Large-Scale Correlation Analysis of Automated Metrics for Topic Models

Jia Peng Lim
Singapore Management University
jiapeng.lim.2021@smu.edu.sg

Hady W. Lauw
Singapore Management University
hadywlauw@smu.edu.sg

Abstract

Automated coherence metrics constitute an important and popular way to evaluate topic models. Previous works present a mixed picture of their presumed correlation with human judgement. In this paper, we conduct a large-scale correlation analysis of coherence metrics. We propose a novel sampling approach to mine topics for the purpose of metric evaluation, and conduct the analysis via three large corpora showing that certain automated coherence metrics are correlated. Moreover, we extend the analysis to measure topical differences between corpora. Lastly, we examine the reliability of human judgement by conducting an extensive user study, which is designed as an amalgamation of different proxy tasks to derive a finer insight into the human decision-making processes. Our findings reveal some correlation between automated coherence metrics and human judgement, especially for generic corpora.

1 Introduction

Topic modelling is an important tool in the analysis and exploration of text corpora in terms of their salient topics (Blei et al., 2003). To evaluate the effectiveness of topic models, the preponderance of topic modeling literature rely on automated coherence metrics. A key benefit is convenience, allowing researchers to sidestep expensive and time-consuming user studies. The basis for this reliance is the assumption that the coherence metrics correlate with human judgement (Mimno et al., 2011; Lau et al., 2014; Röder et al., 2015).

The presumed correlation with human judgement should not be taken for granted. There are recent works that challenge the assumption. Doogan and Buntine (2021) highlight the inconsistencies of automated coherence metrics via correlation analysis within each metric. In Hoyle et al. (2021), they claimed some disagreement between human judgement and automated coherence metrics.

We postulate that the reasons behind such a mixed picture could be the differences in the topic samples as well as the underlying corpora from which the statistics were derived, resulting in localised "biases" that affect the conclusions reached by respective studies. Given their importance, we seek to conduct an extended analysis of automated coherence metrics on a larger scale than anything previously attempted. This study includes orders of magnitudes greater than the number of topics typically analysed, covering three large corpora, employing a comprehensive user study with extensive labels, across most of the widely used metrics.

There is a strong motivation for quantity. Given a vocabulary, a combinatorially large number of possible topics exist. If each topic is a vector of its scores on different metrics, the resulting curse of dimensionality (Bellman and Kalaba, 1959) necessitates a larger sample size. We argue that evaluating thousands of topics might not be sufficient, and a larger sample size is required to approximate a diverse distribution, where sampled topics is representative of the corpus and the metrics.

We surmise that the previous practice of using topic models to generate topics could introduce a bias in the analysis. Firstly, topic models vary in performance, Hoyle et al. (2021) compiled a lengthy list. There is also emerging debate on the performance between traditional and neural topic models (Doogan and Buntine, 2021). Additionally, some neural models might be inconsistent, producing different topic sets in independent runs (Hoyle et al., 2022). Conversely, topic model might be too stable and generate similar topics (Xing and Paul, 2018). To objectively evaluate whether the coherence metrics are usable, we propose to generate candidate topics independently of topic models.

In this paper, our contributions are three-fold. First, we begin by analysing the inter-*metric* correlations (see Section 4). We propose a novel approach to sample "topics" for the purpose of

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13874–13898

evaluating automated coherence metrics (see Section 4.1). Compared to prior works, we sample these topics free from topic model bias, and in a meaningful diverse manner. Evaluated on three large corpora, we reaffirm that certain selected metrics do not contradict each other, and highlight the underestimated effects of ϵ (see Section 4.2).

Second, we extend our analysis to investigate inter-*corpora* correlations (see Section 5). We examine the understated differences of corpora statistics on the metrics by comparing the correlations across corpora. While such correlations do exist to some degree, the metrics are still dependent on each corpus. Thus, any expectation that these metrics would correlate uniformly with human judgement on all possible corpora may be misplaced.

Finally, pivotal to any interpretability research, we design and conduct a user study, which is the keystone of our work (see Section 6). Compared to prior work, its design is more complex as we seek to benchmark human judgement at a finer granularity across different random user study groups (see Section 6.1). We analyse the user study results via a few novel proxy measures, revealing that human judgement is nuanced and varies between individuals, metric correlation to human judgement is corpus-dependant, with the average participant being attuned to the generic corpora (see Section 6.2).

Our implementation and releasable resources can be found here[1], and we hope that it will enable convenient coherence evaluation of topic models and to further advance interpretability research.

2 Related Work

Topic models. There are many approaches for topic modelling Blei et al. (2003), from non-neural based Zhao et al. (2017b); Hoffman et al. (2010), to many other neural-based methods, via auto-encoders (Kingma and Welling, 2014) such as Miao et al. (2016); Srivastava and Sutton (2017); Dieng et al. (2020); Zhang and Lauw (2020); Bianchi et al. (2021), via graph neural networks (Yang et al., 2020; Shen et al., 2021; Zhang and Lauw, 2022), and hierarchical methods (Meng et al., 2020). A common factor is the use of automated coherence metrics to benchmark against baselines. We select several popular metrics for evaluation as listed in Section 3. Topic models are applied in downstream tasks (Lau et al., 2017; Wang et al., 2019, 2020).

User studies in metric evaluation. Mimno et al. (2011) utilize expert annotators to independently label 148 topics, using another 10 expert annotators to evaluate the same topics via intruder word detection tasks. Röder et al. (2015) benchmark topics against different permutations of metrics with the largest evaluation set containing 900 topics with human ratings aggregated from prior works (Aletras and Stevenson, 2013; Lau et al., 2014; Rosner et al., 2014). In Hoyle et al. (2021), a minimum of 15 crowdworkers were employed in simple rating and word intrusion tasks evaluating 40 topic-model-generated (Griffiths and Steyvers, 2004; Burkhardt and Kramer, 2019; Dieng et al., 2020) and 16 synthetic random topics. In Doogan and Buntine (2021), their largest user study required 4 subject matter experts creating 3,120 labels across 390 topics generated via topic models (Blei et al., 2003; Zhao et al., 2017a). In comparison, our study has both large quantities of topics and study participants, annotating 800 unbiased topics split between 40 study participants with at least an undergraduate level of education, generating 180K word-pair labels[2]. Our automated experiments deal with hundreds of thousands of unique topics.

Human involvement. There are many interesting research that examine linguistic problems via the human lens. Card et al. (2020) investigates the number of annotators required to achieve significant statistical power. Plank (2022) examines the variation in human labels. Ethayarajh and Jurafsky (2022) questions the authenticity of annotators. Clark et al. (2021) tests the human ability to learn how to differentiate between machine-generated and human-generated texts. Human-in-the-loop systems or processes, such as Li et al. (2022), are also being actively explored.

3 Preliminaries

In this section, we define the automated coherence metrics that we will be using, and describe the corpora we use to obtain the word probabilities.

3.1 Coherence Metrics

We follow the definition styles of Röder et al. (2015), where direct confirmation measure m is a function of a word-pair statistic. Direct coherence metrics is defined as a mean aggregation of m between word-pairs (Equation 1), where t is a topic which is a k-sized set of words. For our evaluations,

[1] https://github.com/PreferredAI/topic-metrics

[2] Each question has 45 possible combinations of word-pairs, each label is binary, denoting coherence relations.

we set $k = 10$. Within t, the words are arranged based on $P(w|t)$ in descending order. Since our approach does not produce $P(w|t)$, we can locally optimize the word positions within a topic to obtain the best possible score for position-sensitive metrics C_{UMass} and C_P (See Appendix B). We use subscript s to denote alphabetical order and subscript o to denote optimized positions. Let $p = \frac{|t|\cdot|t-1|}{2}$, which represents the number of word-pairs in a topic.

$$C(t, m) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i>j}} m(w_i, w_j) \quad (1)$$

C_{NPMI} (Equation 2) is the mean aggregation of m_{nlr}, defined as Normalised Pointwise Mutual Information (NPMI) (Bouma, 2009) value, between word-pair statistics in a topic. We exclude C_{UCI} as it uses Point-wise Mutual Information (Church and Hanks, 1990; Lau et al., 2014), which is correlated to NPMI.

$$C_{\text{NPMI}}(t) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i>j}} m_{nlr}(w_i, w_j) \quad (2)$$

$$m_{nlr}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \quad (3)$$

C_{UMass} is the mean ordinal aggregation of m_{lc} (Mimno et al., 2011), which measures the log conditional probability between ordered word-pair in a topic:

$$C_{\text{UMass}}(t) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i>j}} m_{lc}(w_i, w_j) \quad (4)$$

$$m_{lc}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \quad (5)$$

C_P is the mean ordinal aggregation of m_f, Fitelson's coherence (Fitelson, 2003), interpreted as the degree to which w_i supports w_j, between ordered word-pairs in a topic:

$$C_P(t) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i>j}} m_f(w_i, w_j) \quad (6)$$

$$m_f(w_i, w_j) = \frac{P(w_i|w_j) - P(w_i|\neg w_j)}{P(w_i|w_j) + P(w_i|\neg w_j)} \quad (7)$$

C_V (Equation 8) is the final metric that we are using. C_V is considered as an indirect coherence metric, as it uses word-group relations as opposed to word-pairs relations like aforementioned direct coherence metrics. Intuitively, it measures the mean cosine similarity (Equation 9) between each word's feature vector and the topic's feature vector represented as the sum of all of its words' feature vectors (Equation 10).

$$C_V(t, \gamma) = \frac{\sum_{w_i \in t} s_{cos}(v(w_i, t, \gamma), \bar{v}(t, \gamma))}{|t|} \quad (8)$$

$$s_{cos}(\vec{v_i}, \vec{v_j}) = \frac{\sum \vec{v_i} \cdot \vec{v_j}}{||\vec{v_i}||_2 \cdot ||\vec{v_j}||_2} \quad (9)$$

$$\bar{v}(t, \gamma) = \sum_{w_j \in t} v(w_j, t, \gamma) \quad (10)$$

$$v(w, t, \gamma) = \{m_{nlr}(w, w_j)^\gamma \ \forall w_j \in t\} \quad (11)$$

For indirect confirmation measure $\tilde{m}$, instead of directly using word-word probabilities, it uses m to create a vector of features v (Aletras and Stevenson, 2013) that represent a word w from the topic t it belongs to, distorted by hyper-parameter γ (Equation 11). We will evaluate γ at 1 and 2[3].

