Optimizing Semantic Coherence in Topic Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large organizations often face the critical challenge of sharing information and maintaining connections between disparate subunits. Tools for automated analysis of document collections, such as topic models, can provide an important means for communication. The value of topic modeling is in its ability to discover interpretable, coherent themes from unstructured document sets, yet it is not unusual to find semantic mismatches that substantially reduce user confidence. In this paper, we first present an expert-driven topic annotation study, undertaken in order to obtain an annotated set of baseline topics and their distinguishing characteristics. We then present a metric for detecting poor-quality topics that does not rely on human feedback or external reference corpora. Next we introduce a new topic model that incorporates salient properties of this metric. We show significant gains in topic quality on a substantial document collection from the National Institutes of Health, measured using both automated evaluation metrics and expert evaluations.

1 Introduction

The proliferation of digital documents is both a challenge and an opportunity. Large institutions such as corporations, universities, and government agencies are increasingly faced with the difficult task of organizing and navigating rapidly-growing and evolving text collections. Although search engines are effective at satisfying specific information needs, they do little to describe overall semantic content or to provide high-level summaries of institutional emphases. Meanwhile, systems that identify common themes and recognize similarities between documents can be a major strategic asset, especially for large, complex organizations: Such institutions typically have many independent departments that may not be aware of developments in other groups. Opportunities for collaboration and strategic changes are easily lost if documents produced by all areas of an institution are not analyzed in aggregate, thereby providing a window into complex intra-institution relationships.

Statistical topic models such as latent Dirichlet allocation (LDA) [2] provide a powerful framework for representing and summarizing the contents of large document collections. In our experience, however, the primary obstacle to acceptance of statistical topic models by users outside of the topic modeling community is the presence of poor quality topics. Topics that mix unrelated or loosely-related concepts substantially reduce users' confidence in the utility of such automated systems.

The evaluation of statistical topic models has traditionally been dominated by either extrinsic methods (i.e., using the inferred topics as to perform some external task such as information retrieval [13]) or quantitative intrinsic methods, such as computing the probability of held-out documents [12]. Comparatively little attention has been given to the "quality" or semantic coherence of the inferred topics (i.e., do the topics contain words that, according to subjective human judgment, are representative of single coherent concept). In fact, even for some external tasks, semantic coherence is crucially important. For example, users will be less likely to trust topic-based navigation tools if they perceive the quality of presented topics as poor. For topic models to be widely adopted by such users, two conditions must be satisfied. First, such models need to be useful—they must

provide users with information that they do not already know. To do this, they must be sufficiently specialized. Models that provide corpus coverage using a large number of fine-grained topics can provide users with a focused view of the data that may reveal surprising insights and connections. In contrast, models that use a smaller number of quite general topics are unlikely to provide users with new information. Second, model output must be perceived as being accurate. Widespread-adoption of topic modeling tools depends on users' perceptions of utility. If users are confident that the information generated by such models is accurate, then the resulting tools are perceived as being more useful. While topic modeling researchers are often comfortable ignoring poor-quality topics and focusing their attention on topics of higher quality, poor-quality topics can cause users outside of the machine learning community to lose confidence in the accuracy of topic models. In order to satisfy these conditions for widespread-adoption, this paper focuses on the task of building fine-grained statistical topic models with high-quality topics from highly domain-specific document collections.

Recent work by Chang et al. [4] and Newman et al. [10] challenged established evaluation methodologies by exploring human-based evaluation, with some surprising results: Chang et al. found that the probability of held-out documents is not always a good predictor of human judgments—human evaluators sometimes preferred models that assigned *lower* probability to held-out documents. Meanwhile, Newman et al. found that for general-purpose topic models (i.e., topic models constructed from news articles, books, and other documents intended for general public consumption), an automated evaluation metric based on word co-occurrence statistics gathered from Wikipedia performed well at predicting human evaluations. Although there has been some work in semi-supervised contexts on constructing models that avoid semantic coherence problems [1], we are not aware of any work that relies only the unstructured documents being modeled. In this paper, we take a different approach to evaluating the semantic coherence of inferred topics, specifically focusing on highly domain-specific data, for which nonexpert evaluations [4] and external reference corpora [10] are inappropriate or unavailable. First, we present an expert-driven topic annotation study, undertaken using grant abstracts from the National Institutes of Health and related journal paper abstracts. Using the findings from this study, we identify salient characteristics of multiple types of poor-quality topics, and design a new *intrinsic* evaluation metric that predicts expert topic annotations without recourse to external reference corpora (which, for many domains, are not readily available). We then develop a novel statistical topic model, based on this metric, intended to yield the types of topics preferred by expert evaluators, without any human intervention. Since this model exhibits significant gains in topic quality, measured using automated metrics and expert evaluations, we recommend it as a replacement for LDA wherever semantic coherence of topics is a priority.

