A Novel Measure for Coherence in Statistical Topic Models

Fred Morstatter and Huan Liu
Arizona State University
Tempe, Arizona, USA
{fred.morstatter, huan.liu}@asu.edu

Abstract

Big data presents new challenges for understanding large text corpora. Topic modeling algorithms help understand the underlying patterns, or "topics", in data. Researchersauthor often read these topics in order to gain an understanding of the underlying corpus. It is important to evaluate the interpretability of these automatically generated topics. Methods have previously been designed to use crowdsourcing platforms to measure interpretability. In this paper, we demonstrate the necessity of a key concept, coherence, when assessing the topics and propose an effective method for its measurement. We show that the proposed measure of coherence captures a different aspect of the topics than existing measures. We further study the automation of these topic measures for scalability and reproducibility, showing that these measures can be automated.

1 Introduction

Big data poses new challenges in analyzing text corpora. Topic modeling algorithms have recently grown to popularity for their ability to help discover the underlying topics in a corpus. Topic words are the words selected to represent a topic. They have been shown to be useful in the areas of machine learning, text analysis (Grimmer and Stewart, 2013), and social media analysis (O'Connor et al., 2010), among others. Topic models can be used as predictive models to classify new documents in the context of the training corpus. They are evaluated by measuring their predictive performance on a held-out set of documents. Topic models can also be inspected manually by a human to understand the themes of the underlying corpus. A widely adopted way is suggested by (Chang et al., 2009): it measures the quality of a topic by inspecting how far topic words are from some random words. The idea is that the quality of a topic can be measured by how far topic words are from some random words. In other words, if human evaluators can consistently separate random words from topic words, these topics are good, otherwise, they are not good. An advantage of this measure is that it can be easily implemented to deploy on a crowd-sourcing platform like Amazon's Mechanical Turk.

Assuming that random words represent random topics, we can name the above method "between-topic" measure. In this paper, we hypothesize that this measure considers just one important aspect in assessing the quality of statistical topics. Specifically, we investigate the topic interpretability by examining the "coherence" of a topic generated by topic modeling algorithms, i.e., how close topic words are within a topic. Thus, this measure is a "within-topic" measure. Two immediate challenging questions are: (1) without knowing ground truth of topic coherence, how can we design an equally effective method like "between-topic" measure for crowd-sourcing evaluation? and (2) how different is this "within-topic" coherence measure from the existing "between-topic" measure? We elaborate how we answer these two challenges by starting with some related work, showing how the "between-topic" measure faces difficulty in measuring coherence, and presenting our proposal of a coherence measure.

2 Related Work

Topic modeling is pervasive, and has been widely accepted across many communities such as machine learning and social sciences (Ramage et al., 2009; Schmidt, 2012; Yang et al., 2011). One of

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 543–548,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

the reasons for the wide appreciation of these algorithms is their ability to find underlying topics in enormous sets of data (Blei, 2012). More recently topic modeling has been widely applied to social media data (Kireyev et al., 2009; Joseph et al., 2012; Morstatter et al., 2013), *e.g.* (Yin et al., 2011; Hong et al., 2012; Pozdnoukhov and Kaiser, 2011) focus on identifying topics in geographical Twitter datasets. In (Kumar et al., 2013; Mimno et al., 2011), the authors had to employ subject-matter experts to assess topic quality. These manual topic labels can be supplemented with automatic labeling algorithms (Maiya et al., 2013). While these works attempt to ensure topic quality by employing domain experts, these are highly domain-specific cases. The measures we discuss going forward are more general, and can be applied to topic models trained with text data.

The most important point of comparison between our work and others lies in the Model Precision measure proposed in (Chang et al., 2009). The insight of this measure is that a good topic is one whose top few words are distant, or highly separate, from randomly-selected words. Their task works by showing several human participants, or Turker, the top 5 words from a topic and one randomly-chosen, low-ranking "intruded" word. The humans are then asked to select the word that they think was intruded. The measure then estimates the topic's quality by calculating the number of times the humans correctly guessed the intruded word. While Word Intrusion provides insight into a topic's interpretability, the key assumption is that topic goodness comes only from the top words being separate from a randomly-selected word. This measure does not offer any insight about the coherence of the top words. We propose a new measure which complements Word Intrusion by measuring distance *within* a topic.

