The Sensitivity of Topic Coherence Evaluation to Topic Cardinality

Jey Han Lau[1,2] and **Timothy Baldwin**[2]
[1] IBM Research
[2] Dept of Computing and Information Systems,
The University of Melbourne
jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

When evaluating the quality of topics generated by a topic model, the convention is to score topic coherence — either manually or automatically — using the top-N topic words. This hyper-parameter N, or the *cardinality* of the topic, is often overlooked and selected arbitrarily. In this paper, we investigate the impact of this cardinality hyper-parameter on topic coherence evaluation. For two automatic topic coherence methodologies, we observe that the correlation with human ratings decreases systematically as the cardinality increases. More interestingly, we find that performance can be improved if the system scores and human ratings are aggregated over several topic cardinalities before computing the correlation. In contrast to the standard practice of using a fixed value of N (e.g. $N = 5$ or $N = 10$), our results suggest that calculating topic coherence over several different cardinalities and averaging results in a substantially more stable and robust evaluation. We release the code and the datasets used in this research, for reproducibility.[1]

1 Introduction

Latent Dirichlet Allocation ("LDA": Blei et al. (2003)) is an approach to document clustering, in which "topics" (multinomial distributions over terms) and topic allocations (multinomial distributions over topics per document) are jointly learned. When the topic model output is to be presented to humans, optimisation of the number of topics is a non-trivial problem. In the seminal paper of Chang et al. (2009), e.g., the authors showed that — contrary to expectations — extrinsically measured topic coherence correlates negatively with model perplexity. They introduced the word intrusion task, whereby a randomly selected "intruder" word is injected into the top-N words of a given topic and users are asked to identify the intruder word. Low reliability in identifying the intruder word indicates low coherence (and vice versa), based on the intuition that the more coherent the topic, the more clearly the intruder word should be an outlier.

Since then, several methodologies have been introduced to automate the evaluation of topic coherence. Newman et al. (2010) found that aggregate pairwise PMI scores over the top-N topic words correlated well with human ratings. Mimno et al. (2011) proposed replacing PMI with conditional probability based on co-document frequency. Aletras and Stevenson (2013) showed that coherence can be measured by a classical distributional similarity approach. More recently, Lau et al. (2014) proposed a methodology to automate the word intrusion task directly. Their results also reveal the differences between these methodologies in their assessment of topic coherence.

A hyper-parameter in all these methodologies is the number of topic words, or its *cardinality*. These methodologies evaluate coherence over the top-N topic words, where N is selected arbitrarily: for Chang et al. (2009), $N = 5$, whereas for Newman et al. (2010), Aletras and Stevenson (2013) and Lau et al. (2014), $N = 10$.

[1] https://github.com/jhlau/topic-coherence-sensitivity

Lau, Jey Han and Timothy Baldwin (to appear) 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 (NAACL HLT 2016), San Diego, USA.

The germ of this paper came when using the automatic word intrusion methodology (Lau et al., 2014), and noticing that introducing one extra word to a given topic can dramatically change the accuracy of intruder word prediction. This forms the kernel of this paper: to better understand the impact of the topic cardinality hyper-parameter on the evaluation of topic coherence.

To investigate this, we develop a new dataset with human-annotated coherence judgements for a range of cardinality settings ($N = \{5, 10, 15, 20\}$). We experiment with the automatic word intrusion (Lau et al., 2014) and discover that correlation with human ratings decreases systematically as cardinality increases. We also test the PMI methodology (Newman et al., 2010) and make the same observation. To remedy this, we show that performance can be substantially improved if system scores and human ratings are aggregated over different cardinality settings before computing the correlation. This has broad implications for topic model evaluation.

2 Dataset and Gold Standard

To examine the relationship between topic cardinality and topic coherence, we require a dataset that has topics for a range of cardinality settings. Although there are existing datasets with human-annotated coherence scores (Newman et al., 2010; Aletras and Stevenson, 2013; Lau et al., 2014; Chang et al., 2009), these topics were annotated using a fixed cardinality setting (e.g. 5 or 10). We thus develop a new dataset for this experiment.