2 Latent Dirichlet Allocation

LDA is a generative probabilistic model for documents $\mathcal{W} = \{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(D)}\}$. Each "topic" t is a discrete probability distribution over words with probability vector $\boldsymbol{\phi}_t$. A Dirichlet prior is placed over $\Phi = \{\boldsymbol{\phi}_1, \ldots \boldsymbol{\phi}_T\}$. This prior is typically assumed to be symmetric (i.e., the base measure is fixed to a uniform distribution $\boldsymbol{u}$ over words) with concentration parameter β:

$$P(\Phi) = \textstyle\prod_t \mathrm{Dir}\left(\boldsymbol{\phi}_t; \beta\boldsymbol{u}\right) = \prod_t \dfrac{\Gamma(\beta)}{\prod_w \Gamma(\frac{\beta}{W})} \prod_w \phi_{w|t}^{\frac{\beta}{W}-1} \delta\left(\textstyle\sum_w \phi_{w|t} - 1\right). \tag{1}$$

Each document, indexed by d, has a document-specific distribution over topics $\boldsymbol{\theta}_d$. The prior over $\Theta = \{\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_D\}$ is also assumed to be a symmetric Dirichlet, this time with concentration parameter α. The tokens in every document $\boldsymbol{w}^{(d)} = \{w_n^{(d)}\}_{n=1}^{N_d}$ are associated with corresponding topic assignments $\boldsymbol{z}^{(d)} = \{z_n^{(d)}\}_{n=1}^{N_d}$, drawn i.i.d. from the document-specific distribution over topics, while the tokens are drawn i.i.d. from the topics' distributions over words $\Phi = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T\}$:

$$P(\boldsymbol{z}^{(d)} \mid \boldsymbol{\theta}_d) = \textstyle\prod_n \theta_{z_n^{(d)}|d} \quad \text{and} \quad P(\boldsymbol{w}^{(d)} \mid \boldsymbol{z}^{(d)}, \Phi) = \prod_n \phi_{w_n^{(d)}|z_n^{(d)}}. \tag{2}$$

Dirichlet–multinomial conjugacy allows Θ and Φ to be marginalized out. Given a corpus of documents $\mathcal{W} = \{\boldsymbol{w}^{(d)}\}_{d=1}^{D}$ and corresponding topic assignments $\mathcal{Z} = \{\boldsymbol{z}^{(d)}\}_{d=1}^{D}$, maximum a posteriori (MAP) estimates of $\boldsymbol{\phi}_t$ and $\boldsymbol{\theta}_d$ are given by the corresponding conditional posterior distributions. The conditional posterior probability of word w occurring in topic t given $\mathcal{W}$ and $\mathcal{Z}$ is given by

$$P(w \mid t, \mathcal{W}, \mathcal{Z}, \beta\boldsymbol{u}) = \int \mathrm{d}\boldsymbol{\phi}_t \, P(w \mid \boldsymbol{\phi}_t) \, P(\boldsymbol{\phi}_t \mid \mathcal{W}, \mathcal{Z}, \beta\boldsymbol{u}) = \frac{N_{w|t} + \frac{\beta}{W}}{N_t + \beta}, \tag{3}$$

where N_t is the number of times topic t occurs in $\mathcal{Z}$ and $N_{w|t}$ is the number of times word w occurs in the context of topic t in $(\mathcal{W}, \mathcal{Z})$. In other words, the conditional posterior distribution over words for topic t is a Pólya conditional distribution, or a simple Pólya urn model [8]. Similarly, the conditional posterior distribution over topics for document d is also a Pólya conditional distribution.