3.2 Corpora

Corpus	#Docs.	Mean Doc. Size	Vocab. Size
ArXiv	2.09M	75	26K
Pubmed	1.07M	1500	39K
Wiki	5.51M	217	40K

Table 1: Numerical descriptions of the corpora used. Lemmatized variants are similar with the exception of ArXiv-lemma where its vocabulary size is 22K.

We use word co-occurrences statistics obtained from three large corpora:

ArXiv. We use ArXiv abstracts dataset[4] where we consider each abstract as a document. These abstracts mainly comprise of research work related to non-medical science disciplines.

Pubmed. We use PubMed Central (PMC) Open Access Subset[5] that contains journal articles and pre-prints related to medical research and information. We consider each article body as a document and we remove citations within it.

[3]Prior to version 0.1.4 (released Sep 21, 2022), Palmetto's (Röder et al., 2015) γ was set to 2.

[4]Kaggle - Cornell-University/ArXiv

[5]ncbi.nlm.nih.gov/pmc/tools/openftlist

Wiki. We use the English-Wikipedia dump[6] of August'22 processed using Attardi (2015). We consider the content of the article as a document. To check for correctness, we also use the popular benchmark Palmetto (Röder et al., 2015), which uses a subset of Wikipedia'11.

For each corpus, we apply processing steps suggested in Hoyle et al. (2021), retaining up to 40K frequently occurring words. Moreover, we generate a lemmatized (denoted with the suffix -lemma) and unlemmatized variant (original) for further analysis. More information on common vocabulary between corpora can be found in Table 14, Appendix C.

4 Examining Inter-Metric Correlations

Intuitively, if two different metrics are to correlate with human judgement, we would expect the scores of these metrics to correlate. However, it is claimed in Doogan and Buntine (2021) that these metrics do not correlate well. For reasons described in Section 1, we propose a new non-topic modelling approach to sample topics to evaluate these metrics.

4.1 Approach: Balanced Sampling

There are few tested methods to generate topics: from topic models (Aletras and Stevenson, 2013; Lau et al., 2014), beam search optimized on coherence (Rosner et al., 2014), random sampling of words (Hoyle et al., 2021). Considering only optimized topics, or completely random topics (mostly bad), would generate a skewed distribution. In contrast, we seek to mine topics that emulates a balanced distribution for a meaningful comparison. We also desire uniqueness among topics, which avoids repetition and is representative of the corpus. Figure 1 illustrates an overview of our approach.

Mining topics of k words can be framed as the classical k-clique listing problem (Chiba and Nishizeki, 1985; Danisch et al., 2018). To generate meaningful topics, we can map the corpus-level information as a graph, treating each word from its vocabulary set V as a vertex. Each word will share an edge with every other word. We choose m_{nlr} to determine the value of the edges between two vertices as its normalised range is intuitive allowing us to easily identify the range of values for sub-graph generation. In contrast, using m_{lc} and m_f increases sampling's complexity as they are order-dependant resulting in bi-directional edges in its sub-graph. Sampling using any m, not only m_{nlr}, might introduce bias, which our approach seeks to mitigate.

Corpus	*neg*	*pos*	*mid*	*random*	*ext*	Total
ArXiv	66,007	2,120	14,436	10,000	49,777	142,340
Pubmed	10,450	3,310	8,218	10,000	61,035	93,013
Wiki	56,903	21,698	35,195	10,000	136,036	259,832

Table 2: Average quantity of topics mined by our balanced sampling approach by segments per corpus from the 5 independent sampling runs. Quantities of lemmatized variants are similar with the exception of *ext* segment, where it has half the numbers.

The initial graph will be a complete graph of $|V|$ vertices. A topic of k words would be a k-sized sub-graph. Combinatorially, there are $|V|$ choose k number of possible unique topics. It is practically infeasible and unnecessary to list all k-cliques. For a more tractable approach, we modify the routine from Yuan et al. (2022) (pseudo-code in Appendix A) to include:

Sub-graphs of varying quality. This routine seeks to generate smaller graphs from the original complete graph to cover the spectrum of topic quality. We eliminate edges conditionally via their value, and the remaining edges and connected vertices constitute the new sub-graph. We generate three different kinds of sub-graphs, *pos* where edge-values are above a given lower-bound, *mid* where edge-values are between threshold values, and *neg* where edges are below an upper-bound[7].

Topic extraction. Inspired by Perozzi et al. (2014), instead of iterating through all the neighbouring nodes or searching for the next best node, we randomly select a neighbour, that has an edge with all explored nodes, to explore. We extract the explored k-path as our sampled topic.

Topic uniqueness. To attain a variety of topics, we remove all edges in a mined clique, making it impossible to sample a similar topic from the same sub-graph. Figure 2 illustrates this feature.

Balance distribution of topics. For a given corpus, we further introduce common topics sampled from a different corpora, which differ in its word distribution. We refer to this segment of external topics as *ext*. Lastly, *random* is a segment, comprising of groups of random words, included to represent topics that might not have been covered via the other segments. Table 2 shows the result from this mining approach. The total would thus be more balanced, comprising topics of varying scores along the spectrum.

[6] dumps.wikimedia.org

[7] Hyper-parameters listed in Table 9, Appendix A

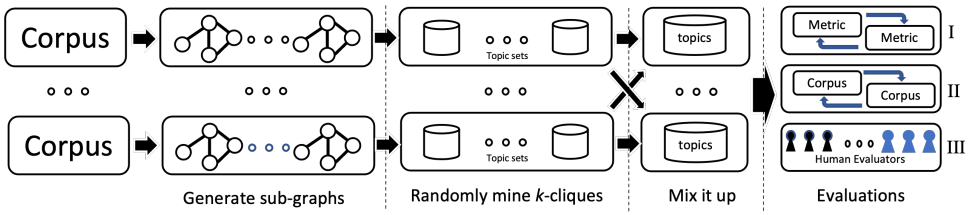

Figure 1: Illustration of our Balanced Sampling

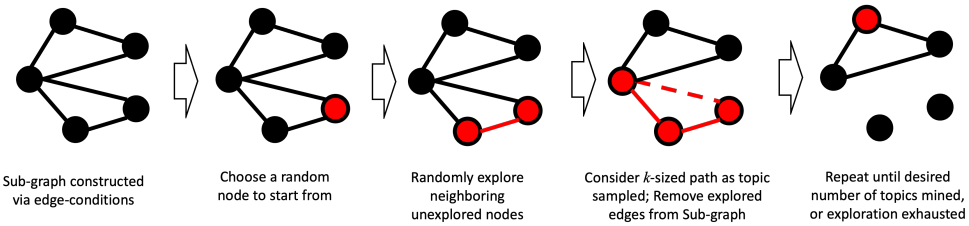

Figure 2: Illustration of the process of sampling a topic from a sub-graph.

4.2 Evaluation: Metric Correlations Analysis

ϵ	$C_V^{\gamma=1}$	$C_V^{\gamma=2}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_V^{\gamma=1}$	-	0.09	0.69	0.64	0.11
$C_V^{\gamma=2}$	0.09	-	-0.59	-0.63	-0.72
C_{NPMI}	0.69	-0.59	-	**0.91**	**0.58**
$C_{P,o}$	0.64	-0.63	**0.91**	-	**0.71**
$C_{\text{UMass},o}$	0.11	-0.72	**0.58**	**0.71**	-

(a) Correlation scores with $\epsilon = 1\text{e}{-12}$

$\not\epsilon$	$C_V^{\gamma=1}$	$C_V^{\gamma=2}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_V^{\gamma=1}$	-	**0.87**	**0.95**	**0.81**	**0.45**
$C_V^{\gamma=2}$	**0.87**	-	**0.94**	**0.66**	**0.28**
C_{NPMI}	**0.95**	**0.94**	-	0.73	0.31
$C_{P,o}$	**0.81**	**0.66**	0.73	-	0.65
$C_{\text{UMass},o}$	**0.45**	**0.28**	0.31	0.65	-

(b) Correlation scores with $\epsilon = 0$

Table 3: Pearson's r scores (Mean of 5 independently sampled sets of topics) between coherence metrics measured on Wiki. Bold indicates the better value across both tables. Error bars omitted as S.D ≤ 0.02.

We evaluate the correlation (Pearson's r[8]) between different automated metrics measured on Wiki (see Table 3), Pubmed, and ArXiv (see Table 10, Appendix C). We expect a high positive correlation score between metrics if they are both purportedly measuring for coherence. Our first inter-metric analysis (see Table 3a), with metrics calculated at $\epsilon = 1\text{e}{-12}$, shows the poor correlation of C_V metrics against other metrics. Theoretically, C_V relies on m_{nlr} as its features, and given an unrelated topic, where word-pair scored on m_{nlr} with $\epsilon = 1\text{e}{-12}$ produces similar m_{nlr} vectors which scores highly on C_V. This phenomenon of high cosine similarity between the equally negative m_{nlr} vectors, results in contradicting scores between C_V and other metrics.

Hence, for our second inter-metric analysis (see Table 3b) we evaluate the metrics at $\epsilon = 0$, denoted with subscript $\not\epsilon$. For the resulting undefined calculations, we default to 0. Intuitively, the purpose of setting $\epsilon = 1\text{e}{-12}$ is to prevent and to penalise word-pairs that produces undefined calculation. In contrast, $\epsilon = 0$ treats these word-pairs neutrally. Comparing the new results in Table 3b to the previous results in Table 3a, we note that correlation scores between C_V metric and other automated coherence metrics improved greatly, suggesting alleviation of the contradicting factor. Additionally, we note that for C_P and C_{UMass}, ϵ is essential. We then examine these metrics with their better ϵ mode (see Table 4a), and most metrics (except C_{UMass}) have a decent correlation with other metrics, implying that they do not contradict each other.

There could be a concern that the *neg* and *random* sampled sections would have an outsized influence in the previous analysis. In this ablation, we restrict the same analysis to only topics where $C_{\text{NPMI}} > 0$. Comparing to the previous results (see Table 4a), we derive a similar interpretation from this constrained results (see Table 4b), suggesting that our balanced sampling approach is effective as the behaviour of the full set of data is similar to its smaller subset.

[8]Based on reasons provided in Doogan and Buntine (2021), with the main argument that datasets (scores) are continuous and have a bi-variate normal distribution.

	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_{V,\notin}^{\gamma=1}$	-	0.87	0.95	0.74	0.81	0.33
$C_{V,\notin}^{\gamma=2}$	0.87	-	0.94	0.56	0.66	0.24
$C_{\text{NPMI},\notin}$	0.95	0.94	-	0.63	0.73	0.25
C_{NPMI}	0.74	0.56	0.63	-	0.91	0.58
$C_{P,o}$	0.81	0.66	0.73	0.91	-	0.71
$C_{\text{UMass},o}$	0.33	0.24	0.25	0.58	0.71	-

(a) Correlation scores of metrics measured on Wiki. Combined results of Table 3 on selected metrics.