Following Lau et al. (2014), we use two domains: (1) WIKI, a collection of 3.3 million English Wikipedia articles (retrieved November 28th 2009); and (2) NEWS, a collection of 1.2 million New York Times articles from 1994 to 2004 (English Gigaword). We sub-sample approximately 50M tokens (100K and 50K articles for WIKI and NEWS respectively) from both domains to create two smaller document collections. We then generate 300 LDA topics for each of the sub-sampled collection.[2]

There are two primary approaches to assessing topic coherence: (1) via word intrusion (Chang et al., 2009); and (2) by directly measuring observed coherence (Newman et al., 2010; Lau et al., 2014). With the first method, Chang et al. (2009) injects an intruder word into the top-5 topic words, shuffles the topic words, and sets the task of selecting the single intruder word out of the 6 words. In preliminary experiments, we found that the word intrusion task becomes unreasonably difficult for human annotators when the topic cardinality is high, e.g. when $N = 20$. As such, we use the second approach as the means for generating our gold standard, asking users to judge topic coherence directly over different topic cardinalities.[3]

To collect the coherence judgements, we used Amazon Mechanical Turk and asked Turkers to rate topics in terms of coherence using a 3-point ordinal scale, where 1 indicates incoherent and 3 very coherent (Newman et al., 2010). For each topic (600 topics in total) we experiment with 4 cardinality settings: $N = \{5, 10, 15, 20\}$. For example, for $N = 5$, we display the top-5 topic words for coherence judgement.

For annotation quality control, we embed a bad topic generated using random words into each HIT. Workers who fail to consistently rate these bad topics low are filtered out.[4] On average, we collected

Domain	N			
	5	10	15	20
WIKI	2.42 (±0.54)	2.37 (±0.53)	2.35 (±0.51)	2.29 (±0.50)
NEWS	2.49 (±0.53)	2.46 (±0.53)	2.42 (±0.51)	2.39 (±0.51)

Table 1: Mean rating across different N (numbers in parentheses denote standard deviations)

Cardinality Pair	WIKI	NEWS
5 vs. 10	0.834	0.849
5 vs. 15	0.777	0.834
5 vs. 20	0.826	0.815
10 vs. 15	0.841	0.876
10 vs. 20	0.853	0.854
15 vs. 20	0.831	0.871
Mean	0.827	0.850

Table 2: Correlation between different pairwise cardinality settings.

[2]The sub-sampled document collections are lemmatised using OpenNLP and Morpha (Minnen et al., 2001) before topic modelling.

[3]This is not a major limitation, however, as Lau et al. (2014) found a strong correlation between the judgements generated by the two methodologies.

[4]We filter workers who rate bad topics with a rating > 1 in more than 30% of their HITs.

