WHAT'S NEW? SUMMARIZING CONTRIBUTIONS IN SCIENTIFIC LITERATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

With thousands of academic articles shared on a daily basis, it has become increasingly difficult to keep up with the latest scientific findings. To overcome this problem we introduce a new task of *disentangled paper summarization*, which seeks to generate separate summaries for the paper contributions and the context of the work, making it easier to identify the key findings shared in articles. For this purpose, we extend the S2ORC corpus of academic articles, which spans a diverse set of domains ranging from economics to psychology, by adding disentangled "contribution" and "context" reference labels. Together with the dataset, we introduce and analyze three baseline approaches: 1) a unified model controlled by input code prefixes, 2) a model with separate generation heads specialized in generating the disentangled outputs, and 3) a training strategy that guides the model using additional supervision coming from inbound and outbound citations. We propose a novel automatic evaluation protocol which reports the *relevance*, *novelty*, and *disentanglement* of generated outputs. We demonstrate zero-shot generalization of our models on the CORD dataset of Covid-related scientific articles.

1 INTRODUCTION

Academic progress in recent era observes a stream of papers, making it harder for researchers to find the right papers to read, to read all the relevant papers of interest, or to explore the pre-requisite knowledge to understand papers. Human efforts on curating and summarizing the important takeaways from papers have been in high demand (Cohen & Lo, 2017) to organize such overwhelming amount of new papers. Since such high quality summaries are mainly produced by community experts voluntarily, however, they do not scale to large collection of papers.

One way to overcome this challenge is to apply state-of-the-art summarization models on those papers. Recent advances in abstractive summarization models have seen a great success using pre-trained language models (Lewis et al., 2020; Zhang et al., 2019) on summarizing papers into highly fluent and relevant abstracts, enabled by the development of large-scale datasets (Cohan et al., 2018). These models are fine-tuned to generate abstracts as the target text, which is a mixture of key information authors put together and relevant contexts for positioning the paper. Presenting these two aspects by controllably generating both the key contributions and the contexts from papers can further help us understand them quickly.

In this paper, we introduce the task of disentangled paper summarization, whose goal is to separately generate contribution and context summaries of scientific documents. To this end, we introduce a new dataset built upon the S2ORC corpus (Lo et al., 2020) which contains rich paper metadata and citation networks among them. Contributions and contexts are inherently relative: one paper's contribution is the future paper's context. We take this into account by leveraging the citation information from S2ORC corpus to allow models for comparing multiple papers to generate more accurate summaries. We introduce three models for this task, a unified model governed by prefixing control codes for switching generation mode inspired by (Fan et al., 2018; Keskar et al., 2019), a one-to-many sequence model with branching decoders for multi-mode generation, and a new objective function that guides the model based on citations to further improve the generation quality. Through the evaluation protocol which measures the degree of relevance, novelty, and disentanglement, we found that the unified model generally outperforms the other and the proposed objective function is effective on the one-to-many sequence model particularly on improving the disentangle-

ment. We analyze the model outputs and apply to a real-world application on COVID-19 papers and discuss its implications and usability.

2 Related Work

Automatic summarization on scientific documents has been investigated at a large scale (PubMed, ArXiv; Cohan et al. 2018), which have become one of the standard benchmark datasets to evaluate summarization systems. Full paper texts and the paper abstracts are usually adopted as input and output text data for the summarization task, respectively, but other output texts have been studied as well (Cachola et al., 2020). General challenge in this domain is the long input length; papers tend to contain several thousands of tokens. Efficient mechanisms to overcome the computational bottleneck of processing long inputs have been proposed (Liu et al., 2018; Zaheer et al., 2020).

One of the unique traits of scientific documents are the citations among documents, which have been provided at various scale within restricted domains (Radev et al., 2013; Yasunaga et al., 2019). Previous work has investigated the use of them as graph-based features (Yasunaga et al., 2019), a guide of focus for the extractive summarization systems (Cohan & Goharian, 2015), or as summaries themselves (Qazvinian & Radev, 2008). In this paper, our model leverage citation information as additional target sequence to guide the summarization systems. We also form a new dataset derived from S2ORC (Lo et al., 2020) at a similar scale to the existing large-scale scientific document summarization datasets like PubMed.

Automatic summarization using large-scale pre-trained (masked) language models have yielded state-of-the-art on many domains, including scientific document summarization tasks. Zhang et al. (2019) proposed PEGASUS which adopted a pre-training strategy of masking and reconstructing important *sentences* (as opposed to spans), achieving the state-of-the-art. BART (Lewis et al., 2020) pre-trains a sequence-to-sequence model with a denoising autoencoder objective. The model has shown effective when adapted to text generation tasks such as summarization.

