BART-TL: Weakly-Supervised Topic Label Generation

Cristian Popa
Universitatea Politehnica of Bucharest
cristian.viorel.popa@gmail.com

Traian Rebedea
Universitatea Politehnica of Bucharest
traian.rebedea@upb.ro

Abstract

We propose a novel solution for assigning labels to topic models by using multiple weak labelers. The method leverages generative transformers to learn accurate representations of the most important topic terms and candidate labels. This is achieved by fine-tuning pre-trained BART models on a large number of potential labels generated by state of the art non-neural models for topic labeling, enriched with different techniques. The proposed *BART-TL* model is able to generate valuable and novel labels in a weakly-supervised manner and can be improved by adding other weak labelers or distant supervision on similar tasks.

1 Introduction

As topic modeling has been used for unsupervised exploration of large text corpora, several topic labeling approaches have been proposed. These range from heuristic-based methods (Mei et al., 2007; Gourru et al., 2018) that focus on the underlying topic distributions to newer methods that use word embeddings (Bhatia et al., 2016). Supervised topic labeling methods (Lau et al., 2011; Bhatia et al., 2016) typically use annotator data with the quality of the labels to train a more accurate ranker than the unsupervised counterpart. Deep learning approaches, which gained quick popularity in NLP, are starting to be used for solving this task as well (Sorodoc et al., 2017; Alokaili et al., 2020).

Recently, transformer models pre-trained on very large amounts of data achieved impressive results on a lot of downstream NLP tasks using fewer resources than previously necessary. We introduce a method of performing a weakly-supervised fine-tuning on these models pre-trained on English data in order to obtain human-comprehensible and meaningful topic labels. We also provide a quality evaluation of the model-generated labels, in addition to an analysis of the contribution gained from using this approach that we ultimately refer to as *BART-TL*, inspired by the name of the original transformer architecture.

2 Related Work

Topic modeling is a popular unsupervised method for exploring large corpora of documents. Topics are represented as distributions over words, while documents as mixtures of topics. Historically, these methods used dimensionality reduction techniques (Deerwester et al., 1990), then migrated to probabilistic-based methods (Hofmann, 1999), with Latent Dirichlet Allocation (Blei et al., 2003) gaining popularity. LDA makes use of variational inference to obtain the distribution matrices. Further developments include hierarchical (Wang et al., 2011) and online (Hoffman et al., 2010) versions of LDA.

While the resulting distributions of topic models are useful for computational purposes, such as measuring the similarity of two documents, these may prove difficult to interpret by humans. Topic labeling aims to solve this issue by computing labels for each topic. Historically, this was achieved by establishing a pool of labels and ranking them using certain scoring functions. First attempts were fully unsupervised, extracting labels from the original corpus (Mei et al., 2007). Later approaches started using external corpora, such as Wikipedia, as candidates for labels and trained supervised rankers (Lau et al., 2011), as well as employed word embeddings (Bhatia et al., 2016) such as *word2vec* (Mikolov et al., 2013) and *doc2vec* (Le and Mikolov, 2014) for computing the similarity between a topic and a candidate label.

Huge progress was made in the NLP field with the introduction of attention models (Bahdanau et al., 2014) and, later on, transformers (Vaswani et al., 2017), which are deep neural networks that

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1418–1425
April 19 - 23, 2021.

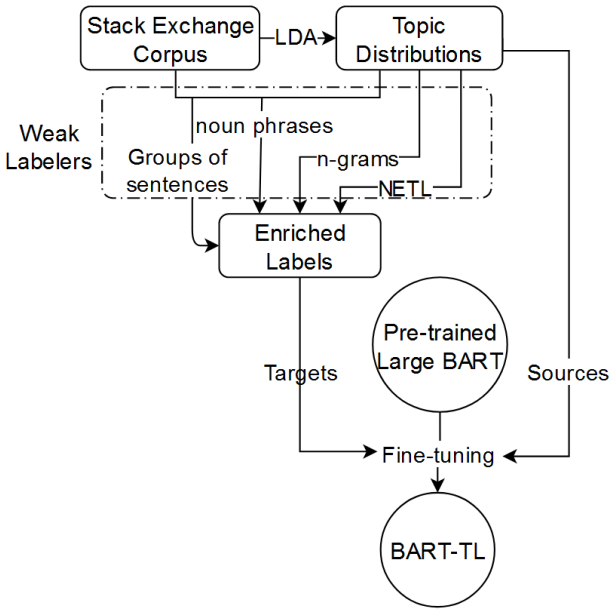

Figure 1: End-to-end training of *BART-TL* for topic labeling using weak supervision.

use an encoder-decoder architecture. A multitude of transformer-based models (Devlin et al., 2018; Radford et al., 2019; Liu et al., 2019; Lewis et al., 2019) emerged that achieved state of the art performance on a large number of NLP tasks through transfer learning. These models are pre-trained on large amounts of data in order to encompass general knowledge of the language to be later fine-tuned on downstream tasks. This allows for better results on small datasets, where deep learning was not a viable option beforehand.