(Lau et al., 2014) built a machine learning algorithm to automatically detect the intruded word in a topic. Methods for evaluating topic models were proposed in (Wallach et al., 2009). We investigate the applicability of this measure in our work.

3 Model Precision Quandary

Model Precision works by asking the user to choose the word that does not fit within the rest of the set. We are measuring the top words in the topic by comparing them to an outlier. While this method has merit, it does not help us understand the coherence *within* the top words for the topic.

A diagram illustrating this phenomenon is shown in Figure 1. In Figure 1(a), we see a coherent topic. This topic is coherent because all 5 of the top words are close together, while the intruded word is far away. In Figure 1(b) we see a topic that is less coherent because the fifth word lies at a distance from the first four. In both cases, Model Precision gives us the intruder word in the topic, as seen in Figures 1(c), and 1(d). While this is the desired performance of Model Precision, it leaves us with no understanding of the coherence of the top words of the topic. Results are masked by the outlier, and do not give information about the intra-cluster distance, or coherence of the topic.

In light of this, we look for a way to separate topics not just by their distance from an outlier, but also by the distance within the top words in the topic. The next section of this paper investigates a method which can measure not just the intruder word, but also the coherence of the top words in the topic. In this way we separate topics such as those shown in Figure 1 based on the coherence of their top words.

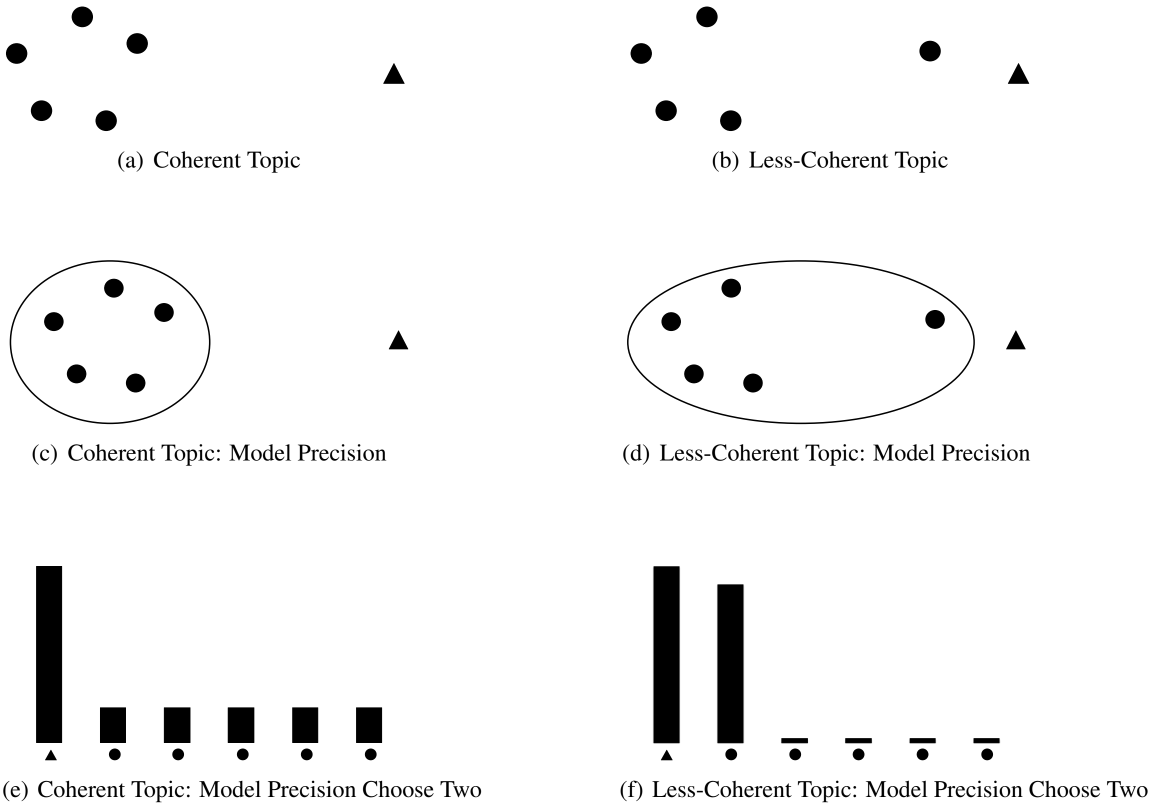
4 Word Intrusion Choose Two

In this section we propose a new experiment that measures the interpretability of the top words of a topic. This experiment sets up the task as before: we select the top five words from a topic, and inject one low-probability word. The key difference is that we ask the Turker to select *two* intruded words among the six.

