Optimizing Semantic Coherence in Topic Models

David Mimno
Princeton University
Princeton, NJ 08540
mimno@cs.princeton.edu

Hanna M. Wallach
University of Massachusetts, Amherst
Amherst, MA 01003
wallach@cs.umass.edu

Edmund Talley **Miriam Leenders**
National Institutes of Health
Bethesda, MD 20892
{talleye,leenderm}@ninds.nih.gov

Andrew McCallum
University of Massachusetts, Amherst
Amherst, MA 01003
mccallum@cs.umass.edu

Abstract

Latent variable models have the potential to add value to large document collections by discovering interpretable, low-dimensional subspaces. In order for people to use such models, however, they must trust them. Unfortunately, typical dimensionality reduction methods for text, such as latent Dirichlet allocation, often produce low-dimensional subspaces (topics) that are obviously flawed to human domain experts. The contributions of this paper are threefold: (1) An analysis of the ways in which topics can be flawed; (2) an automated evaluation metric for identifying such topics that does not rely on human annotators or reference collections outside the training data; (3) a novel statistical topic model based on this metric that significantly improves topic quality in a large-scale document collection from the National Institutes of Health (NIH).

1 Introduction

Statistical topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) 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 the outside machine learning 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.

In general, users prefer models with larger numbers of topics because such models have greater resolution and are able to support finer-grained distinctions. Unfortunately, we have observed that there is a strong relationship between the size of topics and the probability of topics being nonsensical as judged by domain experts: as the number of topics increases, the smallest topics (number of word tokens assigned to each topic) are almost always poor quality. The common practice of displaying only a small number of example topics hides the fact that as many as 10% of topics may be so bad that they cannot be shown without reducing users' confidence.

The evaluation of statistical topic models has traditionally been dominated by either extrinsic methods (i.e., using the inferred topics to perform some external task such as information retrieval (Wei and Croft, 2006)) or quantitative intrinsic methods, such as computing the probability of held-out documents (Wallach et al., 2009). Recent work has focused on evaluation of topics as semantically-coherent concepts. For example, Chang et al. (2009) found that the probability of held-out documents is not always a good predictor of human judgments. Newman et al. (2010) showed that an automated evaluation metric based on word co-occurrence statistics gathered from Wikipedia could predict human evaluations of topic quality. AlSumait et al. (2009) used differences between topic-specific distributions over words and the corpus-wide distribution over words to identify overly-general "vacuous" topics. Finally, Andrzejewski et al. (2009) developed semi-supervised methods that avoid specific user-labeled semantic coherence problems.

The contributions of this paper are threefold: (1) To identify distinct classes of low-quality topics, some of which are not flagged by existing evaluation methods; (2) to introduce a new topic "coherence" score that corresponds well with human coherence judgments and makes it possible to identify

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 262–272,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

specific semantic problems in topic models without human evaluations or external reference corpora; (3) to present an example of a new topic model that learns latent topics by directly optimizing a metric of topic coherence. With little additional computational cost beyond that of LDA, this model exhibits significant gains in average topic coherence score. Although the model does not result in a statistically-significant reduction in the number of topics marked "bad", the model consistently improves the topic coherence score of the ten lowest-scoring topics (i.e., results in bad topics that are "less bad" than those found using LDA) while retaining the ability to identify low-quality topics without human interaction.

2 Latent Dirichlet Allocation

LDA is a generative probabilistic model for documents $\mathcal{W} = \{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(D)}\}$. To generate a word token $w_n^{(d)}$ in document d, we draw a discrete topic assignment $z_n^{(d)}$ from a document-specific distribution over the T topics $\boldsymbol{\theta}_d$ (which is itself drawn from a Dirichlet prior with hyperparameter α), and then draw a word type for that token from the topic-specific distribution over the vocabulary $\boldsymbol{\phi}_{z_n^{(d)}}$. The inference task in topic models is generally cast as inferring the document–topic proportions $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_D\}$ and the topic-specific distributions $\{\boldsymbol{\phi}_1 \ldots, \boldsymbol{\phi}_T\}$.

