Model 2, the seed topic information is first transfered to the document level based on the document tokens and then it is used to improve document-topic distributions. In the final, SeededLDA, model we combine both the models. In Model 1 and SeededLDA, we dropped the dependency of ϕ^s on hyper parameter β_s since it is observed. And, for clarity, we also dropped the dependency of x on π.

the documents that contain them. Then, the document-topic distribution is drawn in a two step process: we sample a seed set (g for group) and then use its group-topic distribution (ψ_g) as prior to draw the document-topic distribution (θ_d). We used this two step process, to allow flexible number of seed and regular topics, and to tie the topic distributions of all the documents within a group. We assume the following generative story and its graphical notation is shown in Fig. 2(b).

1. For each k = 1$\cdots$T,
   (a) Choose $\phi^r_k \sim$ Dir(β_r).
2. For each seed set s = 1$\cdots$S,
   (a) Choose group-topic distribution $\psi_s \sim$ Dir(α). // the topic distribution for s^{th} group (seed set) – a vector of length T.
3. For each document d,
   (a) Choose a binary vector $\vec{b}$ of length S.
   (b) Choose a document-group distribution $\zeta^d \sim$ Dir($\tau\vec{b}$).
   (c) Choose a group variable $g \sim$ Mult(ζ^d)
   (d) Choose $\theta_d \sim$ Dir(ψ_g). // of length T
   (e) For each token $i = 1 \cdots N_d$:
      i. Select a topic $z_i \sim$ Mult(θ_d).
      ii. Select a word $w_i \sim$ Mult($\phi^r_{z_i}$).

We first generate T topic-word distributions (ϕ_k) and S group-topic distributions (ψ_s). Then for each document, we generate a list of seed sets that are allowed for this document. This list is represented using the binary vector $\vec{b}$. This binary vector can be populated based on the document words and hence it is treated as an observed variable. For example, consider the (very short!) document "oil companies have merged". According to the seed sets from Table 2, we define a binary vector that denotes which seed topics contain words in this document. In this case, this vector $\vec{b} = \langle 1, 1, 0, 1, 1 \rangle$, indicating the presence of seeds from sets 1, 2, 4 and 5.[1] As discussed in (Williamson et al., 2010), generating binary vector is crucial if we want a document to talk about topics that are less prominent in the corpus.

The binary vector $\vec{b}$, that indicates which seeds exist in this document, defines a *mean* of a Dirichlet distribution from which we sample a *document-group* distribution, ζ^d (step 3b). We set the concentration of this Dirichlet to a hyperparamter τ, which we set by hand (Sec. 4); thus, $\zeta^d \sim$ Dir($\tau\vec{b}$). From the resulting multinomial, we draw a *group* variable g for this document. This group variable brings clustering structure among the documents by grouping the documents that are likely to talk about same seed set.

Once the group variable (g) is drawn, we choose the document-topic distribution (θ_d) from a Dirichlet distribution with the group's-topic distribution as the prior (step 3d). This step ensures that the topic distributions of documents within each group are related. The remaining sampling

[1] As a special case, if no seed word is found in the document, $\vec{b}$ is defined as the all-ones vector.

process proceeds like LDA. We sample a topic for each word and then generate a word from its corresponding topic-word distribution. Observe that, if the binary vector is all ones and if we set $\theta_d = \zeta^d$ then this model reduces to the LDA model with τ and β_r as the hyperparameters.

2.3 SeededLDA

Both of our models use seed words in different ways to improve topic-word and document-topic distributions respectively. We can combine both the above models easily. We refer to the combined model as SeededLDA and its generative story is as follows (its graphical notation is shown in Fig. 2(c)). The variables have same semantics as in the previous models.

