Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment

Jason Chuang JCCHUANG@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

Sonal Gupta SONAL@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

Christopher D. Manning MANNING@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

Jeffrey Heer JHEER@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

Abstract

The use of topic models to analyze domain-specific texts often requires manual validation of the latent topics to ensure that they are meaningful. We introduce a framework to support such a large-scale assessment of topical relevance. We measure the correspondence between a set of latent topics and a set of reference concepts to quantify four types of topical misalignment: *junk*, *fused*, *missing*, and *repeated* topics. Our analysis compares 10,000 topic model variants to 200 expert-provided domain concepts, and demonstrates how our framework can inform choices of model parameters, inference algorithms, and intrinsic measures of topical quality.

1. Introduction

Data analysts often apply probabilistic topic models to analyze document collections too large for any one person to read. In many real-word applications, latent topics need to be verified by experts to ensure they are semantically meaningful within the domain of analysis (Talley et al., 2011; Hall et al., 2008). Human-in-the-loop supervision may involve inspecting individual latent topics, comparing multiple models, or re-training using different parameter settings. As a result, manual validation can dominate the time and cost of building high-quality topic models.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

Intrinsic evaluation, based on statistical (Blei et al., 2003) or coherence (Newman et al., 2010b) measures, can be problematic in these contexts because these measures do not account for domain relevance. We also currently lack tools that provide diagnostic feedback on how latent topics differ from users' organization of domain concepts during the construction of a topic model. Analysts often resort to spot-checking topics in an ad hoc manner after the model is created.

In response, we introduce a framework to support **large-scale assessment of topical relevance**. We first quantify the **topical alignment** between a set of latent topics and a set of reference concepts. We say a topic *resolves* to a concept if a one-to-one correspondence exists between the two, and recognize four types of misalignment: when models produce *junk* or *fused* topics or when reference concepts are *missing* or *repeated* among the latent topics.

We then introduce a process to automate the calculation of topical alignment, so that analysts can compare any number of models to known domain concepts and examine the deviations. Taking a human-centered approach, we estimate the likelihood of topic-concept pairs being considered equivalent by human judges. Using 1,000 ratings collected on Amazon Mechanical Turk, we find that a *rescaled dot product* outperforms KL-divergence, cosine, and rank-based measures in predicting user-identified topic matches. We estimate and remove topical correspondences that can be attributed to random chance via a generative probabilistic process. Finally, we visualize the results in a *correspondence chart* (Figure 1) to provide detailed diagnostic information.

Our framework is sufficiently general to support the comparison of latent topics to any type of reference concepts, including model-to-model comparisons by treating one model's outputs as the reference. For this work, we demonstrate our approach using expert-generated concepts. We asked ten experienced researchers in information visualization to exhaustively identify domain concepts in their field, and compiled a set of high-quality references.

We show that, in addition to supporting model evaluation, our framework can also provide insights into the following aspects of topic modeling research.

Model Exploration. We construct latent Dirichlet allocation (LDA) models (Blei et al., 2003) using over 10,000 parameter and hyperparameter settings, and compare the resulting 569,000 latent topics to the expert concepts. We observe that a small change in term smoothing prior can significantly alter the ratio of resolved and fused topics. In many cases, increasing the number of latent topics leads to more junk and fused topics with a corresponding reduction in resolved topics. About 10% of the concepts in our dataset are only uncovered within a narrow range of settings.

Evaluation of Inference Algorithms. We examine the effectiveness of parameter optimization and semi-supervised learning. We find that hyperparameter optimization (Wallach et al., 2009a) is generally effective for LDA. Author-topic models (Rosen-Zvi et al., 2004) achieve lower coverage than optimized LDA but favor resolved over fused topics. Partially labeled LDA models (Ramage et al., 2011) also achieve lower coverage but uncover a subset of concepts not resolved by LDA.

Evaluation of Intrinsic Measures. Automatic evaluation is desirable when reference concepts are not available. We examine the ability of ten intrinsic measures (Newman et al., 2010a; Alsumait et al., 2009; Mimno et al., 2011) to identify topical misalignments. We find little correlation between these measures and topics that are identified as junk. While some measures can distinguish junk topics comprised of function words, they are unable to separate junk topics comprised of incoherent content words from useful topics.

In sum, we provide a new visualization tool and techniques for effective human-in-the-loop construction, diagnosis, and repair of domain-relevant topic models.

2. Related Work

Latent Dirichlet allocation (Blei et al., 2003) and variants (Blei et al., 2004; Blei & Lafferty, 2006; Ramage et al., 2009; Steyvers et al., 2004; Wang & McCallum, 2006) have been applied in numerous domains (Talley et al., 2011; Newman et al., 2006; Ramage et al., 2010). While topic models can improve the performance of task-based systems (Wei & Croft, 2006; Titov & McDonald, 2008), they are most frequently used in exploratory text mining and typically evaluated based on statistical measures such as perplexity (Stevens et al., 2012) or held-out likelihood (Wallach et al., 2009b). Such measures, however, do not always correlate with human judgment of topical quality (Budiu et al., 2007) nor capture concepts that people consider to be relevant and interpretable (Chang et al., 2009). Chuang et al. (2012b) emphasize the importance of interpretation and trust in model-driven data analysis.

