Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, **Jordan Boyd-Graber**, and Kevin Seppi. **Automatic and Human Evaluation of Local Topic Quality**. *Association for Computational Linguistics*, 2019, 10 pages.


```
@inproceedings{Lund:Armstrong:Fearn:Cowley:Byun:Boyd-Graber:Seppi-2019,
Title = {Automatic and Human Evaluation of Local Topic Quality},
Author = {Jeffrey Lund and Piper Armstrong and Wilson Fearn and Stephen Cowley and Courtni Byun and Jordan
Booktitle = {Association for Computational Linguistics},
Year = {2019},
Location = {Florence, Italy},
Url = {docs/2019_acl_local.pdf},
}
```


Links:

- Code [http://github.com/jefflund/ankura]

Automatic and Human Evaluation of Local Topic Quality

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun
Computer Science
Brigham Young University
{jefflund, piper.armstrong, wfearn, scowley4, courtni.byun}
@byu.edu

Jordan Boyd-Graber
Computer Science, iSchool,
LSC, and UMIACS
University of Maryland
jbg@umiacs.umd.edu

Kevin Seppi
Computer Science
Brigham Young University
kseppi@byu.edu

Abstract

Topic models are typically evaluated with respect to the global topic distributions that they generate, using metrics such as coherence, but without regard to local (token-level) topic assignments. Token-level assignments are important for downstream tasks such as classification. Recent models, which claim to improve token-level topic assignments, are only validated on global metrics. We elicit human judgments of token-level topic assignments: over a variety of topic model types and parameters, global metrics agree poorly with human assignments. Since human evaluation is expensive we propose automated metrics to evaluate topic models at a local level. Finally, we correlate our proposed metrics with human judgments: an evaluation based on the percent of topic switches correlates most strongly with human judgment of local topic quality. This new metric, which we call consistency, should be adopted alongside global metrics such as topic coherence.

1 Introduction

Topic models such as Latent Dirichlet Allocation (Blei et al., 2003, LDA) automatically discover topics in a collection of documents, giving users a glimpse into themes present in the documents. LDA jointly derives a set of topics (a distribution over words) and token-topic assignments (a distribution over the topics for each token). While the topics by themselves are valuable, the token-topic assignments are also useful as features for document classification (Ramage et al., 2009; Nguyen et al., 2015; Lund et al., 2018) and, in principle, for topic-based document segmentation.

Given the breadth of topic model variants and implementations, the question of algorithm selection and model evaluation can be as daunting as it is important. When the model is used for a downstream evaluation task (e.g., document classification), these questions can often be answered by maximizing downstream task performance. In other cases, automated metrics such as topic coherence (Newman et al., 2010) can help assess topic model quality. Generally speaking, these metrics evaluate topic models *globally*, meaning that the metrics evaluate characteristics of the topics (word distributions) themselves, ignoring the topic assignments of individual tokens.

In the context of human interaction, this means that models produce global topic-word distributions that typically make sense to users and serve to give a good high-level overview of the general themes and trends in the data. However, the local topic assignments can be bewildering. For example, Figure 1 shows typical topic assignments using LDA. Arguably, most, if not all, of the sentence should be assigned to the Music topic—the sentence is about a music video for a particular song. However, parts of the sentence are assigned to other topics: Gaming and Technology, possibly because of other sentences in the same document. Even noun-phrases, such as 'Mario Winans' in Figure 1, which presumably should be assigned to the same topic, are split across topics.

In the context of downstream tasks, global evaluation ignores that local topic assignments are often used as features. If the topic assignments are inaccurate, the accuracy of the classifier may suffer.

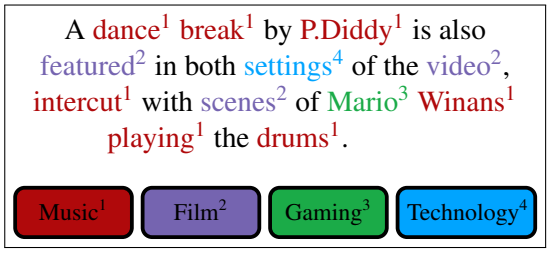

Figure 1: Topic assignments from LDA on a sentence from a Wikipedia document. Notice that even noun-phrases are split in a way which is bewildering to users.