The multinomial topic distributions are usually drawn from a shared symmetric Dirichlet prior with hyperparameter β, such that conditioned on $\{\boldsymbol{\phi}_t\}_{t=1}^T$ and the topic assignments $\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(D)}\}$, the word tokens are independent. In practice, however, it is common to deal directly with the "collapsed" distributions that result from integrating over the topic-specific multinomial parameters. The resulting distribution over words for a topic t is then a function of the hyperparameter β and the number of words of each type assigned to that topic, $N_{w|t}$. This distribution, known as the Dirichlet compound multinomial (DCM) or Pólya distribution (Doyle and Elkan, 2009), breaks the assumption of conditional independence between word tokens given topics, but is useful during inference because the conditional probability of a word w given topic t takes a very simple form: $P(w \mid t, \beta) = \frac{N_{w|t}+\beta}{N_t+|\mathcal{V}|\,\beta}$, where $N_t = \sum_{w'} N_{w'|t}$ and $|\mathcal{V}|$ is the vocabulary size.

The process for generating a sequence of words from such a model is known as the simple Pólya urn model (Mahmoud, 2008), in which the initial probability of word type w in topic t is proportional to β, while the probability of each subsequent occurrence of w in topic t is proportional to the number of times w has been drawn in that topic plus β. Note that this unnormalized weight for each word type depends only on the count of that word type, and is independent of the count of any other word type w'. Thus, in the DCM/Pólya distribution, drawing word type w must *decrease* the probability of seeing all other word types $w' \neq w$. In a later section, we will introduce a topic model that substitutes a *generalized* Pólya urn model for the DCM/Pólya distribution, allowing a draw of word type w to *increase* the probability of seeing certain other word types.

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 (Blei et al., 2003; Teh et al., 2006) or MCMC methods (Griffiths and Steyvers, 2004). Here, we use MCMC methods—specifically Gibbs sampling (Geman and Geman, 1984), which involves sequentially resampling each topic assignment $z_n^{(d)}$ from its conditional posterior given the documents $\mathcal{W}$, the hyperparameters α and β, and $\mathcal{Z}_{\setminus d,n}$ (the current topic assignments for all tokens other than the token at position n in document d).

3 Expert Opinions of Topic Quality

Concentrating on 300,000 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 inferring the kinds of topics desired by domain experts.[1]

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 areas funded by NINDS. Each topic

[1] All evaluated models will be released publicly.

t was presented to the experts as a list of the thirty most probable words for that topic, in descending order of their topic-specific "collapsed" probabilities, $\frac{N_{w|t}+\beta}{N_t+|\mathcal{V}|\,\beta}$. In addition to the most probable 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-occurred 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".[2] 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 clearly contained text referring to the concept inferred from the topic words. Finally, for each topic marked as being either "intermediate" or "bad", one or more of the following problems (defined by the domain experts) was identified, as appropriate:

- **Chained:** every word is connected to every other word through some pairwise word 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 versus the building blocks of DNA and RNA) chained via the word "acids".
- **Intruded:** either two or more unrelated sets of related words, joined arbitrarily, or an otherwise good topic with a few "intruder" words.
- **Random:** no clear, sensical connections between more than a few pairs of words.
- **Unbalanced:** the top words are all logically connected to each other, but the topic combines very general and specific terms (e.g., "signal transduction" versus "notch signaling").

Examples of a good general topic, a good research topic, and a chained research topic are in Table 1.

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 expert always believed the topic to be "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". (The 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 models with large numbers of specific, high-quality topics from domain-specific corpora. We therefore explore the extent to which information already contained in the documents being modeled can be used to assess topic quality.

In this section we evaluate several methods for ranking the quality of topics and compare these rankings to human annotations. No method is likely to perfectly predict human judgments, as individual annotators may disagree on particular topics. For an application involving removing low quality topics we recommend using a weighted combination of metrics, with a threshold determined by users.

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 assigned to that topic) and expert annotations ("good", "intermediate", "bad") for the 148 topics manually labeled by annotators as described above. This figure suggests that topic size is a reasonable predic-

[2] Equivalent to "vacuous topics" of AlSumait et al. (2009).