1. For each k=1$\cdots$T,
   - (a) Choose regular topic $\phi_k^r \sim$ Dir(β_r).
   - (b) Choose *seed* topic $\phi_k^s \sim$ Dir(β_s).
   - (c) Choose $\pi_k \sim$ Beta$(1, 1)$.
2. For each seed set s = 1$\cdots$S,
   - (a) Choose group-topic distribution $\psi_s \sim$ Dir(α).
3. For each document d,
   - (a) Choose a binary vector $\vec{b}$ of length S.
   - (b) Choose a document-group distribution $\zeta^d \sim$ Dir$(\tau\vec{b})$.
   - (c) Choose a group variable $g \sim$ Mult(ζ^d).
   - (d) Choose $\theta_d \sim$ Dir(ψ_g). // of length T
   - (e) For each token $i = 1 \cdots N_d$:
     - i. Select a topic $z_i \sim$ Mult(θ_d).
     - ii. Select an indicator $x_i \sim$ Bern(π_{z_i}).
     - iii. if x_i is 0
       - Select a word $w_i \sim$ Mult$(\phi_{z_i}^r)$.
     - iv. if x_i is 1
       - Select a word $w_i \sim$ Mult$(\phi_{z_i}^s)$.

In the SeededLDA model, the process for generating group variable of a document is same as the one described in the Model 2. And like in the Model 2, we sample a document-topic probability distribution as a Dirichlet draw with the group-topic distribution of the chosen group as prior. Subsequently, we choose a topic for each token and then flip a biased coin. We choose either the seed or the regular topic based on the result of the coin toss and then generate a word from its distribution.

2.4 Automatic Seed Selection

In (Andrzejewski and Zhu, 2009; Andrzejewski et al., 2009), the seed information is provided manually. Here, we describe the use of feature selection techniques, prevalent in the classification literature, to automatically derive the seed sets. If we want the topicality structure identified by the LDA to align with the underlying class structure, then the seed words need to be representative of the underlying topicality structure. To enable this, we first take class labeled data (doesn't need to be multi-class labeled data unlike (Ramage et al., 2009)) and identify the discriminating features for each class. Then we choose these discriminating features as the initial sets of seed words. In principle, this is similar to the prototype driven unsupervised learning (Haghighi and Klein, 2006).

We use Information Gain (Mitchell, 1997) to identify the required discriminating features. The Information Gain (IG) of a word (w) in a class (c) is given by

$$IG(c, w) = H(c) - H(c|w)$$

where $H(c)$ is the entropy of the class and $H(c|w)$ is the conditional entropy of the class given the word. In computing Information Gain, we binarize the document vectors and consider whether a word occurs in any document of a given class or not. Thus obtained ranked list of words for each class are filtered for ambiguous words and then used as initial sets of seed words to be input to the model.

3 Related Work

Seed-based supervision is closely related to the idea of seeding in the bootstrapping literature for learning semantic lexicons (Thelen and Riloff, 2002). The goals are similar as well: growing a small set of seed examples into a much larger set. A key difference is the *type* of semantic information that the two approaches aim to capture: semantic lexicons are based on much more specific notions of semantics (*e.g.* all the country names) than the generic "topic" semantics of topic models. The idea of seeding has also been used in prototype-driven learning (Haghighi and Klein, 2006) and shown similar efficacies for these semi-supervised learning approaches.

LDAWN (Boyd-Graber et al., 2007) models sets of words for the word sense disambiguation task. It assumes that a topic is a distribution over synsets and relies on the Wordnet to obtain the synsets. The most related prior work is that of (Andrzejewski et al., 2009), who propose the use Dirichlet Forest priors to incorporate Must Link and Cannot Link constraints into the topic models. This work is analogous to constrained K-means clustering (Wagstaff et al., 2001; Basu et al., 2008). A must link between a pair word types represents that the model should encourage both the words to have either high or low probability in any particular topic. A cannot link between a word pair indicates both the words should not have high probability in a single topic. In the Dirichlet Forest approach, the constraints are first converted into trees with words as the leaves and edges having pre-defined weights. All the trees are joined to a dummy node to form a forest. The sampling for a word translates into a random walk on the forest: starting from the root and selecting one of its children based on the edge weights until you reach a leaf node.