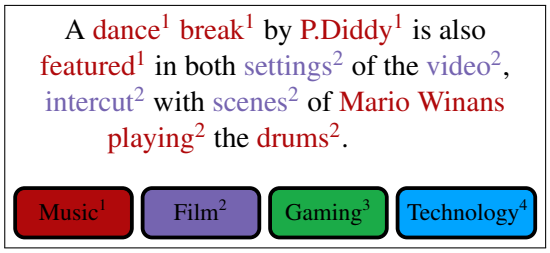

Figure 2: An example of how topics might be assigned if done by a human.

The literature surrounding this issue focuses on improving local topic assignments, but no metrics that specifically assess the quality of these assignments exist. Instead, the literature evaluates models with global metrics or subjective examination.

For example, HMM-LDA (Griffiths et al., 2004) integrates syntax and topics by generating words from a special syntax-specific topic. TagLDA (Zhu et al., 2006) adds a tag-specific word distribution for each topic, allowing syntax to impose local topic structure. The syntactic topic model (Boyd-Graber and Blei, 2009, STM) extends this idea and generates topics using a parse tree. An alternative approach to improving local topic quality is by adding a Markov property: the hidden topic Markov model (Gruber et al., 2007, HTMM) adds a switch variable on each token which determines whether to reuse the previous topic or generate a new topic. More recently, SentenceLDA (Balikas et al., 2016a) assigns each sentence to a single topic. CopulaLDA (Balikas et al., 2016b) supersedes SentenceLDA, instead using copulas to impose topic consistency within each sentence of a document.

This paper evaluates token-level topic assignment quality to understand which topic models produce meaningful local topics for individual documents and proposes metrics that correlate with human judgment of the quality of these assignments.

2 Global Evaluation

Prior work in automated metrics to evaluate topic model quality primarily deals with global evaluations (i.e. evaluations of the topic-word distributions that represent topics). Early topic models such as LDA were typically evaluated using held-out likelihood or perplexity (Blei et al., 2003; Wallach et al., 2009). Indeed, perplexity is still frequently used to evaluate models, and each of the models mentioned in the previous section, including CopulaLDA—designed to improve local topic quality—uses perplexity to evaluate the model. However, while held-out perplexity can test the generalization of predictive models, it is negatively correlated with human evaluations of global topic quality (Chang et al., 2009). This result comes from a topic-word intrusion task, in which human evaluators must identify a randomly chosen 'intruder' word which was injected into the top n most probable words in a topic-word distribution. If a topic is semantically coherent, then the intruder will be easy to identify.

2.1 Coherence

While human evaluation of topic coherence is useful, automated evaluations are easier to deploy. Consequently, Newman et al. (2010) proposed a variety of automated evaluations of topic coherence and correlated these metrics with human evaluations using the topic-word intrusion task mentioned above and showed that an evaluation based on aggregating pointwise mutual information (PMI) scores across the most likely terms in a topic distribution correlates well with human evaluations. In fact, there are multiple metrics referred to as 'coherence', including Newman et al. (2010); Mimno et al. (2011) and Lau et al. (2014), as well as some more recent exploration of coherence (Röder et al., 2015; Lau and Baldwin, 2016). All of these 'coherence' metrics are measures of global topic quality, since they consider only the global topic-word distributions. For consistency with Arora et al. (2013), we use the Mimno et al. (2011) formulation of coherence in our evaluations, and use this automated evaluation as a proxy for human evaluations using topic-intrusion tasks. Because automated evaluation is known to correlate with human evaluations of global topic quality, we do not investigate global topic quality with any additional user evaluations.