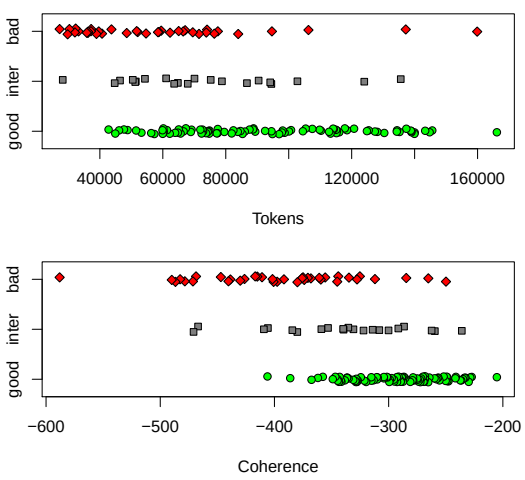

Figure 1: **Topic size is a good indicator of quality; the new *coherence* metric is better**. Top shows expert-rated topics ranked by topic size (AP 0.89, AUC 0.79), bottom shows same topics ranked by coherence (AP 0.94, AUC 0.87). Random jitter is added to the y-axis for clarity.

tor 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 relatively 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 probable words for that topic, in descending order of their topic-specific "collapsed" probabilities. Although there has been previous work on automated generation of labels or headings for topics (Mei et al., 2007), we choose to work only with the 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 described 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 within a single document (e.g., "nucleic" and "acids" in documents about DNA), word pairs belonging to different concepts (e.g., "fatty" and "nucleic") will not. For random topics, it is likely that few words will co-occur.

This insight can be used to design a new metric for assessing topic quality. Letting $D(v)$ be the *document frequency* of word type v (i.e., the number of documents with least one token of type v) and $D(v, v')$ be *co-document frequency* of word types 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

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

where $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ is a list of the M most probable words in topic t. A smoothing count of 1 is included to avoid taking the logarithm of zero.

Figure 1 shows the association between the expert annotations and both topic size (top) and our coherence metric (bottom). We evaluate these results using standard ranking metrics, average precision and the area under the ROC curve. Treating "good" topics as positive and "intermediate" or "bad" topics as negative, we get average precision values of 0.89 for topic size vs. 0.94 for coherence and AUC 0.79 for topic size vs. 0.87 for coherence. We performed a logistic regression analysis on the binary variable "is this topic bad". Using topic size alone as a predictor gives AIC (a measure of model fit) 152.5. Coherence alone has AIC 113.8 (substantially better). Both predictors combined have AIC 115.8: the simpler coherence alone model provides the best performance. We tried weighting the terms in equation 1 by their corresponding topic–word probabilities and and by their position in the sorted list of the M most probable words for that topic, but we found that a uniform weighting better predicted topic quality.

Our topic coherence metric also exhibits good qualitative behavior: 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-occur, but are conceptually distinct). Of the 20 worst scoring topics, 15 are "bad," 4 are "intermediate," and only one (with the 19th worst coherence score) is "good."

Our coherence metric relies only upon word co-occurrence statistics gathered from the corpus being modeled, and does not depend on an external reference corpus. Ideally, all such co-occurrence information would already be accounted for in the model. We believe that one of the main contributions of our work is demonstrating that standard topic models *do not* fully utilize available co-occurrence information, and that a held-out reference corpus is therefore not required for purposes of topic evaluation.

Equation 1 is very similar to pointwise mutual information (PMI), but is more closely associated with our expert annotations than PMI (which achieves AUC 0.64 and AIC 170.51). PMI has a long history in language technology (Church and Hanks, 1990), and was recently used by Newman et al. (2010) to evaluate topic models. When expressed in terms of count variables as in equation 1, PMI includes an additional term for $D(v_m^{(t)})$. The improved performance of our metric over PMI implies that what matters is *not* the difference between the joint probability of words m and l and the product of marginals, but the *conditional* probability of each word given the each of the higher-ranked words in the topic.

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 closely related: 487 documents include "aging" at least once, 122 include "lifespan", and 55 include both. Meanwhile, the third word "globin" occurs with only one of the top seven words—the common word "human".

