Evaluating Topic Quality with Posterior Variability

Linzi Xing† and Michael J. Paul‡ and Giuseppe Carenini†
† University of British Columbia, Vancouver, Canada
‡ University of Colorado, Boulder, Colorado, USA
{lzxing, carenini}@cs.ubc.ca, mpaul@colorado.edu

Abstract

Probabilistic topic models such as latent Dirichlet allocation (LDA) are popularly used with Bayesian inference methods such as Gibbs sampling to learn posterior distributions over topic model parameters. We derive a novel measure of LDA topic quality using the variability of the posterior distributions. Compared to several existing baselines for automatic topic evaluation, the proposed metric achieves state-of-the-art correlations with human judgments of topic quality in experiments on three corpora.1 We additionally demonstrate that topic quality estimation can be further improved using a supervised estimator that combines multiple metrics.

1 Introduction

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic modeling has been widely used for NLP tasks which require the extraction of latent themes, such as scientific article topic analysis (Hall et al., 2008), news media tracking (Roberts et al., 2013), online campaign detection (Paul and Dredze, 2014) and medical issue analysis (Huang et al., 2015, 2017). To reliably utilize topic models trained for these tasks, we need to evaluate them carefully and ensure that they have as high quality as possible. When topic models are used in an extrinsic task, like text categorization, they can be assessed by measuring how effectively they contribute to that task (Chen et al., 2016; Huang et al., 2015). However, when they are generated for human consumption, their evaluation is more challenging. In such cases, interpretability is critical, and Chang et al. (2009); Aletras and Stevenson (2013) have shown that the standard way to evaluate the output of a probabilistic model, by measuring perplexity on held-out data (Wallach et al., 2009), does not imply that the inferred topics are human-interpretable.

A topic inferred by LDA is typically represented by the set of words with the highest probability given the topic. With this characteristic, we can evaluate the topic quality by determining how coherent the set of topic words is. While a variety of techniques (Section 2) have been geared towards measuring the topic quality in this way, in this paper, we push such research one step further by making the following two contributions: (1) We propose a novel topic quality metric by using the variability of LDA posterior distributions. This metric conforms well with human judgments and achieves the state-of-the-art performance. (2) We also create a topic quality estimator by combining two complementary classes of metrics: the metrics that use information from posterior distributions (including our new metric), along with the set of metrics that rely on topic word co-occurrence. Our novel estimator further improves the topic quality assessment on two out of the three corpora we have.

2 Automatic Topic Quality Evaluation

There are two common ways to evaluate the quality of LDA topic models: Co-occurrence Based Methods and Posterior Based Methods.

Co-occurrence Based Methods Most prominent topic quality evaluations use various pairwise co-occurrence statistics to estimate topic’s semantic similarity. Mimno et al. (2011) proposed the Coherence metric, which is the summation of the conditional probability of each topic word given all other words. Newman et al. (2010) showed that the summation of the pairwise pointwise mutual information (PMI) of all possible topic word pairs is also an effective metric to assess topic quality. Later, in Lau et al. (2014), PMI was replaced...
by the normalized pointwise mutual information (NPMI) (Bouma, 2009), which has an even higher correlation with human judgments. Another line of work exploits the co-occurrence statistics indirectly. Aletras and Stevenson (2013) devised a new method by mapping the topic words into a semantic space and then computing the pairwise distributional similarity (DS) of words in that space. However, the semantic space is still built on PMI or NPMI. Roder et al. (2015) studied a unifying framework to explore a set of co-occurrence based topic quality measures and their parameters, identifying two complex combinations, (named CV and CP in that paper\footnote{The framework proposed in Roder et al. (2015) has four stages. Every stage has multiple settings. CV and CP are different at the Confirmation Measure stage, which measures how strongly a set of topic words connect with each other.}), as the best performers on their test corpora.

