Hierarchical Topic Evaluation: Statistical vs. Neural Models

Mayank Kumar Nagda Charu James Marius Kloft Sophie Burkhardt
Department of Computer Science, TU Kaiserslautern, Kaiserslautern 67653, Germany
{mnagda21, charu, kloft, burkhardt} @cs.uni-kl.de

Abstract

Hierarchical topic models (HTMs)—especially those based on Bayesian deep learning—are gaining increasing attention from the ML community. However, in contrast to their flat counterparts, their proper evaluation is rarely addressed. We propose several measures to evaluate HTMs in terms of their (branch-wise and layer-wise) topic hierarchy. We apply these measures to benchmark several HTMs on a wide range of datasets. We compare neural HTMs to traditional statistical HTMs in topic quality and interpretability. Our findings may help better judge advantages and disadvantages in different deep hierarchical topic models and drive future research in this area.

1 Introduction and Related Work

Topic modeling is a statistical technique that helps to discover sets of words (called *topics*) concisely describing a text's semantics [2]. Especially popular are unsupervised hierarchical topic models (HTM), which aim to find a *hierarchy* of topics [1], without requiring to annotate or label documents.

Many HTMs have been created as variations of hierarchical latent Dirichlet allocation (HLDA) [1]. Recently, neural topic models have enabled great advances in hierarchical topic modeling [12, 8].

In particular, several Bayesian deep learning methods exist for hierarchical topic modeling, most of which are based on variational methods [6, 7].

(Unsupervised) topic models learn topics from an unlabelled document collection. Interpreting the topics in absence of labels is challenging. Hence, the evaluation of topic models has attracted great interest from the ML community. Most earlier work considers assessing topics with low holdout perplexity [13]. While perplexity can be an excellent statistical measure of topic models for model selection and parameter tuning, it does not evaluate the topics' semantic quality. Human evaluation of topic models has proved that perplexity-based evaluation can be counter-intuitive: in studies humans have picked models with high perplexity [4].

Previous work has focused on measures such as *topic coherence* [10, 9] to determine the semantic quality of topics. Coherence measures compute the semantic agreement of frequently co-occurring words. The latest trend in semantic evaluation are *topic redundancy measures*, which account for the component collapse in topic coherence measures [3]. Topic redundancy has been combined with coherence scores to give *topic quality* [5].

In contrast to flat topic modeling (where the model learns unstructured topics), HTM learns topic hierarchies. Commonly, HTMs represent the topic hierarchies by directed acyclic graph (DAG) or a tree. This assumes that the topics in the upper layers are more general/abstract than those in the lower layers. Every node represents a topic in a tree, and every edge (branch) represents a hierarchy. Unstructured quantitative evaluation measures (such as those mentioned in the previous paragraph) cannot evaluate the hierarchical relationships of HTMs and are thus inadequate for HTMs. This forces researchers to use qualitative methods and human judgment to assess the hierarchical topics.

Bayesian Deep Learning workshop, NeurIPS 2021.

In this paper, we propose a quantitative evaluation framework for HTMs. Our work focuses on modifying the topic quality measure to account for hierarchical topics, but it could also be adapted to other base measures. For simplicity, we focus on tree-structured HTMs, which account for most neural HTMs. We divide the evaluation of the topic tree into branch- and level-wise assessments. Each branch here represents a hierarchy, and every level covers all its nodes to account for the topic distribution and component collapse. In our experiments, we use the new evaluation technique to benchmark various HTMs. Our study reveals that previous qualitative evaluations of HTMs have lead to deceived comparisons. (Research articles lack the space to present the whole topic hierarchy.) It also demonstrates how neural topic models can discover reasonable flat topics (and document reconstruction), but traditional statistical approaches outperform them in discovering a meaningful hierarchical structure.

2 Proposed Methodology

We propose a novel approach that allows constructing coherence measures for a hierarchical tree structure. The proposed method builds on the intuition that strong coherence in nodes and branches of a tree can result in more interpretable hierarchical topics. We compute the coherence in branches and levels of the tree and use the resulting coherence to calculate branch topic quality (BTQ) and level topic quality (LTQ), the aggregation of which gives us hierarchical topic quality (HTQ). We use BTQ, LTQ, and HTQ for the quantitative assessment and benchmarking of the topic models.