4.3 Comparison to word intrusion

As an additional check for both our expert annotations and our automated metric, we replicated the "word intrusion" evaluation originally introduced by Chang et al. (2009). In this task, one of the top ten most probable words in a topic is replaced with a

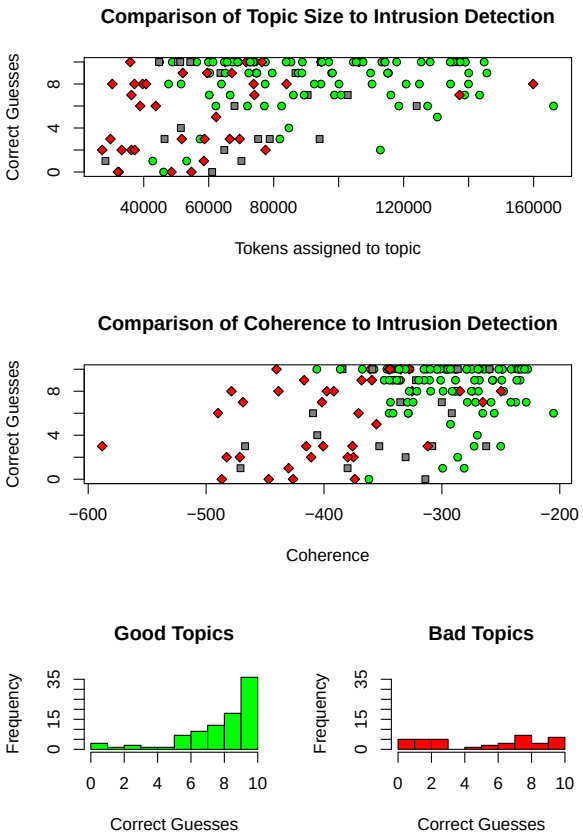

Figure 2: Top: results of the intruder selection task relative to two topic quality metrics. Bottom: marginal intruder accuracy frequencies of good and bad topics.

another word, selected at random from the corpus. The resulting set of words is presented, in a random order, to users, who are asked to identify the "intruder" word. It is very unlikely that a randomly-chosen word will be semantically related to any of the original words in the topic, so if a topic is a high quality representation of a semantically coherent concept, it should be easy for users to select the intruder word. If the topic is not coherent, there may be words in the topic that are also not semantically related to any other word, thus causing users to select "correct" words instead of the real intruder.

We recruited ten additional expert annotators from NINDS, not including our original annotators, and presented them with the intruder selection task, using the set of previously evaluated topics. Results are shown in figure 2. In the first two plots, the x-axis is one of our two automated quality met-

Table 1: Example topics (good/general, good/research, chained/research) with different coherence scores (numbers closer to zero indicate higher coherence). The chained topic combines words related to aging (indicated in plain text) and words describing blood and blood-related diseases (bold). The only connection is the common word *human*.

-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 our coherence metric), shaded to highlight zeros. The diagonal (light gray) 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

rics (topic size and coherence) and the y-axis is the number of annotators that correctly identified the true intruder word (accuracy). The histograms below these plots show the number of topics with each level of annotator accuracy for good and bad topics. For good topics (green circles), the annotators were generally able to detect the intruder word with high accuracy. Bad topics (red diamonds) had more uniform accuracies. These results suggest that topics with low intruder detection accuracy tend to be bad, but some bad topics can have a high accuracy. For example, spotting an intruder word in a chained topic can be easy. The low-quality topic *receptors, cannabinoid, cannabinoids, ligands, cannabis, endocannabinoid, cxcr4, [virus], receptor, sdf1*, is a typical "chained" topic, with *CXCR4* linked to *cannabinoids* only through *receptors*, and otherwise unrelated. Eight out of ten annotators correctly identified "virus" as the correct intruder. Repeating the logistic regression experiment using intruder detection accuracy as input, the AIC value is 163.18—much worse than either topic size or coherence.

5 Generalized Pólya Urn Models

Although the topic coherence metric defined above provides an accurate way of assessing topic quality, *preventing* poor quality topics from occurring in the first place is preferable. Our results in the previous section show that we can identify low-quality topics without making use of external supervision; the training data by itself contains sufficient information at least to reject poor combinations of words.

