Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents

Domenic Rosati
scite.ai / Brooklyn, NY

Abstract

Topic models represent groups of documents as a list of words (the topic labels). This work asks whether an alternative approach to topic labeling can be developed that is closer to a natural language description of a topic than a word list. To this end, we present an approach to generating human-like topic labels using abstractive multi-document summarization (MDS). We investigate our approach with an exploratory case study. We model topics in citation sentences in order to understand what further research needs to be done to fully operationalize MDS for topic labeling. Our case study shows that in addition to more human-like topics there are additional advantages to evaluation by using clustering and summarization measures instead of topic model measures. However, we find that there are several developments needed before we can design a well-powered study to evaluate MDS for topic modeling fully. Namely, improving cluster cohesion, improving the factuality and faithfulness of MDS, and increasing the number of documents that might be supported by MDS. We present a number of ideas on how these can be tackled and conclude with some thoughts on how topic modeling can also be used to improve MDS in general.

1 Introduction

Topic modeling, a common approach for extracting themes from scientific documents, is currently facing many challenges: methodological validity (Shadrova, 2021), validity of automated evaluation (Doogan and Buntine, 2021; Hoyle et al., 2021), and utility of classical approaches (Sia et al., 2020; Zhang et al., 2022). We propose an additional challenge: *are lists of words the best we can do for topic labels?*

Topic models have tended to represent a topic as a list of words. Traditional topic labels are supposed to be "a set of terms, when viewed together, enable human recognition of an identifiable category" (Hoyle et al., 2021). However, a set of terms do not align with our intuitive understandings of what a topic is: a common theme or concept explicated as a word, phrase, or natural language description (Shadrova, 2021). In this paper, we present an exploratory case study using multi-document summaries (MDS) as labels for clusters of citations in order to understand current limitations and future work needed for using abstractive topic labels for human-like topics of scientific documents. To our knowledge, it is the first work that proposes to use MDS for topic labeling on top of topic clusters constructed with contextualized embeddings.

In addition to word lists not aligning with natural understanding of what a topic is, Shadrova (2021) has presented an extensive criticism of why traditional topic models based on lexical overlap measures lead to problematic topic models. Namely that they they *fail to understand word sense and capture context*. Recent approaches have relaxed these restrictions when constructing topic clusters (Bianchi et al., 2021; Grootendorst, 2022) by using contextualized word embeddings. However topic labels in those models are still constructed as word lists drawn from documents such as through TF-IDF.

Some work has anticipated this challenge by developing topic representations with phrases (Popa and Rebedea, 2021) and summaries (Basave et al., 2014; Gourru et al., 2018; Wan and Wang, 2016). But those works tend to be extractive, drawing the phrase or summary from a single document in the cluster[1]. In the extractive setting, *there may be no existing and fluent phrase or sentence that is capable of describing all documents in the cluster* or there may be multiple and even conflicting subtopics in the cluster that require a longer abstractive representation for producing a factual summary.

[1] see Alokaili et al. (2020); Popa and Rebedea (2021) for recent abstractive works.

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 91–99

2 Proposed Method

2.1 Topic modeling as clustering and MDS

In order to address the issues presented above, we propose using abstractive MDS as an approach to topic labeling. Topic modeling can be reframed as a set of two tasks: (1) finding meaningful clusters for documents (Sia et al., 2020; Zhang et al., 2022) and (2) performing MDS on those individual clusters to find meaningful topic labels. In this framework, LDA (Blei et al., 2003) uses document-word distributions to construct clusters and word lists drawn from those clusters as a form of MDS. Since we are looking at abstractive MDS that moves beyond word lists, we propose that the *topic representation be a sentence or paragraph* but there is no reason why an abstractive MDS can't be trained to generate phrases or even word lists (see Alokaili et al. (2020)) since word lists may still be appropriate in some situations.

In order to accomplish this, one can first use an approach for document clustering that uses contextualized word embeddings to avoid the issues mentioned above. By separating the clustering step from the representation step, we can use separate measures of cluster coherence to evaluate the quality of document clusters before we proceed to topic representation. We can also use evaluations of resulting topic representations later as an additional step to inspect the quality of our topic clusters.