	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_{V,\notin}^{\gamma=1}$	-	0.92	0.98	0.95	0.99	-0.14
$C_{V,\notin}^{\gamma=2}$	0.92	-	0.95	0.94	0.90	-0.02
$C_{\text{NPMI},\notin}$	0.98	0.95	-	0.98	0.98	-0.14
C_{NPMI}	0.95	0.94	0.98	-	0.95	-0.09
$C_{P,o}$	0.99	0.90	0.98	0.95	-	-0.20
$C_{\text{UMass},o}$	-0.14	-0.02	-0.14	-0.09	-0.20	-

(b) Correlation scores of metrics on subsection of data used in Table 4a where $C_{\text{NPMI}} > 0$.

Table 4: Comparing correlations (Mean of 5 independently sampled sets of topics) between selected automated coherence metrics with their better mode of ϵ measured on Wiki. Error bars omitted as S.D ≤ 0.02. The results on ArXiv and Pubmed are similar.

corpus-pairs	$\lvert T\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
ArXiv/Pubmed	267K	0.55	0.55	0.63	0.77	0.66	0.63
ArXiv/Wiki	338K	0.58	0.55	0.60	0.73	0.63	0.49
Pubmed/Wiki	341K	0.67	0.65	0.62	0.74	0.75	0.70

Table 5: Pearson's r between exact automated coherence metric measured on different corpus-pairs (independent samples aggregated totalling $\lvert T\rvert$ topics). See Table 13, Appendix C for complete results.

5 Examining Inter-Corpus Correlations

A natural extension after inter-metrics comparison, is to compare metrics measured on different corpora. It is a common expectation that research works would employ multiple corpora, with the differences between corpora quantified superficially (such as in Section 3.2). We propose an alternative approach to quantify the differences, at a topical level, using common topics measured using automated coherence metrics. If the corpora are thematically similar, we would expect a high correlation.

Analysis. Using the common topics from the paired corpora, we conduct a correlation analysis on the scores measured on each corpus per metric. Table 5 shows decent correlations between each corpus. However, even as they are positive, these correlations do not imply identical statistics in various corpora. Assuming that human judgement is constant for a given topic, we posit that variance in scores measured on different corpora could result in a lower correlation due to the missing themes

Corpus	$\lvert\bar{T}\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
ArXiv	80K	0.98	0.98	0.98	0.98	0.97	0.92
Pubmed	27K	0.94	0.97	0.94	0.92	0.93	0.94
Wiki	143K	0.99	0.99	0.99	0.98	0.96	0.95

(a) Comparison of scores from selected topics measured on both lemmatized and unlemmatized corpus.

Corpus	$\lvert\bar{T}\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
ArXiv	111K	0.97	0.98	0.95	0.94	0.94	0.95
Pubmed	60K	0.97	0.98	0.98	0.92	0.95	0.97
Wiki	150K	0.99	0.98	0.98	0.98	0.98	0.98

(b) Selected topics compared to its lemmatized variants, scores from both variants are measured on unlemmatized corpus.

Corpus	$\lvert\bar{T}\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
ArXiv	126K	0.94	0.95	0.92	0.84	0.85	0.88
Pubmed	68K	0.93	0.95	0.91	0.82	0.83	0.82
Wiki	245K	0.98	0.98	0.97	0.92	0.93	0.92

(c) Selected topics, measured on the unlemmatized corpus, are compared to its lemmatized variants, which are measured on the lemmatized corpus.

Table 6: Pearson's r (mean from 5 independently sampled sets of size $\lvert\bar{T}\rvert$) of automated coherence metric measured on different scenarios. Each selected topic will have two variants that will produce two scores for each metric. We compare the correlation of the two set of scores for a set of topics. Error bars omitted as S.D ≤ 0.01. See Table 12 and Table 11, Appendix C for additional quantitative data on the topics.

within the shared vocabulary space in either corpus.

We conduct a control analysis on pairs of similar corpus differing in lemmatization, originating from the same documents, in Table 6a. These corpora would be thematically similar whilst being superficially different. Our previous analysis in Table 5, comparing to the control analysis in Table 6a, shows lower correlation scores suggesting some topical difference between the various corpora. This difference highlights the metrics' strong dependency on the corpus used, with a subset of common topics disagreeing on the scores, revealing that these metrics are not a one-size-fits-all solution for coherence evaluation.

Ablations. While we know how lemmatization affects topic modelling (Schofield and Mimno, 2016), its effect on evaluation is unclear. We carried out two additional ablations simulating lemmatizing topics post-training. For the first ablation, we shortlist topics that contain at least one unlemmatized word, where if lemmatized, the lemmatized word can be found in the same unlemmatized corpus. We compare the correlation of the original and lemmatized topic, with their scores measured on the same unlemmatized corpus. Their scores have a strong correlation (see Table 6b), suggesting that the difference between lemmatized topics and

unlemmatized topics is small. For the second ablation, the shortlisting process is similar, however, with lemmatized topics measured on the lemmatized corpus. Our results (see Table 6c) show a strong correlation across the various metrics and imply that post-processing topics for evaluation is a viable option.

6 User Study

Previous works measure human judgement through simple evaluation tasks such as rating the coherence of a topic on a few-point ordinal scale (Mimno et al., 2011; Aletras and Stevenson, 2013), identifying the intruder word that introduced into the topic (Chang et al., 2009), or both (Lau et al., 2014; Hoyle et al., 2021). For word intrusion, the detection of outliers signals the cohesiveness of the topic, which is similar to rating topics on an ordinal scale. However for both tasks, qualitative gaps might exist. In word intrusion, study participants are restricted to just one outlier per topic, assuming perfect coding, it results in exponential drop in scoring, i.e. 100% detection for a perfect topic, 50% for a topic with a clear outlier, and so forth. For topic ratings, topics of differing qualities might get the same score, i.e. a perfect topic and a topic with a clear outlier might both get the same scores.

Additionally, while the decisions between human annotators might be equivalent, it is not evident if their thought processes are similar. The key reason for this line of inquiry stems from the observation that everyone is different in some aspects, such as knowledge, culture, and experiences. Assuming our understanding of words is influenced by our prior beliefs, what and how we perceive similarity and coherence might differ from person to person.

For these reasons, we decide to design a user study that combines both word intrusion and topic rating tasks but measured at a finer granularity such that we can quantify the decision-making process. Users are tasked to cluster word-groups which indicate coherent and outlier word-groups. We then examine the relationships between automated coherence metrics and different proxy tasks derived from the user study.

6.1 User Study Design

For our study S, we recruit 8 user study groups U, $S = \{U_1, \ldots, U_8\}$, 5 study participants per group. Majority of the participants recruited have at least a graduate degree or are undergraduates. For each study group, we prepared 8 unique question sets $Q = \{T_1, \ldots, T_8\}$, each containing 100 10-word topics, $T_i = \{t_{1,i}, \ldots, t_{100,i}\}$ and $t = \{w_{0,j,i}, \ldots, w_{10,j,i}\}$. For each participant $u \in U_i$, we present each $t_{j,i} \in T_i$ individually sorted alphabetically. We ask participants to cluster words in $t_{j,i}$ that they deem similar to form coherent word groups g, where their response $R_{u,j,i}$ to $t_{j,i}$ is a set of unique g. We constrain each word to only belong to one coherent word group to limit the task complexity. Additionally, a word considered to be unrelated may form its group of one. We use Likert matrix[9] as the response format (see Figure 3), mandating a response for each word $w_{k,j,i} \in t_{j,i}$. Actual instructions are shown in Appendix E.

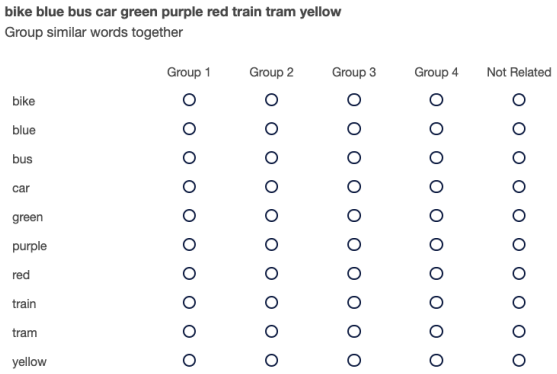

Figure 3: Format of question that is presented to study participants. Each word is to be assigned to only one group whose members are deemed coherent together. The topic displayed in this example is manually created to serve as a verification question and is not included in the evaluation. Refer to Appendix D for a sample of actual examples used.

Topic selection. We construct an initial pool of 1000 topics. To achieve comparability between corpus, we randomly sample 400 common topics from Wiki, ArXiv, and Pubmed. To represent non-scientific topics, we randomly sample 200 topics from Wiki that do not appear in ArXiv/Pubmed. For ArXiv/Pubmed exclusive topics, we randomly sample 200 topics each, with these topics also appearing in Wiki. We sample in a 7:1:1:1 ratio of *pos*/*mid*/*neg*/*random* segments of the corpus, seeking to emulate a uniform score distribution. To account for word familiarity, we select lemmatized topics with words found in 20K most frequently used words[10]. For each user study, we randomly

[9] There is no scaling.

[10] Corpus of Contemporary American English

sampled 100 topics from the pool without replacement. For topics not found in ArXiv or Pubmed, we exclude them during evaluation of those corpus.

Proxy Tasks. Representing coherence as word-clusters allows us to derive a deeper insight into what we perceive as human judgement. From our user study task, we further decompose this study into a few proxy tasks, where we measure the correlation (Spearman's ρ[11]) of its results to automated coherent metrics. We propose three topic-level human coherence measures. Using density of human agreement, we define P_1 as the mean agreement of U_i on all possible word-pairs on any topic $t_{j,i}$:

$$P_1(t_{j,i}) = \frac{\sum_{u \in U_i} \sum_{g \in R_u} |g|(|g|-1)}{|U_i| \frac{|t_{j,i}|(|t_{j,i}|-1)}{2}} \quad (12)$$

If $t_{j,i}$ has perfect agreement on coherence, we expect $P_1(t_{j,i})$ to have a value of 1, and for incoherence, a value of 0.

Subsequently, we consider the largest selected word group within $t_{j,i}$, and define P_2 as the mean of this measure amongst U_i:

$$P_2(t_{j,i}) = \frac{1}{|U_i|} \sum_{u \in U_i} \max(\{|g| | g \in R_u\}) \quad (13)$$

A value of 1 will suggest that each word in $t_{j,i}$ have no relations to each other and a value of $|t_{j,i}|$ suggest perfect agreement on coherence.