3 Task

Given a scientific paper D, the task is to summarize the *contribution* y_{con} and *context* y_{ctx} of the paper both at the same time. Here, the contribution refers to the elements of work newly provided by the target paper D, and the context is otherwise. As the name suggests, this task inherently asks for relative comparisons of D against other papers to identify what is newly provided and what has been done. Therefore we also consider two sets of citations: inbound citations C_I and outbound citations C_O as the inputs to inform these comparisons more accurately. Inbound citations refer to the set of papers that cites D, *i.e.* relevant future papers, while outbound citations are the set of papers that D cites, *i.e.* papers in the references section of D.

This task poses unique challenges to overcome: successful model needs to 1) identify salient content of D and related papers from C_I and C_O, 2) compare the content between D and each document from the citations, 3) summarize along the two axes: contributions and context. In this work, to focus on the content composition and generation, we compose a dataset with already-filtered sets of inbound and outbound citations.

3.1 Dataset

Key datasets used as the benchmark in this domain either lacks in variety of domains or the essential citation information we require to train the models. Thus we construct our dataset from S2ORC (Lo et al., 2020), a dataset with a large collection of papers from multiple domains along with rich citation-related metadata, such as citation links between papers and annotated citation spans.

Identifying Related Papers. Not all the papers in S2ORC contain complete information we need for our summarization task: paper text, references, citation information, and the abstract. From S2ORC's first release[1], we filter out incomplete entries and construct a paper summarization dataset

[1]Release ID: 20190928.

Split	#Examples	Avg. #Tokens				
		Paper D	Inbound C_I	Outbound C_O	Contribution y_C	Background y_B
Train	805152	6351	925	877	136	236
Valid	36129	6374	922	875	135	236
Test	54242	6350	927	892	136	237

Table 1: Various statistics of our dataset.

that a) has an abstract and body text, and b) has at least 5 or more linked inbound and outbound citations respectively. Here, inbound and outbound refer to the direction of citation from the perspective of the target paper, and are equivalent to future papers and previous work, respectively. When a paper has more than 20 references, we sort them by the number of times the paper cites in descending order and keep the top 20 most relevant items.

Extracting Citation Features. For each citation, different features can be considered such as abstracts or other metadata. In this study, we follow (Qazvinian & Radev, 2008; Cohan & Goharian, 2015) and harvest snippets where a paper is cited (by different papers), which gives the reader's perspectives of which the paper to obtain features.[2] Obtained snippets directly refer to citations and highlight the content in relation to the target paper. Specifically for inbound citations, we take the mentions about the target paper made by future work at sentence-level. For outbound citations, they are the mentions about previous work made in the *target* paper.

Extracting Contributions Our task requires the gold standard contribution sentences and context sentences for each paper, which are not explicitly available or identifiable due to a wide variety of domains and their writing styles in the dataset. We therefore approximate the full contribution statements with relevant sentences in the paper's abstract. Specifically, we manually annotate 400 abstracts at sentence-level from the training set with binary labels indicating *contribution* and *context*[3]. From the abstracts, we obtain 3341 sentences with the these binary labels[4]. We fine-tune SciBERT (Beltagy et al., 2019) on this task and obtained 86.3 % accuracy in classifying *contribution* and *context* on the held-out test set. Finally, the fine-tuned model was used to populate labels in all the papers.

The statistics on the resulting dataset is shown in Table 1. We also show the domain statistics on the test set to illustrate the diverse collection of domains in Figure 1a.

3.2 Where in paper are contributions mentioned?

To understand and motivate the need for contribution summarization, we use the fine-tuned contribution classifier to classify the papers themselves. While not perfect, analyzing where the contributions locate in papers can illustrate the importance of summarizing contributions. We applied the classifier on the full training set and obtain the frequency of contribution sentences in various parts of paper, binned evenly to 10 parts. We show the results in Figure 1b.