We compute coherence in the tree structure by extending the existing unifying framework proposed by Röder et al. [11]. The framework represents a coherence measure as a composition of four sets: segmentation (S), confirmation measure (M), probability estimation (P), and aggregation (Σ). The authors then group these sets in dimensions that span a configuration space (C). It is described as C = S x M x P x Σ. In summary, the S set represents ways to divide a word set into smaller pieces. The set M describes different kinds of confirmation measures that score the agreement of a word pair, e.g., the normalized pointwise mutual information (NPMI) of two words. The set P comprises different methods to estimate word probabilities in the set M. Finally, the set Σ consists of methods used to aggregate the scalar values computed by a confirmation measure. We add one more dimension to the configuration space in our approach, representing hierarchical word sets (H_W). We define the modified configuration space as $C^{(h)}$ = H_W x S x M x P x Σ.

The hierarchical word set (H_W) modifies the input of the segmentation set to account for the hierarchical structure. (H_W) collectively represents all the word sets formed by topics in branches and levels of a tree. For a particular tree structure, the word set of a specific branch (W_B^b) includes words from both the parent node and the child node, representing a hierarchy. Similarly, the word set (W_L^l) accounts for all the nodes at a particular level in a tree. We use word sets W_B and W_L for computing BTQ and LTQ, respectively.

For computing coherence from the word sets, we use elements from the configuration space of C_V because it correlates well to human topic ratings. The C_V measure combines the indirect cosine measure with the Normalized Pointwise Mutual Information (NPMI) and the boolean sliding window [11]. The elements which we use directly from the C_V measure are for the segmentation (S_{one}^{set}), the probability estimation (P_{sw}), and the confirmation measure (ϕ). S_{one}^{set} compares every single word to the whole word set and forms a set of pairs of word sets. It is defined as: $S_{one}^{set} = \{(W_x, W_y)|W_x = W_i; W_i \in W; W_y = W\}$. The Boolean sliding window(P_{sw}) determines word counts using a sliding window to form new virtual documents, which are used for computing word probabilities. We use a sliding window of size 110. The confirmation measure ϕ takes word subsets (S_i) with corresponding probabilities to compute a similarity measure using context vectors. We can exemplify one such context vector as: $\vec{v}(W^{'}) = \{\Sigma_{w_i \in W'} NPMI(w_i, w_j)\}_{j=1,..,|W|}$. These context vectors then combine the indirect cosine measure with NPMI as described in (1a) and (1b).

$$NPMI = \frac{\log \frac{P(W_x, W_y)+\epsilon}{P(W_x)P(W_y)}}{-\log P(W_x, W_y) + \epsilon} \quad \text{(1a)} \qquad \phi(\vec{u}, \vec{w}) = \frac{\Sigma_{i=1}^{|W|} u_i \cdot w_i}{||\vec{u}||_2 ||\vec{w}||_2} \quad \text{(1b)}$$

As for aggregation, we introduce a diversity term following [3, 5]. All confirmation measures $\phi_N = \{\phi_1, \phi_2 ... \phi_{|s|}\}$ of all subset pairs S_i are aggregated using the arithmetic mean and are multiplied with the topic diversity (d) to produce the topic quality index. We compute BTQ and LTQ for all hierarchies and levels in the tree. It is formulated as: $BTQ = \frac{\sum_{b=1}^{B} \phi_i^{W_B} \cdot d_b}{B}$ and $LTQ = \frac{\sum_{l=1}^{L} \phi_i^{W_L} \cdot d_l}{L}$. The arithmetic mean of BTQ and LTQ is our final hierarchical topic quality (HTQ) of a complete topic model.

