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

[1] Our code and data are available here.

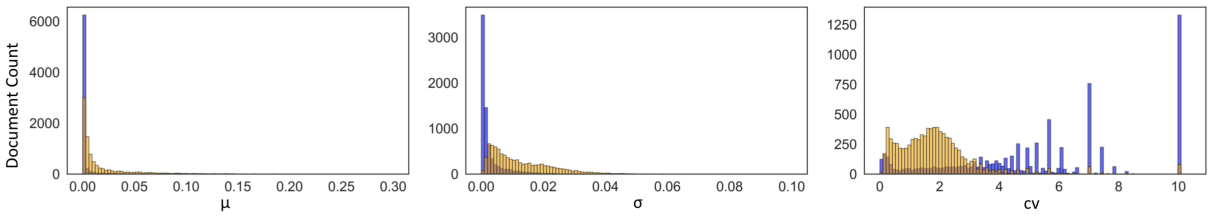

Figure 1: Two example topics and their distributions of μ, σ and cv from the NYT corpus. Two topics are: Topic1 (in blue): {*financial, banks, bank, money, debt, fund, loans, investors, funds, hedge*}. Topic2 (in orange): {*world, one, like, good, even, know, think, get, many, got*}. Their human rating scores are 3.4 and 1.0 respectively.

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[2]), 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 (θ), and for words given a topic (ϕ). Topic stability considers ϕ and is defined as:

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

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 ($\mathbf{\Phi_k}$) and topic k's estimates from each Gibbs sampler (ϕ_k). Fared against the co-occurrence based methods, topic stability is parameter-free and needs no external corpora to infer the word co-occurrence.

[2]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.

Metric	20NG	Wiki	NYT	Mean
μ	0.185	0.030	0.148	0.121
σ	0.480	0.295	0.600	0.458
cv	**0.679**	**0.703**	**0.774**	**0.719**

Table 1: Pearson's r of each potential metric of posterior variability with human judgments

However, due to the high frequency of common words across the corpus, low quality topics may also have high stability, and this undermines the performance of this method.

3 Variability Driven from Topic Estimates

In this paper, we also use Gibbs sampling to infer the posterior distribution over LDA parameters. Yet, instead of ϕ, our new topic evaluation method analyzes estimates of θ, the topic distribution in documents. Let $\mathbf{\Theta}$ be a set of different estimates of θ, which in our experiments will be a set of estimates from different iterations of Gibbs sampling. Traditionally, the final parameter estimates are taken as the mean of all the sampled estimates, $\hat{\theta}_{dk} = \frac{1}{|\mathbf{\Theta}|} \sum_{\theta \in \mathbf{\Theta}} \theta_{dk}$. In this paper, we use the shorthand μ_{dk} to denote $\hat{\theta}_{dk}$ for a particular document d and topic k.

In the rest of this section, we first discuss what types of information can be derived from the topic posterior estimates from different Gibbs samplers. Then, we examine how the corpus-wide variability can be effectively captured in a new metric for topic quality evaluation.

Two types of information can be derived from the topic posterior estimates: (1) the mean of estimates, μ_{dk}, as discussed above, and (2) the variation of estimates. For variation of estimates, we considered using the standard deviation σ_{dk}. However, this measure is too sensitive to the order-of-magnitude differences of μ_{dk}, that typically occur

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

Table 2: Pearson's r correlation with human judgments for metrics.

in different documents. So, in order to capture a 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 μ_{dk} and cv_{dk} can arguably help distinguish high and low quality topics because:

- High quality topics will have high μ_{dk} for related documents and low μ_{dk} for unrelated documents. But low quality topics will have relatively close μ_{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 (μ) and variation of estimates (σ, 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 μ and σ 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:

$$variability(k) = std(cv_{1k}, cv_{2k}, \cdots, cv_{Dk}) \quad (2)$$

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 μ or σ 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

Test	Train			Mean	Variability
20NG	Wiki	NYT	Wiki+NYT	**0.801**	0.679
	0.790	0.804	0.810		
Wiki	20NG	NYT	20NG+NYT	**0.716**	0.703
	0.707	0.731	0.710		
NYT	20NG	Wiki	20NG+Wiki	0.770	**0.774**
	0.762	0.775	0.773		

Table 3: Pearson's r correlation with human judgments for the topic quality estimator.

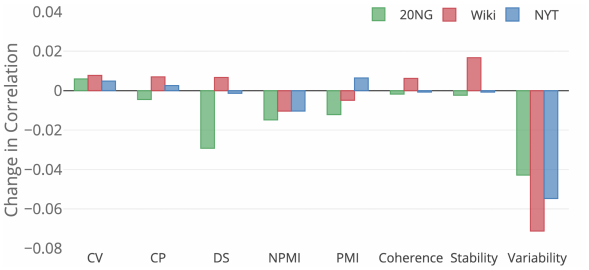

Figure 2: The ablation analysis.