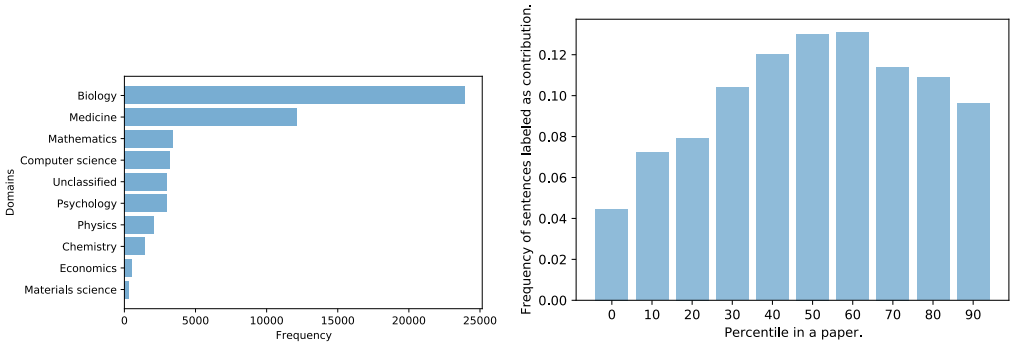
4 Models

Our aim is to provide a summarization model with the ability to selectively generate contributions or background information. We propose two methods based on pre-trained language models to achieve this controlability. Specifically, BART (Lewis et al., 2020) is adopted in two approached described below.

[2]Our initial experimetns sought for different features directly drived from the papers (*e.g.* abstracts), but showed worse performance.

[3]This implies that every statement that is not about contributions is context information. We leave modeling with finer-grained labels for future work

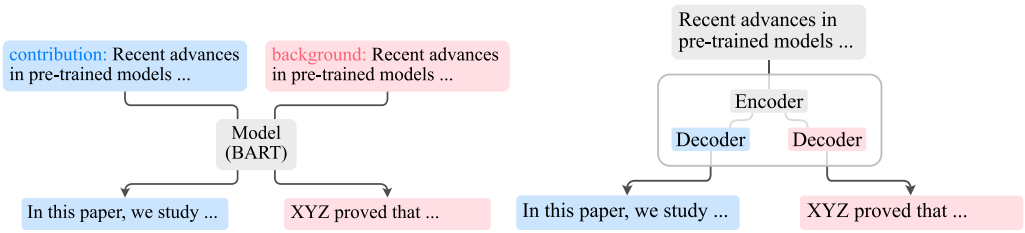
[4]The ratio of the number of examples with contribution to those with context is 1827 to 1514.



(a) Top-10 frequent domains in our dataset.

(b) Frequency of contribution mentions in different parts of papers.

Figure 1: Data statistics. Left: Top-10 domain frequency on the test set. Right: Frequency of contributions appearing in various parts of a paper in percentile.



(a) CONTROLCODE model.

(b) MULTIHEAD model.

Figure 2: Two approaches for controllable generation.

CONTROLCODE (CC) One of the popular approaches two control the outputs is to condition on the controlling indicator when generating the outputs. With enough data, previous work on controllable generation and summarization (Fan et al., 2018; Keskar et al., 2019) showed that it is effective to prepend an indicator code to the input and train the model normally. At test time, desired output mode is fed as the control code before the input which allows the model to condition on it throughout decoding. Following this, we modify the training instances by prepending control codes to the input documents and fine-tune BART. The model is expected to associate the provided control code to the desired output mode and learn to generate such outputs within the same parameters.

MULTIHEAD (MH) We also propose a model that learns output distribution more explicitly by allocating mode-specific parameters within the model. In this model, the decoder of the model shares all the paramters but the last layer at which point the decoder branches into two final layers. Each branch is dedicated to each output mode and thus only get updated with gradients calculated from the losses in the same mode. We keep the following linear layer for the token prediction shared because of the assumption that token-level vocabulary distributions are similar in two modes due to being in the same topic.

4.1 INFORMATION THEORETIC GUIDANCE ON INFORMATIVENESS

Peyrard (2019) proposed a information theoretic interpretation on summarization which decomposes the criteria of a good summay into redundancy, relevance, and informativenes. Among these criteria, *informativeness* measures the degree of surprisal by the readers when reading a summary given their background knowledge. Formally, given a primitive semantic unit ω_i, the probability over the semantic units under reader's knowledge P_K, and the probability over them with respect to the paper content P_D, the following quantity defines the informativeness:

$$Inf(D, K) = -\sum_i P_D(\omega_i) \log P_K(\omega_i). \quad (1)$$

Informativeness offers a direct correspondence to contribution summarization because paper contributions are novel contents introduced to the community, which causes surprisal given the researchers' knowledge about the state of research. Therefore in this work, we explore the possiblity of utilizing this quantity in the form of an auxiliary objective to maximize. We assume that the semantic unit ω_i is the summary itself[5], which enables simpler interpretation on each of the probabilities. We regard P_D as the likelihood of the summary given the paper: $P_D(\omega_i) = p(y \mid D)$. Since each paper has its own context and required background knowledge, we treat the previous knowledge for the paper as the relevant papers published before the target paper, *i.e.* outbound citations C_O. Therefore, P_K is interpreted as the likelihood of the summary given the description about the previous work: $P_K(\omega_i) = p(y \mid C_O)$. By collapsing the probabilities for all possible semantic units to the target summary to optimize for, we obtain the following objective function for contribution summary generation:

$$L_{Inf} = -\log p(y_{con} \mid D) - Inf(D, K) \tag{2}$$

$$= -\log p(y_{con} \mid D) + p(y_{con} \mid D) \log p(y_{con} \mid C_O). \tag{3}$$

We apply the same argument in a reverse direction for contexte summary generation and obtain the following final loss:

$$L_{Inf} = -\log p(y_{ctx} \mid D) + p(y_{ctx} \mid D) \log p(y_{ctx} \mid C_I), \tag{4}$$

where the second term maximizes for the surprisal by reading background information given the target paper.

5 Experiments and Results

In this section, we report experimental settings and the automatic evaluation results. We consider four model variants:

- **CC**, **CC+INF**: ControlCode model and the same model with informativeness objective from Section 4.1
- **MH**, **MH+INF**: MultiHead model and the same model with informativeness objective from Section 4.1

5.1 Evaluation

We first perform automatic evaluation using the system outputs (s_{con}, s_{ctx}) and the gold standard references (y_{con}, y_{ctx}) as shown in Figure 3. Given four texts, we consider three axes to evaluate as listed below.

Relevance. Most importantly, generated summaries should be as close as possible to corresponding references. Using ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), we measure the similiarity betweent (s_{con}, y_{con}) and (s_{ctx}, y_{ctx}), respectively.

Novelty. Another aspect to consider is we measure the linear combination of ROUGE scores against multiple references. Typically, update summarization has one reference y_{con} to target and another reference y_{ctx} to avoid. Suppose ROUGE-i score is represnted with $R^i(s, y)$, Nouveau-ROUGE is defined as $N_i(s_{con}, y_{con}, y_{ctx}) = \alpha_0^i + \alpha_1^i R^i(s_{con}, y_{con}) + \alpha_2^i R^i(s_{con}, y_{ctx}), (i = 1, 2)$, where the coefficients are determined such that the score maximizes the correlation against human annotation. We adopt corresponding coefficients α^i for N_i from (Conroy et al., 2011).

Disentanglement. To measure the degree of disentanglement, we also measure textual similarities between the two generated summaries (s_{con}, s_{ctx}). For all the metrics M used in Relevance, we define the disentanglement metrics as $100 - M$: D-i $= 100 -$ R-i, $DisBertScore = 100 - BERTScore$.

[5]For simplicity in modeling, we chose the entire summary. However, this goes against the requirement set by Peyrard (2019) that ω_i is a *primitive* semantic unit, because a paragraph's meaning can be decomposed into higher granular units. We will investigate on different choices of the units in the future work.

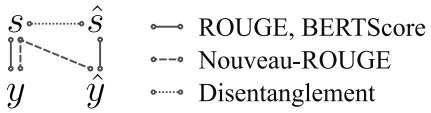


Figure 3: Evaluation scheme. We compare decoded summaries and references in three ways.

5.2 Implementation Details

We build upon BART (Lewis et al., 2020), a pre-trained sequence-to-sequence model suitable for text generation. Specifically, we fine-tune DistilBART on our dataset, a down-sized BART where the model weights were intialized from BART-large fine-tuned on CNN/DM summarization task[6]. For MultiHead model, the final layer in the BART decoder was duplicated and initialized with identical weights. We fine-tuned on the training set for 80,000 gradient steps with a fixed learning rate of 3.0 $\times 10^{-5}$. Evaluation was performed using SummEval toolkit Fabbri et al. (2020).

5.3 Results

We report results for the respective evaluation criterion in Table 2.