Posterior Based Method Recently, Xing and Paul (2018) analyzed how the posterior of LDA parameters vary during Gibbs sampling inference (Geman and Geman, 1984; Griffiths and Steyvers, 2004) and proposed a new topic quality measurement named Topic Stability. The Gibbs sampling for LDA generates estimates for two distributions: for topics given a document ($\theta$), and for words given a topic ($\phi$). Topic stability considers $\phi$ and is defined as:

$$stability(\Phi_k) = \frac{1}{|\Phi_k|} \sum_{\phi_k \in \Phi_k} sim(\phi_k, \bar{\phi}_k)$$  

The stability of topic $k$ is computed as the mean cosine similarity between the mean ($\bar{\phi}_k$) of all the sampled topic $k$’s distribution estimates ($\Phi_k$) and topic $k$’s estimates from each Gibbs sampler ($\phi_k$). Eased against the co-occurrence based methods, topic stability is parameter-free and needs no external corpora to infer the word co-occurrence. However, due to the high frequency of common

<table>
<thead>
<tr>
<th>Metric</th>
<th>20NG</th>
<th>Wiki</th>
<th>NYT</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0.185</td>
<td>0.030</td>
<td>0.148</td>
<td>0.121</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.480</td>
<td>0.295</td>
<td>0.600</td>
<td>0.458</td>
</tr>
<tr>
<td>$cv$</td>
<td>0.679</td>
<td>0.703</td>
<td>0.774</td>
<td>0.719</td>
</tr>
</tbody>
</table>

Table 1: Pearson’s $r$ of each potential metric of posterior variability with human judgments
more stable dispersion of estimates from different iterations of Gibbs sampling, we propose to compute the variation of topic $k$’s estimates in document $d$ as its coefficient of variance ($cv$) (Everitt, 2002), which is defined as: $cv_{dk} = \sigma_{dk}/\mu_{dk}$

Notice that both $\mu_{dk}$ and $cv_{dk}$ can arguably help distinguish high and low quality topics because:

- High quality topics will have high $\mu_{dk}$ for related documents and low $\mu_{dk}$ for unrelated documents. But low quality topics will have relatively close $\mu_{dk}$ throughout the corpus.

- High quality topics will have low $cv_{dk}$ for related documents and high $cv_{dk}$ for unrelated ones. But low quality topics will have relatively high $cv_{dk}$ throughout the corpus.

Now, focusing on the design of the new metric, we consider using the corpus-wide variability of topics’ estimates as our new metric. Figure 1 shows a comparison of the distributions of mean of estimates ($\mu$) and variation of estimates ($\sigma$, $cv$) for two topics across the NYT corpus (Section 5.1). We can see the $cv$ distributions of good (Topic1) and bad (Topic2) topics are the most different. The $cv$ distribution of Topic1 covers a large span and has a heavy head and tail, while $cv$ values of Topic2 are mostly clustered in a smaller range. In contrast, the difference between Topic1 and Topic2’s distributions of $\mu$ and $\sigma$ throughout the corpus appears to be less pronounced. Taking the corpus-wide variability difference between good and bad topics observed in Figure 1, we choose $cv$ to measure the corpus-wide variability of topic $k$’s estimates as our new metric. Formally, it can be defined as:

$$variability(k) = std(cv_{1k}, cv_{2k}, \ldots, cv_{Dk})$$  \hspace{1cm} (2)$$

where $D$ is the size of the corpus. High quality topics will have higher variability and low quality topics will have lower variability. Table 1 shows a comparison in performance (correlation with human judgment) of our variability defined by $cv$ with the variability defined by $\mu$ or $\sigma$ on three commonly used datasets (Section 5.1). The variability defined by $cv$ is a clear winner.

### 4 Topic Quality Estimator

Our new method, like all other methods driven from the posterior variability, does not use any information from the topic words, which is in contrast the main driver for co-occurrence methods. Based on this observation, posterior variability and co-occurrence methods should be complementary to each other. To test this hypothesis, we investigate if a more reliable estimator of topic quality can be obtained by combining these two classes of methods in a supervised approach. In particular, we train a support vector regression (SVR) estimator (Joachims, 2006) with the estimations of these methods as features, including all the topic quality measures introduced in Section 2 along with our proposed variability method.