After obtaining document clusters, MDS models such as (Lu et al., 2020) can be used to produce natural language summaries that synthesize common themes from documents. Recent work on MDS within the scientific and biomedical domain (DeYoung et al., 2021; Lu et al., 2020; Shen et al., 2022) show good results in producing both single sentence (extreme) summaries as well as long form summaries over many scientific documents.

2.2 Evaluation

Topic model evaluation is challenging (see Chang et al. (2009); Hoyle et al. (2021); Doogan and Buntine (2021)). Traditional metrics like coherence (NPMI), perplexity, and diversity scores are studied in the context of topic word lists and validated with correlation to human ratings of the utility or coherence of those topic word lists. Since we suggest developing abstractive topic representations, we want a way to compare various forms of both abstractive and extractive topic representations presented by the model. Since we are treating representation as a summarization task and this task includes measures that work across extractive and abstractive settings, we suggest that we start with standard summarization metrics such as overlap metrics like Rogue (as used in Cui and Hu (2021) or semantic metrics such as BERTScore (Zhang et al., 2022) (as used in Alokaili et al. (2020)).

Model	Source
multi-lexsum-long	Shen et al. (2022)
multi-lexsum-tiny	Shen et al. (2022)
ms2	DeYoung et al. (2021)
multixscience	Lu et al. (2020)
topic lists	Bianchi et al. (2021)

Table 1: Generative models used for abstractive MDS topic representations.

3 Case study: how has a scientific document been cited?

To evaluate our proposed method, we chose topic modeling over scientific documents as a setting. While several methods exist for determining citation intent function (Basuki and Tsuchiya, 2022; Nicholson et al., 2021) and the relationship between two papers (Luu et al., 2021), there is very little work on topic models over citations (for some representative work on "citation summary" see Elkiss et al. (2008); Wang et al. (2021); Zou et al. (2021)). Topic representations of citations are interesting for characterizing trends in how a paper has been cited or helping researchers identify relevant citations to read among potentially thousands of other citations. In this work, we treat topic labels as a "citation intent" label and use the proposed approach to understand the utility of MDS for topic modeling in this setting.

4 Experimental Setup

We apply the method described in section 2 in order to identify clusters of citations and provide labels for those clusters without any supervision. Specifically, we present a case study of what this looks like on a single paper to illustrate the potential of our approach and try to assess future work needed in order to make MDS a good solution for topic labeling in general and citation summarization in particular.

For this study, we used scite.ai (Nicholson et al., 2021) to extract in-text passages which contained citations (citation statements) to the paper (Lau

Model	R-1	BERTScore
multi-lexsum-long	38	85
multi-lexsum-tiny	3	81
ms2	3	81
multixscience	15	80
topic lists	1	76

Table 2: Rouge-1 (R-1) and BERTScore (F1) results for each models topic representations measured against.

et al., 2014), a well known paper that introduces the NPMI metric in topic modeling. This resulted in 183 citation statements which is the corpus we will use for topic modeling.

In order to identify meaningful groups of clusters we use contextualized topic models (CTM) (Bianchi et al., 2021) since this method uses contextualized word embeddings (we used SPECTER for constructing embeddings (Cohan et al., 2020)). We selected CTM since we still get word lists as topic labels which we used for evaluation. In order to select the number of topics hyperparameter, we trained CTM several times steadily increasing the number of topics from 3 to 50 and selected the best model according to coherence (NPMI) resulting in a 10 topic model (see Appendix A for more details) over 183 citation statements.

The models selected for generating abstractive MDS are outlined in Table 1. All MDS models used are based on the longformer architecture (Beltagy et al., 2020) and used beam search (5 beams) with greedy decoding.