While the Dirichlet Forest method requires supervision in terms of Must link and Cannot link information, the Topics In Sets (Andrzejewski and Zhu, 2009) model proposes a different approach. Here, the supervision is provided at the *token* level. The user chooses specific tokens and restrict them to occur only with in a specified list of topics. While this needs minimal changes to the inference process of LDA, it requires information at the level of tokens. The word type level seed information can be converted into token level information (like we do in Sec. 4) but this prevents their model from distinguishing the tokens based on the word senses.

Several models have been proposed which use supervision at the document level. Supervised LDA (Blei and McAuliffe, 2008) and DiscLDA (Lacoste-Julien et al., 2008) try to predict the category labels (*e.g.* sentiment classification) for the input documents based on a document labeled data. Of these models, the most related one to SeededLDA is the LabeledLDA model (Ramage et al., 2009). Their model operates on multi-class labeled corpus. Each document is assumed to be a mixture over a known subset of topics (classes) with each topic being a distribution over words. The process of generating document topic distribution in LabeledLDA is similar to the process of generating group distribution in our Model 2 (Sec. 2.2). However our model differs from LabeledLDA in the subsequent steps. Rather than using the group distribution directly, we sample a group variable and use it to constrain the document-topic distributions of all the documents within this group. Moreover, in their model the binary vector is observed directly in the form of document labels while, in our case, it is automatically populated based on the document tokens.

Interactive topic modeling brings the user into the loop, by allowing him/her to make suggestions on how to improve the quality of the topics at each iteration (Hu et al., 2011). In their approach, the authors use Dirichlet Forest method to incorporate the user's preferences. In our experiments (Sec. 4), we show that SeededLDA performs better than Dirichlet Forest method, so SeededLDA when used with their framework can allow an user to explore a document collection in a more meaningful manner.

4 Experiments

We evaluate different aspects of the model separately. Our experimental setup proceeds as follows: a) Using an existing model, we evaluate the effectiveness of automatically derived constraints indicating the potential benefits of adding seed words into the topic models. b) We evaluate each of our proposed models in different settings and compare with multiple baseline systems.

Since our aim is to overcome the dominance of majority topics by encouraging the topicality structure identified by the topic models to align with that of the document corpus, we choose extrinsic evaluation as the primary evaluation method. We use document clustering task and use frequent-5 categories of Reuters-21578 corpus (Lewis et al., 2004) and four classes from the 20 Newsgroups data set (*i.e.*'rec.autos', 'sci.electronics', 'comp.hardware' and 'alt.atheism'). For both the corpora we do the standard preprocessing of removing stopwords and infrequent words (Williamson et al., 2010).

For all the models, we use a Collapsed Gibbs sampler (Griffiths and Steyvers, 2004) for the inference process. We use the standard hyperparameters values $\alpha = 1.0$, $\beta = 0.01$ and $\tau = 1.0$ and run the sampler for 1000 iterations, but one can use techniques like slice sampling to estimate the hyperparameters (Johnson and Goldwater, 2009).

	Reuters		20 Newsgroups	
	F-measure	VI	F-measure	VI
LDA	0.64 ($\pm.05$)	1.26 ($\pm.16$)	0.77 ($\pm.06$)	0.9 ($\pm.13$)
Dirichlet Forest	**0.67*** ($\pm.02$)	**1.17** ($\pm.11$)	**0.79**($\pm.01$)	**0.83***($\pm.03$)
Δ over LDA	(*+4.68%*)	(*-7.1%*)	(*+2.6%*)	(*-7.8%*)

Table 3: The effect of adding constraints by Dirichlet Forest Encoding. For Variational Information (VI) a lower score indicates a better clustering. * indicates statistical significance at $p = 0.01$ as measured by the t-test. All the four improvements are significant at $p = 0.05$.

We run all the models with the same number of topics as the number of clusters. Then, for each document, we find the topic that has maximum probability in the posterior document-topic distribution and assign it to that cluster. The accuracy of the document clustering is measured in terms of F-measure and Variation of Information. F-measure is calculated based on the pairs of documents, *i.e.* if two documents belong to a cluster in both ground truth and the clustering proposed by the system then it is counted as correct, otherwise it is counted as wrong. Variational Information (VI) of two clusterings X and Y is given as (Meilă, 2007):

$$\text{VI}(X, Y) = H(X) + H(Y) - 2I(X, Y)$$

where $H(X)$ denotes the entropy of the clustering X and $I(X, Y)$ denotes the mutual information between the two clusterings. For VI, a lower value indicates a better clustering. All the accuracies are averaged over 25 different random initializations and all the significance results are measured using the t-test at $p = 0.01$.