However, research on using deep learning methods for topic labeling is scarce. A very recent study proposes an RNN-based encoder-decoder architecture (Alokaili et al., 2020) trained with distant supervision using Wikipedia page titles and employing BERTScore (Zhang et al., 2019) for evaluation.

3 Method

Our method utilizes a pre-trained BART (Lewis et al., 2019) transformer model, with a denoising autoencoder architecture, as we adopt a sequence-to-sequence approach for the task of topic labeling.

3.1 Building a Weakly Supervised Dataset

Topic labeling is generally performed in two steps: establishing a pool of candidate labels and then ranking them appropriately. This workflow is also adopted by a state of the art labeler that we will refer to as *NETL* (Bhatia et al., 2016).[1] This method uses names of Wikipedia articles as candidate labels and trains word2vec and doc2vec models on Wikipedia dumps. Preliminary filtering is done by selecting the labels with the highest embedding similarity scores to the topic terms, while the remaining labels are ranked in an unsupervised manner using letter trigrams. The authors also explore training a supervised ranker after obtaining feedback from annotators, incorporating PageRank (Page et al., 1999) and lexical features.

We build a dataset for fine-tuning BART starting from the **NETL labeler**. We extract the initial candidate labels for each topic after the embeddings similarity filtering but modify this process by assigning a greater weight in the scoring based on the importance of the word in the topic distribution. To avoid overfitting the most important word, we equalize the weights of the top-5 terms. The labels that consist only of stopwords are removed. We make these changes to be able to use a larger number of highest-rated topic terms in extracting labels than the standard 10 employed by NETL, expecting a better performance given a more ample context. Finally, we construct a one-to-many sequence mapping from topics, represented as a concatenation of the top-20 terms separated by spaces, to the corresponding labels. This represents the *baseline dataset*.

We also propose adding several enrichment approaches for this dataset, using other weak labelers as follows. The first additions are entries consisting of space-separated **n-grams** sampled from the most important words in the topic. The sampling is weighted by the underlying probability distribution and these do not have to be consecutive. Inspired by the work of Gourru et al. (2018), **groups of sentences** are added as targets using a variant of the COS10 technique for sentence extraction. The best sentences are joined one-by-one into a short paragraph until a minimum character threshold is met. One last idea for improving the baseline dataset is including popular **noun phrases** from the corpus. They are ranked based on the relevance to the topic and must appear at least a certain number of times in the corpus.

[1]The code is open-source: https://github.com/sb1992/NETL-Automatic-Topic-Labelling-.

3.2 Fine-tuning BART-TL

Pre-trained BART models are fine-tuned on the resulting datasets. The final *BART-TL* models are able to make predictions on sequences of topic terms. Output labels are generated as sequences and beam search is used to extract multiple ranked labels for a single topic. This strategy joins the extensive knowledge about language encompassed in the original transformer layers with traditional topic labeling techniques. The final models are fine-tuned based on unsupervised labelers and are, thus, weakly-supervised. A detailed representation of the end-to-end process can be seen in Figure 1.

4 Experiments

4.1 Baseline Dataset

We conduct experiments on corpora crawled from Stack Exchange[2] on 5 different subjects: English, Biology, Economics, Law, and Photography. These are preprocessed by removing XML artifacts, stopwords, and individual numbers. Documents with fewer than 20 words are removed from the corpus, along with words that occur less than 10 and more than 50,000 times. A total of 419,189 documents remain in the corpus. We apply LDA (Blei et al., 2003) on each corpus and obtain 100 topics for each subject. This choice for the number of topics is based on the prior work of Bhatia et al. (2016) where the authors generate 100 topics for each domain. These are filtered based on coherence (Röder et al., 2015), removing topics with a C_V score under 0.30, leaving a total of 303 topics. With the probability distributions of topics over the top-100 words, we generate 100 candidate labels for each topic using the NETL approach described in Section 3.

For the weak labelers, we choose to extract 5 n-grams with a n varying between 2 and 4, 5 groups of sentences with a character threshold of 120 and 10 noun phrases with a length of 2 to 4 words that have at least 25 occurrences. We experimented with each strategy individually but provided results for a model employing only the n-grams enrichment, *BART-TL-ng*, and one using all of them, *BART-TL-all*.