Lastly, we define P_3 as the mean number of annotated word groups amongst U_i:

$$P_3(t_{j,i}) = \frac{1}{|U_i|} \sum_{u \in U_i} |R_u| \quad (14)$$

The interpretation of P_3 is the inverse of P_2. While these group-wise measures might seem similar, they measure different nuances of human-annotated data. P_1 evaluates the sizes of multi-word groups, weighted towards larger groups. P_2 only accounts for the largest word group, which ignores the properties of the other remaining group. P_3 ignores group sizes to a certain extent and includes single-word "outlier" groups. We evaluate these measures' correlation against various $C(t_{j,i})$.

6.2 User Study Results

We find that the three different proxy tasks produce similar results[12], shown in Table 7a, 7b, and 7c respectively, indicating correlations between human judgement and some automated coherence metrics. Since most of our study participants have some science-related background, we are surprised by ArXiv's lower correlation scores relative to Wiki in each proxy task. These results imply that our perception of coherence might be biased towards the word distribution of a generic corpus such as Wiki. Lastly, in each proxy task, the higher variances in ArXiv's and Pubmed's correlation scores compared to Wiki's might imply increased subjectivity.

Inter-rater reliability (IRR). There are many factors that wil affect the variation for IRR (Belur et al., 2021). For our user study, we attempted to mitigate some of these factors. In terms of fram-

	ArXiv	Pubmed	Wiki
$C_{V,\notin}^{\gamma=1}$	0.319 ± 0.152	0.516 ± 0.067	0.651 ± 0.099
$C_{V,\notin}^{\gamma=2}$	0.356 ± 0.146	0.510 ± 0.095	0.652 ± 0.119
$C_{\text{NPMI},\notin}$	0.366 ± 0.136	0.521 ± 0.064	0.664 ± 0.094
C_{NPMI}	0.304 ± 0.169	0.428 ± 0.111	0.624 ± 0.087
$C_{P,o}$	0.266 ± 0.178	0.459 ± 0.093	0.634 ± 0.091
$C_{\text{UMass},o}$	0.243 ± 0.176	0.183 ± 0.161	0.329 ± 0.066

(a) Proxy Task I: Density of agreement among study participants. Full Breakdown in Table 16, Appendix C.

	ArXiv	Pubmed	Wiki
$C_{V,\notin}^{\gamma=1}$	0.316 ± 0.159	0.511 ± 0.053	0.643 ± 0.110
$C_{V,\notin}^{\gamma=2}$	0.355 ± 0.153	0.507 ± 0.080	0.648 ± 0.130
$C_{\text{NPMI},\notin}$	0.369 ± 0.135	0.517 ± 0.049	0.654 ± 0.104
C_{NPMI}	0.303 ± 0.175	0.421 ± 0.094	0.615 ± 0.090
$C_{P,o}$	0.260 ± 0.182	0.454 ± 0.081	0.624 ± 0.103
$C_{\text{UMass},o}$	0.232 ± 0.182	0.170 ± 0.152	0.320 ± 0.060

(b) Proxy Task II: Mean of maximum coherent group between study participants. Full Breakdown in Table 17, Appendix C.

	ArXiv	Pubmed	Wiki
$C_{V,\notin}^{\gamma=1}$	-0.382 ± 0.164	-0.547 ± 0.109	-0.645 ± 0.085
$C_{V,\notin}^{\gamma=2}$	-0.415 ± 0.168	-0.541 ± 0.135	-0.648 ± 0.100
$C_{\text{NPMI},\notin}$	-0.434 ± 0.171	-0.549 ± 0.118	-0.660 ± 0.084
C_{NPMI}	-0.342 ± 0.195	-0.453 ± 0.118	-0.627 ± 0.085
$C_{P,o}$	-0.320 ± 0.200	-0.484 ± 0.107	-0.631 ± 0.082
$C_{\text{UMass},o}$	-0.277 ± 0.172	-0.202 ± 0.126	-0.354 ± 0.053

(c) Proxy Task III: Mean of coherent group counts between study participants. For this task, stronger negative score is better as a completely coherent topic gets $P_3(t) = 1$ and an incoherent topic gets $P_3(t) = 10$. Hence, this proxy measure is inversely related to the coherence metric score where a larger score indicates coherence. Full Breakdown in Table 18, Appendix C.

Table 7: Average Spearman's ρ between automated coherence metrics and respective proxy measure. The values shown are the mean correlation scores from the 8 study groups with error bars. The lemmatized version of corpus are ommitted as its values are similar to the original. $C_{\text{UMass},s}$ and $C_{P,s}$ ommited as they are almost identical to their o variant.

[11]We use Spearman's ρ instead of Pearson's r, as we generally obtain a better r (than ρ shown) through distortion of scores. To ensure parity, we use ρ instead.

[12]We note that these results include outlier U_3, whose negative results differ radically from other groups. Individual results detailed in Appendix C.

ing and education, study participants were given a short introductory primer as well as some example questions prior to starting the tasks (Appendix E). To mitigate fatigue effect, we allowed the study participants a week to work on the task, pausing and continuing at their own pace. We were not concerned about learning effect, as our presented topics spans across a plethora of themes and the correctness of the task is subjective to their own personal preference. As our objective is to poll for their beliefs, with many possible valid answers, there is not a need to review and enforce consistency between study participants.

We use Krippendorf's α (Krippendorff, 2011) , defining pair-wise rater similarity as Jaccard distance measuring common answers between raters. We treat each $w_{k,j,i} \in t_{j,i}$ as a multi-classification question, comprising of other words (in $t_{j,i}$) and "not related" as categories, producing boolean vector representations. The mean $\bar{\alpha}$ is 0.366 with a standard deviation of 0.04, lowest α at 0.325 and highest α at 0.464 (see Table 15, Appendix C). A completely random study response will have an α of 0.12, being significantly less than the study's $\bar{\alpha}$, giving us some confidence about the reliability of the responses. Overall, considering that there are many possible combinations for each topic response, the α reported suggests some degree of similarity between different responses.

	ArXiv	Pubmed	Wiki
$C_{P,s}$	0.115 ± 0.062	0.139 ± 0.043	0.285 ± 0.091
$C_{P,o}$	0.201 ± 0.066	0.269 ± 0.036	0.447 ± 0.072
$C_{\text{UMass},s}$	0.119 ± 0.057	0.072 ± 0.039	0.128 ± 0.043
$C_{\text{UMass},o}$	0.185 ± 0.068	0.101 ± 0.037	0.209 ± 0.037

Table 8: Average Spearman's ρ between automated coherence metrics pair-wise proxy measure, similar in evaluation and interpretation to Table 7. This table shows the difference in correlation results between sorted (s) and optimal (o) position-dependent metrics. Full Breakdown in Table 19, Appendix C.

User study ablations. We examine if positioning affects position-dependent automated coherence metrics via human pair-wise agreement proxy task P_4. We detail our optimizing approach in Appendix B. We define P_4 as the percentage of agreement between any word-pairs w_a and w_b from $t_{j,i}$ from T_i evaluated by its corresponding U_i:

$$P_4(w_a, w_b) = \frac{1}{|U_i|} \sum_{u \in U_i} \sum_{g \in R_u} w_a \in g \wedge w_b \in g \tag{15}$$

We measure the correlation of $P_4(w_a, w_b)$ in a group to its pair-wise automated coherence metric score via $m(w_a, w_b)$ from different orderings. Our results in Table 8 show some non-significant differences in correlation on the pair-wise level. However, that difference disappears when we evaluate the topics as a group, with the sorted and optimized variant achieving similar correlations (see Table 7). Furthermore, this difference of coherence at the pair-wise and group-wise levels, suggests that the presence of other words in the topic has an influence on the human perception of word-pair coherence. Finally, we replicate most experiments with the corpus statistics from Palmetto (Röder et al., 2015), which produced similar correlation results to Wiki.

7 Conclusion

Our large-scale analysis reaffirms that these automated coherence metrics are still meaningful. We are confident in using these metrics measured on generic corpus such as Wiki, and specialised corpora, Arxiv and Pubmed, for nicher tasks. Our user study empirically supports this conclusion, as our participants' collective response correlates well to metrics measured on Wiki, albeit weaker but meaningful correlation on the specialized corpora. This work shows that popular automated coherence metrics, C_{NPMI} , C_V, and C_P, are alive and well, and works regardless of lemmatization. Furthermore, we stress that the selection of the reference corpus is just as important as the selection of the metric, with Wiki being the best reference corpus that correlates with human perception of coherence. Moving forward, when evaluating for coherence aligned towards human interpretability, we recommend future topic models to be evaluated against Wiki-variants. We also recommend calculating C_V with $\epsilon = 0$, to avoid the confusion from its contradiction of other metrics at $\epsilon = 1\text{e}{-12}$.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-020). Hady W. Lauw gratefully acknowledges the support by the Lee Kong Chian Fellowship awarded by Singapore Management University. We extend our gratitude to our user study participants for their efforts, as well as, our reviewers for their kind feedback.

Limitations

User Study. Most, if not all, of the participants are pursuing or have obtained at least a university degree/bachelor's. While we attempted to recruit widely, majority of our participants' education background is science-related, with strong leanings towards technology. Furthermore, we assume that our participants are proficient in English from their education level and the fact that they are based in a city that uses English as the common language. It is possible that there are some unknown common bias such as culture or knowledge that might affect the results. The tie-breaking constrain in our study, where study participants are required to assign one word to its most coherent group, might affect the correlation scores for the user study.

Corpora. The selected corpora are constructed from documents that are formal in prose, with the purpose of being informative and instructional. We do not know if the user study results are applicable to a corpus with documents that are informal in prose, such as that of a conversational nature. However, one can always evaluate topics on a large external generic corpus to determine coherence relative to human judgement.

Ethics Statement

User Study. Prior to carrying out our user study, the survey methodology was reviewed and approved by our Institutional Review Board for ethical compliance. While unlikely, we examined each question for its appropriateness. To ensure participants' anonymity, the responses are anonymized and aggregated, and it is extremely unlikely that a participant can be identified via their response. In terms of fair compensation, we paid S$15 for each complete response of 100 questions, assuming an hour's worth of work, it is higher than our institution's prevailing rate for undergraduate student work. To ensure their well-being, study participants are allowed up to a week to complete the tasks, at their own preferred pace and place.

Corpora. We select corpora that have open licensing agreements that allows for non-profit academic use, and the permissions allowing us to transform and re-distribute the processed corpora as word-pair counts.