4.1 Seed Extraction

The seeds were extracted automatically (Sec. 2.4) based on a small sample of labeled data other than the test data. We first extract 25 seeds words per each class and then remove the seed words that appear in more than one class. After this filtering, on an average, we are left with 9 and 15 words per each seed topic for Reuters and 20 Newsgroups corpora respectively.

We use the existing Dirichlet Forest method to evaluate the effectiveness of the automatically extracted seed words. The Must and Cannot links required for the supervision (Andrzejewski et al., 2009) are automatically obtained by adding a must-link between every pair of words belonging to the same seed set and a split constraint between every pair of words belonging to different sets. The accuracies are averaged over 25 different random initializations and are shown in Table 3. We have also indicated the relative performance gains compared to LDA. The significant improvement over the plain LDA demonstrates the effectiveness of the automatic extraction of seed words in topic models.

4.2 Document Clustering

In the next experiment, we compare our models with LDA and other baselines. The first baseline (maxCluster) simply counts the number of tokens in each document from each of the seed topics and assigns the document to the seed topic that has most tokens. This results in a clustering of documents based on the seed topic they are assigned to. This baseline evaluates the effectiveness of the seed words with respect to the underlying clustering. Apart from the maxCluster baseline, we use LDA and z-labels (Andrzejewski and Zhu, 2009) as our baselines. For z-labels, we treat all the tokens of a seed word in the same way. Table 4 shows the comparison of our models with respect to the baseline systems.[2] Comparing the performance of maxCluster to that of LDA, we observe that the seed words themselves do a poor job in clustering the documents.

We experimented with two variants of Model 1. In the first run (Model 1) we sample the π_k value, *i.e.* the probability of choosing a seed topic for each topic. While in the 'Model 1 ($\hat{\pi} = 0.7$)' run, we fix this probability to a constant value of 0.7 irrespective of the topic.[3] Though both the models

[2] The code used for LDA baseline in Tables 3 and 4 are different. For Table 3, we use the code available from http://pages.cs.wisc.edu/∼andrzeje/research/df_lda.html. We use our own version for Table 4. We tried to produce a comparable baseline by running the former for more iterations and with different hyperparameters. In Table 3, we report their best results.

[3] We chose this value based on intuition; it is *not* tuned.

	Reuters		20 Newsgroups	
	F-measure	VI	F-measure	VI
maxCluster	0.53	1.75	0.58	1.44
LDA	0.66 ($\pm$.04)	1.2 ($\pm$.12)	0.76 ($\pm$.06)	0.9 ($\pm$.14)
z-labels	0.73 ($\pm$.01)	1.04 ($\pm$.01)	0.8 ($\pm$.00)	0.82 ($\pm$.01)
Δ over LDA	(*+10.6%*)	(*-13.3%*)	(*+5.26%*)	(*-8.8%*)
Model 1	0.69 ($\pm$.00)	1.13 ($\pm$.01)	0.8 ($\pm$.01)	0.81 ($\pm$.02)
Model 1 ($\hat{\pi}$ = 0.7)	0.73 ($\pm$.00)	1.09 ($\pm$.01)	0.8 ($\pm$.01)	0.81 ($\pm$.02)
Model 2	0.66 ($\pm$.04)	1.22 ($\pm$.1)	0.77 ($\pm$.07)	0.85 ($\pm$.12)
SeededLDA	**0.76*** ($\pm$.01)	**0.99*** ($\pm$.03)	**0.81*** ($\pm$.01)	**0.75*** ($\pm$.02)
Δ over LDA	(*+15.5%*)	(*-17.5%*)	(*+6.58%*)	(*-16.7%*)

Table 4: Accuracies on document clustering task with different models. * indicates significant improvement compared to the z-labels approach, as measured by the t-test with $p = 0.01$. The relative performance gains are with respect to the LDA model and are provided for comparison with Dirichlet Forest method (in Table 3.)

performed better than LDA, fixing the probability gave better results. When we attempt to learn this value, the model chooses to explain some of the seed words by the regular topics. On the other hand, when π is fixed, it explains almost all the seed words based on the seed topics. The next row (Model 2) indicates the performance of our second model on the same data sets. The first model seems to be performing better than the second model, which is justifiable since the latter uses seed topics indirectly. Though the variants of Model 1 and Model 2 performed better than the LDA, they fell short of the z-labels approach.