For real-world data, documents $\mathcal{W}$ are observed, while the corresponding topic assignments $\mathcal{Z}$ are unobserved and may be inferred using either variational methods [2, 11] or MCMC methods [7]. Here, we use MCMC methods—specifically Gibbs sampling [6], which involves sequentially re-sampling each topic assignment $z_n^{(d)}$ from its conditional posterior given $\mathcal{W}$, $\alpha\boldsymbol{u}$, $\beta\boldsymbol{u}$ and $\mathcal{Z}_{\setminus d,n}$ (the current topic assignments for all tokens other than the token at position n in document d):

$$\begin{aligned} P(z_n^{(d)} \mid \mathcal{W}, \mathcal{Z}_{\setminus d,n}, \alpha\boldsymbol{u}, \beta\boldsymbol{u}) &\propto P(w_n^{(d)} \mid z_n^{(d)}, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta\boldsymbol{u})\, P(z_n^{(d)} \mid \mathcal{Z}_{\setminus d,n}, \alpha\boldsymbol{m}) \\ &\propto \frac{N^{\setminus d,n}_{w_n^{(d)}|z_n^{(d)}} + \frac{\beta}{W}}{N^{\setminus d,n}_{z_n^{(d)}} + \beta} \, \frac{N^{\setminus d,n}_{z_n^{(d)}|d} + \frac{\alpha}{T}}{N^{\setminus d,n}_{d} + \alpha}, \end{aligned} \tag{4}$$

where sub- or super-script "$\setminus d, n$" denotes a quantity excluding data from position n in document d.

3 Expert Opinions of Topic Quality

Concentrating on 300000 grant and related journal paper abstracts from the National Institutes of Health (NIH), we worked with two experts from the National Institute of Neurological Disorders and Stroke (NINDS) to collaboratively design an expert-driven topic annotation study. The goal of this study was to develop an annotated set of baseline topics, along with their salient characteristics, as a first step towards automatically identifying and producing the kinds of topics desired by experts.

3.1 Expert-Driven Annotation Protocol

In order to ensure that the topics selected for annotation were within the NINDS experts' area of expertise, they selected 148 topics (out of 500), all associated with NINDS funding. Each topic t was presented to the experts as a list of the thirty most common words for that topic, in descending order of their topic-specific MAP probabilities, computed using (3). In addition to the most common words, the experts were also given metadata for each topic: the most common sequences of two or more consecutive words assigned to that topic, the four topics that most often co-occur with that topic, the most common IDF-weighted words from titles of grants, thesaurus terms, NIH institutes, journal titles, and finally a list of the highest probability grants and PubMed papers for that topic.

The experts first categorized each topic as one of three types: "research", "grant mechanisms and publication types" or "general". The quality of each topic ("good", "intermediate", or "bad") was then evaluated using criteria specific to the type of topic. In general, topics were only annotated as "good" if they contained words that could be grouped together as a single coherent concept. Additionally, each "research" topic was only considered to be "good" if, in addition to representing a single coherent concept, the aggregate content of the set of documents with appreciable allocations to that topic was also largely consistent and coherent. Finally, for each topic marked as being either "intermediate" or "bad", one or more of the following problems was identified, as appropriate:

- **Chained:** every word is connected to every other word through some pairwise chain, but not all word pairs make sense. For example, a topic whose top three words are "acids", "fatty" and "nucleic" consists of two distinct concepts (i.e., acids produced when fats are broken down vs. the building blocks of DNA and RNA) chained via the word "acids".
- **Intruded:** either a) two or three unrelated sets of related words, joined arbitrarily, or b) an otherwise good quality topic with a few "intruder" words.
- **Random:** no clear connections between more than a few pairs of words.
- **Unbalanced:** the top words are are all logically connected, but the topic combines very general and specific terms (e.g., "signal transduction" vs. "notch signaling").