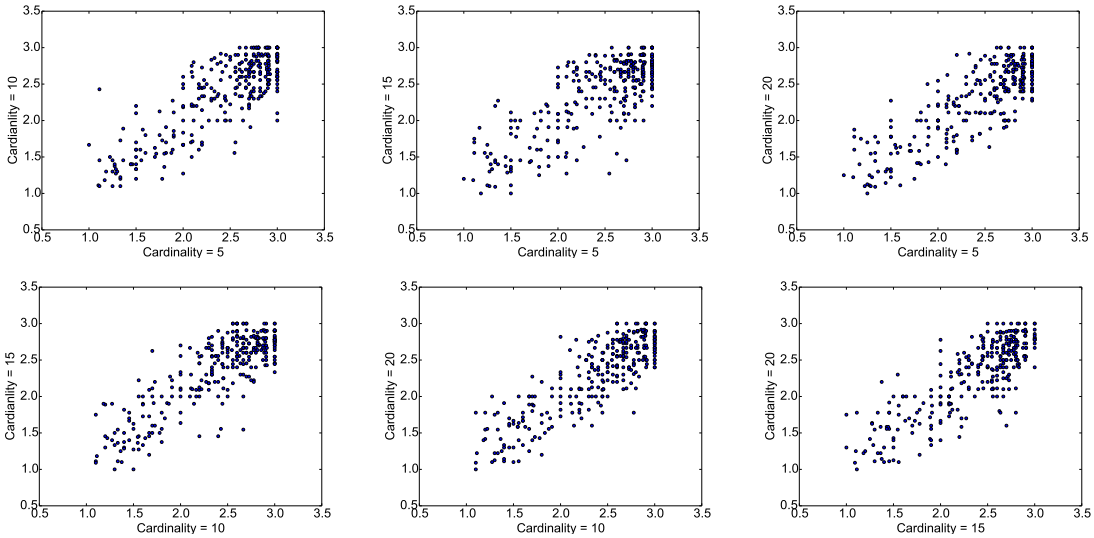

Figure 1: Scatter plots of human ratings for different pairwise cardinality settings for the WIKI topics.

approximately 9 ratings per topic in each cardinality setting (post-filtered), from which we generate the gold standard via the arithmetic mean.

To understand the impact of cardinality (N) on topic coherence, we analyse: (a) the mean topic rating for each N (Table 1), and (b) the pairwise Pearson correlation coefficient between the same topics for different values of N (Table 2).

Coherence decreases slightly but systematically as N increases, suggesting that users find topics less coherent (but marginally more consistently interpretable, as indicated by the slight drop in standard deviation) when more words are presented in a topic. The strong pairwise correlations, however, indicate that the ratings are relatively stable across different cardinality settings.

To better understand the data, in Figure 1 we present scatter plots of the ratings for all pairwise cardinality settings (where a point represents a topic). Note the vertical lines for $x = 3.0$ (cf. the weaker effect of horizontal lines for $y = 3.0$), in particular for the top 3 plots where we are comparing $N = 5$ against higher cardinality settings. This implies that topics that are rated as perfectly coherent (3.0) for $N = 5$ exhibit some variance in coherence ratings when N increases. Intuitively, it means that a number of perfectly coherent 5-word topics become less coherent as more words are presented.

3 Automated Method — Word Intrusion

Lau et al. (2014) proposed an automated approach to the word intrusion task. The methodology computes pairwise word association features for the top-N words, and trains a support vector regression model to rank the words. The top-ranked word is then selected as the predicted intruder word. Note that even though it is supervised, no manual annotation is required as the identity of the true intruder word is known. Following the original paper, we use as features normalised PMI (NPMI) and two conditional probabilities (CP1 and CP2), computed over the full collection of WIKI (3.3 million articles) and NEWS (1.2 million articles), respectively. We use 10-fold cross validation to predict the intruder words for all topics.

To generate an intruder for a topic, we select a random word that has a low probability ($P < 0.0005$) in the topic but high probability ($P > 0.01$) in another topic. We repeat this ten times to generate 10 different intruder words for a topic. The 4 cardinalities of a given topic share the same set of intruder words.

To measure the coherence of a topic, we compute *model precision*, or the accuracy of intruder word prediction. For evaluation we compute the Pearson correlation coefficient r of model precisions and human ratings for each cardinality setting. Results are summarised in Table 3.

Domain	N	In-domain Features	Out-of-Domain Features
WIKI	5	0.46	0.66
	10	0.41	0.54*
	15	0.32*	0.51*
	20	0.33*	0.43*
	Avg	0.46	0.65
NEWS	5	0.45*	0.65
	10	0.40*	0.60*
	15	0.38*	0.54*
	20	0.43*	0.47*
	Avg	0.50	0.65

Table 3: Pearson correlation between system model precision and human ratings across different values of N for word intrusion. '*' denotes statistical significance compared to aggregate correlation.

Domain	N	In-domain Features	Out-of-Domain Features
WIKI	5	0.02	0.59*
	10	−0.05*	0.58*
	15	0.00	0.56*
	20	0.06	0.55*
	Avg	0.00	0.63
NEWS	5	0.22*	0.62*
	10	0.27*	0.68*
	15	0.35	0.68*
	20	0.35	0.65*
	Avg	0.31	0.71

Table 4: Pearson correlation between system topic coherence and human ratings across different values of N for NPMI. "*" denotes statistical significance compared to aggregate correlation.