Table 4 also shows the performance of our combined model (SeededLDA) on both the corpora. When the models are combined, the performance improves over each of them and is also better than the baseline systems. As explained before, our individual models improve both the topic-word and document-topic distributions respectively. But it turns out that the knowledge learnt by both the individual models is complementary to each other. As a result the combined model performed better than the individual models and other baseline systems. Comparing the last rows of Tables 4 and 3, we notice that the relative performance gains observed in the case of SeededLDA is significantly higher than the performance gains obtained by incorporating the constraints using the Dirichlet Forest method. Moreover, as indicated in the Table 4, SeededLDA achieves significant gains over the z-labels approach as well.

We have also provided the standard intervals for each of the approaches. A quick inspection of these intervals reveals the superior performance of SeededLDA compared to all the baselines. The standard deviation of the F-measures over different random initializations of our our model is about 1% for both the corpora while it is 4% and 6% for the LDA on Reuters and 20 Newsgroups corpora respectively. The reduction in the variance, across all the approaches that use seed information, shows the increased robustness of the inference process when using seed words. From the accuracies in both the tables, it is clear that SeededLDA model out-performs other models which try to incorporate seed information into the topic models.

4.3 Effect of Ambiguous Seeds

In the following experiment we study the effect of ambiguous seeds. We allow a seed word to occur in multiple seed sets. Table 6 shows the corresponding results. The performance drops when we add ambiguous seed words, but it is still higher than that of the LDA model. This suggests that the quality of the seed topics is determined by the discriminative power of the seed words rather than the number of seed words in each seed topic. The topics identified by the SeededLDA on Reuters corpus are shown in the Table 5. With the help of the seed sets, the model is able to split the ‘Grain’ and ‘Crude’ into two separate topics which were merged into a single topic by the plain LDA.

4.4 Qualitative Evaluation on NIPS papers

We ran LDA and SeededLDA models on the NIPS papers from 2001 to 2010. For this corpus, the seed words are chosen from the call for proposal.

group, offer, common, cash, agreement, shareholders, acquisition, stake, merger, board, sale
oil, price, prices, production, lt, gas, crude, 1987, 1985, bpd, opec, barrels, energy, first, petroleum
0, mln, cts, net, loss, 2, dlrs, shr, 3, profit, 4, 5, 6, revs, 7, 9, 8, year, note, 1986, 10, 0, sales
tonnes, wheat, mln, grain, week, corn, department, year, export, program, agriculture, 0, soviet, prices
bank, market, pct, dollar, exchange, billion, stg, today, foreign, rate, banks, japan, yen, rates, trade

Table 5: Topics identified by SeededLDA on the frequent-5 categories of Reuters corpus

	Reuters		20 Newsgroups	
	F	VI	F	VI
LDA	0.66	1.2	0.76	0.9
SeededLDA	**0.76**	**0.99**	**0.81**	**0.75**
SeededLDA (amb)	0.71	1.08	0.79	0.78

Table 6: Effect of ambiguous seed words on SeededLDA.

There are 10 major areas with sub areas under each of them. We ran both the models with 10 topics. For SeededLDA, the words in each of the areas are selected as seed words and we filter out the ambiguous seed words. Upon a qualitative observation of the output topics, we found that LDA has identified seven major topics and left out “Brain Imaging”, “Cognitive Science and Artificial Intelligence” and “Hardware Technologies” areas. Not surprisingly, but reassuringly, these areas are underrepresented among the NIPS papers. On the other hand, SeededLDA successfully identifies all of the major topics. The topics identified by LDA and SeededLDA are shown in the supplementary material.