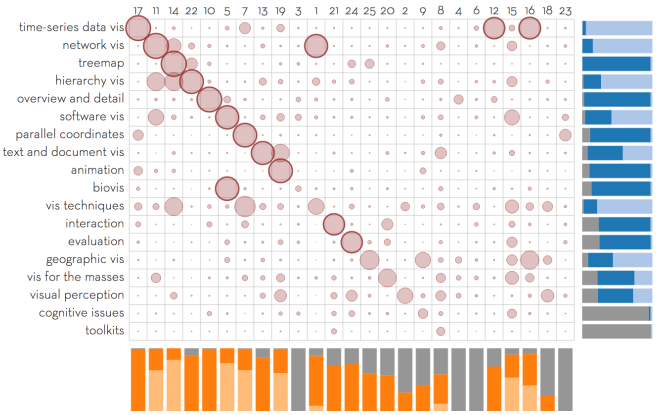


Figure 1. Correspondence chart between latent topics and reference concepts. Area of circles represents the matching likelihoods; likelihoods exceeding random chance are marked with a bold border. Bars on the right show the probability that a concept is *missing* (grey), *resolved* (blue), or *repeated* (light blue). Bars on the bottom indicate whether a topic is *junk* (grey), *resolved* (orange), or *fused* (light orange). This visual analysis tool is available online at: `http://vis.stanford.edu/topic-diagnostics`

More recently, Chang et al. (2009) introduced human validation of topical coherence via intrusion tests, but the process requires manual verification of every model built. Automatic measures of topical coherence have been proposed using word co-occurrence within the corpus (Mimno et al., 2011), Wikipedia articles or Google search results (Newman et al., 2010a), or WordNet (Musat et al., 2011). Alsumait et al. (2009) introduced heuristic measures of topical significance.

Researchers have also explored the use of visualizations for interactive inspections of topic models. The Topic Browser (Chaney & Blei, 2012) and Termite (Chuang et al., 2012a) focus on the exploration of a single topic model while TopicNets (Gretarsson et al., 2012) allows users to adaptively generate new models. However, none have the ability to measure deviations from user-defined reference concepts, nor provide diagnostic information on *why* models may be under-performing.

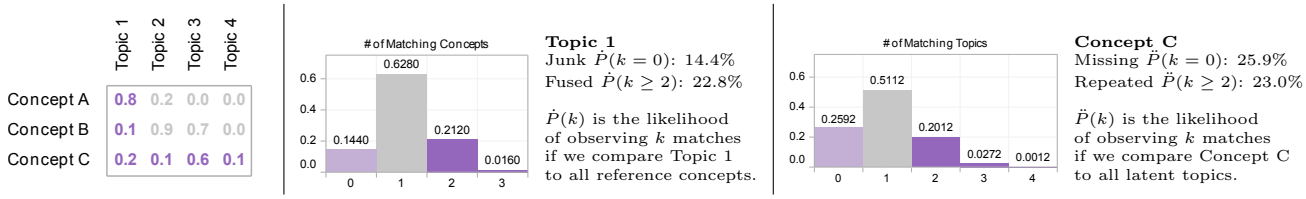


Figure 2. Correspondence Chart Construction. In a correspondence chart, each entry $p_{s,t}$ represents the probability that a user considers the word distributions associated with concept s and topic t as equivalent. Misalignment scores measure how much topical alignment deviates from an optimal one-to-one correspondence. Comparing a topic to all concepts, *junk* and *fused* scores measure how likely the topic matches exactly zero, or more than one reference concept. *Missing* and *repeated* scores measure how likely a concept matches exactly zero, or more than one latent topic.

3. Topical Alignment

We present our method for aligning latent topics with reference concepts, where each topic or concept is a multinomial distribution over words. At the heart of our method is the calculation of *matching likelihoods* for topic-concept pairs: the probability that a human judge will consider a latent topic and a reference concept to be equivalent. Based on human-subjects data, we examine how well various similarity measures predict topic matches and describe how we transform similarity scores into matching likelihoods. To improve robustness when making a large number of comparisons, we introduce a method to account for correspondences that occur due to random chance. We also introduce the *correspondence chart* which visualizes the alignment between latent topics and reference concepts.

3.1. Correspondence Chart and Misalignments

The correspondence chart is an $n \times m$ matrix of all possible pairings among n reference concepts and m latent topics. We treat each entry $p_{s,t}$ as an independent Bernoulli random variable representing the matching likelihood that a user examining the word distributions associated with concept s and topic t would respond that the two are equivalent.

We consider a correspondence optimal when every latent topic maps one-to-one to a reference concept. Deviations from an optimal arrangement lead to four types of misalignment, as shown in Figure 2. We treat entries $\{p_{i,t}\}_{i=1}^{n}$ corresponding to topic t as a Bernoulli-like process: a series of independent events that can take on different probabilities. In this framework, $\dot{P}_t(k)$ is the likelihood that a user responds with exactly k matches after comparing topic t to all n reference concepts. Similarly, $\ddot{P}_s(k)$ is the likelihood of observing exactly k positive outcomes after comparing concept s to all m latent topics. The **junk** score for topic t is the probability $\dot{P}_t(0)$; the topic has no matching concept. The **fused** score for topic t is the likelihood $\sum_{k=2}^{m} \dot{P}_t(k)$; the topic matches two or more concepts. Similarly, the **missing** score for concept s is $\ddot{P}_s(0)$, and the **repeated** score is $\sum_{k=2}^{n} \ddot{P}_s(k)$.