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, Potsdam, Germany. Association for Computational Linguistics.

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Amotz Bar-Noy, Reuven Bar-Yehuda, Ari Freund, Joseph (Seffi) Naor, and Baruch Schieber. 2001. A unified approach to approximating resource allocation and scheduling. *J. ACM*, 48(5):1069–1090.

Richard Bellman and Robert Kalaba. 1959. A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, 45(8):1288–1290.

Jyoti Belur, Lisa Tompson, Amy Thornton, and Miranda Simon. 2021. Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, 50(2):837–865.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Norishige Chiba and Takao Nishizeki. 1985. Arboricity and subgraph listing algorithms. *SIAM J. Comput.*, 14:210–223.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Maximilien Danisch, Oana Balalau, and Mauro Sozio. 2018. Listing k-cliques in sparse real-world graphs*. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 589–598, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 6056–6070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Branden Fitelson. 2003. A probabilistic theory of coherence. *Analysis*, 63(3):194–199.

Thomas Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1:5228–35.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken?: The incoherence of coherence. In *Neural Information Processing Systems*.

Alexander Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

K. Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada. Association for Computational Linguistics.

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, Gothenburg, Sweden. Association for Computational Linguistics.

Raymond Li, Wen Xiao, Linzi Xing, Lanjun Wang, Gabriel Murray, and Giuseppe Carenini. 2022. Human guided exploitation of interpretable attention patterns in summarization and topic segmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10189–10204, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1908–1917, New York, NY, USA.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA.

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, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA. ACM.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, page 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. 2014. Evaluating topic coherence measures.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *WSDM*, pages 399–408.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Dazhong Shen, Chuan Qin, Chao Wang, Zheng Dong, Hengshu Zhu, and Hui Xiong. 2021. Topic modeling revisited: A document graph-based neural network perspective. In *Advances in Neural Information Processing Systems 34 - 35th Conference on Neural Information Processing Systems, NeurIPS 2021*, Advances in Neural Information Processing Systems, pages 14681–14693. Neural information processing systems foundation.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR (Poster)*.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

Linzi Xing and Michael Paul. 2018. Diagnosing and improving topic models by analyzing posterior variability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph attention topic modeling network. In *Proceedings of The Web Conference 2020*, WWW '20, page 144–154, New York, NY, USA. Association for Computing Machinery.

Zhirong Yuan, You Peng, Peng Cheng, Li Han, Xuemin Lin, Lei Chen, and Wenjie Zhang. 2022. Efficient $k-$ clique listing with set intersection speedup. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1955–1968.

Ce Zhang and Hady W Lauw. 2020. Topic modeling on document networks with adjacent-encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6737–6745.

Delvin Ce Zhang and Hady W Lauw. 2022. Variational graph author topic modeling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2429–2438.

He Zhao, Lan Du, Wray Buntine, and Gang Liu. 2017a. Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 635–644.

Renbo Zhao, Vincent Tan, and Huan Xu. 2017b. Online Nonnegative Matrix Factorization with General Divergences. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 37–45.

A Algorithm Pseudocode

	ArXiv	Pubmed	Wiki
pos (>)	$0.05, 0.1, 0.15$		
mid	$(-0.05, 0.15)$	$(0, 0.15)$	$(-0.05, 0.15)$
neg (<)	$-0.2, -0.4$	-0.2	$-0.1, -0.4$

Table 9: Hyper-parameter threshold for different sub-graphs. Multiple thresholds are indicative of multiple runs. *random* and *ext* are not hyper-parameter dependant. When possible, hyper-parameters were chosen to produce to control sub-graph density.

Pre-processing steps to reduce complexity, Algorithm 1 and Algorithm 2, remain unchanged from Yuan et al. (2022). These steps can be skipped when the graph is large and dense, such as during *neg* sub-graphs generation. Our modification in Algorithm 3 and Algorithm 4 introduces randomness via permutations and early stopping, when a k-clique is found in Algorithm 3 and a desired number of k-cliques found in Algorithm 4. The sub-graph reduction is implemented in Algorithm 3.

Algorithm 1 PRE-CORE(G, k)
Prune vertices with less than k edges from G

Input: A graph G and a positive integer k
$Q \leftarrow \emptyset, F \leftarrow \emptyset$
for $u \in G$ **do**
  if $d_u < k - 1$ **then**
    Q.push(u)
    $F \leftarrow F \cup \{u\}$
  end if
end for
while $Q \neq \emptyset$ **do**
  $u \leftarrow Q$.pop()
  for node $v \in$ neighbours N_u **do**
    $d_v \leftarrow d_v - 1$
    if $d_v < k - 1 \wedge v \notin F$ **then**
      $F \leftarrow F \cup \{v\}$
      Q.push(v)
    end if
  end for
end while

Algorithm 2 PRE-LIST(G,k)
Find exact k-cliques and remove them from G

for each connected components $C \in G$ **do**
  $m_c \leftarrow |E(C)|, n_c \leftarrow |V(C)|$
  if $m_c = (n_c - 1)n_c$ **then**
    remove C from G
    output k-cliques C
  end if
end for

A set of connected components refers to a set of nodes where each node shares an edge with all other nodes in the set. Finding next connected components $\hat{C}$, requires a set intersection operation between all possible neighbours of randomly selected node u, denoted N_u^+, and current connected components C.

Algorithm 3 SDegreeList(k, R, C, $\vec{G}$)

for $u \in$ Permutate(C) **do**
  if $|C| \leq l - 2$ **then**
    continue
  end if
  if $k < 2$ **then**
    return $\emptyset$
  end if
  $\hat{C} \leftarrow N_u^+ \cap C$
  if $k = 2 \wedge |\hat{C}| > 0$ **then**
    $O \leftarrow R \cup \{u\}$
    remove (u_i, u_j) from $\vec{G}$ $\forall u_i, u_j \in O$
    return O
  end if
  if $|\hat{C}| > l - 2$ **then**
    return SDegreeList($k - 1$, R, $\hat{C}$, $\vec{G}$)
  end if
end for

Algorithm 4 Main(G, k, target)

$G \leftarrow$ PRE-CORE(G, k)
$G \leftarrow$ PRE-LIST(G, k)
Generate DAG $\vec{G}$
$O \leftarrow \emptyset$
for $u \in$ Permutate($\vec{G}$) **do**
  $r \leftarrow$ SDegreeList($k - 1, \{u\}, N_u^+, \vec{G}$)
  if $|r| ==$ target **then**
    $O = O \cup \{r\}$
  end if
  if target $== |O|$ **then**
    return O
  end if
end for

The main algorithm gets invoked once per sub-graph, we can generate multiple sub-graphs by selecting a set of words that neighbours a randomly chosen word. We then truncate the edges that do not fulfill the edge-conditions.

B Optimizing Position-Based Scoring

Given a set of k words as a topic, our goal is to optimize the position-based score. We can reduce this problem to a weighted activity selection prob-

lem, which is equivalent to finding a max-weight independent set in an interval graph and can be solved in polynomial time (Bar-Noy et al., 2001).

Consider a word w at the j^{th} position, index starting from 0, we can visualize the ordering as having j incoming edges, indicating precedence of other words, and $k - j + 1$ outgoing edges, indicating w precedence to other ensuing words. An activity will be defined by its start-time (position) and its preceding and ensuing activities. Each activity has an equal interval and the weight of the activity is determined by the difference of outgoing and incoming edges to all other words scored via m. We can transform the activities into an interval graph, with $|C_j^l| \cdot |C_{l-j+1}^l|$ combinatorial number of possible instances for each word per time slot in the schedule.

Our transformation will result in an interval graph of k disjoint graphs. While the number of activities might seem to be combinatorially explosive, selecting the first activity at $T = 0$, only involves k activities, and upon selection prunes multiple branches, resulting in $k - j$ choices at $T = j$. Hence, we are only required to select the best activity within each disjoint graph conditioned on availability (word not selected before).

C Supplementary Tables

This section lists tables with quantitative supplementary information.

Table 10 details the results for ArXiv and Pubmed corpus for inter-metric correlation analysis in Section 4.2.

Table 11 provides additional information on the similarity between control and treated topics for the lemmatization effect ablation in Section 5.

Table 12 provides a detailed breakdown of subgraph segments that is shortlisted for the lemmatization effect ablation in Section 5.

Table 13 details the full complete results for inter-corpus correlation analysis, its partial table can be found in Table 5, Section 5.

Table 14 has additional quantitative information regarding the quantity of common topics in corpus-pairs used in the inter-corpus experiments of Section 5.

Table 15 has the individual Krippendorf's α for each user study group U for the user study in Section 6.

Tables 16, 17, 18, and 19 has the individual correlation scores of each user study group U to the various coherence metrics for Proxy Task I, II, III, and pair-wise ablation respectively. Its averages are tabled in Tables 7a, 7b, 7c, and 8 in Section 6.

ϵ	$C_V^{\gamma=1}$	$C_V^{\gamma=2}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_V^{\gamma=1}$	-	0.90	-0.87	-0.72	-0.42
$C_V^{\gamma=2}$	0.90	-	-0.93	-0.81	-0.52
C_{NPMI}	-0.87	-0.93	-	**0.91**	**0.60**
$C_{P,o}$	-0.72	-0.81	**0.91**	-	**0.83**
$C_{\text{UMass},o}$	-0.42	-0.52	**0.60**	**0.83**	-

(a) Correlation scores measured on ArXiv with $\epsilon =$ 1e-12

$\not\epsilon$	$C_V^{\gamma=1}$	$C_V^{\gamma=2}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_V^{\gamma=1}$	-	0.84	**0.85**	**0.75**	0.06
$C_V^{\gamma=2}$	0.84	-	**0.90**	**0.51**	0.08
C_{NPMI}	**0.85**	**0.90**	-	0.47	0.07
$C_{P,o}$	**0.75**	**0.51**	0.47	-	-0.10
$C_{\text{UMass},o}$	0.06	0.08	0.07	-0.10	-

(b) Correlation scores measured on ArXiv with $\epsilon = 0$

ϵ	$C_V^{\gamma=1}$	$C_V^{\gamma=2}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_V^{\gamma=1}$	-	0.21	0.60	0.40	-0.16
$C_V^{\gamma=2}$	0.21	-	-0.56	-0.66	-0.81
C_{NPMI}	0.60	-0.56	-	**0.85**	**0.54**
$C_{P,o}$	0.40	-0.66	**0.85**	-	**0.81**
$C_{\text{UMass},o}$	-0.16	-0.81	**0.54**	**0.81**	-

(c) Correlation scores measured on Pubmed with $\epsilon =$ 1e-12

$\not\epsilon$	$C_V^{\gamma=1}$	$C_V^{\gamma=2}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
$C_V^{\gamma=1}$	-	**0.78**	**0.94**	**0.67**	0.02
$C_V^{\gamma=2}$	**0.78**	-	**0.85**	**0.54**	0.02
C_{NPMI}	**0.94**	**0.85**	-	0.56	-0.02
$C_{P,o}$	**0.67**	**0.54**	0.56	-	-0.13
$C_{\text{UMass},o}$	0.02	0.02	-0.02	-0.13	-

(d) Correlation scores measured on Pubmed with $\epsilon = 0$

Table 10: Pearson's r scores (Mean of 5 independently sampled sets of topics) between automated coherence metrics within ArXiv/Pubmed corpus. Bold indicates better correlation score across both tables. Error bars omitted as S.D $\leq$ 0.02.