2.2 Beyond the Top Words

While topics are typically summarized by their top n most probable words, the entire distribution is important for downstream tasks, like classification. Consider two topics which rank the words of the vocabulary by probability in the same order. Suppose that one of these distributions is more uniform than the other (i.e., has higher entropy). While both ranked word lists are identical, the topic-word distribution with lower entropy places more weight on the high-rank words and is much more specific.

Using this intuition, AlSumait et al. (2009) developed metrics for evaluating topic significance. While this work was originally used to rank topics, it also characterizes entire models by measuring average significance across all topics in a single model (Lund et al., 2017).

Topic significance is the distance between a topic distribution and a background distribution—for instance, either the uniform distribution (SIGUNI) or the empirical distribution of words in the corpus, which we call the vacuous distribution (SIGVAC).

Like coherence, topic significance is a global measure of topic quality; it considers topic-word distributions only and ignores local topic assignments. However, unlike topic coherence, it considers the *entire* topic distribution. When topics are features for document classification, topics with similar coherence can evince disparate downstream classification accuracy (Lund et al., 2017). However, significant topics are consistently more accurate.

Despite the proven success of automated *global metrics*, no automatic metric evaluates *local* topic quality. Before directly addressing this need we will first obtain human judgements of local topic quality and use them to assess existing global metrics of topic quality. We obtain these judgments through the crowdsourcing task described below.

3 Crowdsourcing Task

Following the general design philosophy in developing the coherence metric in Newman et al. (2010), we train a variety of models on various datasets to obtain data with varying token-level topic quality. We then evaluate these models using crowdsourcing on a task designed to elicit human evaluation of local topic model quality. By correlating the human evaluation with existing, global metrics, we identify the deficiencies of global metrics and propose new metrics to better measure local topic quality.

Dataset	Documents	Tokens	Vocabulary
Amazon	39388	1389171	3406
Newsgroups	18748	1045793	2578
New York Times	9997	2190595	3328

Table 1: Statistics on datasets used in user study and metric evaluation.

3.1 Datasets and Models

We choose three datasets from domains with different writing styles. These datasets include: Amazon product reviews[1], free-form discussion from the well-known Twenty Newsgroups dataset (Lang, 2007), and formal news reporting from the New York Times (Sandhaus, 2008). We apply stopword removal and also remove any token which does not appear in at least 100 documents within a given dataset. Statistics for these three datasets can be found in Table 1.

Once again aiming for a wide variety of topic models for our evaluation, for each of these datasets, we train three types of topic models. As a baseline, we train Latent Dirichlet Allocation (Blei et al., 2003) on each of the three datasets using gensim defaults.[2] CopulaLDA (Balikas et al., 2016b) is the most recent and reportedly best on local topic quality; we use the authors' implementation and parameters. Finally, Anchor Words algorithm (Arora et al., 2013) is a fast and scalable alternative to traditional inference techniques based on non-negative matrix factorization. Our implementation of Anchor Words only considers words as candidate anchors if they appear in at least 500 documents, the dimensionality of the reduced space is 1000, and the threshold for exponentiated gradient descent is 1e-10. By itself, Anchor Words only recovers the topic-word distributions; we follow Nguyen et al. (2015) and use variational inference for LDA with fixed topics to assign each token to a topic.

In addition to varying the datasets and topic modeling algorithms, we also vary the number of topics. For both LDA and Anchor Words, we use 20, 50, 100, 150, and 200 topics. For CopulaLDA, we use 20, 50, and 100 topics.[3] Small