3.2 Annotation Results

The experts annotated the topics independently and then aggregated their results. Interestingly, no topics were ever considered "good" by one expert and "bad" by the other—when there was disagreement between the experts, one label was always "intermediate." In such cases, the experts discussed the reasons for their decisions and came to a consensus. Of the 148 topics selected for annotation, 90 were labeled as "good," 21 as "intermediate," and 37 as "bad." Of the topics labeled as "bad" or "intermediate," 23 were "chained," 21 were "intruded," 3 were "random," and 15 were "unbalanced". (Annotators were permitted to assign more than one problem to any given topic.)

4 Automated Metrics for Predicting Expert Annotations

The ultimate goal of this paper is to develop methods for building fine-grained, high-quality topic models from domain-specific corpora. In this section, we therefore explore the extent to which information already contained in the documents being modeled can be used to assess topic quality.

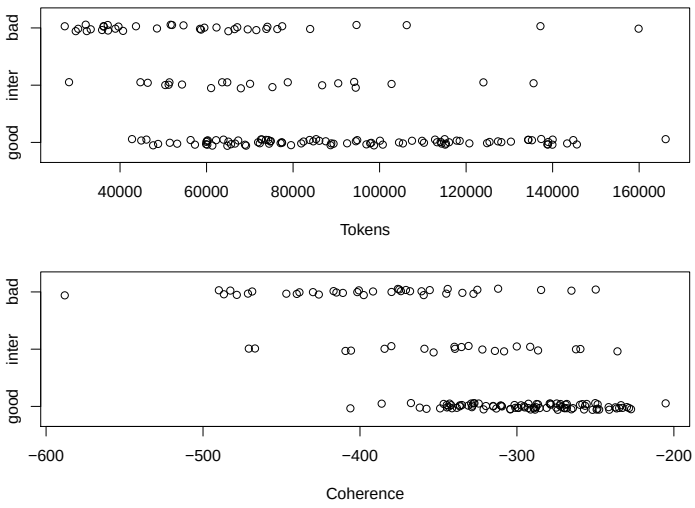

Figure 1: Association of expert annotations with the new coherence metric (top) and topic size (bottom).

4.1 Topic Size

As a simple baseline, we considered the extent to which topic "size" (as measured by the number of tokens assigned to each topic via Gibbs sampling) is a good metric for assessing topic quality. Figure 1 (top) displays the topic size (number of tokens) and expert annotations ("good", "intermediate", "bad") for the 148 topics manually labeled by expert annotators as described above. This figure suggests that topic size is a reasonable predictor of topic quality—although there is some overlap, "bad" topics are generally smaller than "good" topics. Unfortunately, this observation conflicts with the goal of building highly specialized, domain-specific topic models with many high-quality, fine-grained topics—in such models the majority of topics will have few tokens assigned to them.

4.2 Topic Coherence

When displaying topics to users, each topic t is generally represented as a list of the $M = 5, \ldots, 20$ most common words for that topic, in descending order of their topic-specific MAP probabilities. Although there has been work on automated generation of labels or headings for topics [9], we choose to work only with the typical (ordered list) representation. Labels may obscure or detract from fundamental problems with topic coherence, and better labels don't make bad topics good.

The expert-driven annotation study in section 3 suggests that three of the four types of poor-quality topics ("chained," "intruded" and "random") could be detected using a metric based on the co-

occurrence of words within the documents being modeled. For "chained" and "intruded" topics, it is likely that although pairs of words belonging to a single concept will co-occur in a single document (e.g., "nucleic" and "acids" in documents about DNA), pairs belonging to different concepts (e.g., "fatty" and "nucleic") will not. For random topics, it is likely that few words will co-occur at all.

This insight can be used to design a new metric for assessing topic quality. Letting $D(v)$ be the *document frequency* of word v (i.e., the number of documents with least one token of type v) and $D(v, v')$ be *co-document frequency* of words v and v' (i.e., the number of documents containing one or more tokens of type v and at least one token of type v'), we define *topic coherence* as

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)})}{D(v_l^{(t)})}, \tag{5}$$

where $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ is the list of the M most probable words in topic t. This metric, which relies upon word co-occurrence statistics gathered from the corpus being modeled rather than an external reference corpus, is both domain-specific, and does not require additional reference data.