[2]The corpus can be found at https://archive.org/download/stackexchange.

4.2 Fine-tuning Details

We fine-tune the large BART model[3] for 2 epochs using an Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, 0.1 weight decay, 0.1 dropout, 0.1 attention dropout, 0.1 label smoothing, 6% warmup steps and a learning rate of 3e-5. The final labels are generated using beam search with a beam size of 25. These values follow the fine-tuning approach suggested for RoBERTa (Liu et al., 2019) and, by extension, BERT (Devlin et al., 2018), since the BART fine-tuning experiments do not explicitly specify different values for the hyper-parameters.

5 Results

We gather annotations in the form of surveys with 7 questions, one per topic, on the quality of topic labels on a scale from 0 to 3. The annotators have varying backgrounds, including computer science, medicine, law, and economics. For each of the 5 subjects in the corpus, we select 6 coherent topics for evaluation. The labels are taken from the unsupervised and supervised versions of the original NETL method, along with *BART-TL-ngram*, and *BART-TL-all*. For each method, only top-10 labels are considered for evaluation. An extra stopword label is introduced as a distractor, removing answers from annotators with over 25% of these scores ≥ 1. A topic is presented using its top-10 terms, along with 2 relevant short paragraphs, to offer additional context when the topic is unclear. Each survey has balanced topics based on the 5 subjects and each question contains 9 balanced labels based on the models. We gathered a total of 35 survey responses and filtered out the labels that had only a single annotation. This annotation was performed pro-bono and we estimate that the average time per annotated survey was 10 minutes. There is no bias in the annotations for certain models, as the average standard deviation for rating of individual labels is between 0.42 and 0.44 for all of them.

The results of this study are presented in Table 1. We focus on both the overall quality of the labels through top-k average rating, as well as how well the labels are ordered through normalized discounted cumulative gain (Järvelin and Kekäläinen, 2002). The two *BART-TL* models additionally feature statistics of the same labels reordered by the supervised and unsupervised ranking methods of

[3]https://github.com/pytorch/fairseq/tree/master/examples/bart.

Table 1: Qualitative comparison of labels between *NETL* and *BART-TL* models.

Models	All						English					
	Top-k Avg.			nDCG-k			Top-k Avg.			nDCG-k		
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5
NETL (U)	2.66	**2.59**	**2.50**	0.83	**0.85**	0.87	2.19	2.46	**2.38**	0.57	0.78	0.84
NETL (S)	**2.74**	2.57	2.49	**0.88**	**0.85**	**0.88**	2.63	2.47	2.28	0.84	0.86	0.86
BART-TL-all (U)	2.64	2.52	2.43	0.83	0.84	0.87	2.58	2.33	2.20	0.81	0.83	0.89
BART-TL-all (S)	2.64	2.55	2.42	0.81	0.84	0.87	2.58	2.36	2.15	0.81	0.86	0.89
BART-TL-ng (U)	2.62	2.50	2.33	0.82	0.84	0.85	2.58	**2.49**	2.26	0.81	**0.91**	**0.93**
BART-TL-ng (S)	2.73	2.46	2.25	0.87	0.83	0.83	**2.75**	2.40	2.21	**0.91**	0.88	0.91

Table 2: Samples of good and bad quality *new* labels generated by *BART-TL* models.

Top-10 topic terms	Good new labels	Bad new labels
crime center institution chain prison facility prisoner transformation jail custody	criminal justice system administrative court	guarantee principle
plate vehicle state license motor shall registration law apostille issued	driver's license license plate law	no matter what vehiclelicense *(no space)*
rate interest price inflation bond increase real money supply nominal	investment rate discount rate	rate interest rate principle

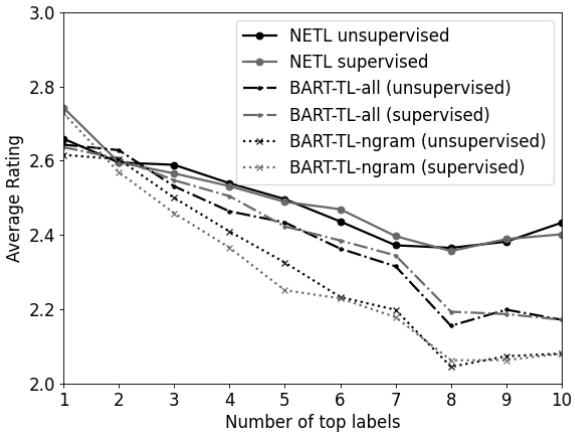

Figure 2: Evolution of average rating considering top-k labels.