3 Results and Discussion

We ran our experiments on a range of datasets: the 20NG dataset[1] (18,846 documents), two subsets of Reuters 21578 dataset[2] (R8 with 7674 documents and R52 with 9100 documents), and the (very large) NYT Corpus[3], which serves to replicate real-world usage as it contains news articles from the past century. We benchmark multiple HTMs: nCRP [1] (which is a statistical model), TSNTM [7] (a hybrid of a statistical and a neural model), and SawETM [6] (a neural model). We also use a baseline model consisting of a neural flat topic model combined with agglomerative topic clustering (based on the cosine similarity of TF-IDF vectors). For benchmarking, we restrict all the models to level 3 in terms of tree height. We use the same hyperparameters for the models as provided by the authors. Table 1 shows the results. We observe that the neural models have good document reconstruction and topic quality in layers (LTQ) but lack behind in hierarchical representation.

With the help of our evaluation measures, we can pick branches and levels with low topic quality in a topic model. It enables us to perform an in-depth qualitative study of a specific region in a tree structure[A]. This is helpful as it saves researchers and practitioners from the tedious and inaccurate human evaluation of a relatively large tree structure (recent neural HTMs can have a depth of 15-layers [6]).

Table 1: Performance of Models

Models	*nCRP*			*TSNTM*			*SawETM*			*baseline*		
Datasets	BTQ	LTQ	HTQ	BTQ	LTQ	HTQ	BTQ	LTQ	HTQ	BTQ	LTQ	HTQ
20NG	**0.387**	**0.456**	**0.421**	0.381	0.365	0.373	0.312	0.390	0.351	0.241	0.373	0.307
R8	0.331	**0.382**	**0.356**	**0.366**	0.347	**0.356**	0.237	0.345	0.291	0.217	0.326	0.271
R52	**0.417**	0.264	0.341	0.336	**0.375**	**0.355**	0.241	0.363	0.302	0.203	0.315	0.259
NYT NC	**0.531**	**0.472**	**0.501**	0.523	0.460	0.491	0.389	0.407	0.398	0.253	0.386	0.319

4 Conclusion

We proposed a novel evaluation measures for hierarchical topic models based on the topic coherence measure (C_V). The evaluation measures help in benchmarking various topic models on the hierarchical topic quality and provides an in-depth analysis of the produced tree structure. Experimental results demonstrate that traditional statistical hierarchical models discover better hierarchical representation of documents as compared to newer Bayesian deep learning methods. The proposed evaluation measures can serve as a tool for future research in the field to perform comparative studies on hierarchical topic models.

Acknowledgement

The authors acknowledge support by the Carl-Zeiss Foundation, the BMBF awards 01|S20048, 01IS18051A, 031B0770E, and 01IS21010C, and the DFG awards KL 2698/2-1 and KL 2698/5-1.

[1] `http://qwone.com/~jason/20Newsgroups/`

[2] `https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection`

[3] `https://www.kaggle.com/tumanovalexander/nyt-articles-data`

References

[1] David M Blei, Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, et al. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, volume 16, 2003.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27, 2019.

[4] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

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

[6] Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. Sawtooth factorial topic embeddings guided gamma belief network. In *International Conference on Machine Learning*, pages 2903–2913. PMLR, 2021.

[7] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Tree-Structured Neural Topic Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 800–806, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.73. URL `https://aclanthology.org/2020.acl-main.73`.

[8] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2017.

[9] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

[10] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108, 2010.

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

[12] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[13] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

A The curse of component collapse

Topic models are among the most valuable methods for learning hidden text representations in an extensive collection of documents. Neural topic models rely on Bayesian neural techniques such as variational autoencoders (VAEs). The ML community has long reported that neural topic models suffer from a well-known problem of component collapse, which results in similar topics and hinders the overall representation of documents.

The results of our quantitative analysis are backed by qualitative results. Figures 1 and 2 visualize topics from the last two layers of the neural methods SawETM and TSNTM as obtained in our experiments. Figure 3 shows topics from the traditional nCRP model. All the models are trained on the same version of the R8 dataset with the same pre-processing techniques. These figures reveal the presence of component collapse in the neural SawETM and TSNTM. We can observe the presence of similar topics within these models. Whereas, in the statistical nCRP model, the topics are more diverse and discover a better document representation.