5 Results

Table 2 shows the Rogue-1 (R-1) and BERTScore (average F1 across topics) for each of the models selected for generating topic representations using MDS as well as the topic lists generated by CTM. It is important to underscore that R-1 and BERTScore are not validated against human studies for topic representations and this is simply a small case study on what an approach might look like. In spite of this, our results paint an initial picture of how these methods perform, especially when compared to model outputs (see Appendix B for samples). Topic word lists have the worst R-1 and BERTScore. The MDS models do a little bit better with multi-lexsum-long having the best overall score. multixscience also does well with regards to R-1. Since multixscience and multi-lexsum-long are long form summaries, it appears that R-1 is potentially biased towards longer summaries and may not be a good measure across representations, in particular it may be uninformative for evaluating the performance of topic lists. ms2 and multi-lexsum-tiny are smaller and have better BERTScore then multixscience indicating they might provide more semantically similar representations. We are also not sure whether BERTScore suffers from the same bias towards longer or more sentence-like inputs.

We randomly sampled 3 topics to explore their representations. As an example, table 3 shows representations using the multi-lexsum-tiny model (full details are available in Appendix B. In representations for topic 0 (Table 5), we see there is a general agreement across models that the citing documents are discussing measurement. We can see that the topic representations appear to be split between measuring interpretability (multixscience, multi-lexsum-long) and those discussing the correlation between measures (ms2, multi-lexsum-long) or even potentially an additional topic of describing measures used (multi-lexsum-tiny). Conflicting summaries are not surprising given issues in MDS with regards to summarizing diverse and potentially conflicting documents (DeYoung et al., 2021). Table 5 shows a diversity of topic labels that might be appropriate under different scenarios of applying topic models. Labels like the ones in Table 3 might be useful for labels that are easy and fast to read while longer summaries in multixscience and multi-lexsum-long might be useful for users who want to engage deeper.

6 Discussion

In order to ensure downstream topic labels are coherent, document clusters must represent meaningful and well separated clusters. Grootendorst (2022); Sia et al. (2020); Zhang et al. (2022) have shown that traditional clustering methods might provide good candidates for moving beyond topic models like LDA that suffer from lack of contextualized natural language understanding due to their use of word co-occurence statistics for constructing topic clusters. However in order to fully replace traditional methods we would like to see: (1) the demonstration of effective mixed-membership approaches in abstractive topic modeling to recover the ability for documents to belong to multiple topic clusters, (2) the demonstration of cluster evaluation measures that correlate well with how hu-

Topic 0
NPMI and Topic Coherence are measures used to measure the semantic coherence of topics.
Topic 4
Topic model quality and interpretability are two different metrics used to measure the semantic interpretability of a topic.
Topic 2
Evaluation metrics: Log predictive probability (LPP) and topic interpretability

Table 3: Topic representations produced by multi-lexsum-tiny. Compared to word lists they are much more readable and closer to everyday notions of topics.

mans might group documents and possibly (3) the development of fully learnable architectures where clustering might be learned with feedback from topic representation quality.

DeYoung et al. (2021) has shown that MDS struggles with factual consistency. We see an opportunity for topic clustering as a step before performing MDS as a potential method for improving factual consistency since a contradicting source document that would normally be in the document set might be separated out with initial topic clustering. Furthermore, initial topic clustering might provide a way for developing more granular multi-aspect summarization techniques by clustering documents by aspect. Either way, we are weary of the known issues with factuality in MDS (DeYoung et al. (2021)) especially in the scientific domain where factual consistency is critical. To develop our approach along these lines, we suggest continuing to extend evaluation of factuality and faithfulness to the MDS setting (as identified in (DeYoung et al., 2021)).

In order to make this approach work for a wide variety of application and analysis scenarios, controllable summarization (such as Keskar et al. (2019)) should be investigated so that users can control for length of summaries (such as question, phrase, sentence, or paragraph) or style of summary (such as in the style of a paper title, abstract, citation, or literature review). Additional controls such as the ones suggested in Shadrova (2021) like granularity of topic label can also be developed in a controllable summarization framework in such a way as to make topic representations better fit for user's needs.

Finally while methods like longformer (Beltagy et al., 2020) enable the use of transformers with multiple documents as input, more research needs to be done to enable a method like the one we proposed on large sets of documents. In the scientific domain, where we might want to model hundreds or even thousands of full-text articles belonging to a single cluster, the approaches presented would be intractable without further development of long-attention transformer models.