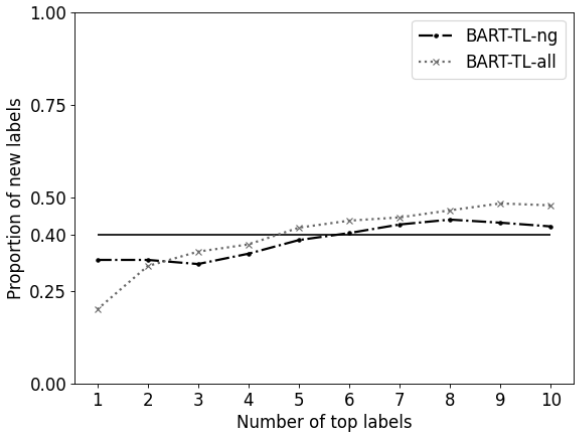

Figure 3: Average proportion of new labels in top-k.

NETL, as these usually perform better than the raw beam search results. The supervised variant of NETL uses the pre-trained ranker from the original paper. An extended version of this table is available in Appendix A.

To further investigate the results, we plot the evolution of the average rating in relation to the number of top labels considered. This can be seen in Figure 2. We study the capacity for novelty of the models in Figure 3, which outlines the proportion of new labels never encountered in the fine-tuning dataset or NETL top-10 predicted labels, as well as Figure 4, which illustrates the average rating of these labels. We observe a significant loss of up to 0.20 in rating, but even larger variations in rating are frequent in Table 1. That said, the novel labels would still be considered relevant with a

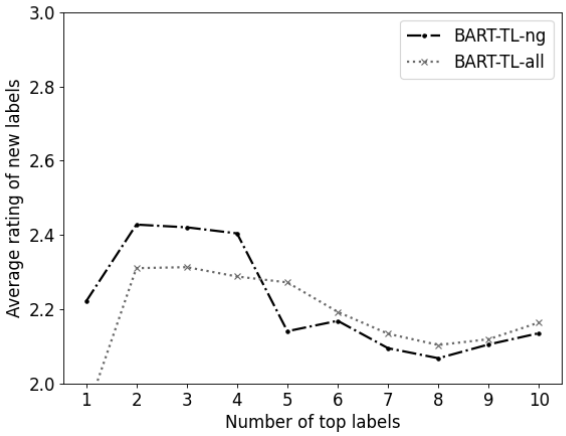

Figure 4: Average rating of new labels in top-k.

rating between 2.0 and 2.5. Table 2 showcases a few samples of original labels.

The results highlight that generative *BART-TL* models produce similar quality labels as the NETL methods when considering the top 1-2 labels. However, the quality of the generated labels degrades as their number increases. There is also no clear winner between the supervised and unsupervised versions of the proposed models, as they have similar trends. At the same time, the novelty tends to improve slightly with the number of considered labels. On average, 40% of the labels were never provided when fine-tuning the models. While novelty is an important feature for *BART-TL*, it can further be conditioned to generate labels with specific characteristics (Keskar et al., 2019).

The *BART-TL* models outperform the NETL methods on the English corpus, the largest of the five. At the same time, they achieve similar results on the Law and Biology corpora, that have the least amount of topics and are outperformed on the rest. Therefore, there was no correlation found between corpus size and the quality of the generated labels.

6 Conclusion

We introduced the *BART-TL* model that builds upon previous topic labeling solutions by adopting a generative deep learning strategy. Large pre-trained transformer models are fine-tuned in a weakly-supervised manner using unsupervised labelers to obtain meaningful labels. While current results have varying quality compared NETL, *BART-TL* is able to generate novel labels of similar quality. Although BART-TL experiments have been carried out for English, our generative methodology can be applied to any language if a pre-trained BART model is available.

Acknowledgment

This research has been partially supported by EEA Grants 2014-2021 and UEFISCDI, under project contract EEA-RO-NO-2018-0496.

References

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2020. Automatic generation of topic labels. *arXiv preprint arXiv:2006.00127*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. "Automatic Labelling of Topics with Neural Embeddings". In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963, Osaka, Japan. The COLING 2016 Organizing Committee.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Antoine Gourru, Julien Velcin, Mathieu Roche, Christophe Gravier, and Pascal Poncelet. 2018. United we stand: Using multiple strategies for topic labeling. In *International Conference on Applications of Natural Language to Information Systems*, pages 352–363. Springer.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online Learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. "Automatic Labelling of Topic Models". In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic Labeling of Multinomial Topic Models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 490–499, New York, NY, USA. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Ionut Sorodoc, Jey Han Lau, Nikolaos Aletras, and Timothy Baldwin. 2017. Multimodal topic labelling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 701–706.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chong Wang, John Paisley, and David Blei. 2011. Online Variational Inference for the Hierarchical Dirichlet Process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 752–760, Fort Lauderdale, FL, USA. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix A. Extended Results

We present an extended version of the results showcased previously in Table 1. While the target metrics remain the same, the additions are the raw BART models, with the labels retaining the order that they were generated in using beam search, as well as all of the 5 different subjects. These were added in Table 3.