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[6]. 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 α (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 $\mathbf{\Theta}$ 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 (**one-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 Stevenson, 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

[6] https://www.figure-eight.com/

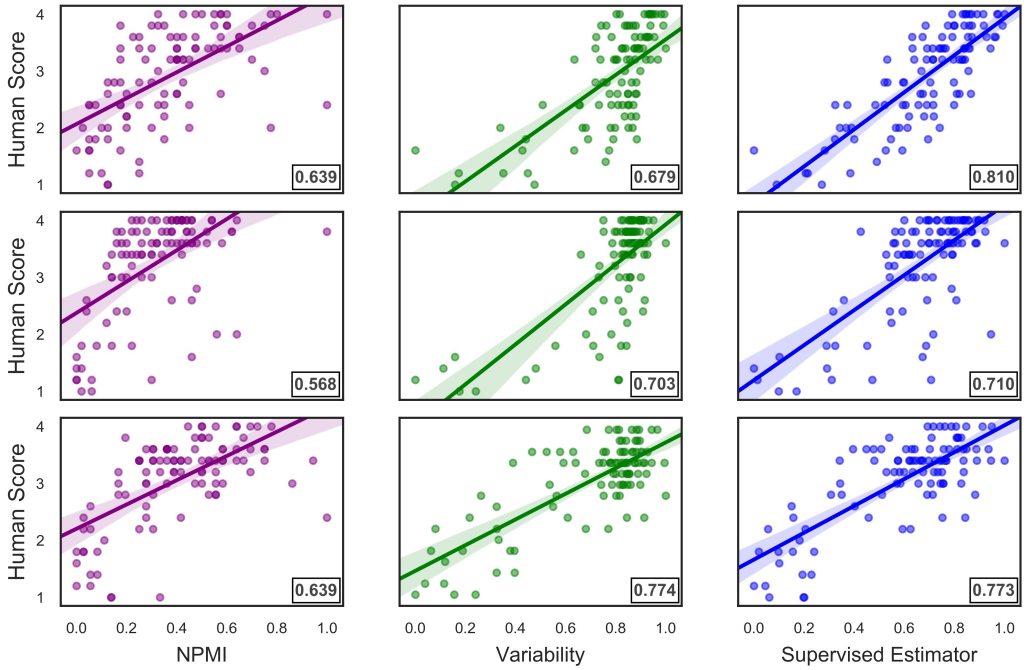

Figure 3: Scatter plots illustrating the correlation between human rating scores and the three metrics: *NPMI*, *Variability* and *Supervised Estimator* on three datasets: *20NG* (top row), *Wiki* (middle row) and *NYT* (bottom row). The numerical Pearson's r correlations are shown in the bottom-right corner.

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 Universitat Paderborn for making their topic quality evaluation toolkit publicly available.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2018. Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference'09*, pages 31–40.

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

Qiuxing Chen, Lixiu Yao, and Jie Yang. 2016. Short text classification based on lda topic model. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)*.

Brian Everitt. 2002. *The Cambridge dictionary of statistics*. Cambridge University Press.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 503–512.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National academy of Sciences*, pages 5228–5235.

David Hall, Daniel Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 363–371.

Shudong Hao, Jordan Boyd-Graber, and Michael J. Paul. 2018. Lessons from the Bible on modern topics: Low-resource multilingual topic model evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1090–1100.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in Chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562.

Xiaolei Huang, Linzi Xing, Jed R. Brubaker, and Michael J. Paul. 2017. Topic model for identifying suicidal ideation in Chinese microblog. In *IEEE International Conference on Healthcare Informatics (ICHI)*, pages 470–477.

Thorsten Joachims. 1997. A probabilistic analysis of the Rocchio algorithm with TF-IDF for text categorization. In *International Conference on Machine Learning (ICML)*, pages 143–151.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226.

Klaus Krippendorff. 2007. Computing krippendorffs alpha reliability. *Departmental papers (ASC) (2007)*, 43.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Michael J Paul and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PLOS ONE*, page 9(8):e103408.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, and Edoardo M Airoldi. 2013. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

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.

Hanna M. Wallach, David M. Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *Proceedings of the Neural Information Processing Systems*, pages 1973–1981.

Linzi Xing and Michael Paul. 2018. Diagnosing and improving topic models by analyzing posterior variability. In *AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 6005–6012.

A Guidelines for Human Annotation

Overview

In this job, you will be presented with topics represented by top10 topic words. Review the sets of topic words to determine the quality (coherence) of topics. Do you think these topics are useful (interpretable), relatively useful (interpretable but with some noise), relatively useless (combination of more than one topic) or useless (guess no one would have a sure idea what this topic is about).

Steps

- Make sure you read and understand the rules and follow the given examples.
- Read the 10 topic words presented.
- Evaluate the topic coherence (useful / relatively useful / relatively useless / useless).

Rules Tips

- **Useful(4) :**
  - The set of words is semantically coherent, meaningful and interpretable.
  - You can easily summarize the topic that these words are describing.
- **Relatively useful(3) :**
  - Most of words can represent the same topic but there are few words not.
  - Even with some noise, you can still get the general idea of what these words are describing.
- **Relatively useless(2) :**
  - A few words are coherent and informative but more others are not.
  - There are some words used in specific domains, but it's hard to come up with the idea what these words are describing.
- **Useless(1) :**
  - Words appear randomly and unrelated to each other.
  - Most of the words are common used words with no specific meaning, like got, went, new...

Examples

Useful (4)

eg. **space orbit mission launch lunar spacecraft surface missions satellite shuttle**
Note: From this set of words, you can very easily summarize the topic **aerospace**.

Relatively useful (3)

eg. **police said officers man year shot two shooting authorities arrested**
Note: You can still summarize that the topic is about **criminal**, but there are some noisy words like 'said', 'year', 'two' which are not directly related to this topic.

Relatively useless (2)

eg. **housing store homeless home food christmas animals shopping video city**
Note: Most of the words here are still informative individually. However, it is hard to quickly get a clear topic from this set of words.

Useless (1)

eg. **said world us things new people next decision two make**
Note: Almost all the words presented here are not informative at all. They are "common words" which are possible to be used anywhere. You have no idea what topic these words refer to...

Figure 4: The details of the guidelines we provided to the topic quality annotators. All topics were rated on a 4-point Likert scale. In particular, we provided the descriptions of the meaning of the four criteria, as well as the made-up examples (one for each score) that were created to help raters understand the criteria.