	ArXiv	Pubmed	Wiki
Table 6b	7.2	7.9	7.7
Table 6c	7.7	8.5	8.6

Table 11: Accompanying statistics for respective lemmatization effect ablation experiments (see Section 5). Value indicates mean number of similar words per topic. While the variants contain similar words, we note that the word probabilities differ and reflects the composition of lemmatized and base words in the vocabulary.

	anti	pos	middle	random	ext	Total
ArXiv	63,648	1,262	12,055	9,169	25,546	111,680
Pubmed	7,675	2,161	6,839	9,616	33,776	60,067
Wiki	52,867	15,074	27,638	8,811	45,194	149,584

(a) Accompanying details for experiment results in Table 6b.

	anti	pos	middle	random	ext	Total
ArXiv	58,274	1,449	13,559	7,833	44,446	125,561
Pubmed	9,857	2,396	119	2,025	53,751	68,148
Wiki	52,435	16,965	33,788	8,967	132,840	244,995

(b) Accompanying details for experiment results in Table 6c.

Table 12: Quantity of segmentation of sampled topics for respective lemmatization effect ablation experiments (see Section 5).

corpus-pairs	$\lvert T\rvert$	$C_{V,\not\in}^{\gamma=1}$	$C_{V,\not\in}^{\gamma=2}$	$C_{\text{NPMI},\not\in}$	C_{NPMI}	$C_{P,o}$	$C_{\text{UMass},o}$
ArXiv/Pubmed	267K	0.55	0.55	0.63	0.77	0.66	0.63
ArXiv/Wiki	338K	0.58	0.55	0.60	0.73	0.63	0.49
ArXiv/Palmetto	114K	0.51	0.54	0.57	0.50	0.44	0.44
Pubmed/Wiki	341K	0.67	0.65	0.62	0.74	0.75	0.70
Pubmed/Palmetto	130K	0.67	0.67	0.65	0.69	0.69	0.55
Wiki/Palmetto	447K	0.98	0.98	0.98	0.98	0.95	0.84
Wiki-l/ArXiv-l	114K	0.54	0.55	0.60	0.60	0.47	0.70
Pubmed-l/ArXiv-l	101K	0.59	0.57	0.70	0.76	0.59	0.78
Pubmed-l/Wiki-l	125K	0.70	0.68	0.71	0.78	0.74	0.78
Pubmed-l/Palmetto	125K	0.70	0.67	0.69	0.77	0.74	0.59
ArXiv-l/Palmetto	114K	0.54	0.55	0.58	0.58	0.49	0.49
Wiki-l/Palmetto	447K	0.99	0.99	0.99	0.99	0.97	0.91

Table 13: Pearson's r (independent samples were aggregated) between exact automated coherence metric measured on different corpus-pairs (independent samples were aggregated). Suffix -l. short form for -lemma.

corpus	ArXiv	ArXiv-l.	Pubmed	Pubmed-l.	Wiki	Wiki-l.	Palmetto
Total	26,620	22,184	38,829	39,997	40003	40,009	16,567
ArXiv	-	19,637	13,138	10,527	12,955	10,230	6,827
ArXiv-l	19,637	-	9,636	11,015	9,563	10,504	7,130
Pubmed	13,138	9,636	-	23,328	15,459	12,565	8,006
Pubmed-l	10,527	11,015	23,328	-	12,637	14,112	8,932
Wiki	12,955	9,563	15,459	12,637	-	31,047	13,136
Wiki-l	10,230	10,504	12,565	14,112	31,047	-	14,392
Palmetto	6,827	7,130	8,006	8,932	13,136	14,392	-

Table 14: Quantity of common vocabularies between corpus. Suffix -l. short form for -lemma. Palmetto was re-constructed using 20K most frequent words excluding stop words.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D)
Kripp's α	0.463	0.391	0.323	0.376	0.325	0.366	0.333	0.347	0.366 (0.04)

Table 15: Detailed Krippendorf's α for each user study.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D)
ArXiv									
$C_{V,\notin}^{\gamma=1}$	0.464	0.448	-0.021	0.330	0.399	0.437	0.218	0.281	0.319 ± 0.152
$C_{V,\notin}^{\gamma=2}$	0.503	0.469	0.030	0.281	0.459	0.462	0.344	0.300	0.356 ± 0.146
$C_{\text{NPMI},\notin}$	0.475	0.426	0.073	0.392	0.516	0.470	0.304	0.270	0.366 ± 0.136
C_{NPMI}	0.368	0.490	-0.110	0.309	0.386	0.394	0.251	0.348	0.304 ± 0.169
$C_{P,o}$	0.372	0.455	-0.157	0.285	0.355	0.383	0.208	0.231	0.266 ± 0.178
$C_{\text{UMass},o}$	0.348	0.476	-0.162	0.256	0.309	0.261	0.152	0.305	0.243 ± 0.176
Pubmed									
$C_{V,\notin}^{\gamma=1}$	0.609	0.560	0.372	0.550	0.462	0.511	0.526	0.535	0.516 ± 0.067
$C_{V,\notin}^{\gamma=2}$	0.662	0.622	0.356	0.465	0.415	0.543	0.492	0.521	0.510 ± 0.095
$C_{\text{NPMI},\notin}$	0.574	0.605	0.396	0.534	0.453	0.498	0.548	0.560	0.521 ± 0.064
C_{NPMI}	0.479	0.447	0.165	0.531	0.442	0.368	0.453	0.537	0.428 ± 0.111
$C_{P,o}$	0.519	0.511	0.231	0.531	0.482	0.409	0.502	0.488	0.459 ± 0.093
$C_{\text{UMass},o}$	0.252	0.177	-0.115	0.327	0.280	0.043	0.087	0.417	0.183 ± 0.161
Wiki									
$C_{V,\notin}^{\gamma=1}$	0.692	0.715	0.413	0.758	0.607	0.670	0.692	0.664	0.651 ± 0.099
$C_{V,\notin}^{\gamma=2}$	0.719	0.739	0.348	0.727	0.631	0.673	0.702	0.678	0.652 ± 0.119
$C_{\text{NPMI},\notin}$	0.737	0.718	0.445	0.760	0.608	0.670	0.706	0.664	0.664 ± 0.094
C_{NPMI}	0.718	0.679	0.451	0.734	0.556	0.582	0.641	0.630	0.624 ± 0.087
$C_{P,o}$	0.658	0.695	0.422	0.737	0.585	0.671	0.684	0.621	0.634 ± 0.091
$C_{\text{UMass},o}$	0.405	0.322	0.226	0.427	0.381	0.272	0.272	0.326	0.329 ± 0.066
Palmetto									
$C_{V,\notin}^{\gamma=1}$	0.696	0.690	0.401	0.740	0.614	0.715	0.696	0.668	0.653 ± 0.101
$C_{V,\notin}^{\gamma=2}$	0.726	0.705	0.363	0.739	0.646	0.726	0.706	0.685	0.662 ± 0.116
$C_{\text{NPMI},\notin}$	0.721	0.694	0.439	0.734	0.613	0.722	0.719	0.654	0.662 ± 0.093
C_{NPMI}	0.647	0.610	0.464	0.697	0.562	0.666	0.699	0.638	0.623 ± 0.073
$C_{P,o}$	0.635	0.628	0.404	0.703	0.573	0.690	0.663	0.656	0.619 ± 0.089
$C_{\text{UMass},o}$	0.409	0.205	0.210	0.324	0.290	0.201	0.200	0.317	0.269 ± 0.073

Table 16: Detailed breakdown of Proxy Task I, values are Spearman's ρ of density of agreement and coherence scores. $C_{\text{UMass},s}$ and $C_{P,s}$ ommited as they are almost identical to their o variant.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D)
ArXiv									
$C_{V,\not\in}^{\gamma=1}$	0.497	0.418	-0.028	0.308	0.383	0.463	0.200	0.289	0.316 ± 0.159
$C_{V,\not\in}^{\gamma=2}$	0.534	0.438	0.027	0.263	0.448	0.497	0.332	0.305	0.355 ± 0.153
$C_{\text{NPMI},\not\in}$	0.524	0.383	0.094	0.372	0.488	0.509	0.298	0.283	0.369 ± 0.135
C_{NPMI}	0.400	0.465	-0.130	0.282	0.361	0.425	0.266	0.353	0.303 ± 0.175
$C_{P,o}$	0.401	0.420	-0.175	0.260	0.315	0.415	0.209	0.235	0.260 ± 0.182
$C_{\text{UMass},o}$	0.352	0.469	-0.189	0.215	0.284	0.278	0.150	0.298	0.232 ± 0.182
Pubmed									
$C_{V,\not\in}^{\gamma=1}$	0.607	0.530	0.408	0.529	0.470	0.520	0.510	0.514	0.511 ± 0.053
$C_{V,\not\in}^{\gamma=2}$	0.663	0.574	0.399	0.444	0.431	0.538	0.486	0.520	0.507 ± 0.080
$C_{\text{NPMI},\not\in}$	0.579	0.572	0.432	0.505	0.456	0.516	0.534	0.546	0.517 ± 0.049
C_{NPMI}	0.468	0.446	0.190	0.482	0.453	0.374	0.454	0.498	0.421 ± 0.094
$C_{P,o}$	0.518	0.504	0.256	0.502	0.492	0.409	0.492	0.456	0.454 ± 0.081
$C_{\text{UMass},o}$	0.234	0.196	-0.130	0.280	0.290	0.028	0.096	0.367	0.170 ± 0.152
Wiki									
$C_{V,\not\in}^{\gamma=1}$	0.682	0.701	0.367	0.754	0.624	0.683	0.678	0.657	0.643 ± 0.110
$C_{V,\not\in}^{\gamma=2}$	0.715	0.726	0.310	0.724	0.652	0.695	0.690	0.675	0.648 ± 0.130
$C_{\text{NPMI},\not\in}$	0.729	0.706	0.397	0.749	0.625	0.682	0.689	0.658	0.654 ± 0.104
C_{NPMI}	0.708	0.672	0.413	0.712	0.568	0.594	0.635	0.616	0.615 ± 0.090
$C_{P,o}$	0.645	0.679	0.373	0.733	0.598	0.677	0.670	0.613	0.624 ± 0.103
$C_{\text{UMass},o}$	0.397	0.311	0.210	0.398	0.365	0.278	0.288	0.311	0.320 ± 0.060
Palmetto									
$C_{V,\not\in}^{\gamma=1}$	0.680	0.679	0.364	0.736	0.629	0.722	0.690	0.661	0.645 ± 0.111
$C_{V,\not\in}^{\gamma=2}$	0.716	0.692	0.328	0.735	0.663	0.742	0.700	0.680	0.657 ± 0.127
$C_{\text{NPMI},\not\in}$	0.706	0.685	0.397	0.728	0.630	0.725	0.712	0.651	0.654 ± 0.103
C_{NPMI}	0.630	0.605	0.428	0.688	0.577	0.662	0.707	0.633	0.616 ± 0.081
$C_{P,o}$	0.617	0.618	0.373	0.695	0.591	0.691	0.662	0.649	0.612 ± 0.096
$C_{\text{UMass},o}$	0.392	0.206	0.218	0.283	0.289	0.194	0.245	0.310	0.267 ± 0.061