Table 3: Extended quality comparison of labels between *NETL* and *BART-TL* models.

Models	**All**						**English**					
	Top-k Avg.			**nDCG-k**			**Top-k Avg.**			**nDCG-k**		
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5
NETL (U)	2.66	**2.59**	**2.50**	0.83	**0.85**	0.87	2.19	2.46	**2.38**	0.57	0.78	0.84
NETL (S)	**2.74**	2.57	2.49	**0.88**	**0.85**	**0.88**	2.63	2.47	2.28	0.84	0.86	0.86
BART-TL-all	2.41	2.38	2.28	0.73	0.75	0.79	1.89	1.94	2.00	0.52	0.60	0.74
BART-TL-all (U)	2.64	2.52	2.43	0.83	0.84	0.87	2.58	2.33	2.20	0.81	0.83	0.89
BART-TL-all (S)	2.64	2.55	2.42	0.81	0.84	0.87	2.58	2.36	2.15	0.81	0.86	0.89
BART-TL-ng	2.31	2.28	2.16	0.67	0.72	0.75	1.71	2.05	1.98	0.39	0.60	0.68
BART-TL-ng (U)	2.62	2.50	2.33	0.82	0.84	0.85	2.58	**2.49**	2.26	0.81	**0.91**	**0.93**
BART-TL-ng (S)	2.73	2.46	2.25	0.87	0.83	0.83	**2.75**	2.40	2.21	**0.91**	0.88	0.91
	Biology						**Economics**					
	Top-k Avg.			**nDCG-k**			**Top-k Avg.**			**nDCG-k**		
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5
NETL (U)	2.57	2.26	2.18	0.86	0.80	0.83	2.83	**2.73**	2.71	0.88	**0.85**	**0.89**
NETL (S)	2.57	2.27	2.16	**0.87**	0.77	0.82	**2.89**	2.68	**2.72**	**0.92**	0.83	**0.89**
BART-TL-all	2.63	**2.51**	2.23	**0.87**	**0.89**	0.84	2.59	2.52	2.41	0.75	0.76	0.76
BART-TL-all (U)	2.42	2.43	2.37	0.75	0.82	**0.85**	2.66	2.62	2.55	0.83	0.82	0.83
BART-TL-all (S)	2.38	2.34	**2.41**	0.73	0.77	**0.85**	2.74	2.60	2.56	0.85	0.81	0.82
BART-TL-ng	**2.66**	2.42	2.09	**0.87**	0.84	0.80	2.53	2.64	2.62	0.72	0.78	0.82
BART-TL-ng (U)	2.42	2.47	2.27	0.72	0.81	0.83	2.66	2.63	2.65	0.83	0.83	0.86
BART-TL-ng (S)	2.48	2.18	2.14	0.76	0.68	0.74	2.77	2.68	2.57	0.87	**0.85**	0.83
	Law						**Photography**					
	Top-k Avg.			**nDCG-k**			**Top-k Avg.**			**nDCG-k**		
	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5	k=1	k=3	k=5
NETL (U)	2.81	**2.78**	2.52	0.88	**0.91**	0.87	**2.88**	2.61	**2.61**	**0.97**	0.89	0.92
NETL (S)	2.79	2.71	**2.59**	0.87	0.89	0.89	2.81	**2.66**	**2.61**	0.92	**0.91**	**0.93**
BART-TL-all	2.29	2.48	2.44	0.67	0.73	0.79	2.71	2.49	2.30	0.85	0.80	0.82
BART-TL-all (U)	2.86	2.67	2.57	0.91	0.86	0.88	2.67	2.59	2.46	0.82	0.85	0.88
BART-TL-all (S)	2.70	2.77	2.61	0.80	0.89	**0.91**	2.73	2.64	2.38	0.85	0.87	0.86
BART-TL-ng	2.17	2.16	1.98	0.60	0.65	0.67	2.53	2.12	2.09	0.81	0.76	0.77
BART-TL-ng (U)	**2.97**	2.39	2.29	**0.98**	0.82	0.86	2.42	2.52	2.11	0.72	0.84	0.79
BART-TL-ng (S)	2.86	2.54	2.14	0.91	0.87	0.82	2.74	2.46	2.15	0.91	0.85	0.83