[1] http://jmcauley.ucsd.edu/data/amazon/

[2] https://radimrehurek.com/gensim

[3] Unfortunately, CopulaLDA does not scale beyond 100

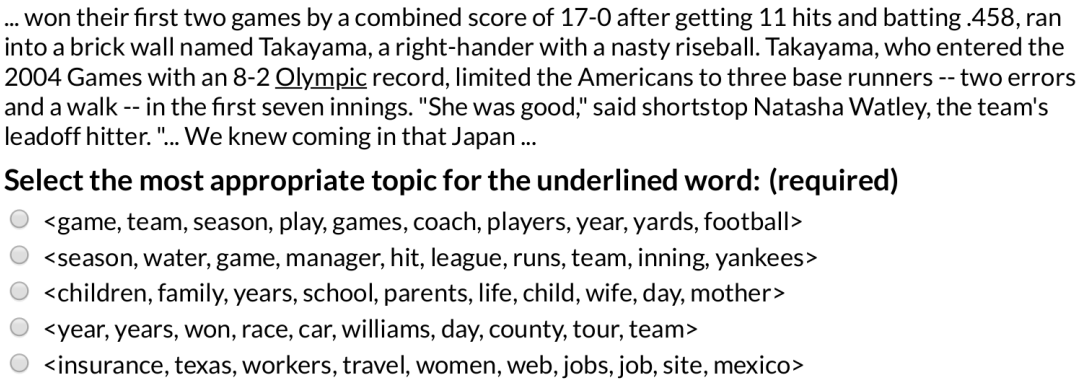

Figure 3: Example of the topic-word matching task. Users are asked to select the topic which best explains the underlined token ("Olympic").

topic models have a few coherent—albeit less significant—topics, while large topic models have many significant topics. Since each model includes non-determinism, we train five instances of each dataset, model, and topic cardinality and average our results (Nguyen et al., 2014, "Multiple Final").

In the interest of reproducibility, the data, the scripts for importing and preprocessing the data, and the code for training and evaluating these topic models are available.[4]

3.2 Task Design

The goal for our crowdsourcing task is to have human annotators evaluate local topic quality. Not only will this task allow us to evaluate and compare topic models themselves, but it will also allow us to determine the effectiveness of automated metrics. Because local topic quality is subjective, directly asking annotators to judge assignment quality can result in poor inter-annotator agreement. Instead, we prefer to ask users to perform a task which illuminates the underlying quality indirectly. This parallels the reliance on the word intrusion task to rate topic coherence and topic intrusion to rate document coherence (Chang et al., 2009).

We call this proposed task 'topic-word matching'. Like Chang (2010), we show the annotator a short snippet from the data with a single token underlined along with five topic summaries (i.e., the 10 most probable words in the topic-word distribution). We then ask the user to select the topic which best fits the underlined token (Figure 3). One of the five options is the topic that the model actually assigns to the underlined token. The intuition is that the annotator will agree more often with a topic model which makes accurate local topic assignments. As alternatives to the model-selected topic for the token, we also include the three most probable topics in the document, excluding the topic assigned to the underlined token. A model which gives high quality token-level topic assignments should consistently choose the best possible topic for each individual token, even if these topics are closely related. Finally, we include a randomly selected intruder topic as a fifth option. This fifth option is included to help distinguish between an instance where the user sees equally reasonable topics for the underlined token (in which case, the intruding topic will not be selected), and when there are no reasonable options for the underlined token (in which case, all five topics are equally likely to be chosen).

We note the similarity between the topic-word matching task and the task of constructing lexical chains (Hirst et al., 1998). While the relationship between topic modeling and lexical chains has been explored (Chiru et al., 2014; Joty et al., 2010), our task is unique in that it asks users to consider a single word in isolation, rather than to consider any relationship between words in a chain.

topics. In contrast to LDA and Anchor Words, which run in minutes and seconds respectively, CopulaLDA takes days to run using the original authors' implementation. Our attempts to run it with 150 and 200 topics never finished and were finally killed due to excessive memory consumption on 32GB systems.

[4] https://github.com/jefflund/ankura

For each of our 39 trained models (i.e., for each model type, dataset, and topic cardinality), we randomly select 1,000 tokens to annotate. For each of the 39,000 selected tokens, we obtain five judgments. We aggregate the five judgments by selecting the contributor response with the highest confidence, with agreement weighted by contributor trust. Contributor trust is based on accuracy on test questions.