Each domain has 2 sets of correlation figures, based on in-domain and out-of-domain features. In-domain (out-of-domain) features are word association features computed using the same (different) domain as the topics, e.g. when we compute coherence of WIKI topics using word association features derived from WIKI (NEWS).

The correlations using in-domain features are in general lower than for out-of-domain features. This is due to idiosyncratic words that are closely related in the collection, e.g. remnant Wikipedia markup tags. The topic model discovers them as topics and the word statistics derived from the same collection supports the association, but these topics are generally not coherent, as revealed by out-of-domain statistics. This result is consistent with previous studies (Lau et al., 2014).

We see that correlation decreases systematically as N increases, implying that N has high impact on topic coherence evaluation and that if a single value of N is to be used, a lower value is preferable.

To test whether we can leverage the additional information from the different values of N, we aggregate the model precision values and human ratings per-topic before computing the correlation (Table 3: Cardinality = "Avg"). We also test the significance of difference for each N with the aggregate correlation using the Steiger Test (Steiger, 1980); they are marked with '*' in the table.[5]

[5]The test measures if the aggregate correlation is significantly higher ($p < 0.1$) than a non-aggregate correlation using a one-tailed test.

The correlation improves substantially. In fact, for NEWS using in-domain features, the correlation is higher than that of any individual cardinality setting. This observation suggests that a better approach to automatically computing topic coherence is to aggregate coherence scores over different cardinality settings, and that it is sub-optimal to evaluate a topic by only assessing a single setting of N. Instead, we should repeat it several times, varying N.

4 Automated Method — NPMI

The other mainstream approach to evaluating topic coherence is to directly measure the average pairwise association between the top-N words. Newman et al. (2010) found PMI to be the best association measure, and later studies (Aletras and Stevenson, 2013; Lau et al., 2014) found that normalised PMI (NPMI: Bouma (2009)) improves PMI further.

To see if the benefit of aggregating coherence measures over several cardinalities transfers across to other methodologies, we test the NPMI methodology. We compute the topic coherence using the full collection of WIKI and NEWS, respectively, for varying N. Results are presented in Table 4.

The in-domain features perform much worse, especially for the WIKI topics. NPMI assigns very high scores to several incoherent topics, thereby reducing the correlation to almost zero. These topics consist predominantly of Wikipedia markup tags, and the high association is due to word statistics idiosyncratic to the collection.

Once again, aggregating the topic coherence over

multiple N values boosts results further. The correlations using aggregation and out-of-domain features again produce the best results for both WIKI and NEWS.

It is important to note that, while these findings were established based on manual annotation of topic coherence, for practical applications, topic coherence would be calculated in a fully-unsupervised manner (averaged over different topic cardinalities), without the use of manual annotations.

5 Conclusion

We investigate the impact of the cardinality of topic words on topic coherence evaluation. We found that human ratings decrease systematically when cardinality increases, although pairwise correlations are relatively high. We discovered that the performance of two automated methods — word intrusion and pairwise NPMI — can be substantially improved if the system scores and human ratings are aggregated over several cardinality settings before computing the correlation. Contrary to the standard practice of using a fixed cardinality setting, our findings suggest that we should assess topic coherence using several cardinality settings and then aggregate over them. The human-judged coherence ratings, along with code to compute topic coherence, are available online.

6 Acknowledgements

This research was supported in part by funding from the Australian Research Council.

References

Nikos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the Tenth International Workshop on Computational Semantics (IWCS-10)*, pages 13–22, Potsdam, Germany.

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

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

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 21 (NIPS-09)*, pages 288–296, Vancouver, Canada.

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 EACL (EACL 2014)*, pages 530–539, Gothenburg, Sweden.

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 (EMNLP 2011)*, pages 262–272, Edinburgh, UK.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 100–108, Los Angeles, USA.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87:245–251.