3.2. Human Judgment of Topic Matches

We conducted a study to acquire data on when topics (probability distributions over terms) are considered matching by people. We trained two LDA topic models on a corpus of information visualization publications and sampled pairs of topics, one from each model. The texts were chosen to be consistent with the corpus of the expert-generated concepts that we collected (details in §4). Preliminary analysis suggested that the corpus contained about 28 domain concepts, and thus we trained the two models with 40 and 50 latent topics using priors $\alpha = 0.01$ and $\beta = 0.01$.

We presented study subjects with topical pairs, one at a time in a webpage. Each topic was displayed as a list of words, sorted by frequency, where the height of each word was scaled proportional to its frequency in the topic's distribution. We asked the subjects whether the two topics match ("*represent the same meaningful concept*"), partially match, or do not match ("*represent different concepts or meaningless concepts*"). We conducted our study using Amazon Mechanical Turk. We included five topical pairs in each task, posted 200 tasks with a US$0.25 reward per task in December 2012, and received 1,000 ratings for 167 topical pairs.

3.3. Evaluating Topical Similarity Measures

We evaluated how well similarity measures predict human judgment in terms of precision and recall. For each topical pair, we assign it a rating of $\{1, 0.5, 0\}$ for each {match, partial match, no match} response, and consider a pair as matching if it has an average rating above 0.5. We computed the similarity between topics using four measures. Cosine, Spearman rank coefficient, and KL-divergence are three common similarity measures. We also introduce a *rescaled dot product* to improve upon cosine (Table 1).

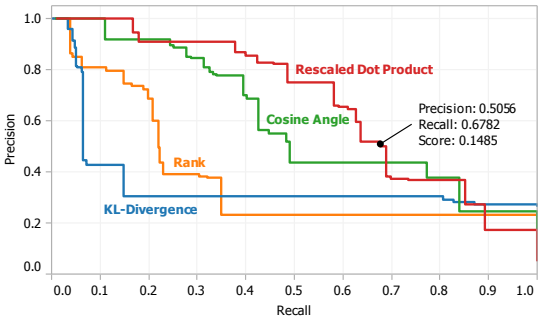


Figure 3. Precision and recall. Predicting human judgment of topic matches using topical similarity measures.

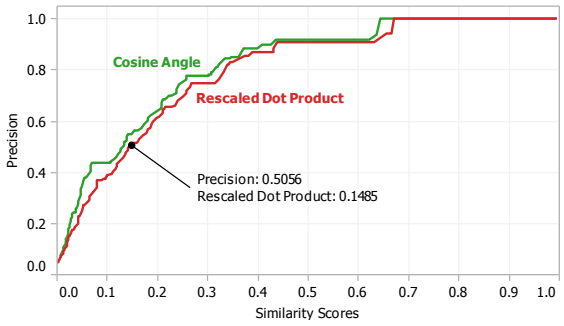


Figure 4. Similarity score vs. precision. Topical pairs with a rescaled dot product score greater than 0.148 were considered matching by human judges over 50% of the time.

Precision-recall scores in Figure 3 compare user-identified matches to the ordering of topical pairs induced by the similarity measures. Our rescaled dot product achieved the highest AUC, F1, F0.5, and F2 scores. We found that KL-divergence did a poor job of predicting human judgment; topical pairs ranked in the 90th percentile (among the 10% of most divergent pairs) still contained matches. Spearman rank correlation was concentrated in a narrow range $(-0.04, 0.16)$ for 96% of our data points. We observed that L_2 normalization in the cosine calculation was largely ineffective when applied to (L_1 normalized) probability distributions. Instead, given two word distributions we rescaled their dot product to the range of minimum and maximum possible similarities, and found that this outperformed the other measures.

3.4. Mapping Similarity Scores to Likelihoods

While the rescaled dot product is predictive of human judgment, the actual similarity values deviate from our definition of matching likelihood. Figure 4 plots precision against the similarity score at which that precision is achieved. By definition, topical pairs ranked above a precision of 0.5 are considered matching by human judges over 50% of the time. For the rescaled dot product, this threshold occurs at 0.1485 instead of the desired value of 0.5.

Linear transformation in log-ratio likelihood space performs well for correcting this deviation. We convert both similarity scores and precision values to log-ratio likelihoods, and apply linear regression to determine optimal mapping coefficients (Table 2). For the rescaled dot product, the transformed scores deviate from average user ratings by 0.0650. Transformed cosine angles deviate from user ratings by 0.1036. Provided with sets of reference concepts and latent topics, we can now populate entries of a correspondence chart using the transformed rescaled dot product scores.

Table 1. Rescaled dot product. Given a word probability distribution X, the scalar x_i denotes the probability for term i in topic X. $\overrightarrow{X}$ is a vector consisting of all x_i ordered in descending values; $\overleftarrow{X}$ is a vector of x_i in ascending order.

$$\text{Rescaled Dot Product} = \frac{P \cdot Q - d_{\text{Min}}}{d_{\text{Max}} - d_{\text{Min}}} \qquad \begin{aligned} d_{\text{Max}} &= \overrightarrow{P} \cdot \overrightarrow{Q} \\ d_{\text{Min}} &= \overrightarrow{P} \cdot \overleftarrow{Q} \end{aligned}$$

3.5. Estimating Random Chance of Matching

Matching likelihoods determined from human judgments are rarely exactly zero. As a topic model may contain hundreds of latent topics, even a small chance probability of matching can accumulate and bias misalignment scores toward a high number of repeated concepts or fused topics. To ensure our framework is robust for large-scale comparisons, we introduce a method to estimate and remove topical correspondences that can be attributed to random chance.