Model		Relevance				Novelty		Disentanglement			
		R-1	R-2	R-L	BS	N_1	N_2	D-1	D-2	D-L	DBS
CC	Con	39.16	15.96	24.65	63.22	2.49	3.44	52.95	72.18	69.12	33.62
	Ctx	41.84	17.24	24.55	63.78						
CC+INF	Con	38.92	15.95	24.65	62.94	2.47	3.44	53.68	71.97	68.46	34.09
	Ctx	41.49	17.03	24.50	63.40						
MH	Con	39.20	15.98	24.72	63.04	2.46	3.42	50.89	69.51	65.97	32.51
	Ctx	41.67	17.23	24.65	63.77						
MH+INF	Con	38.74	15.90	24.59	62.70	2.43	3.42	53.35	71.47	67.20	33.86
	Ctx	40.39	16.31	23.83	62.85						

Table 2: Automatic evaluation results on the test set. For all metrics, higher values indicate better results.

Relevance. A general trend of achieving higher textual similarities on summarizing context information is observed on all models, whereas the two models do not show significant difference when using the same output mode. As it is considered separately from the relevance criterion by Peyrard (2019), the informativeness objective does not explicitly ask the model to improve the *relevance*. Hence, we observe automatic evaluation scores on par with the base models when using informativeness objective (+INF), but not better than the base models.

Novelty. Despite sharing the entire model parameters, ControlCode model shows slightly better performance overall. While the informativeness objective was expected to improve novelty (or surprisal), the results showed otherwise. We suspect that this is due to the context summaries y_{ctx} not being contexts discussing the state of the research *before* the target paper, but rather discussing the paper content irrelevant to the contribution. We look into this more in the analysis.

Disentanglement. ControlCode model performs generally better than MultiHead model in terms of generating dissimilar summaries, despite completely sharing the parameters between the two generation modes. This is probably due to the two modes being very relevant everything else except for the "theme", making it easier for the model to share paremeters and learn to partition the probability space than leaarning separate layers which is more costly in terms of number of parameters.

[6]We did not observe a performance drop when using DistilBART compared to BART in our preliminary experiments. Initial model weights are available at https://huggingface.co/sshleifer/student_cnn_6_6.

Abstract: Energy optimization in buildings by controlling the Heating Ventilation and Air Conditioning (HVAC) system is being researched extensively. In this paper, a model-free actor-critic Reinforcement Learning (RL) controller is designed using a variant of artificial recurrent neural networks called Long-Short-Term Memory (LSTM) networks. Optimization of thermal comfort alongside energy consumption is the goal in tuning this RL controller. The test platform, our office space, is designed using SketchUp. Using OpenStudio, the HVAC system is installed in the office. . . .

Contribution: In this work, we introduce novel control algorithms from a branch of machine learning called reinforcement learning. In our current approach, the impetus is thermostat control. Instead of traditional on/off heating and cooling control, reinforcement learning is utilized to set this schedule to obtain improved Predicted Mean Vote (PMV)-based thermal comfort at an optimal energy expenditure. Hence, a thermostats schedule is computed using an RL controller. The results show that the Q-learning algorithm can learn to adapt to time-varying and nonlinear system dynamics without explicit identification of the plant model in both systems and controls.
Context: The Heating, Ventilation and Air Conditioning (HVAC) systems can account for up to 50% of total building energy demand. In the hopes of moving toward a greener, more energy-efficient future, a significant improvement in energy efficiency is needed to achieve this goal. Despite the advances in research on HVAC control algorithms, most field equipment is controlled using classical methods that include hysteresis/on/off and Proportional Integral and Derivative controllers. . . .

Table 3: Generated Samples.

6 Analysis

6.1 Qualitative Analysis

We also perform qualitative analyses on generated summaries. Table 6 show generated summaries and the original abstracts of the papers (truncated for space-saving purpose). More summaries without truncation are included in the appendices. In the first example, our model is successfully focusing on both of the target modes respectively, making them easy to follow. The original abstract does not mention the context information except for the first sentence. However, the generated context summary from the trained model includes more in-depth contexet information, successfully balancing the insufficient description by the abstract.

6.2 Per-domain Performance

Next, we take the advantage of rich metadata information from S2ORC and analyze the model performance on each domain. We show the results of contribution summarization using CONTROLCODE model on Top-10 frequent domains in Table 4[7]. Summarization on Medicine domain seems to be particularly difficult, achieving about 7 points lower than others on ROUGE-1. Through manual inspection on instances with low ROUGE-1 scores (R-1 < 20), we found that the summaries in Medicine domain are highly quantiative and descriptive of the results (*e.g.* "Among these treated . . . retinopathy was noted in X%"). Since the findings to report on summaries are unique to each paper, models have hard times learning to generate exact results for each paper differently. While other domains such as Biology also suffer from the same phenomenon, low ROUGE summaries of this type (*i.e.* results are discussed quantitatively) in Medicine was 1.9 times as frequent as Biology.