Equation 5 is very similar to pointwise mutual information (PMI). PMI has a long history in language technology [5], and was recently used by Newman et al. [10] to evaluate topic models. An important difference between our coherence metric and Newman et al.'s approach is that we do not compute a weighted average of the log values—we simply add them. We tried weighting the terms by their corresponding topic–word probabilities and and by their position in the sorted list of the M most probable words, but we found that a uniform weighting resulted in a better predictor of topic quality.

In order to provide intuition for the behavior of our topic coherence metric, table 1 shows three example topics and their topic coherence scores. The first topic, related to grant-funded training programs, is one of the best-scoring topics. All pairs of words have high co-document frequencies. The second topic, on neurons, is more typical of quality "research" topics. Overall, these words occur less frequently, but generally occur moderately interchangeably: there is little structure to their covariance. The last topic is one of the lowest-scoring topics. Its co-document frequency matrix is shown in table 2. The top two words are very closely related: 487 documents include "aging" at least once, 122 include " lifespan", and 55 include both. Meanwhile, the third word " globin" co-occurs with only one of the top seven words—the very common word "human".

Figure 1 shows the association between the expert annotations and both topic size (top) and our coherence metric (bottom). By itself, topic size is a good predictor of topic quality. To further investigate this relationship, we performed a logistic regression analysis on the binary variable "is this topic bad" given topic size. This analysis finds a coefficient of -3.98×10^{-5}, or a change in log-odds ratio of being "bad" of roughly -0.4 for each additional 10000 tokens. This coefficient is significant at least at the $p = 0.001$ level, and gives an AIC value of 142.16. When we include topic coherence in the analysis, however, the coefficient for topic coherence is highly significant and the coefficient for topic size drops to 5.9×10^{-7}—essentially zero. The AIC value improves to 116.48.

The topic coherence metric is also very good qualitatively: of the 20 best scoring topics, 18 are labeled as "good," one is "intermediate" ("unbalanced"), and one is "bad" (combining "cortex" and "fmri", words that commonly co-ocur, but are conceptually distinct). Of the 20 worst scoring topics, 15 are "bad," 4 are "intermediate," and one (with the second from highest coherence score) is "good."

Table 1: Example topics with different coherence scores. (Numbers closer to zero represent higher coherence.) For the bottom topic, the words highlighted in bold belong to a concept different to that of the other words.

-167.1	students, program, summer, biomedical, training, experience, undergraduate, career, minority, student, careers, underrepresented, medical_students, week, science
-252.1	neurons, neuronal, brain, axon, neuron, guidance, nervous_system, cns, axons, neural, axonal, cortical, survival, disorders, motor
-357.2	aging, lifespan, **globin**, age_related, longevity, human, age, **erythroid**, **sickle_cell**, **beta_globin**, **hb, senescence, adult, older, lcr**

Table 2: Co-document frequency matrix for the top words in a low-quality topic (according to the topic coherence metric). The diagonal shows the overall document frequency for each word w. The column on the right is $N_{w|t}$. Note that "Globin" and "erythroid" do not co-occur with any of the aging-related words.

aging	487	53	**0**	65	42	**0**	51	**0**	138	**0**	914
lifespan	53	122	**0**	15	28	**0**	15	**0**	44	**0**	205
globin	**0**	**0**	39	**0**	**0**	19	**0**	15	27	3	200
age_related	65	15	**0**	119	12	**0**	25	**0**	37	**0**	160
longevity	42	28	**0**	12	73	**0**	6	**0**	20	1	159
erythroid	**0**	**0**	19	**0**	**0**	69	**0**	8	23	1	110
age	51	15	**0**	25	6	**0**	245	1	82	**0**	103
sickle_cell	**0**	**0**	15	**0**	**0**	8	1	43	16	2	93
human	138	44	27	37	20	23	82	16	4347	157	91
hb	**0**	**0**	3	**0**	1	1	**0**	2	5	15	73

5 Generalized Pólya Urn Models

Although the topic coherence metric defined in section 4.2 provides an accurate way of assessing the quality of inferred topics, *preventing* poor quality topics from occurring in the first place is clearly preferable. One way of doing this is to incorporate the corpus-specific word co-occurrence information used in our coherence metric directly into the statistical topic modeling framework.