Given a correspondence matrix, we treat it as a linear combination of two sources: a *definitive* matrix whose entries are either 0 or 1; and a *noise* matrix representing some chance probability. We assume that matching likelihoods are randomly drawn from the definitive matrix $(1-\gamma)$ of the time and from the noise matrix γ of the time, where γ is a noise factor between $[0, 1]$.

Without explicitly specifying the values of the entries in the definitive matrix, we can still construct $P^k_{\text{definitive}}$ if we know it contains k non-zero values. We compute the average row and column matching likeli-

Table 2. Transformed similarity score. We fit similarity scores s to empirically obtained precisions based on linear regression in log-ratio likelihood space. f denotes the logit function; $\hat{f}$ denotes the inverse logit function. The coefficients are $a = 1.567$ and $b = 2.446$ for rescaled dot product, and are $a = 1.970$ and $b = 4.163$ for cosine.

$$\text{Transformed Similarity Score} = \hat{f}\left(af(s) + b\right)$$

hoods, and create a noise matrix whose entries equal $\hat{p}_{s,t} = 0.5\sum_{i=1}^{n} p_{i,t}/n + 0.5\sum_{j=1}^{m} p_{s,j}/m$. The action of sampling from the two source charts produces a corresponding $P_{\text{combined}} = P_{\text{definitive}}^{k(1-\gamma)} * P_{\text{noise}}^{\gamma}$ where $*$ is the convolution operator; mathematical derivations are provided in the supplementary materials. We compute γ by solving the convex optimization:

$$\underset{\gamma}{\operatorname{argmin}}\, \mathrm{KL}(P_{\text{definitive}}^{k(1-\gamma)} * P_{\text{noise}}^{\gamma} || P)$$

The optimal γ value represents the estimated amount of matches that can be attributed to noise. We then estimate the most likely distribution of topical matches P_{denoised} without the chance matches, by solving the following constrained optimization:

$$\underset{P_{\text{denoised}}}{\operatorname{argmin}}\, \mathrm{KL}(P_{\text{denoised}} * P_{\text{noise}}^{\gamma} || P)$$

subject to P_{denoised} being a proper probability distribution whose entries sum to 1 and are in the range $[0, 1]$. We apply the above process to each row and column in the correspondence matrix, to obtain $\dot{P}_{\text{denoised}}$ and $\ddot{P}_{\text{denoised}}$ from which we estimate topical misalignment scores as described previously.

4. Reference Concepts

Reference concepts in our diagnostic framework can be determined from various sources: elicited from domain experts, derived from available metadata, or based on the outputs of other topic models. For this work, we focus on the construction and use of high-quality expert-authored concepts to demonstrate our framework.

We conducted a survey in which we asked ten experienced researchers in information visualization (InfoVis) to exhaustively enumerate research topics in their field. Human topical organization can depend on factors such as expertise (Johnson & Mervis, 1997) where prior research finds that experts attend to more subtle features and recognize more functionally important concepts. To ensure quality of our reference concepts, we designed our survey to meet the following three criteria: (1) admitting only expert participants, (2) eliciting an exhaustive instead of partial list of concepts from each participant, and (3) collecting reference concepts from multiple subjects instead of a single source.

Survey responses consisted of manually-constructed topics comprising a title, a set of descriptive keyphrases, and a set of exemplary documents. Respondents authored these topical descriptions using a web-based interface with a searchable index of all 442 papers published at IEEE Information Visualization (details in supplementary materials). We received a total of 202 reference concepts from the experts.

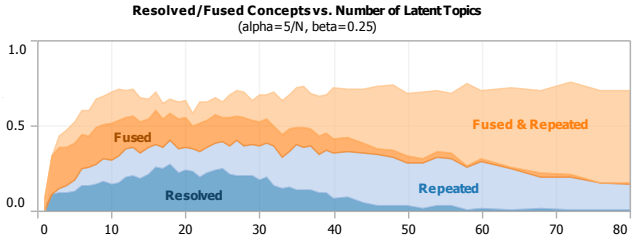


Figure 5. Topical alignment. LDA models for $N \in [1, 80]$, $\alpha = 5/N$, $\beta = 0.25$. The y-axis shows the fraction of reference concepts that have a single matching topic (*resolved*), multiple matching topics (*repeated*) or are subsumed by one (*fused*) or multiple fused topics (*fused & repeated*). These models uncover up to 75% of the reference concepts, but coverage increases only marginally for $N \geq 10$.

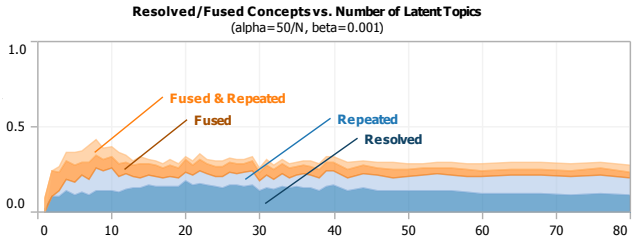


Figure 6. Topical alignment. LDA models for $N \in [1, 80]$, $\alpha = 50/N$, $\beta = 0.001$. This series of models uncovers up to 40% of the reference concepts. Coverage peaks at N=8. The proportion of resolved and fused topics remains stable for $N \geq 15$; increasing N produces only more junk topics.

We map survey responses to reference concepts as follows. For each expert-authored topic, we construct two term frequency counts: one consisting of provided title and keyphrase terms, and another consisting of the terms found in the representative documents. We perform TF.IDF weighting, normalize, and average the two distributions to produce a reference concept.