One advantage of our approach is that since we are breaking topic models out into clustering and MDS as separate steps we can rely on a established work for evaluation of document clusters and summaries to assess models performance. While we'd need to validate the application of these metrics in end-to-end topic modeling scenarios, if text clustering and summarization metrics do correlate with human judgements of topic cluster and representation quality then we can avoid using topic modeling metrics which have come into question repeatedly Chang et al. (2009); Hoyle et al. (2021); Doogan and Buntine (2021)). However, we will not know this until we design robust human studies to validate the approach we have proposed above.

7 Conclusion

In this paper, we presented a reframing of topic modeling as document clustering with MDS applied to produce topic representations that might (1) align more intuitively with what humans understand as topics and (2) overcome some of the issues with topic models using bag of word assumptions such as inability to capture context. An initial case study on using this approach for unsupervised discovery of citation intents was explored. We found that while cohesive alternatives to topic representations can be produced using MDS in a variety of styles (short and long summaries), there are still many obstacles that need to be overcome before we can fully evaluate whether this approach could provide a viable alternative to traditional topic modeling and representation. Namely, improving cluster cohesion, improving the factuality and faithfulness of MDS, and increasing the num-

ber of documents that might be supported by MDS. While there might be an advantage in utilizing well validated approaches for evaluating clustering and summarization as measures of our approach, future studies will need to validate those with human studies. It is our hope that further work in this area can use our discussion as a roadmap towards what needs to be done if we want to move past word lists as topic representations.

References

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2020. Automatic Generation of Topic Labels. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. 2014. Automatic Labelling of Topic Models Learned from Twitter by Summarisation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Setio Basuki and Masatoshi Tsuchiya. 2022. SDCF: semi-automatically structured dataset of citation functions. *Scientometrics*, 127(8):4569–4608.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

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

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.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Peng Cui and Le Hu. 2021. Topic-Guided Abstractive Multi-Document Summarization. *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MS^2: Multi-Document Summarization of Medical Studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Caitlin Doogan and Wray Buntine. 2021. Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua*.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20707.

Antoine Gourru, Julien Velcin, Mathieu Roche, Christophe Gravier, and Pascal Poncelet. 2018. United we stand: Using multiple strategies for topic labeling. In *NLDB: Natural Language Processing and Information Systems*, volume LNCS, pages 352–363, Paris, France. Issue: 10859.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv:2203.05794 [cs].

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. ArXiv:2107.02173 [cs].

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation.

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.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles. ArXiv:2010.14235 [cs].

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining Relationships Between Scientific Documents. 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 2130–2144, Online. Association for Computational Linguistics.

Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3):882–898.

Cristian Popa and Traian Rebedea. 2021. BART-TL: Weakly-Supervised Topic Label Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1418–1425, Online. Association for Computational Linguistics.

Anna Shadrova. 2021. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining &Amp; Digital Humanities*, 2021.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. MultiLexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Xiaojun Wan and Tianming Wang. 2016. Automatic Labeling of Topic Models Using Text Summaries. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Mingyang Wang, Dongtian Leng, Jinjin Ren, and Peng Yu. 2021. Generating a Citation Summary Based on Cited Sentences and the Implied Citation Emotions. *IEEE Access*, 9:18042–18051. Conference Name: IEEE Access.

Zihan Zhang, Meng Fang, Ling Chen, and M. Namazi-Rad. 2022. Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. *undefined*.

Lixue Zou, Xiwen Liu, Wray Buntine, and Yanli Liu. 2021. Citation context-based topic models: discovering cited and citing topics from full text. *Library Hi Tech*, 39(4):1063–1083. Publisher: Emerald Publishing Limited.