In this section, we describe a new topic model that incorporates the corpus-specific word co-occurrence information used in our coherence metric directly into the statistical topic modeling framework. It is important to note that simply disallowing words that never co-occur from being assigned to the same topic is not sufficient. Due to the power-law characteristics of language, most words are rare and will not co-occur with most other words regardless of their semantic similarity. It is rather the *degree* to which the most prominent words in a topic do not co-occur with the other most prominent words in that topic that is an indicator of topic incoherence. We therefore desire models that guide topics towards semantic similarity without imposing hard constraints.

As an example of such a model, we present a new topic model in which the occurrence of word type w in topic t increases not only the probability of seeing that word type again, but also increases the probability of seeing other related words (as determined by co-document frequencies for the corpus being modeled). This new topic model retains the document–topic component of standard LDA, but replaces the usual Pólya urn topic–word component with a *generalized* Pólya urn framework (Mahmoud, 2008).

A sequence of i.i.d. samples from a discrete distribution can be imagined as arising by repeatedly drawing a random ball from an urn, where the number of balls of each color is proportional to the probability of that color, replacing the selected ball after each draw. In a Pólya urn, each ball is replaced *along with another ball of the same color*. Samples from this model exhibit the "burstiness" property: the probability of drawing a ball of color w increases each time a ball of that color is drawn. This process represents the marginal distribution of a hierarchical model with a Dirichlet prior and a multinomial likelihood, and is used as the distribution over words for each topic in almost all previous topic models. In a *generalized* Pólya urn model, having drawn a ball of color w, A_{vw} additional balls of each color $v \in \{1, \ldots, W\}$ are returned to the urn. Given $\mathcal{W}$ and $\mathcal{Z}$, the conditional posterior probability of word w in topic t implied by this generalized model is

$$P(w \mid t, \mathcal{W}, \mathcal{Z}, \beta, \boldsymbol{A}) = \frac{\sum_v N_{v|t}\, A_{vw} + \beta}{N_t + |\mathcal{V}|\beta}, \quad (2)$$

where $\boldsymbol{A}$ is a $W \times W$ real-valued 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. Unlike the simple Pólya distribution, we do not know of a representation of the generalized Pólya urn distribution that can be expressed using a concise set of conditional independence assumptions. A standard graphical model with plate notation would therefore not be helpful in highlighting the differences between the two models, and is not shown.

Algorithm 1 shows pseudocode for a single Gibbs sweep over the latent variables $\mathcal{Z}$ in standard LDA. Algorithm 2 shows the modifications necessary to

for $d \in \mathcal{D}$ **do**
  for $w_n \in \boldsymbol{w}^{(d)}$ **do**
    $N_{z_i|d_i} \leftarrow N_{z_i|d_i} - 1$
    $N_{w_i|z_i} \leftarrow N_{w_i|z_i} - 1$
    `sample` $z_i \propto (N_{z|d_i} + \alpha_z) \frac{N_{w_i|z} + \beta}{\sum_{z'} (N_{w_i|z'} + \beta)}$
    $N_{z_i|d_i} \leftarrow N_{z_i|d_i} + 1$
    $N_{w_i|z_i} \leftarrow N_{w_i|z_i} + 1$
  end for
end for

Algorithm 1: One sweep of LDA Gibbs sampling.

for $d \in \mathcal{D}$ **do**
  for $w_n \in \boldsymbol{w}^{(d)}$ **do**
    $N_{z_i|d_i} \leftarrow N_{z_i|d_i} - 1$
    for all v **do**
      $N_{v|z_i} \leftarrow N_{v|z_i} - A_{vw_i}$
    end for
    `sample` $z_i \propto (N_{z|d_i} + \alpha_z) \frac{N_{w_i|z} + \beta}{\sum_{z'} (N_{w_i|z'} + \beta)}$
    $N_{z_i|d_i} \leftarrow N_{z_i|d_i} + 1$
    for all v **do**
      $N_{v|z_i} \leftarrow N_{v|z_i} + A_{vw_i}$
    end for
  end for
end for

Algorithm 2: One sweep of gen. Pólya Gibbs sampling, with differences from LDA highlighted in red.

support a generalized Pólya urn model: rather than subtracting exactly one from the count of the word given the old topic, sampling, and then adding one to the count of the word given the new topic, we subtract a column of the schema matrix from the entire count vector over words for the old topic, sample, and add the same column to the count vector for the new topic. As long as A is sparse, this operation adds only a constant factor to the computation.