We deploy this task on a popular crowdsourcing website[5] and pay contributors $0.12 USD per page, with 10 annotations per page. For quality control on this task, each page contains one test question. The test questions in our initial pilot study are questions we hand-select with an obvious correct answer. For our test questions in the final study, we use the ones mentioned above in addition to questions from the pilot studies with both high annotator confidence and perfect agreement. We require that contributors maintain at least a 70% accuracy on test questions throughout the job. We also require that they spend at least 30 seconds per page. This restriction is simply to prevent contributors from blindly completing the task; we expect that most contributors will require more than 30 seconds per page. We impose no other constraints on contributors.

3.3 Agreement Results

We first measure inter-annotator agreement using Krippendorff's alpha with a nominal level of measurement (Krippendorff, 2013). Generally , $\alpha = 1$ indicates perfect reliability, while $\alpha < 0$ indicates systematic disagreement. Over all the judgments we obtain, we compute a value of $\alpha = 0.44$, which indicates a moderate level of agreement.

When using crowdsourcing, particularly with subjective tasks such as topic-word matching, we expect somewhat lower inter-annotator agreement. However, previous work indicates that when properly aggregated, we can still filter out noisy judgments and obtain reasonable opinions (Nowak and Rüger, 2010).

Figure 4 summarizes the human agreement with the three different model types. Surprisingly, despite claiming to produce superior local topic quality, CopulaLDA actually has lower agreement than LDA on the topic-word matching task.

Users agree with Anchor Words more often than LDA by a wide margin. However, in terms of global topic quality, Anchor Words is roughly similar to LDA (Arora et al., 2013). It is important to note that Anchor Words only discovers the global topics, while variational inference assigns those topics to each token. We discuss this further in Section 6.

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

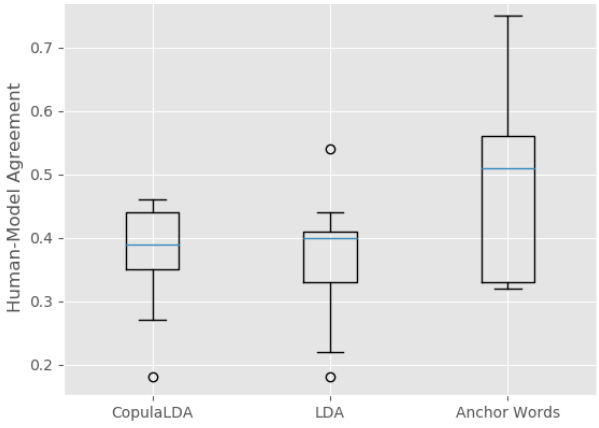

Figure 4: Human agreement with each model type. CopulaLDA performs slightly worse than LDA. Humans preferred topic assignments from Anchor Words by a wide margin.

Metric	Amazon	Newsgroups	New York Times
SIGVAC	0.6960	0.6081	0.6063
SIGUNI	0.6310	0.4839	0.4935
COHERENCE	0.4907	0.4463	0.3799

Table 2: Coefficient of determination (r^2) between global metrics and crowdsourced topic-word matching annotations.

3.4 Global Metrics Correlation

For coherence and significance human-model agreement on the topic-word matching task, Table 2 reports the coefficient of determination (r^2) for each global metric and dataset. While global metrics do correlate somewhat with human judgment of local topic quality, the correlation is moderate to poor, especially for coherence, and we propose new metrics that will achieve greater correlation with human evaluations.

4 Proposed Metrics

We develop an automated methodology for evaluating local topic model quality. Following the pattern used by Newman et al. (2010) to develop coherence, we propose potential metrics to better reflect token-level topic quality, such as that in Figure 2. As with coherence, we correlate these automated metrics with human evaluations in order to

determine which automated metric yields the most accurate estimate of local topic quality, as judged by human annotators.