A Topic Model Selection

Table 4 describes the evaluation of all the CTM (Bianchi et al., 2021) models we trained by coherence (NPMI), diversity, and topic number. We trained the model on a dataset of 183 citation statements which are in-text passages from publications citing (Lau et al., 2014) retrieved from scite.ai (Nicholson et al., 2021). This dataset was also used as the reference corpus for evaluating NPMI in this initial case study. We acknowledge that when building robust topic models a standard reference corpus should be used so results can be comparable in future works.

coherence (NPMI)	diversity	topics
-0.25	0.97	10
-0.27	0.97	20
-0.32	0.96	25
-0.32	0.97	15
-0.35	1.00	5
-0.38	1.00	3
-0.38	0.97	50

Table 4: Selecting CTM topic model by evaluating CTM coherence (NPMI) and diversity on different topic numbers parameters.

B Topic Representations

Three randomly sampled topics are presented here so readers can inspect the quality of the topic labels as well as the original topic labels produced by CTM.

Model: topic lists
measure perplexity npmi probability paper correlation high log_p ranking correspond
Model: ms2
In conclusion, NPMI has a positive correlation with the human evaluation of topic coherence.
Model: multixscience
NPMI has been shown to correlate better with human evaluation of topic coherence than PMI cite cite cite. The NPMI measure was first introduced by Lau et. al cite and has been used extensively in the literature since cite. It is used to evaluate the coherence of the top @math words of a topic and to measure the difficulty of associating an individual topic with a single semantic concept, and consequently, evaluating topic models by their interpretability. It has also been used as a measure of the performance of automatic topic evaluation methods, and has shown to be more correlated with human performance than other traditional topic measures such as perplexity cite, co-occurrence cite cite, and held-out likelihood cite. In this paper, we focus on NPM I since it has been found to have a higher correlation with the human evaluation than other topic measures, such as PMI, and will be used as the evaluation measure in this paper. We refer the reader to cite cite for more details.
Model: multi-lexsum-tiny
NPMI and Topic Coherence are measures used to measure the semantic coherence of topics.
Model: multi-lexsum-long
For simplicity, we will refer to this measure as NPMI since it has been shown to have a higher correlation with the human evaluation of topic coherence than PMI (Lau et al, 2014). For more on this topic, see our related article. Introduction to Topic Coherence. In the literature, human-annotated topic models are usually measured using the Coherence and Normalized Pointwise Mutual Information (NPMI), which consider the probability that both terms in a particular term pair, where frequent terms will be downweighted with respect to those that occur less frequently, such as education from this example. However, in this paper, the authors adapted it to measure the difficulty of associating an individual topic with a single semantic concept, and consequently, evaluating topic models by their interpretability. The authors of this paper also used the predictive Probability and Predictive Probability to see the performance of the learning methods. The results of this study were shown to be significantly improved compared with the counterparts without PPMI, especially in terms of perplexity and topic-coherence and with low proportion of the training words. The performance measure was estimated using 2 measures to evaluate the methods: Log predictive probability (LPP) which considers the generalization of a model and Normalization pointwise mutual information (NPI) which evaluates the coherence in the most significant words (top words) of a topic. The coherence score was estimated through the point-wars mutual information. The Coherence Score was calculated using the standard linear predictive probability method and was estimated to correlate well with human judging.