The intuition behind this experiment is that the Turkers' first choice will be the intruded word, just as in Model Precision. However, their second choice is what makes the topic's quality clear. In a coherent topic the Turkers won't be able to distinguish a second word as all of the words will seem similar. A graphical representation of this phenomenon is shown in Figure 1(e). In the case of an incoherent, a strong "second-place" contender will emerge as the Turkers identify a 2nd intruder word, as in Figure 1(f).

4.1 Experimental Setup

To perform this experiment, we inject *one* low-probability word for each topic, and we ask the Turkers to select *two* words that do not fit within the group. We show the six words to the Turker in random order with the following prompt:



(a) Coherent Topic

(b) Less-Coherent Topic

(c) Coherent Topic: Model Precision

(d) Less-Coherent Topic: Model Precision

(e) Coherent Topic: Model Precision Choose Two

(f) Less-Coherent Topic: Model Precision Choose Two

Figure 1: Comparison between Model Precision, and Model Precision Choose Two for a toy topic. Circles represent the top words and triangles represent intruded words. Model Precision Choose Two can distinguish the less-coherent topic.

You will be shown six words. Four words belong together, and two of them do not. Choose two words that do **not** belong in the group.

Coherent topics will cause the Turkers' responses regarding the second intruded word to be unpredictable. Thus, our measure of the goodness of the topic should be the predictability of the Turkers' second choice. We propose a new measure called "Model Precision Choose Two" to measure this. Model Precision Choose Two (MPCT) measures this spread as the peakedness of the probability distribution. We define $MPCT_k^m$ for topic k on model m as:

$$MPCT_k^m = H(p_{turk}(\mathbf{w}_{k,1}^m), ..., p_{turk}(\mathbf{w}_{k,5}^m)), \tag{1}$$

where $H(\cdot)$ is the Shannon entropy (Cover and Thomas, 2006), $\mathbf{w}_k^m$ is the vector of the top words in topic k generated by model m, and $p_{turk}(\mathbf{w}_{k,i}^m)$ is the probability that a Turker selects $\mathbf{w}_{k,i}^m$. This measures the strength of the second-place candidate, with higher values indicating a smoother, more even distribution, and lower values indicating Turkers gravitation towards a second word.

The intuition behind choosing entropy is that it will measure the unpredictability in the Turker selections. That is, if the Turkers are confused about which second word to choose, then their answers will be scattered amongst the remaining five words. As a result, the entropy will be *high*. Conversely, if the second word is obvious, the Turkers will begin to congregate around that second choice, meaning that their answers will be focused. As a result, the entropy will be *low*. Because entropy is able to measure the confusion of the Turkers responses about the second word, we use it directly in the design of our measure.

4.2 Data

The data used in this study consists of articles from English Wikipedia. We sample 10,000 articles uniformly at random from across the dataset. We selected articles containing more than 50 words. In preprocessing we stripped case, removed punctuation, stopwords, and words consisting entirely of numbers. This process yields a corpus containing 10,000 documents, 4,200,174 tokens, and

Table 1: Example topics showing the variance of $MPCT$ when $MP = 1.0$.

MPCT	Top Five Words	Intruded Word
0.202	canada, canadian, north, ontario, http	shipping
0.373	language, century, word, english, greek	drew
0.407	river, highway, road, north, route	berea
0.569	born, children, family, life, father	boatsman
0.795	design, engine, model, power, system	resynthesized
0.946	railway, station, road, line, route	anagarika
1.000	film, series, show, television, films	bubblegrunge

196,219 types.

The topic modeling algorithm used is latent Dirichlet allocation (LDA) (Blei et al., 2003). To build the models used in the experiments, we run LDA on the Wikipedia corpus using values of $K = \{10, 25, 50, 100\}$ with the Mallet package (McCallum, 2002). This yields 4 models and 185 total topics. The model generated by each value of K is denoted by m in the equations.