We chose InfoVis because of our familiarity with the community, which allowed us to contact experts capable of exhaustively enumerating research topics. The survey responses, though specific to a domain, constitute a rare and large-scale collection of manual topical categorization. The dataset provides us with a concrete baseline for assessing how machine-generated latent topics correspond to trusted concepts identified by experts, and enables comparisons with future studies. We are currently collecting topical categorizations in other domains and for larger corpora.

5. Applications

All results in this section are based on the InfoVis corpus. All models are built using Mallet (McCallum, 2013) unless otherwise stated. N denotes the number of latent topics in a topic model; α and β denote topic and term smoothing hyperparameters, respectively.

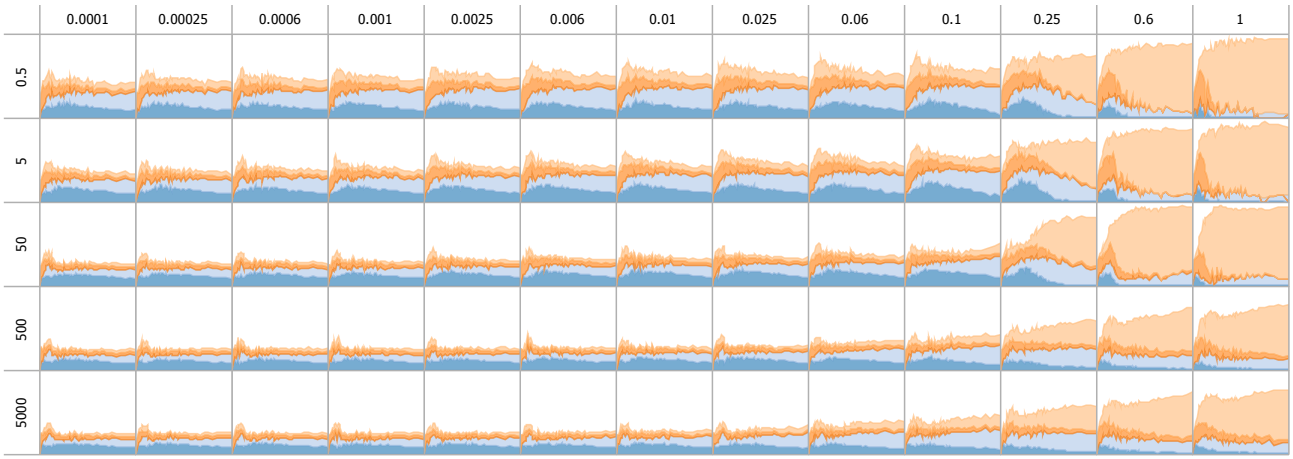


Figure 7. Exhaustive grid search. Topical alignment for LDA models over a grid of parameter/hyperparameter settings: $N \in [1, 80]$ (horizontal axis across subgraphs), 13 values of $\alpha \in [0.5/N, 5000/N]$ (vertical axis; only 5 shown due to page limitation), and 13 values of $\beta \in [0.0001, 1]$ (horizontal axis). We observe a qualitative shift in topical composition around β=0.25. For $\beta > 0.25$, the models generate fused topics that uncover but do not fully resolve a majority of the reference concepts as N increases. For $\beta < 0.25$, the proportion of resolved and fused topics remain stable regardless of N. Overall, decreasing β or increasing α leads to a decrease in coverage. See supplementary materials for a more detailed figure.

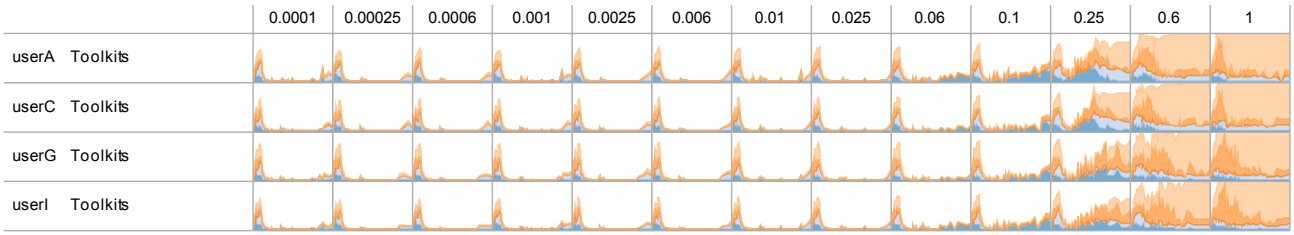


Figure 8. Uncovering the toolkit concept. The vertical axis within each subgraph shows the proportion of LDA models, trained using 10,000 parameter settings, that contained the concept *toolkits*. We find that the *toolkit* concept is uncovered only if β lies within a narrow range of values between $[0.06, 0.6]$. We also observe that PLDA models (Ramage et al., 2011) are more likely to uncover this class of topics than LDA models trained with hyperparameter optimization.

5.1. Exploration of Topic Models

We experiment with an exploratory process of topic model construction, in which users specify reference concepts a priori and utilize alignment scores to analyze the parameter space of models. We first examine the effects of varying $N \in [1, 80]$, and then perform an exhaustive grid search over N, α, and β.

Talley et al. (2011) found that N affects concept resolution and the number of poor quality topics. They arrived at this conclusion only after building a large number of models and performing an extensive manual review. In contrast, our framework allows users to map a large number of models onto predefined concepts and immediately inspect model qualities. In Figure 5, our misalignment measures indicate that the number of resolved topics peaks at $N = 18$. While the ratio of fused topics dips at $N = 20$, the proportion of fused topics increases again for $N \geq 30$. Trends in Figure 6 suggest that for a different set of hyperparameters, increasing N produces only more junk topics.