Another property 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 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 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. There are several ways around this problem. The first is to use sequential Monte Carlo methods, which have been successfully applied to topic models previously (Canini et al., 2009). The second approach is to approximate the true Gibbs sampling distribution by treating each token as if it were the last, ignoring implications for subsequent tokens and their topic assignments. We find that this approximate method performs well empirically.

5.1 Setting the Schema A

Inspired by our evaluation metric, we define A as

$$\begin{aligned} \boldsymbol{A}_{vv} &\propto \lambda_v D(v) \\ \boldsymbol{A}_{vw} &\propto \lambda_v D(w, v) \end{aligned} \tag{3}$$

where each element is scaled by a row-specific weight λ_v and each column is normalized to sum to 1. Normalizing columns makes comparison to standard LDA simpler, because the relative effect of smoothing parameter $\beta = 0.01$ is equivalent. We set $\lambda_v = \log(D \,/\, D(v))$, the standard IDF weight used in information retrieval, which is larger for less frequent words. The column for word type w can be interpreted as word types with significant association with w. The IDF weighting therefore has the effect of increasing the strength of association for rare word types. We also found empirically that it is helpful to remove off-diagonal elements for the most common types, such as those that occur in more than 5% of documents (IDF < 3.0). Including nonzero off-diagonal values in $\boldsymbol{A}$ for very frequent types causes the model to disperse those types 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 word

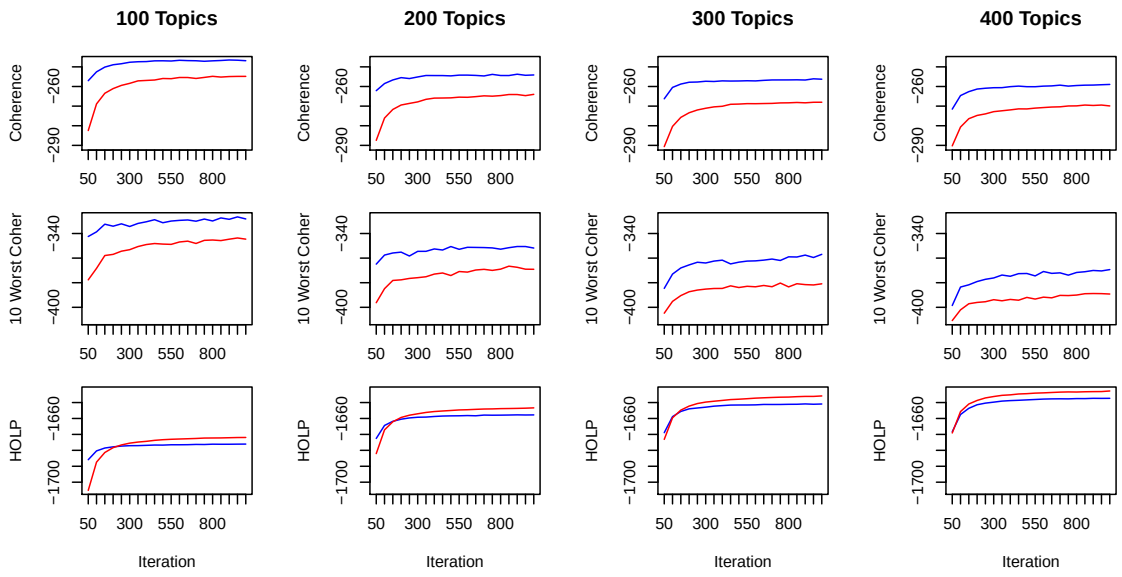

Figure 3: **Pólya urn topics (blue) have higher average coherence and converge much faster than LDA topics (red)**. The top plots show topic coherence (averaged over 15 runs) over 1000 iterations of Gibbs sampling. Error bars are not visible in this plot. The middle plot shows the average coherence of the 10 lowest scoring topics. The bottom plots show held-out log probability (in thousands) for the same models (three runs each of 5-fold cross-validation).