4.3 Experimental Results

The results of this experiment, aggregated by model, are shown in Figure 2. We see that as the value of K increases, the median score for MPCT stays roughly the same. We compute the Spearman's ρ correlation coefficient (Spearman, 1904) between the MP and $MPCT$ measures, and find that the measures have $\rho = 0.09$. This lack of correlation indicates that this measure is assessing a different dimension of the topics.

To help explain these results, we provide some examples of topics that received different MPCT scores with a perfect separateness (MP) score in Table 1. We see that although all of the topics have perfect scores along this dimension, their cohesiveness score varies. This is due to the Turkers' agreement about the second intruded word.

5 Automating Model Precision Choose Two

The crowdsourced experiments carried out in this paper provide a complementary understanding of how humans understand the topics that are generated using statistical topic models. One drawback of these methods lies in the difficulty of reproducing these experiments. This difficulty comes from two sources: 1) the monetary cost of employing the Turkers to solve the HITs, and 2) the time cost to build the surveys and to collect the results. To overcome these issues, we propose automated methods that can estimate the topics' performance

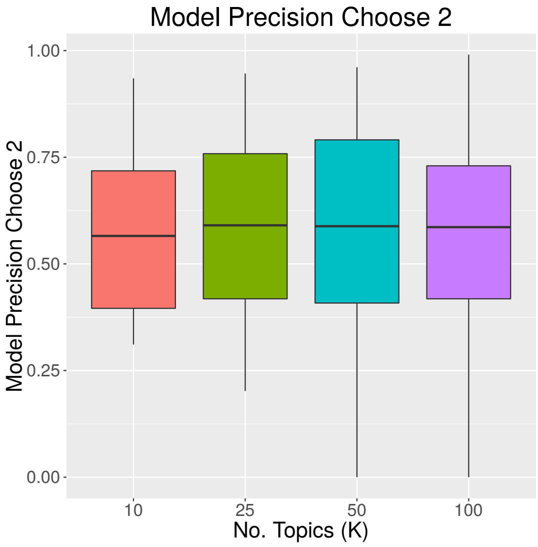


Figure 2: Model Precision Choose Two across the four models used in this work. Higher scores are better. We see that as K increases, the median score does not improve noticeably.

along these different dimensions. These measures can be used by future researchers to automatically gauge their topics.

We test several automated measures for their ability to predict the outcome of the crowdsourced measures. To test these measures, we calculate the Spearman's ρ between the automated measure of the topic and the crowdsourced measure. The automated measures we propose are as follows:

1. **Topic Size:** LDA assigns a topic label to each *token* in the dataset. Topic size measures the number of tokens assigned to the topic by the LDA model, where more tokens indicates a larger topic. This has been tested in (Mimno et al., 2011).

2. **Topic Entropy:** The entropy of the entire probability distribution for the topic. High entropy indicates a flat distribution of probabilities, while low entropy indicates a peaked distribution around the first few words.

3. **Mimno Co-Occurrence:** Measures the frequency of the top words co-occurring within the same document. Proposed in (Mimno et al., 2011), and measured as:

$$MCO(\mathbf{w}) = \sum_{j=2}^{|\mathbf{w}|} \sum_{k=1}^{j-1} log \frac{D(\mathbf{w}_j, \mathbf{w}_k) + 1}{D(\mathbf{w}_k)}, \tag{2}$$

Table 2: Performance of automated measures in approximating the crowdsourced experiments. All values are Spearman's ρ correlation coefficients with the crowdsourced measure.

	Automated Measure	MPCT
1.	**Topic Size**	-0.572
2.	**Topic Entropy**	-0.539
3.	**Mimno**	-0.438
4.	**No. Word Senses**	-0.456
5.	**Avg. Pairwise JCD**	**-0.844**
6.	**Mean-Link JCD**	-0.434
7.	**NPMI**	-0.582

where $\mathbf{w}$ is the vector of the top 20 words in the topic, and $D(\cdot)$ returns the number of times the words co-occur in any document in the corpus.

4. **No. Word Senses:** The total number of word senses, according to WordNet, of the top five words in the topic. This varies slightly from the measure proposed in (Chang et al., 2009), where the authors also consider the intruded word. Because the intruded word is generally far away, we exclude it from our calculation.