In Figure 7, we extend the space of models to over 10,000 parameter settings by searching 13 values of α and β. The set of hyperparameter values are chosen so they center at the default setting ($\alpha = 50/N$ and $\beta = 0.01$) and cover a range across 4 orders of magnitude. We observe additional qualitative changes in topic composition, such as the transition between fused and resolved concepts around $\beta = 0.25$.

5.2. Evaluation of Inference Algorithms

We analyze three categories of models to examine the effectiveness of hyperparameter optimization (Wallach et al., 2009a) for LDA, and the inclusion of metadata for author-topic models (Steyvers et al., 2004) and partially labeled LDA (PLDA) (Ramage et al., 2011).

We built 176 LDA models with hyperparameter optimization using Mallet (McCallum, 2013) for a grid of 11 values of $N \in [5, 80]$ and 4 initial values for each of α and β. We manually resolved all author names in all papers in the InfoVis corpus, and built

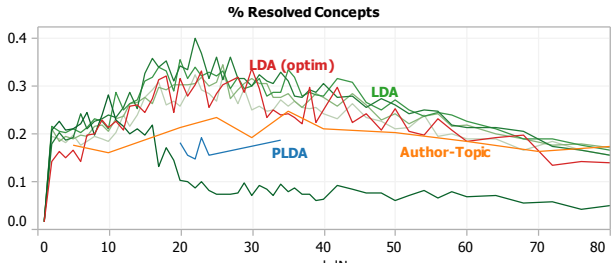


Figure 9. Hyperparameter optimization and inference algorithms. Green lines denote LDA models trained at $\beta = \{1, 0.1, 0.01, 0.001, 0, 0001\}$. LDA models with hyperparameter optimization (red) track with the highest-performing LDA model for most parameter settings.

10 author-topic models using the Matlab Topic Modeling Toolbox (Steyvers & Griffiths, 2013) by varying $N \in [5, 80]$. Finally, we built 11 PLDA models using the Stanford Topic Modeling Toolbox (Ramage, 2013), without hyperparameter optimization to isolate the effects of learning from metadata. We trained 10 models corresponding to concepts identified by each of the 10 experts. We then manually identified 28 concepts provided by at least 3 experts, and built an additional PLDA model containing these 28 concepts.

Figure 9 shows the number of resolved concepts for the best performing model from each category. The graph also includes the best performing LDA models for five values of β. We find that hyperparameter optimization is often able to select a β value comparable to the best-performing LDA model among our set of 10,000 from Section 5.1. Both author-topic models and PLDA without optimization uncover fewer resolved topics. Qualitatively, we note that author-topic models generally exhibit a higher proportion of resolved topics than fused topics. Examining individual concepts, we find that approximately 10% are uncovered by LDA only within narrow range of N, α, β values. An example of such a topic is *toolkit* (Figure 8), provided by eight of our experts. We find that PLDA was able to consistently recover the *toolkit* concept.

5.3. Evaluation of Intrinsic Measures

We apply our measures of *junk* and *resolved* topics to assess existing intrinsic measures for topical quality. We first describe the intrinsic measures under consideration, and then present our comparison results.

Alsumait et al. (2009) proposed three classes of topical significance measures. The authors describe a latent topic as uninformative if it consists of a uniform word distribution (*UniformW*), a word distribution matching the empirical term frequencies in the corpus (*Vacuous*), or uniform weights across documents (*Background*). The significance of a topic is its distance from one of these uninformative attributes; the exact definition of a "distance" was left open by the authors. For this work, we evaluated six significance measures based on KL-divergence (*UniformW-KL*, *Vacuous-KL*, *Background-KL*) and cosine dissimilarity (*UniformW-Cos*, *Vacuous-Cos*, *Background-Cos*). We also examined Pearson rank correlation (*Vacuous-Cor*).[1]

Newman et al. (2010a) measured topical coherence based on word co-occurrence in WordNet, Wikipedia articles, or Google search results. We examined their two top performing measures: word co-occurrence in the titles of search engine results[2] (*BingTitles-10*) and pointwise mutual information in Wikipedia text (*WikiPMI-10*). Mimno et al. (2011) measured topical coherence based on word co-occurrence in the document corpus (*MimnoCo-10*). These three coherence scores examine only the top k most probable words belonging to a topic. We experimented with various values up to $k \leq 30$, but report only $k = 10$ which is representative of the overall results.

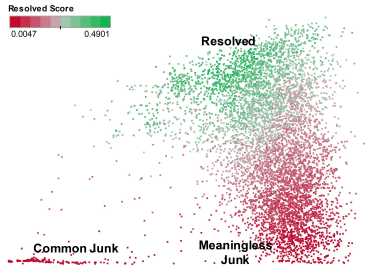
We computed the topical significance and coherence scores for each of the 176 LDA models with hyperparameter optimization built in Section 5.2. Figure 10 shows the correlation between the *Vacuous-KL* score and our junk measure. We observe that a small set of topics (bottom left) are marked as problematic by both measures. We also find, however, a large number of discrepancies (bottom right): junk topics without a meaningful corresponding expert-identified concept but marked as significant by *Vacuous-KL*.