### 5 Experiments

#### 5.1 Datasets

We evaluate our topic quality estimator on three datasets: 20NG, Wiki and NYT. 20NG is the 20 Newsgroup dataset (Joachims, 1997) which consists of 9,347 paragraphs categorized into 20 classes\(^3\). Wiki consists of 10,773 Wikipedia articles written in simple English\(^4\). NYT consists of 8,764 New York Times articles collected from April to July, 2016\(^5\).

\(^3\)http://archive.ics.uci.edu/ml/datasets/twenty+newsgroups  
\(^4\)http://simple.wikipedia.org/  
\(^5\)https://www.kaggle.com/nzalake52/new-york-times-articles

---

<table>
<thead>
<tr>
<th>Method</th>
<th>20NG</th>
<th>Wiki</th>
<th>NYT</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV (Roder et al., 2015)</td>
<td>0.129</td>
<td>0.385</td>
<td>0.248</td>
<td>0.254</td>
</tr>
<tr>
<td>CP (Roder et al., 2015)</td>
<td>0.378</td>
<td>0.403</td>
<td>0.061</td>
<td>0.280</td>
</tr>
<tr>
<td>DS (Aletras and Stevenson, 2013)</td>
<td>0.461</td>
<td>0.423</td>
<td>0.365</td>
<td>0.416</td>
</tr>
<tr>
<td>NPMI (Lau et al., 2014)</td>
<td>0.639</td>
<td>0.568</td>
<td>0.639</td>
<td>0.615</td>
</tr>
<tr>
<td>PMI (Newman et al., 2010)</td>
<td>0.602</td>
<td>0.550</td>
<td>0.623</td>
<td>0.591</td>
</tr>
<tr>
<td>Coherence (Mimno et al., 2011)</td>
<td>0.280</td>
<td>0.102</td>
<td>0.535</td>
<td>0.305</td>
</tr>
<tr>
<td>Stability (Xing and Paul, 2018)</td>
<td>0.230</td>
<td>0.137</td>
<td>0.322</td>
<td>0.230</td>
</tr>
<tr>
<td>Variability</td>
<td>0.679</td>
<td>0.703</td>
<td>0.774</td>
<td>0.719</td>
</tr>
</tbody>
</table>

Table 2: Pearson’s $r$ correlation with human judgments for metrics.
Table 3: Pearson’s $r$ correlation with human judgments for the topic quality estimator.

<table>
<thead>
<tr>
<th>Test</th>
<th>Train</th>
<th>Mean Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG Wiki</td>
<td>0.790</td>
<td>0.810</td>
</tr>
<tr>
<td>Wiki NYT</td>
<td>0.707</td>
<td>0.731</td>
</tr>
<tr>
<td>NYT Wiki+</td>
<td>0.762</td>
<td>0.775</td>
</tr>
</tbody>
</table>

We removed stop words, low-frequency words (appearing less than 3 times), proper nouns and digits from all the datasets, following Chang et al. (2009), so the topic modeling can reveal more general concepts across the corpus.

Following the common setting shared by most of the papers we compared with, for each dataset we built an LDA model which consists of 100 topics represented by the 10 most probable words. The gold-standard annotation for the quality of each topic is the mean of 4-scale human rating scores from five annotators, which were collected through a crowdsourcing platform, Figure Eight. In order to obtain more robust estimates given the variability in human judgments, we removed ratings from annotators who failed in the test trail and recollected those with additional reliable annotators. To verify the validity of the collected annotations, we computed the Weighted Krippendorff’s $\alpha$ (Krippendorff, 2007) as the measure of Inter-Annotator Agreement (IAA) for three datasets. The average human rating score/IAA for 20NG, Wiki and NYT are 2.91/0.71, 3.23/0.82 and 3.06/0.69, respectively.