5. **Avg. Pairwise Jiang-Conrath Distance:** The Jiang-Conrath (Jiang and Conrath, 1997) distance (JCD) is a measure of *semantic similarity*, or coherence, that considers the lowest common subsumer according to WordNet. Here we compute the average JCD of all $\binom{5}{2} = 10$ pairs of the top five words of the topic. This approach was introduced by (Chang et al., 2009), however we modify it slightly to only consider the top five words in the topic.

6. **Mean-Link JCD:** Using the JCD measure as before, we compute the average distance from the intruded word to each of the top 5 words from the topic.

7. **Normalized Pointwise Mutual Information (NPMI):** NPMI measures the association between the top words in a topic. It is normalized to yield a score of 1 in the case of perfect association. This measure was first introduced by (Bouma, 2009). We use the calculation adapted for the problem of estimating a topic's performance introduced in (Lau et al., 2014).

We calculate the correlation between all automated methods and MPCT, shown in Table 2. MPCT is best predicted using the Avg. Pairwise JCD measure. The implications of this result are important: MPCT is best predicted by JCD, a measure that approximates the *coherence* of topics. Furthermore the correlations are negative, indicating that a low average distance (and thus, a high semantic similarity) indicates a high performance along this automated measure.

6 Conclusion and Future Work

In this work we define a new measure for the performance of statistical topic models. We show that this measure gauges a different aspect of the topics than the traditional model precision measure. Finally, we identify automated measures that can approximate the crowdsourced measures for both interpretability and coherence. This measure can be used by future researchers to complement their analysis of statistical topics. The results from our experiments indicate that Word Intrusion Choose Two is different from Word Intrusion, with almost no correlation between the two measures.

Furthermore, we propose automatic measures that can replace the crowdsourced measures. This is important as it allows for both scalability and reproducibility, as experiments using crowdsourcing are costly in terms of both time and money. We find that measures based on the interpretability of topics can best approximate the Model Precision Choose Two measure, indicating that this measure favors topics whose top words are more semantically similar, furthering our claim that this measure is assessing the coherence of the topic. Code and data to reproduce Model Precision Choose Two can be found at `http://bit.ly/mpchoose2`.

While model precision choose two offers a new way to understand topics, there may be others that could help to reveal other dimensions of topic quality. Future work is to find other measures for the semantic properties of topic modeling algorithms. Furthermore, the automated measures we discover to approximate the crowdsourced ones may be incorporated into a topic modeling algorithm that can better produce interpretable topics.

Acknowledgments

This work is sponsored, in part, by Office of Naval Research (ONR) grant N000141410095.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

David Blei. 2012. Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1):8–11.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296.

T. M. Cover and J. A. Thomas. 2006. *Elements of Information Theory*. Wiley InterScience, Hoboken, New Jersey.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 769–778, New York, NY, USA. ACM.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Kenneth Joseph, Chun How Tan, and Kathleen M. Carley. 2012. Beyond "local", "categories" and "friends": clustering foursquare users with latent "topics". In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 919–926, New York, NY, USA. ACM.

K Kireyev, L Palen, and K Anderson. 2009. Applications of Topics Models to Analysis of Disaster-Related Twitter Data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, volume 1.

Shamanth Kumar, Fred Morstatter, Reza Zafarani, and Huan Liu. 2013. Whom Should I Follow?: Identifying Relevant Users During Crises. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 139–147, New York, NY, USA. ACM.

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 European Chapter of the Association for Computational Linguistics*.

Arun S Maiya, John P Thompson, Francisco Loaiza-Lemos, and Robert M Rolfe. 2013. Exploratory analysis of highly heterogeneous document collections. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1375–1383. ACM.

Andrew McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose. *Proceedings of ICWSM*.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Alexei Pozdnoukhov and Christian Kaiser. 2011. Space-time dynamics of topics in streaming text. In *Proc. of the 3rd ACM SIGSPATIAL Int'l Workshop on Location-Based Social Networks*, LBSN '11, pages 1–8, New York, NY, USA. ACM.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. 2009. Topic modeling for the social sciences. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, volume 5.

Benjamin M Schmidt. 2012. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1):49–65.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

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

Tze-I Yang, Andrew J Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104. Association for Computational Linguistics.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 247–256, New York, NY, USA. ACM.