Figure 11 repeats the graph for all ten intrinsic measures. As a whole, we observe little correlation across the graphs. *Background-KL* and *Background-Cos* exhibit a similar pattern as *Vacuous-KL* with some shared junk topics but a large number of discrepancies. *WikiPMI-10* performs poorly because many domain-specific terms do not co-occur in Wikipedia articles. *BingTitles-10* can separate topics comprising of functional words but otherwise lacks discriminative power.

We also examined the ranking of topics within each model. We computed the Spearman rank correlation between the ranking of topics by the intrinsic scores and by our junk measure. The median values are shown in the chart titles in Figure 11. We find low

[1] Applying Pearson rank correlation to calculate deviations from *UniformW* or *Background* leads to mathematically undefined significance values, due to zero variance in uniform word and topic distributions. We contacted the authors, but they were unable to explain or reconstruct how they computed these measures in their original paper. By the same reason, the 4-phase weighted combination described in the same paper is also ill-defined.

[2] Google search API was no longer publicly available.



Common	Junk topics marked as significant			Resolved topics		
the	and	highlighting	electronic	text	layout	tree
of	of	wordle	signal	documents	graph	trees
a	analysis	layout	test	document	algorithm	nodes
in	visual	stereoscopic	measurement	mining	graphs	hierarchy
to	interactive	results	hotmap	tools	algorithms	space
we	exploration	maniwordle	rate	graphical	layouts	hierarchical
is	system	both	imagery	content	drawing	hierarchies
our	users	route	entire	interface	node	displayed
that	from	composition	tracks	collections	directed	down
by	multiple	queries	ev	java	network	drill

Figure 10. Topical significance vs. our junk measure. The plot shows topical quality according to *Vacuous-KL* scores (horizontal axis; right is significant) and using our junk measure (vertical axis; downward is junk). While the two measures agree on a small set of problematic topics ("common" at the bottom left), we observe a large number of discrepancies ("meaningless" at the bottom right) that are considered junk by experts but marked as significant. Color represents how well a topic is considered resolved and free from misalignments (green for resolved; red otherwise).

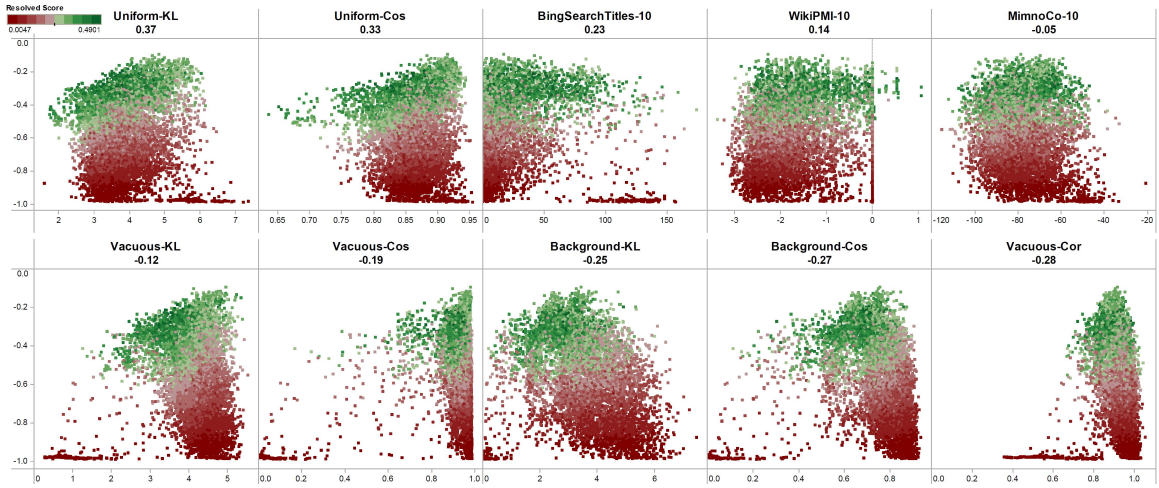


Figure 11. Intrinsic measures vs. our junk measure. Within each subgraph, right is significant/coherent and downward is junk; topical rank correlation score is shown under the subgraph label. Color represents how well a topic is resolved.

levels of correlations, indicating discrepancies between topics considered meaningful by experts and those marked significant/coherent by existing measures.

6. Discussion

For many domain-specific tasks, applying topic modeling requires intensive manual processing. In this paper, we introduce a framework in which analysts can express their domain knowledge and assess topic diagnostics. We quantify four types of topical misalignment, and introduce a process to automate the calculation of topical correspondences. Our technique enables large-scale assessment of models. Our applications suggest that diagnostic information can provide useful insights to both end-users and researchers.

Our long-term research goal is to support a human-in-the-loop modeling workflow. While recent work has contributed learning techniques for incorporating user inputs to aid the construction of domain-specific models (Hu et al., 2011; Ramage et al., 2011), we believe empirical studies of human topical organization and the design of user-facing tools can be equally important in supporting effective interactive topic modeling.

For this work, we elicited high-quality expert-authored concepts for evaluating topic models. In various other domains, reference concepts may exist but can be of differing levels of quality or coverage. An open research question is how semi-supervised learning algorithms and automatic measures of topical quality would perform under noisy or incomplete user inputs. Our dataset and results provide a benchmark and a point of comparison for future research in these areas.

Another research question is how topical misalignment affects a user's ability to interpret and work with topic models. Our results suggest that different models produce different types of misalignment. A better understanding may lead to improved detection of problematic latent topics and more informed decisions on how to match models to analysis tasks. Certain misalignments may be more easily remedied by people; we can then design interactive diagnostic tools to elicit corrective actions from the users accordingly.