5.2 Experimental Design

Topic Modeling Following the settings in Xing and Paul (2018), we ran the LDA Gibbs samplers for 2,000 iterations (Griffiths and Steyvers, 2004) for each datasets, with 1,000 burn-in iterations, collecting samples every 10 iterations for the final 1,000 iterations. The set of estimates $\Theta$ thus contains 100 samples.

Estimator Training was performed following the cross-domain training strategy (Bhatia et al., 2018). With the ground truth (human judgments), we train the estimator on all topics over one dataset, and test it on another (two-to-one). To enlarge the training set, we also train the estimator on two datasets merged together and test it on the third one (two-to-one). Given the limited amount of data and the need for interpretability, we experimented only with non-neural classifiers, including linear regression, nearest neighbors regression, Bayesian regression, and Support Vector Regression (SVR) using sklearn (Pedregosa et al., 2011); we report the results with SVR, which gave the best performance. We also experimented with different kernels of SVR and rbf kernel worked best.

Baselines include seven commonly adopted measures for topic quality assessment: CV, CP (Roder et al., 2015), DS (Aletras and Steven-son, 2013), PMI (Newman et al., 2010), NPMI (Mimno et al., 2011), Coherence (Newman et al., 2010) and Stability (Xing and Paul, 2018). All of them are introduced in Section 2.

5.3 Results

Following (Roder et al., 2015), we use Pearson’s $r$ to evaluate the correlation between the human judgments and the topic quality scores predicted by all the automatic metrics. The higher is the Pearson’s $r$, the better the metric is at measuring topic quality. Table 2 shows the Pearson’s $r$ correlation with human judgments for all the metrics. Our proposed variability-based metric substantially outperforms all the baselines.

Table 3 shows the Pearson’s $r$ correlation with our proposed topic quality estimator trained and tested on different datasets. The average correlations of the estimator dominates our proposed variability-based metric on two out of three datasets, and on one of them by a wide margin.

Additionally, to better investigate how well the metrics align with human judgments, in Figure 3 we use scatter plots to visualize their correlations and make the following observations. The top performer co-occurrence based metric, NPMI, tends to underestimate topic quality by giving low ratings to relatively high-quality topics (dots with
high human scores tend to be above the purple line), but it performs relatively well for low-quality topics. On the contrary, the top performer posterior based metric, variability, is more likely to overestimate topic quality by assigning high ratings to relatively bad topics (dots with low human scores tend to be below the green line), but it performs relatively well for high-quality topics. Thus, when we combine all the metrics in a supervised way, the topic quality estimation becomes more accurate, especially on 20NG corpus (i.e. the top row).

Ablation Analysis: Since some features in the topic quality estimator are closely related, their overlap/redundancy may even hurt the model’s performance. To better understand the contributions of each feature in our proposed estimator, we conduct ablation analysis whose results are illustrated in Figure 2. We track the change of performance by removing one feature each time. The more significant drop in performance indicates that the removed feature more strongly contributes to the estimator’s accuracy. By training on two datasets and testing on the third dataset, we find that only Variability and NPMI consistently contributes to accurate predictions on all three datasets. This indicates that our new Variability metric and NPMI are the strongest ones from the two families of Posterior-based and Co-occurrence-based metrics, respectively.

6 Conclusion and Future Work

We propose a novel approach to estimate topic quality grounded on the variability of the variance of LDA posterior estimates. We observe that our new metric, driven by Gibbs sampling, is more accurate than previous methods when tested against human topic quality judgment. Additionally, we propose a supervised topic quality estimator that by combining multiple metrics delivers even better results. For future work, we intend to work with larger datasets to investigate neural solutions to combine features from different metrics, as well as to apply our findings to other variants of LDA models trained on low-resource languages, where high-quality external corpora are usually not available (Hao et al., 2018).

Acknowledgments

We would like to thank the anonymous reviewers for their valuable suggestions and comments, as well as Jey Han Lau and Data Science Group at Universität Paderborn for making their topic quality evaluation toolkit publicly available.
References


Michael Roder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In the Eighth ACM International Conference on Web Search and Data Mining, pages 39–408.