Table 5: Outputs for topic #0

Model: topic lists
automate measure mutual correlate röderal coherence draw recent pointwise topic
Model: ms2
The meta- analysis showed that topic coherence was positively correlated with the interpretability of the topic models. In contrast, the meta-analyses did not find a significant association between topic co-occurrence and interpretability.
Model: multixscience
Topic coherence is a measure of the interpretability of a topic model. It is based on the coherence of the words assigned to a topic and has been studied extensively in the context of topic modeling. Various measures have been proposed to measure topic coherence, such as the pointwise mutual information (PMI) between the topic words and the co-occurrence frequency of these words in the reference corpus, as well as the number of topics in the model. The PMI-based methods have been widely used in the evaluation of topic models (see for example cite cite and the references therein). However, these methods do not take into account the internal representation of the topic models. To the best of our knowledge, there is no prior work that evaluates the topic interpretability by measuring the PMI. However, there has been a large body of work on evaluating topic models by measuring their interpretability, including methods based on model perplexity, coherence, predictiveness cite, NPMI, topic diversity, and distributional semantics
Model: multi-lexsum-tiny
Topic model quality and interpretability are two different metrics used to measure the semantic interpretability of a topic.
Model: multi-lexsum-long
More specifically, Chang et al showed that models that fare better in predictive perplexity often have less interpretable topics, suggesting that evaluation should consider the internal representation of topic models and aim to quantify their interpretability. The idea soon gave rise to a new family of methods (Newman et al, 2010). Auto-Auto-NPMI that evaluate the semantic interpretability by measuring the number of chosen topics. These methods assume that topic coherence correlates with the coherence of the words assigned to that topic and thus quantify topic model quality. The resulting output does not explicitly provide an interpretation to human users, but remains a list of words with associated topic probabilities, which humans must interpret (as rated by human annotators) despite the observation that evaluation methods such as perplexity are often not correlated with human judgements of topic quality. According to the most recent work in topic design evaluation, indirect approach with word intrusion, and direct approach with observed coherence are two major approaches to measure the human interpretability in ML algorithms. The direct approach is to choose the model with the lowest perplexity on unseen data, which may generate topics that are hardly interpretable, and the direct approach uses the automated evaluation measure OC-Auto NPMI in Grieser et al. that evaluates the co-occurrence frequencies of terms within a reference corpus. Both methods have been developed to provide empirical evidence for the number chosen topics, because they are based on topic diversity; it is the percentage of unique words in the top 25 words of all topics, without mention of obstruction 5/10/20 topic words. The second method is to use coherence, which is to compute topic diversity by measuring topic diversity. The third method uses coherence to offer empirical evidence.

Table 6: Outputs for topic #4

Model: topic lists
compute word mutual coherence draw interpretability similarity assess pointwise human
Model: ms2
Meta-analyses of topic coherence showed that the coherence of topics was in line with human evaluation of the interpretability of the topics.
Model: multixscience
Topic coherence is a quantitative measure of the interpretability of individual topics. It is the average pointwise mutual information of two words drawn randomly from the same document cite. The coherence between top words within a topic is estimated using the PMI between topic words cite cite cite. Various formulations have been proposed to compute topic coherence, including those based on the NPMI cite cite, PMI and its variations cite, the Normalised PMI cite, and the Point-wise Mutual Information (PMI) cite. Topic coherence scores judged by human annotators cite cite are used as a measure of topic interpretability. The most popular evaluation metrics are LPP cite, which measures the generalization of a topic model on unseen data, and NPMI cite, that measures the coherence of the topics. However, LPP is not the best measure for evaluating topic coherency.
Model: multi-lexsum-tiny
Evaluation metrics: Log predictive probability (LPP) and topic interpretability
Model: multi-lexsum-long
Evaluation metrics: Log predictive probability (LPP) and Normalized pointwise mutual information (NPMI) are used. While LPP measures the generalization of a model on unseen data, NPMI examines the coherence and interpretability of the learned topics. For each topic t, Experiments show topic coherence (TC), which is in line with human evaluation of topic interpretability, and Experiments ShowTopic Coherence Experiments (TC) computed with the Coherence between a topic's most representative words (e.g., top 10 words) is inline with human eval of topic interpretationability. As the reference corpus for computing word occurrences, we use the English Wikipedia. As various formulations have been proposed to compute TC, we refer readers to Röder et al. (2015) for more concrete ways to see how the topic models interact with each other. To quantitatively measure the interpretability or the semantic quality of individual topics, we used the observed coherence measure from (Lau et al., 2014), which was adopted from psychology theory and showed better topic interpretation compared with other measures [1, 2]. In addition to the above measures, we looked for the observed relationship between the topic and human interpretation of topic models. The observed correlation between the top N words within a topic and its coherence between the bottom 10 words was inline with the human evaluation in evaluations 2-5 8. It is a preferred method for such tasks (Aletras and Stevenson, 2013;Newman and al, 2010a) as it is unaffected by variability in the range for each dataset.

Table 7: Outputs for topic #2