Topic Switch Percent (SWITCHP) It is a platitude of good writing that a sentence expresses one idea (Williams, 1990), and by this logic we would expect the topic assignments in a sentence or local token cluster to be consistent. Our first metric measures the percentage of times a topic switch occurs relative to the number of times a switch could occur (and a switch is possible after every token but the last). The intuition is that tokens near each other should switch infrequently, and thus be consistent in expressing a single idea. In a corpus with n tokens, with z_i the topic assignment of the i^{th} token in the corpus, and $\delta(i, j)$ being the Kronecker delta function, we measure this consistency with

$$\frac{1}{n-1}\sum_{i=1}^{n-1}\delta(z_i, z_{i+1}). \tag{1}$$

Topic Switch Variation of Information (SWITCHVI) SWITCHP penalizes all switches equally, but intuitively there are probably times when a sentence or local cluster expresses multiple ideas. Figure 2 has a noun phrase at the beginning referencing P. Diddy, but then switches to talking about music videos, a reasonable switch in this case. This would be penalized by metrics like SWITCHP, but SWITCHVI focuses on whether the *distribution over topics* is different when switches happen.

To capture this, we build two partitions: source topics S and target topics T. These partitions encode the difference between distributions. Source captures what topics change *from*—the empirical distribution over topics in the document, and target captures what topics change *to*—the distribution over topics of token j given that $z_{j-1} \neq z_j$. SWITCHVI measures the difference between the distributions over topics in these two partitions by measuring mutual information.

We use variation of information (or VI) to measure the amount of information lost in changing from one partition to another (Meilă, 2003). Assuming that our model has K topics, and once again using z_i as the topic assignment for token w_i, we consider two partitions $S = \{S_1, ..., S_K\}$ and $T = \{T_1, ..., T_K\}$ of the set of tokens w, such that $S_i = \{w_j \,|\, z_j = i\}$ and $T_i = \{w_j \,|\, z_{j+1} = i\}$.

Variation of information is

$$\mathbb{H}_z\left[S\right] + \mathbb{H}_z\left[T\right] - 2\mathbb{MI}\left(S, T\right), \tag{2}$$

where $\mathbb{H}_z\left[\cdot\right]$ is entropy with respect to topic distribution and $\mathbb{MI}\left(S, T\right)$ is the mutual information between S and T. In other words, we measure how much information we lose in our topic assignments if we reassign every token to the topic of the token that follows.

Window Probabilities (WINDOW) Modifying slightly the intuition behind SWITCHP, WINDOW rewards topic models which have topic assignments which not only explain *individual* tokens, but also the tokens within a *window* around the assignment. This will give a high score if the words surrounding word i have a high probability in the topic z_i (regardless of the topic assignments of those surrounding words).

Consider a topic model with K topics, V token types, and D documents with topic-word distributions given by a $K \times V$ matrix ϕ such that $\phi_{i,j}$ is the conditional probability of word j given topic i. Given a window size s, we compute:

$$\frac{1}{n(2s+1)}\sum_{i}^{n}\sum_{j=i-s}^{i+s}\phi_{z_i, w_j}\ . \tag{3}$$

Our experiments use a window size of three ($s = 1$), meaning that for each token we consider the probability of seeing it in the assigned topic z_i, as well as the probabilities of seeing the tokens immediately preceding and following the target token in topic z_i. This maintains consistency while allowing for topics to switch mid-sentence.

Topic-Word Divergence (WORDDIV) Stepping away from human intuition about the structure of sentences and topics, we imagine a statistical metric that resembles traditional likelihood metrics for topic models.[6] A reminder that $\phi_{i,j}$ is the topics (K) by vocabulary (V) matrix representing the conditional probability of word j given topic i. Furthermore, let θ_d be the K-dimension document-topic distribution for the dth document and ψ_d be the V-dimensional distribution of words for document d. This metric measures how well the topic-word probabilities explain the tokens

[6]The connection to likelihood via a matrix factorization perspective (Arora et al., 2012).

which are assigned to those topics:

$$\frac{1}{D}\sum_{d}^{D} \textsc{js}\left(\theta_d \cdot \phi \,||\, \psi_d\right) \tag{4}$$

where $\textsc{js}\left(P \,||\, Q\right)$ is the Jensen-Shannon divergence between the distributions P and Q. Like traditional likelihood metrics, this evaluation scores high on a document when the topics used in that document explain the overall topic document distribution, regardless of the local topic assignments.