Name	Docs	Avg. Tok.	Tokens	Vocab
NIH	18756	114.64 ± 30.41	2150172	28702

Table 3: Data set statistics.

types, 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 standard 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 word types with high document frequency is 0.822 ± 0.09.

Setting the off-diagonal elements of the schema $\boldsymbol{A}$ to zero for the most common word types also has the fortunate effect of substantially reducing preprocessing time. We find that Gibbs sampling for the generalized Pólya model takes roughly two to three times longer than for standard LDA, depending on the sparsity of the schema, due to additional bookkeeping needed before and after sampling topics.

5.2 Experimental Results

We evaluated the new model on a corpus of NIH grant abstracts. Details are given in table 3. Figure 3 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. (2009), with each cross-validation fold using its own schema $\boldsymbol{A}$. The generalized Pólya model performs very well in average 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 (Chang et al., 2009). Results are consistent over $T \in \{100, 200, 300\}$.

In section 4.2, we demonstrated that our topic coherence metric correlates with expert opinions of topic quality for standard LDA. The generalized

Pólya urn model was therefore designed with the goal of directly optimizing that metric. It is possible, however, that optimizing for coherence directly could break the association between coherence metric and topic quality. We therefore repeated the expert-driven evaluation protocol described in section 3.1. We trained one standard LDA model and one generalized Pólya urn model, each with $T = 200$, and randomly shuffled the 400 resulting topics. The topics were then presented to the experts from NINDS, with no indication as to the identity of the model from which each topic came. As these evaluations are time consuming, the experts evaluated the only the first 200 topics, which consisted of 103 generalized Pólya urn topics and 97 LDA topics. AUC values predicting bad topics given coherence were 0.83 and 0.80, respectively. Coherence effectively predicts topic quality in both models.

Although we were able to improve the average overall quality of topics and the average quality of the ten lowest-scoring topics, we found that the generalized Pólya urn model was less successful reducing the overall number of bad topics. Ignoring one "unbalanced" topic from each model, 16.5% of the LDA topics and 13.5% from the generalized Pólya urn model were marked as "bad." While this result is an improvement, it is not significant at $p = 0.05$.

6 Discussion

We have demonstrated the following:

- There is a class of low-quality topics that cannot be detected using existing word-intrusion tests, but that can be identified reliably using a metric based on word co-occurrence statistics.
- It is possible to improve the coherence score of topics, both overall and for the ten worst, while retaining the ability to flag bad topics, all without requiring semi-supervised data or additional reference corpora. Although additional information may be useful, it is not necessary.
- Such models achieve better performance with substantially fewer Gibbs iterations than LDA.

We believe that the most important challenges in future topic modeling research are improving the semantic quality of topics, particularly at the low end, and scaling to ever-larger data sets while ensuring high-quality topics. Our results provide critical insight into these problems. We found that it should be possible to construct unsupervised topic models that do not produce bad topics. We also found that Gibbs sampling mixes faster for models that use word co-occurrence information, suggesting that such methods may also be useful in guiding online stochastic variational inference (Hoffman et al., 2010).

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the CIA, the NSA and the NSF under NSF grant # IIS-0326249, in part by NIH:HHSN271200900640P, and in part by NSF # number SBE-0965436. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

Loulwah AlSumait, Daniel Barbara, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *ECML*.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January.

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

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pages 288–296.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 6(1):22–29.

Gabriel Doyle and Charles Elkan. 2009. Accounting for burstiness in topic models. In *ICML*.

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

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5228–5235.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In *NIPS*.

Hosan Mahmoud. 2008. *Pólya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Yee Whye Teh, Dave Newman, and Max Welling. 2006. A collapsed variational Bayesian inference algorithm for lat ent Dirichlet allocation. In *Advances in Neural Information Processing Systems 18*.

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Interational Conference on Machine Learning*.

Xing Wei and Bruce Croft. 2006. LDA-based document models for ad-hoc retrival. In *Proceedings of the 29th Annual International SIGIR Conference*.