Table 17: Detailed breakdown of Proxy Task II, values are Spearman's ρ of mean of maximum group counts and coherence scores. $C_{\text{UMass},s}$ and $C_{P,s}$ ommited as they are almost identical to their o variant.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D)
ArXiv									
$C_{V,\not\epsilon}^{\gamma=1}$	-0.533	-0.529	0.007	-0.350	-0.485	-0.447	-0.336	-0.384	-0.382 ± 0.164
$C_{V,\not\epsilon}^{\gamma=2}$	-0.563	-0.577	-0.026	-0.319	-0.520	-0.470	-0.454	-0.391	-0.415 ± 0.168
$C_{\text{NPMI},\not\epsilon}$	-0.562	-0.584	-0.019	-0.428	-0.556	-0.499	-0.433	-0.388	-0.434 ± 0.171
C_{NPMI}	-0.429	-0.546	0.144	-0.330	-0.457	-0.376	-0.340	-0.405	-0.342 ± 0.195
$C_{P,o}$	-0.448	-0.536	0.169	-0.290	-0.446	-0.364	-0.325	-0.320	-0.320 ± 0.200
$C_{\text{UMass},o}$	-0.387	-0.442	0.129	-0.299	-0.419	-0.229	-0.214	-0.352	-0.277 ± 0.172
Pubmed									
$C_{V,\not\epsilon}^{\gamma=1}$	-0.608	-0.649	-0.298	-0.636	-0.459	-0.589	-0.579	-0.556	-0.547 ± 0.109
$C_{V,\not\epsilon}^{\gamma=2}$	-0.652	-0.720	-0.248	-0.549	-0.430	-0.586	-0.576	-0.565	-0.541 ± 0.135
$C_{\text{NPMI},\not\epsilon}$	-0.594	-0.705	-0.280	-0.609	-0.474	-0.577	-0.591	-0.563	-0.549 ± 0.118
C_{NPMI}	-0.506	-0.457	-0.179	-0.590	-0.416	-0.434	-0.480	-0.560	-0.453 ± 0.118
$C_{P,o}$	-0.519	-0.562	-0.225	-0.589	-0.438	-0.492	-0.548	-0.499	-0.484 ± 0.107
$C_{\text{UMass},o}$	-0.277	-0.155	0.004	-0.327	-0.234	-0.105	-0.114	-0.408	-0.202 ± 0.126
Wiki									
$C_{V,\not\epsilon}^{\gamma=1}$	-0.713	-0.655	-0.473	-0.756	-0.561	-0.691	-0.680	-0.632	-0.645 ± 0.085
$C_{V,\not\epsilon}^{\gamma=2}$	-0.751	-0.679	-0.410	-0.722	-0.602	-0.686	-0.697	-0.641	-0.648 ± 0.100
$C_{\text{NPMI},\not\epsilon}$	-0.759	-0.661	-0.496	-0.755	-0.572	-0.699	-0.693	-0.646	-0.660 ± 0.084
C_{NPMI}	-0.727	-0.623	-0.496	-0.764	-0.523	-0.627	-0.645	-0.608	-0.627 ± 0.085
$C_{P,o}$	-0.684	-0.636	-0.483	-0.742	-0.538	-0.697	-0.675	-0.596	-0.631 ± 0.082
$C_{\text{UMass},o}$	-0.387	-0.358	-0.276	-0.455	-0.371	-0.342	-0.285	-0.357	-0.354 ± 0.053
Palmetto									
$C_{V,\not\epsilon}^{\gamma=1}$	-0.698	-0.641	-0.454	-0.745	-0.572	-0.739	-0.667	-0.637	-0.644 ± 0.089
$C_{V,\not\epsilon}^{\gamma=2}$	-0.734	-0.648	-0.420	-0.736	-0.600	-0.745	-0.681	-0.644	-0.651 ± 0.100
$C_{\text{NPMI},\not\epsilon}$	-0.733	-0.649	-0.489	-0.737	-0.582	-0.755	-0.684	-0.638	-0.658 ± 0.084
C_{NPMI}	-0.647	-0.579	-0.497	-0.719	-0.550	-0.720	-0.647	-0.625	-0.623 ± 0.073
$C_{P,o}$	-0.635	-0.587	-0.447	-0.718	-0.537	-0.714	-0.632	-0.625	-0.612 ± 0.084
$C_{\text{UMass},o}$	-0.387	-0.242	-0.214	-0.365	-0.296	-0.267	-0.176	-0.340	-0.286 ± 0.070

Table 18: Detailed breakdown of Proxy Task III, values are Spearman's ρ of mean of group counts and coherence scores. $C_{\text{UMass},s}$ and $C_{P,s}$ ommited as they are almost identical to their o variant. For this task, a stronger negative value is better as a completely coherent topic have a group count of 1 and an incoherent topic will have a group count of 10. Hence, the proxy measure is inversely related to the coherence metric score where a larger score indicates coherence.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D)
ArXiv									
$C_{V,\not\in}^{\gamma=1}$	0.262	0.232	0.051	0.170	0.211	0.219	0.072	0.183	0.175 ± 0.071
$C_{V,\not\in}^{\gamma=2}$	0.287	0.224	0.038	0.166	0.203	0.219	0.079	0.208	0.178 ± 0.076
$C_{\text{NPMI},\not\in}$	0.257	0.215	0 .104	0.176	0.254	0.221	0.105	0.188	0.190 ± 0.056
C_{NPMI}	0.272	0.231	0.110	0.209	0.262	0.225	0.123	0.259	0.211 ± 0.058
$C_{P,s}$	0.299	0.238	0.079	0.193	0.230	0.242	0.120	0.202	0.201 ± 0.066
$C_{P,o}$	0.218	0.152	-0.019	0.101	0.124	0.125	0.091	0.126	0.115 ± 0.062
$C_{\text{UMass},s}$	0.280	0.213	0.061	0.140	0.228	0.228	0.111	0.220	0.185 ± 0.068
$C_{\text{UMass},o}$	0.193	0.146	-0.007	0.118	0.133	0.155	0.076	0.137	0.119 ± 0.057
Pubmed									
$C_{V,\not\in}^{\gamma=1}$	0.328	0.321	0.221	0.335	0.269	0.256	0.280	0.340	0.294 ± 0.041
$C_{V,\not\in}^{\gamma=2}$	0.314	0.281	0.213	0.295	0.272	0.235	0.261	0.331	0.275 ± 0.037
$C_{\text{NPMI},\not\in}$	0.240	0.259	0.205	0.269	0.242	0.184	0.229	0.291	0.240 ± 0.032
C_{NPMI}	0.274	0.261	0.188	0.305	0.257	0.201	0.225	0.305	0.252 ± 0.041
$C_{P,s}$	0.294	0.286	0.206	0.306	0.261	0.225	0.256	0.316	0.269 ± 0.036
$C_{P,o}$	0.183	0.160	0.063	0.140	0.109	0.112	0.134	0.210	0.139 ± 0.043
$C_{\text{UMass},s}$	0.114	0.086	0.087	0.132	0.116	0.061	0.044	0.167	0.101 ± 0.037
$C_{\text{UMass},o}$	0.078	0.090	0.009	0.111	0.098	0.056	0.016	0.121	0.072 ± 0.039
Wiki									
$C_{V,\not\in}^{\gamma=1}$	0.560	0.527	0.300	0.547	0.406	0.494	0.422	0.485	0.468 ± 0.082
$C_{V,\not\in}^{\gamma=2}$	0.543	0.518	0.299	0.527	0.399	0.484	0.405	0.470	0.455 ± 0.077
$C_{\text{NPMI},\not\in}$	0.524	0.495	0.295	0.510	0.397	0.433	0.405	0.440	0.437 ± 0.070
C_{NPMI}	0.518	0.498	0.297	0.507	0.396	0.429	0.395	0.454	0.437 ± 0.069
$C_{P,s}$	0.526	0.503	0.299	0.517	0.393	0.469	0.410	0.460	0.447 ± 0.072
$C_{P,o}$	0.384	0.338	0.094	0.379	0.218	0.336	0.265	0.269	0.285 ± 0.091
$C_{\text{UMass},s}$	0.243	0.257	0.159	0.217	0.199	0.202	0.149	0.243	0.209 ± 0.037
$C_{\text{UMass},o}$	0.165	0.163	0.058	0.173	0.103	0.126	0.070	0.163	0.128 ± 0.043
Palmetto									
$C_{V,\not\in}^{\gamma=1}$	0.553	0.503	0.292	0.542	0.398	0.516	0.428	0.496	0.466 ± 0.083
$C_{V,\not\in}^{\gamma=2}$	0.538	0.491	0.299	0.515	0.398	0.509	0.418	0.486	0.457 ± 0.075
$C_{\text{NPMI},\not\in}$	0.524	0.479	0.294	0.508	0.394	0.472	0.424	0.454	0.444 ± 0.069
C_{NPMI}	0.526	0.479	0.295	0.514	0.391	0.472	0.416	0.468	0.445 ± 0.071
$C_{P,s}$	0.516	0.466	0.291	0.504	0.378	0.484	0.406	0.479	0.441 ± 0.072
$C_{P,o}$	0.411	0.325	0.104	0.354	0.209	0.342	0.261	0.325	0.291 ± 0.091
$C_{\text{UMass},s}$	0.217	0.203	0.136	0.172	0.166	0.181	0.146	0.209	0.179 ± 0.028
$C_{\text{UMass},o}$	0.155	0.145	0.070	0.145	0.103	0.110	0.080	0.153	0.120 ± 0.032

Table 19: Detailed breakdown of Pair-wise Proxy Task. Values are Spearman's ρ.