Average Rank (AVGRANK) As an alternative to traditional likelihood metrics, which examine the fitness of specific model parameters, AVGRANK looks at the relative rank of words in their topics; a common way of presenting topics to humans is as a set of related words (the most probable words in the topic-word distributions).

Rather than WORDDIV's focus on specific word probabilities, this metric rewards word types that are probable in the topic (regardless of the absolute probability of the type). Leveraging this intuition, where $\text{rank}(w_i, z_i)$ is the rank of i^{th} word w_i in its assigned topic z_i when sorted by probability, we define AVGRANK as

$$\frac{1}{n}\sum_{i=1}^{n} \text{rank}(w_i, z_i). \tag{5}$$

With this evaluation the lower bound is 1, although this would require that every token be assigned to a topic for which its word is the mode. However, this is only possible if the number of topics is equal to the vocabulary size.

5 Automated Evaluations

As before, for each of our proposed metrics, we compute a least-squares regression for both the proposed metric and the human-model agreement on the topic-word matching task (Table 3).

Humans agree more often with models from Amazon reviews than on New York Times. This likely reflects the underlying data: Amazon product reviews are highly focused on specific products and features, and the generated topics naturally reflect these. In contrast, New York Times data deal with a much wider array of subjects and treats them with nuance and detail—if for no other reason than that the articles are longer—not typically found in product reviews. This makes the judgment of topic assignment more difficult and subjective.

	Metric	Amazon	Newsgroups	New York Times
Local	SWITCHP	0.9077	0.8737	0.7022
	SWITCHVI	0.8485	0.8181	0.6977
	AVGRANK	0.5103	0.5089	0.4473
	WINDOW	0.4884	0.3024	0.1127
	WORDDIV	0.3112	0.2197	0.0836
Global	SIGVAC	0.6960	0.6081	0.6063
	SIGUNI	0.6310	0.4839	0.4935
	COHERENCE	0.4907	0.4463	0.3799

Table 3: Coefficient of determination (r^2) between automated metrics and crowdsourced topic-word matching annotations. We include metrics measuring both local topic quality and global topic quality. The global values are included for comparisons from Table 2. SWITCHP often has a higher correlation with human annotations.

Despite differences across datasets, SWITCHP most closely approximates human judgments of local topic quality, with an r^2 which indicates a strong correlation. This suggests that when humans examine token-level topic assignments, they are unlikely to expect topic switches from one token to the next (Figure 2). As evidenced by the lower r^2 for SWITCHVI, even switching between related topics does not seem to line up with human judgments of local topic quality.

Again, there is a correlation between coherence and the topic-word matching task, although the correlation is only moderate. Similarly, word-based significance metrics have a moderate correlation with topic-word matching. We maintain that these global topic metrics are important measures for topic model quality, but they fail to capture local topic quality as SWITCHP does.

6 Discussion

Considering the intuition gained from the motivating example in Figure 1, it is not surprising that humans would prefer topic models which are locally consistent. Thus, our result that SWITCHP is correlated with human judgments of local topic quality best parallels that intuition.

However, our annotators are only shown the potential topic assignments for a single token and do not know what topics have been assigned to the surrounding tokens. This is in contrast to Chang (2010), who use richer interactions—going from documents to topic assignments—to *build* models; our focus is instead on evaluation. Despite this, our annotators apparently prefer models which are consistent. While the result is intuitive, it is surprising a tasks that asks for a single token can discover it.

Given our results, we recommend that topic switch percent be adopted as an automated metric to measure the quality of token-level topic assignments. We would refer to this metric colloquially as 'consistency' in the same way that PMI scores on the top n words of a topic are referred to as 'coherence'. We advocate that future work on new topic models include validation with respect to topic consistency, just as recent work has included evaluation of topic coherence.