5 Discussion

In traditional topic models, a symmetric Dirichlet distribution is used as prior for topic-word distributions. A first attempt method to incorporate seed words into the model is to use an asymmetric Dirichlet distribution as prior for the topic-word distributions (also called as Informed priors). For example, to encourage Topic 5 to align with a seed set we can choose an asymmetric prior of the form $\vec{\beta}_5 = \{\beta, \cdots, \beta + c, \cdots, \beta\}$, *i.e.* we increase the component values corresponding to the seed words by a positive constant value. This favors the desired seed words to be drawn with a higher probability from this topic. But, it is argued elsewhere that words drawn from such distributions rarely pick words other than the seed words (Andrzejewski et al., 2009). Moreover, since, in our method each seed topic is a distribution over the seed words, the convex combination of regular and seed topics can be seen as adding different weights (c_i) to different components of the prior vector. Thus our Model 1 can be seen as an asymmetric generalization of the Informed priors.

For comparability purposes, in this paper, we experimented with same number of regular topics as the number of seed topics. But as explained in the modeling part, our model is general enough to handle situation with unequal number of seed and regular topics. In this case, we assume that the seed topics indicate a higher level of topicality structure of the corpus and associate each seed topic (or group) with a distribution over the regular topics. On the other hand, in many NLP applications, we tend to have *only* a partial information rather than high-level supervision. In such cases, one can create some empty seed sets and tweak the model 2 to output a 1 in the binary vector corresponding to these seed sets. In this paper, we used information gain to select the discriminating seed words. But in the real world applications, one can use publicly available ODP categorization data to obtain the higher level seed words and thus explore the corporal in a more meaningful way.

In this paper, we have explored two methods to incorporate lexical prior into the topic models, combining them into a single model that we call SeededLDA. From our experimental analysis, we found that automatically derived seed words can improve clustering performance significantly. Moreover, we found out that allowing a seed word to be shared across multiple sets of seed words degrades the performance.

6 Acknowledgments

We thank the anonymous reviewers for their helpful comments. This material is partially supported by the National Science Foundation under Grant No. IIS-1153487.

References

Andrzejewski, D. and Zhu, X. (2009). Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, pages 43–48, Morristown, NJ, USA. Association for Computational Linguistics.

Andrzejewski, D., Zhu, X., and Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32, New York, NY, USA. ACM.

Basu, S., Ian, D., and Wagstaff, K. (2008). *Constrained Clustering : Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC Pres.

Blei, D. and McAuliffe, J. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems 20*, pages 121–128, Cambridge, MA. MIT Press.

Blei., D. M. and Lafferty., J. (2009). Topic models. In *Text Mining: Theory and Applications*. Taylor and Francis.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Maching Learning Research*, 3:993–1022.

Boyd-Graber, J., Blei, D. M., and Zhu, X. (2007). A topic model for word sense disambiguation. In *Empirical Methods in Natural Language Processing*.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

Griffiths, T., Steyvers, M., and Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244.

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

Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Advances in Neural Information Processing Systems*, volume 17, pages 537–544.

Haghighi, A. and Klein, D. (2006). Prototype-driven learning for sequence models. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 320–327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, Y., Boyd-Graber, J., and Satinoff, B. (2011). Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 248–257, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johnson, M. and Goldwater, S. (2009). Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 317–325, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lacoste-Julien, S., Sha, F., and Jordan, M. (2008). DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Proceedings of NIPS '08*.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

Meilă, M. (2007). Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98:873–895.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Paul, M. and Girju, R. (2010). A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Morristown, NJ, USA. Association for Computational Linguistics.

Thelen, M. and Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *In Proc. 2002 Conf. Empirical Methods in NLP (EMNLP)*.

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 577–584, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Wallach, H. M. (2005). Topic modeling: beyond bag-of-words. In *NIPS 2005 Workshop on Bayesian Methods for Natural Language Processing*.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010). The IBP compound dirichlet process and its application to focused topic modeling. In *ICML*, pages 1151–1158.