D Topic Examples (User Study)

This set of 100 topics belongs to T_1, and were shown to U_1:

1.	ethic humanities intellectual interdisciplinary journal philosophical scientific social society sociology
2.	automate behavior check computation correct fluid limitation numerical processing specify
3.	behavioral differ differentiation extent furthermore interaction neural overlap similarity trait
4.	accountant archdiocese citizenship compile cultivate enlarge ferry grab interim wield
5.	care educate educational engage life pandemic participation preparedness social support
6.	advent anatomy enhance harmless interfere mortality psychiatrist swallow terminate urine
7.	agent buy buyer maximize profit risk sell seller social utility
8.	benchmark effectiveness experiment extensive indoor outdoor performance real-world synthetic validate
9.	bandwidth beam conversion generation laser photon pulse pump purity silicon
10.	anxiety child depression distress illness mental parent parenting social stress
11.	account activity audit employment fund provision public purpose resource security
12.	acidity alcoholic biochemical compete fuse insulin pathological short-term smell spontaneous
13.	access communication device hardware infrastructure management resource secure technology wireless
14.	bladder blood cardiac cavity congenital gastrointestinal intestinal obstruction procedure surgical
15.	assess assessment company industry maturity methodology organization quality research software
16.	building conditional embryonic glacial hair multiplicity overly programming questionnaire renewable
17.	adoption encryption insurance job minimal native nowadays predictor resilience visit
18.	continued doubling feedback growing guideline hypothetical induction pad readiness worth
19.	automated detect detection measurement observation optical radar real-time sensor spacecraft
20.	advent bald deficiency household liquid museum parasite physique qualify rude
21.	control evaluation framework implement level monitor optimal regulation response specific
22.	dose gland hormone inflammation inject muscle secretion serum stimulate toxin
23.	creative family handy lie mold rank residual semantic transmission weaken
24.	broad hair irregular length longitudinal mature somewhat spore tooth yellow
25.	appropriate behavioral combination condition define evaluate prescribe specify substance weight
26.	astronomical binary celestial galactic gravitational orbital radiation stellar telescope velocity
27.	cheese dish egg fruit layer leaf meat oven rice tray
28.	appropriate authority case document guideline investigation legal necessary regulation submit
29.	acid biological chromosome cluster determine interact observe similarity structure visualize
30.	affect concern cost development environmental provision quality relate reproductive resource
31.	care health licensed medical nurse provider qualified skilled specialty technician
32.	binary decomposition infinite molecule parameter possible radiation ratio sphere unstable
33.	attacker contract ensure identity malicious protect protection provider trust user
34.	container functionality handle item lock normal optional slot thread type
35.	acquire appraisal author baby device plentiful poor sandwich schizophrenia tailor
36.	cancer cause genetic immune likely malignant occur patient syndrome viral
37.	concern government information legal political public regard society technology topic
38.	bubble gas interstellar medium outflow shock supersonic turbulence turbulent wind
39.	cool heat load plate roof rotate stack tray underneath wrap
40.	attempt collapse crush escape knock push save ultimate unable unconscious
41.	automate benefit health human infrastructure life online public quality user
42.	academic career degree graduate medicine nursing program science student university
43.	amp award consultant deliver new radio scientist staff technology visual
44.	abolish administer annex autonomy dominion mandate sovereign statute territorial treaty
45.	duct ear genital insert lip muscle nerve nipple tissue vagina
46.	align architecture benign command embryonic legal population strange superficial team
47.	historical news perspective reader recommendation researcher science summarize summary try
48.	barrel bolt flame knife metal needle rod rope thread wire
49.	barbecue cuisine dish grill lamb meal pork potato spicy stew
50.	adverse benefit decrease efficacy long-term prevent short-term stress surgery sustain
51.	aftermath avalanche blast collapse damage earthquake explosion landslide massive tsunami
52.	application component design different handle process quality technique typical use

53.	aim automation community document effort expert goal language machine vision
54.	bring challenge engineering functionality practical protection safety threat usage vulnerability
55.	abdominal anemia condition disorder liver lung pain suffer syndrome ulcer
56.	book brother child early finally fine originally piano queen sir
57.	attacker choose client cost decision game maximize objective selfish strategy
58.	accept associate book early inscription middle parish queen seven valuable
59.	build business company engineering intelligence methodology practitioner predictive student tool
60.	atmospheric barrier conventional electron interference internal layer noise radiation thermal
61.	act allow ban discrimination government legislation permit prevent refusal removal
62.	chassis conventional diesel driver fit gear manual maximum speed vehicle
63.	accelerator advance advanced facility offer optic physics promise science versatile
64.	abdominal abnormality blood cardiovascular diagnostic gastrointestinal pain respiratory surgery tissue
65.	argument civilization critique emphasize idea knowledge linguistic phenomenon religious understanding
66.	behavioral institutional intervention nurse occupational practitioner prevention provider rehabilitation specialist
67.	apt bother bounce catalog excuse portrayal respectable royalty smoke strive
68.	drug fever lung paralysis polio prevent recover recovery suffer victim
69.	apologize honest quote remark respond sad smile surprised tell truth
70.	adversary broadcast internet node protocol route send service traffic transmission
71.	expert health participant peer people preference public receive share topic
72.	design enable equipment output package provide quality tool validate verification
73.	atom decomposition determine energy fluid mechanism observe phase ratio substrate
74.	contain core critical date distinct effectively hard mercury method true
75.	billion corporate equity finance financial invest investor portfolio retail telecom
76.	liver lung medication metabolism reduce renal respiratory secretion toxic urine
77.	amphitheater bog combustion construction install lowering parachute populous successive youthful
78.	automate detection electronic equipment measurement optical retrieval scan signal spacecraft
79.	bread fry meat menu onion pie pizza potato specialty vegetable
80.	broad irregular measure slight specimen spherical spore texture tip typical
81.	acknowledge astronomy baseline chapter climate economics explosion movement prize thing
82.	definition french industrial micro percentage post purity spot superior supplement
83.	advance communication computing development device industry platform promising sensor thing
84.	clean drink flush fresh kitchen pipe recycle supply wash waste
85.	algorithm bit detect fast feedback hardware implementation minimize mode slow
86.	characteristic characterize chemical condition diagnostic essential organism plasma precise understanding
87.	adverse brain complication induce muscle pain pregnancy sleep spontaneous surgical
88.	aesthetic criticism interpretation introduction lecture philosophy psychology study theoretical thesis
89.	algorithm arithmetic binary cpu logic manipulate output processing processor programmer
90.	application autonomous capability computing delivery modern networking resource smart software
91.	final finish goal injury preseason raider regular score season squad
92.	application capability desktop enable encryption hardware networking software technology wireless
93.	asleep bed morning notice sleep sneak wake walk watch worry
94.	advance analysis clinical develop high-quality method objective patient provide tool
95.	care health healthy nurse quarantine sanitary sanitation surgeon vaccination veterinary
96.	affordable availability development device hardware internet mobile need platform software
97.	application automate component display install integrate menu monitor server window
98.	aspect auditory behaviour emotional interaction learner psychology relate researcher understand
99.	advantage allow collaboration collaborative construction facilitate open opportunity platform sharing
100.	advantage analog camera card compatible converter modular processor storage use

E User Study Instructions

E.1 Primer on Task

Evaluating the relations between words from a computational lens serves to further the research and understanding of artificial intelligence linguistic research.

A group of words can be considered coherent if they share a similar theme. For example, the group "apples banana coconut durian" can be considered coherent as most people would identify "fruit", "food" or "tree" as the common theme or link.

However, some group of words might be more ambiguous and the common theme might not be as straightforward. For example, "trees ore corn hydrogen" might be considered incoherent to some, while others might identify the common theme as "resources".

Ultimately, it is up to one's personal preferences and experiences to decide on whether a group of words are coherent.

E.2 Task Instructions

You will be presented with 10 English words. These words belongs to the 20,000 most frequently used words, so it is unlikely that you will encounter strange words. If you do encounter words that you have never seen before, you are free to use a dictionary or search engine (e.g. Google).

You will then be asked to assign each word to groups, where each group contains words that you think are coherent when grouped together.

Given an example: alcohol athlete breakfast drink eat habit intake meal obesity sleep

Some might divide the words into two groups identifying Group 1 is "alcoholic"-themed and Group 2 is "healthy"-themed.

	Group 1	Group 2	Group 3	Group 4	Not Related
alcohol	O				
athlete		O			
breakfast		O			
drink		O			
eat		O			
habit		O			
intake		O			
meal		O			
obesity	O				
sleep		O			

In another example given: atom calcium component material reduction temperature titanium typical weight yield

Some might group most of the words as "chemistry"-themed.

	Group 1	Group 2	Group 3	Group 4	Not Related
atom	O				
calcium	O				
component	O				
material	O				
reduction	O				
temperature	O				
titanium	O				
typical					O
weight	O				
yield	O				

If you believe that certain word(s) do not belong in any group, select the "Not Related" option in the last column. There can be multiple words that are not related to each other.

For example: animal bed carrot fungible great osmosis paradise star telcommunication water

	Group 1	Group 2	Group 3	Group 4	Not Related
animal					O
bed					O
carrot					O
fungible					O
great					O
osmosis	O				
paradise					O
star					O
telcommunication					O
water	O				

We want to emphasise that there are no right or wrong answers for the tasks, we wish to capture your beliefs on what you think is "correct". We understand that at times, you might encounter words that belong to multiple groups, however to simplify the tasks, we ask that you be the tiebreaker and assign it to the word-group with the strongest similarity.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3, 4

☑ B1. Did you cite the creators of artifacts you used?
3 - cited 4 - ours

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Ethics

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3, Limitations, Ethics

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3, 4, 5, 6, Appendix C, D

C ☑ Did you run computational experiments?

4, 5, 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3, Appendix C

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4,5,6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

6

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix E, Institute Review Board application withheld for anonymity

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Ethics

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Ethics