However, topic consistency should not be used to the exclusion of other measures of topic model quality. After all, topic consistency is trivially maximized by minimizing topic switches without regard to the appropriateness of the topic assignment. Instead, we advocate that future models be evaluated with respect to global topic quality (e.g., coherence, significance, perplexity) as well as local topic quality (i.e., consistency). These measures, in addition to evaluation of applicable downstream tasks (e.g., classification accuracy), will give practitioners the information necessary to make informed decisions about topic model selection.

Moreover, our work leaves open questions on which models best satisfy local consistency. For instance, Anchor Words finds topics but assigns local topics with variational inference; a natural question is whether variational inference by itself finds locally consistent topics.

7 Conclusion

We develop a novel crowdsourcing task, which we call topic-word matching, to illicit human judgments of local topic model quality. We apply this human evaluation to a wide variety of models, and find that topic switch percent (or SWITCHP) correlates well with this human evaluation. We propose that this new metric, which we colloquially refer to as consistency, be adopted alongside evaluations of global topic quality for future work with topic model comparison.

Acknowledgements

This work was supported by the collaborative NSF Grant IIS-1409287 (UMD) and ISS-1409739 (BYU). Boyd-Graber is also supported by NSF grant IIS-1822494 and IIS-1748663. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *Proceedings of European Conference of Machine Learning*.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference of Machine Learning*.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models—going beyond svd. In *Proceedings of Foundations of Computer Science*.

Georgios Balikas, Massih-Reza Amini, and Marianne Clausel. 2016a. On a topic model for sentences. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Georgios Balikas, Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. 2016b. Modeling topic dependencies in semantically coherent text spans with copulas. In *Proceedings of International Conference on Computational Linguistics*.

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

Jordan L Boyd-Graber and David M Blei. 2009. Syntactic topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Jonathan Chang. 2010. Not-so-latent Dirichlet allocation: Collapsed Gibbs sampling using human judgments. In *NAACL Workshop: Creating Speech and Language Data With Amazon'ss Mechanical Turk*.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*.

Costin-Gabriel Chiru, Traian Rebedea, and Silvia Ciotec. 2014. Comparison between lsa-lda-lexical chains. In *WEBIST (2)*, pages 255–262.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating topics and syntax. In *Advances in neural information processing systems*.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Artificial intelligence and statistics*.

Graeme Hirst, David St-Onge, et al. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond T Ng. 2010. Exploiting conversation structure in unsupervised topic segmentation for emails. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 388–398. Association for Computational Linguistics.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*, 3rd edition, pages 221–250. Thousand Oaks.

Ken Lang. 2007. 20 newsgroups data set. Http://www.ai.mit.edu/people/jrennie/20Newsgroups/.

Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–487.

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

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the Association for Computational Linguistics*.

Jeffrey Lund, Stephen Cowley, Wilson Fearn, Emily Hales, and Kevin Seppi. 2018. Labeled anchors and a scalable, transparent, and interactive classifier. In *Proceedings of Empirical Methods in Natural Language Processing*.

Marina Meilă. 2003. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer.

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

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the Association for Computational Linguistics*.

Thang Nguyen, Jordan Boyd-Graber, Jeffrey Lund, Kevin Seppi, and Eric Ringger. 2015. Is your anchor going up or down? Fast and accurate supervised topic models. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2014. Sometimes average is best: The importance of averaging for prediction using mcmc inference in topic modeling. In *Proceedings of Empirical Methods in Natural Language Processing*.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Evan Sandhaus. 2008. The New York Times annotated corpus. Http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the International Conference of Machine Learning*. ACM.

Joseph M. Williams. 1990. *Style: Toward Clarity and Grace*. University of Chicago Press.

Xiaojin Zhu, David Blei, and John Lafferty. 2006. TagLDA: Bringing document structure knowledge into topic models. Technical report, Technical Report TR-1553, University of Wisconsin.