In this section, we describe a new topic model that incorporates salient properties of our coherence metric, ensuring that the occurrence of word w in topic t increases not only the probability of seeing that word again, but also the probability of seeing other related words (according to co-document frequencies for the corpus). The new model retains the document–topic component of LDA, but replaces the topic–word component with a *generalized* Pólya urn framework [8]. This replacement is best explained in terms of the conditional posterior distribution over words for topic t.

In LDA, the conditional posterior distribution over words for topic t is a simple Pólya urn model, characterized by (3). Under this urn interpretation, the process of drawing a token from the conditional posterior (and incrementing the counts accordingly) is equivalent to imagining a topic-specific urn consisting of N_t balls of W different colors, drawing a ball from the urn uniformly at random, noting its color, and returning the ball to the urn along with an additional ball of the same color (fractional balls representing the prior proportion $\frac{\beta}{W}$ of each color also present but are not essential to the description). If, having drawn a ball of color v, A_{vw} additional balls of each color $w \in 1, \ldots, W$ are returned to the urn, then the resultant model is a *generalized* Pólya urn. Given $\mathcal{W}$ and $\mathcal{Z}$, the conditional posterior probability of word w in topic t implied by this generalized model is

$$P(w \mid t, \mathcal{W}, \mathcal{Z}, \beta \boldsymbol{u}, \boldsymbol{A}) = \frac{\sum_v N_{v|t} A_{vw} + \frac{\beta}{W}}{N_t + \beta}, \tag{6}$$

where $\boldsymbol{A}$ is a $W \times W$ matrix, known as the *addition matrix* or *schema*. The simple Pólya urn model (and hence the conditional posterior probability of word w in topic t under LDA) can be recovered by setting the schema $\boldsymbol{A}$ to the identity matrix. It is worth noting that β and each element of $\boldsymbol{A}$ can be scaled by a constant without changing the distribution over words. For comparison with standard LDA we therefore normalize each row of $\boldsymbol{A}$ to sum to one, and set $\frac{\beta}{W}$ to 0.01 for both models.

One interesting aspect of generalized Pólya urn models is that it is possible to include negative entries in the schema $\boldsymbol{A}$, thus it is possible to metaphorically remove balls from the urn rather than adding them. Although this property could potentially be useful in representing negative correlations between words, it results in a model that is not *tenable*. A tenable urn model is one that can support any sequence of samples, of any length. Negative weights can cause the urn to "run out" of a particular color, thereby preventing the model from being able to support particular sequences.

Another implication of the generalized Pólya urn model is that it is nonexchangeable—the joint probability of the tokens in any given topic is not invariant to permutation of those tokens. Inference of $\mathcal{Z}$ given $\mathcal{W}$ via Gibbs sampling involves repeatedly cycling through the tokens in $\mathcal{W}$ and, for each one, resampling its topic assignment conditioned on $\mathcal{W}$ and the current topic assignments for all tokens other than the token of interest. For LDA, the sampling distribution for each topic assignment is given by (4)—i.e., due to exchangeability, the sampling distribution is simply the product of two predictive probabilities, obtained by treating the token of interest as if it were the last. For a topic model with a generalized Pólya urn for the topic–word component, the sampling distribution is more

Data set	D	$\bar{N}_d$	N	W
NIH	18756	114.64 ± 30.41	2150172	28702

Table 3: Data set statistics. D is the number of documents in each data set, $\bar{N}_d$ is the mean document length plus/minus one standard deviation, N is the total number of tokens, and W is the vocabulary size.

complicated. Specifically, the topic–word component of the sampling distribution is no longer a simple predictive distribution—when sampling a new value for $z_n^{(d)}$, the implication of each possible value for subsequent tokens and their topic assignments must be considered. Unfortunately, this can be very computationally expensive, particularly for large corpora. However, there are several ways around this problem. The first is to use sequential Monte Carlo methods, which have been successfully applied to topic models [3]. In these methods, multiple "particles" make exactly one pass through the data, sampling topic assignments in a left-to-right fashion (i.e., considering only the assignments of previous tokens) and occasionally resampling a small window of previously-sampled topic assignments. The second approach is to approximate the true Gibbs sampling distribution by simply treating each token as if it were the last. While this approximate method does not share the same theoretical guarantees as a sequential Monte Carlo method or the true Gibbs sampling algorithm, it does yield topics that perform well under our empirical evaluation metrics.