References

Alsumait, Loulwah, Barbará, Daniel, Gentle, James, and Domeniconi, Carlotta. Topic significance ranking of LDA generative models. In *ECML*, pp. 67–82, 2009.

Blei, David M. and Lafferty, John D. Dynamic topic models. In *ICML*, pp. 113–120, 2006.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *J of Machine Learning Research*, 3(1):993–1022, 2003.

Blei, David M., Griffiths, Thomas L., Jordan, Michael I., and Tenenbaum, Joshua B. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2004.

Budiu, Raluca, Royer, Christiaan, and Pirolli, Peter. Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In *RIAO*, pp. 314–332, 2007.

Chaney, Allison and Blei, David. Visualizing topic models. In *AAAI*, pp. 419–422, 2012.

Chang, Jonathan, Boyd-Graber, Jordan, Wang, Chong, Gerrish, Sean, and Blei, David M. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

Chuang, Jason, Manning, Christopher D., and Heer, Jeffrey. Termite: Visualization techniques for assessing textual topic models. In *AVI*, pp. 74–77, 2012a.

Chuang, Jason, Ramage, Daniel, Manning, Christopher D., and Heer, Jeffrey. Interpretation and trust: Designing model-driven visualizations for text analysis. In *CHI*, pp. 443–452, 2012b.

Gretarsson, Brynjar, O'Donovan, John, Bostandjiev, Svetlin, Höllerer, Tobias, Asuncion, Arthur, Newman, David, and Smyth, Padhraic. TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Trans on Intelligent Systems and Technology*, 3(2):23:1–23:26, 2012.

Hall, David, Jurafsky, Daniel, and Manning, Christopher D. Studying the history of ideas using topic models. In *EMNLP*, pp. 363–371, 2008.

Hu, Yuening, Boyd-Graber, Jordan, and Satinoff, Brianna. Interactive topic modeling. In *ACL-HLT*, 2011.

Johnson, K. E. and Mervis, C. B. Effects of varying levels of expertise on the basic level of categorization. *J of Experimental Psychology: General*, 126(3):248–277, 1997.

McCallum, Andrew Kachites. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2013. Accessed 2013.

Mimno, David, Wallach, Hanna, Talley, Edmund, Leenders, Miriam, and McCallum, Andrew. Optimizing semantic coherence in topic models. In *EMNLP*, 2011.

Musat, Claudiu Cristian, Velcin, Julien, Trausan-Matu, Stefan, and Rizoiu, Marian-Andrei. Improving topic evaluation using conceptual knowledge. In *IJCAI*, 2011.

Newman, David, Chemudugunta, Chaitanya, Smyth, Padhraic, and Steyvers, Mark. Analyzing entities and topics in news articles using statistical topic models. In *ISI*, pp. 93–104, 2006.

Newman, David, Lau, Jey Han, Grieser, Karl, and Baldwin, Timothy. Automatic evaluation of topic coherence. In *HLT*, pp. 100–108, 2010a.

Newman, David, Noh, Youn, Talley, Edmund, Karimi, Sarvnaz, and Baldwin, Timothy. Evaluating topic models for digital libraries. In *JCDL*, pp. 215–224, 2010b.

Ramage, Daniel. Stanford topic modeling toolbox 0.4. http://nlp.stanford.edu/software/tmt, 2013. Accessed 2013.

Ramage, Daniel, Hall, David, Nallapati, Ramesh, and Manning, Christopher D. Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. In *EMNLP*, pp. 248–256, 2009.

Ramage, Daniel, Dumais, S., and Liebling, D. Characterizing microblogs with topic models. In *ICWSM*, pp. 130–137, 2010.

Ramage, Daniel, Manning, Christopher D., and Dumais, Susan. Partially labeled topic models for interpretable text mining. In *KDD*, pp. 457–465, 2011.

Rosen-Zvi, Michal, Griffiths, Thomas, Steyvers, Mark, and Smyth, Padhraic. The author-topic model for authors and documents. In *UAI*, pp. 487–494, 2004.

Stevens, Keith, Kegelmeyer, Philip, Andrzejewski, David, and Buttler, David. Exploring topic coherence over many models and many topics. In *EMNLP-CoNLL*, pp. 952–961, 2012.

Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. Probabilistic author-topic models for information discovery. In *KDD*, 2004.

Steyvers, Mark and Griffiths, Tom. Matlab topic modeling toolbox 1.4. http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm, 2013. Accessed 2013.

Talley, Edmund M., Newman, David, Mimno, David, Herr, Bruce W., Wallach, Hanna M., Burns, Gully A. P. C., Leenders, A. G. Miriam, and McCallum, Andrew. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8 (6):443–444, 2011.

Titov, Ivan and McDonald, Ryan. A joint model of text and aspect ratings for sentiment summarization. In *ACL-HLT*, pp. 308–316, 2008.

Wallach, Hanna M., Mimno, David, and McCallum, Andrew. Rethinking LDA: Why priors matter. In *NIPS*, 2009a.

Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *ICML*, pp. 1105–1112, 2009b.

Wang, Xuerui and McCallum, Andrew. Topics over time: A non-Markov continuous-time model of topical trends. In *KDD*, pp. 424–433, 2006.

Wei, Xing and Croft, W. Bruce. Lda-based document models for ad-hoc retrieval. In *SIGIR*, pp. 178–185, 2006.