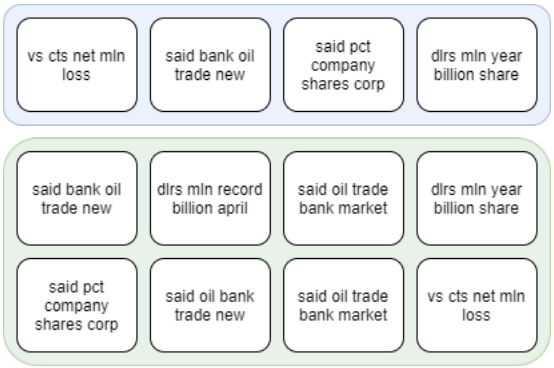

Figure 1: Topics by SawETM in last two layers

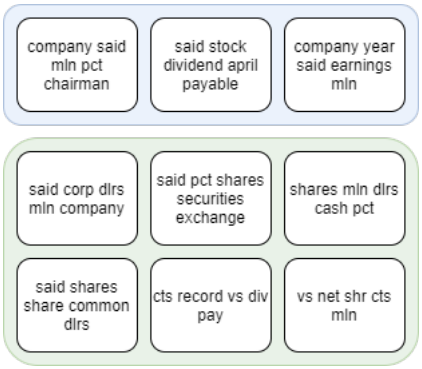

Figure 2: Topics by TSNTM in last two layers

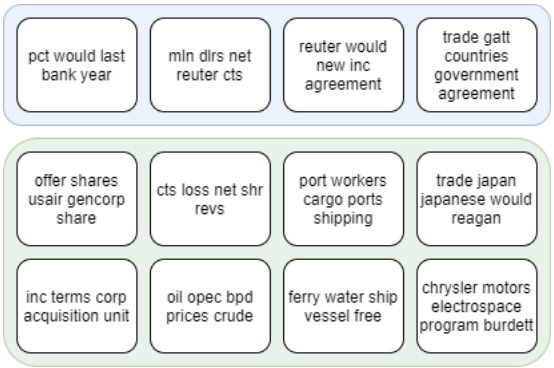

Figure 3: Topics by nCRP in last two layers

B Qualitative Analysis

The proposed hierarchical evaluation measures computes the topic quality score across all the branches and levels of a tree structure. These scores serve as a tool to analyze the topic hierarchies. Figure 4 demonstrates topic quality scores in a sub-tree formed by TSNTM on 20NG dataset.

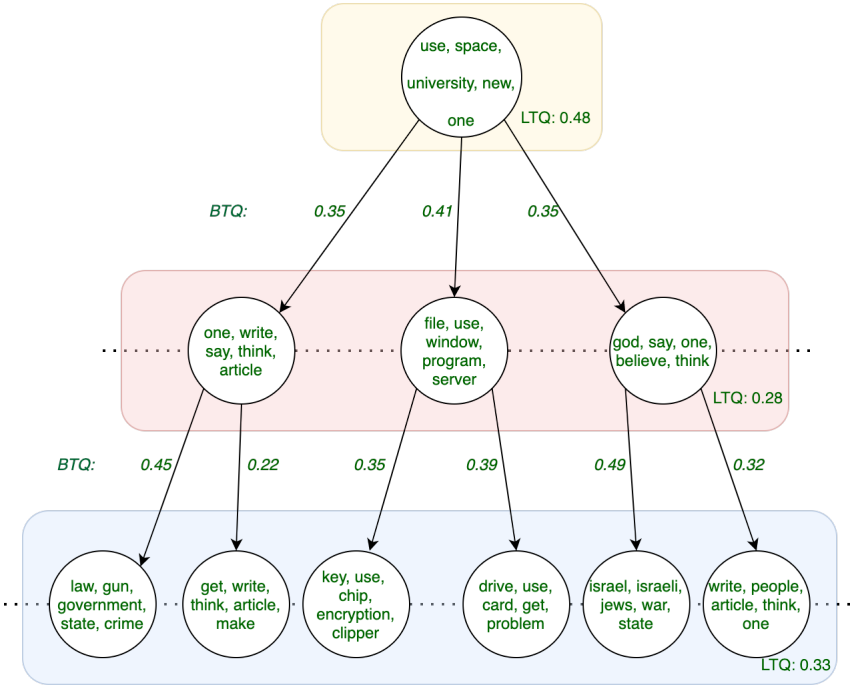

Figure 4: A topic sub-tree with labeled BTQs and LTQs