5.1 Experimental Results

For the values of the schema $\boldsymbol{A}$, we set each row to be proportional to the co-document frequencies used in our coherence metric, multiplied by word-specific scaling parameter, such that $\boldsymbol{A}_{vw} \propto \lambda_w D(w, v)$. In practice, we found that setting λ_w to the inverse document frequency of w improved performance, as did removing off-diagonals for rows corresponding to words with high document frequency (e.g., $> \frac{1}{3}$). Including nonzero off-diagonal values in $\boldsymbol{A}$ for very frequent words causes the model to disperse those words over many topics, which leads to large numbers of extremely similar topics. To measure this effect, we calculated the Jensen-Shannon divergence between all pairs of topic–word distributions in a given model. For a model using off-diagonal weights for all words, the mean of the 100 lowest divergences was $0.29 \pm .05$ (a divergence of 1.0 represents distributions with no shared support) at $T = 200$. The average divergence of the 100 most similar pairs of topics for LDA (i.e., $\boldsymbol{A} = \boldsymbol{I}$) is $0.67 \pm .05$. The same statistic for the generalized Pólya urn model without off-diagonal elements for words with high document frequency is 0.822 ± 0.09.

Storing only the diagonal elements of the schema $\boldsymbol{A}$ for the most common words also has the fortunate effect of substantially reducing preprocessing time, which we find is roughly proportional to three or four iterations of Gibbs sampling. Although we have not made any strong effort to optimize our Gibbs sampling code, we find that inference for the generalized Pólya model takes roughly two to three times longer than for standard LDA, although this varies somewhat with the sparsity of the schema due to additional bookkeeping needed before and after sampling topic assignments.

We evaluate the model on a corpus of NIH grant abstracts. Details are given in Table 3. Figure 2 shows the performance of the generalized Pólya urn model relative to LDA. Two metrics—our new topic coherence metric and the log probability of held-out documents—are shown over 1000 iterations at 50 iteration intervals. Each model was run over five folds of cross validation, each with three random initializations. For each model we calculated an overall coherence score by calculating the topic coherence for each topic individually and then averaging these values. We report the average over all 15 models in each plot. Held-out probabilities were calculated using the left-to-right method of Wallach et al. [12], with each cross-validation fold using its own schema $\boldsymbol{A}$. The generalized Pólya model performs very well in overall topic coherence, reaching levels within the first 50 iterations that match the final score. This model has an early advantage for held-out probability as well, but is eventually overtaken by LDA. This trend is consistent with Chang et al.'s observation that held-out probabilities are not always good predictors of human judgments [4].

In section 4.2, we demonstrated that our topic coherence metric is a good predictor of expert opinions of topic quality for LDA. However, simply incorporating the salient characteristics of this metric into a topic model does not guarantee that the resultant topics will indeed be of higher quality, as judged by experts. We therefore repeated the expert-driven evaluation protocol described in section 3.1 using a total of 100 topics, randomly selected from $T = 200$ topics inferred by the generalized Pólya

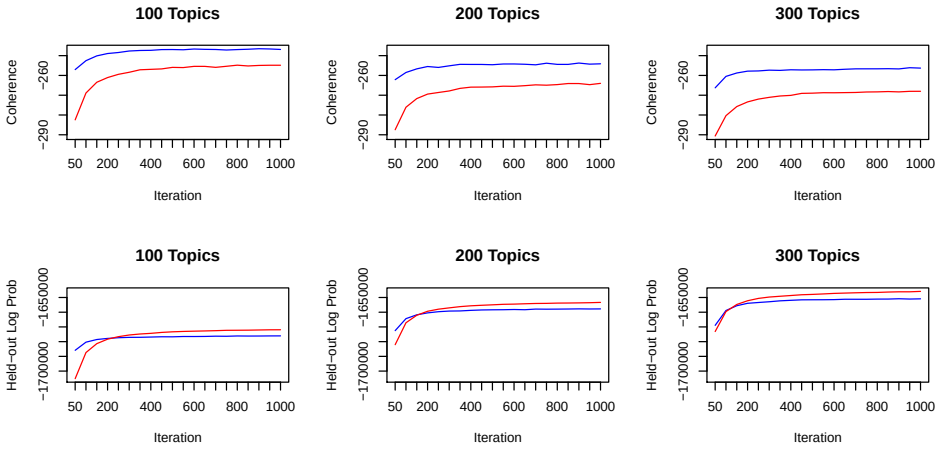

Figure 2: The top plots show topic coherence (averaged over 15 runs) over 1000 iterations of Gibbs sampling. In each case, the generalized Pólya urn model (blue) outperforms LDA (red). Error bars are not visible in this plot. The bottom plots show the log probability of held-out documents for the same models (three runs each of 5-fold cross-validation). LDA gives very slightly higher log probability than the generalized Pólya model.

urn model and LDA (50 from each). These topics were randomly shuffled and presented to the experts from NINDS, with no indication as to the identity of the model from which each topic came.

One topic from each model was labeled as being both "bad" and "unbalanced". Since our coherence metric and generalized Pólya urn model were specifically designed to reduce the other three types of "bad" topics ("chained," "intruded" and "random"), we therefore ignore these two "unbalanced" topics in our subsequent analyses. Of the remaining 49 topics from each model, 12 of the LDA topics were marked as "bad," in contrast with only 5 of the topics from the generalized Pólya urn model. These numbers are encouraging, although they make it somewhat difficult to establish the significance of the relationship between our topic coherence metric and the generalized Pólya urn model. Of the three lowest-scoring topics, two were marked as "bad." The other bad topics were closer to the middle of the range of coherence scores. In the generalized Pólya urn model, the "unbalanced" topic was the single highest-scoring topic in the model. In a logistic regression analysis, the coefficient for the coherence metric is -0.035 for LDA and -0.022 for the generalized Pólya model. Interestingly, we can use this logistic regression framework to take the topics inferred by any topic model and estimate the number of them that are "bad." Given a topic coherence score for any topic, we can estimate the probability that that topic is "bad." Summing these probabilities for all topics in a model yields the expected number of "bad" topics inferred by that model.

6 Discussion

Large-scale, institution-specific topic models can be extremely useful for identifying trends and building connections between disparate groups. However, our experience in deploying such models has indicated that the primary obstacle to their widespread adoption is the presence of semantically incoherent (poor-quality) topics. In this paper, we therefore focused on the task of building fine-grained statistical topic models with high-quality topics from highly-specialized, domain-specific document collections. We presented a new intrinsic evaluation metric that predicts expert topic annotations using only information contained in the documents being modeled. We then developed a novel topic model, based on this metric, intended to avoid the "chained" or "intruded" topics commonly inferred by LDA. This model uses word co-occurrence information from the documents being modeled to discourage inference of topics that do not represent a single concept, thereby avoiding many of the quality problems that plague existing topic models. One avenue for future work is to use domain-specific external information when available, such as controlled vocabularies and manually-curated ontologies, in addition to the corpus-specific word co-occurrence statistics.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, in part under subcontract #B582467 from Lawrence Livermore National Security, LLC, prime contractor to U.S. DOE/NNSA contract #DE-AC52-07NA27344, and in part by NIH:HHSN271200900640P. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[1] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[3] K. Canini, L. Shi, and T. Griffiths. Online inference of topics with latent Dirichlet allocation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009.

[4] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pages 288–296, 2009.

[5] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 6(1):22–29, 1990.

[6] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence 6*, pages 721–741, 1984.

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

[8] H. Mahmoud. *Pólya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science, 2008.

[9] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499, 2007.

[10] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[11] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for lat ent Dirichlet allocation. In *Advances in Neural Information Processing Systems 18*, 2006.

[12] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Interational Conference on Machine Learning*, 2009.

[13] X. Wei and B. Croft. LDA-based document models for ad-hoc retrival. In *Proceedings of the 29th Annual International SIGIR Conference*, 2006.