Metric	R-1	R-2	R-L	BS
Biology	40.63	17.01	25.59	64.23
Medicine	33.97	13.08	21.73	61.75
Mathematics	40.13	15.56	24.42	61.58
Computer science	43.54	16.41	25.86	63.43
None	40.31	18.14	26.68	64.00
Psychology	39.51	15.56	24.34	62.95
Physics	40.09	15.85	24.89	62.10
Chemistry	40.44	17.77	26.14	63.93
Economics	39.56	14.25	23.41	60.91
Materials science	42.52	18.96	27.57	65.25

Table 4: Automatic relevance evaluation on contribution summarization for each of Top-10 domains using CONTROLCODE model.

[7] MULTIHEAD model and the the introduction of informativeness objective exhibited the same pattern.

6.3 EVALUATION AGAINST GOLD ANNOTATIONS

As discussed in Section 3.1, labels for contribution or context are populated automatically using a classifier, which is expected to contain mistakes. Therefore, we created gold standard evaluation set by manually annotating 100 samples in the test set and report the evaluation results in Table 5. A sharp drop in ROUGE scores for the context summaries are due to some examples receiving zero scores for generating context summaries when the manual annotation judged that there are not existent. The overall trend of CONTROLCODE model outpeforming MULTIHEAD model is still observed in the evaluation. More noticeably, we observe a reverse tendency when the two models are applied with the informativeness objective. MULTIHEAD model specifically enjoyed significant improvement in terms of novelty and disentanglement.

Model		Relevance				Novelty		Disentanglement			
		R-1	R-2	R-L	BS	N_1	N_2	D-1	D-2	D-L	DBS
CC	Con	39.37	15.86	24.73	63.28	3.23	3.84	52.81	71.52	68.36	33.05
	Ctx	30.59	11.22	19.08	55.76						
CC+INF	Con	38.38	15.21	23.47	62.59	3.12	3.69	52.49	69.64	66.60	32.76
	Ctx	30.14	11.10	19.00	55.55						
MH	Con	38.63	15.53	24.68	62.84	3.13	3.78	49.62	67.45	64.43	31.39
	Ctx	29.82	10.61	18.51	55.24						
MH+INF	Con	39.43	15.75	24.77	63.11	3.29	3.84	51.56	68.57	64.97	32.35
	Ctx	29.14	10.25	18.48	54.92						

Table 5: Automatic evaluation results on 100 samples from the test set with manual contribution annotations. For all metrics, higher values indicate better results.

6.4 HUMAN EVALUATION ON DISENTANGLEMENT

In addition to various automatic evaluation, we perform human evaluation on disentanglement to understand which models human annotators prefer. We perform best-worst scaling evaluation on the 50 random samples from the test set and asked 4 annotators the best and the worst contribution and context summary pairs in terms of disentanglement.

6.5 USEFULNESS OF SEPARATED SUMMARIES

6.6 APPLICATION: GENERATING CONTRIBUTIONS OF RESEARCH PAPERS ON COVID-19

We demonstrate the use of our model on a new domain, particularly focused on the research of COVID-19. To do this, we took 100 samples from a subset of CORD dataset (Wang et al., 2020) by filtering papers published after the date of release of the origina S2ORC dataset. Models were directly applied without fine-tuning due to the two dataset being similar in domains (*i.e.*, CORD datasets consist of Medicine domain, and our data's second most frequent domain is Medicine).

7 CONCLUSIONS

In this paper, we propose a new task of *disentangled paper summarization*, a task of summarizing both contributions and contexts separately from a paper. We introduce two baseline approaches for the controllable abstractive summarization for the task and a citation-driven objective function to guide the model towards moore disentanglement. Experimental results and human studies indicate the importance and real world applicability of models developed for the task.

REFERENCES

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. TLDR: Extreme summarization of scientific documents. *arXiv:2004.15011*, 2020.

Arman Cohan and Nazli Goharian. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 390–400. Association for Computational Linguistics, September 2015. doi: 10.18653/v1/D15-1045. URL https://www.aclweb.org/anthology/D15-1045.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2097. URL https://www.aclweb.org/anthology/N18-2097.

Joseph Paul Cohen and Henry Z Lo. Shortscience. org-reproducing intuition. *arXiv preprint arXiv:1707.06684*, 2017.

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. Nouveau-ROUGE: A novelty metric for update summarization. *Computational Linguistics*, 37(1):1–8, 2011. doi: 10.1162/coli_a_00033. URL https://www.aclweb.org/anthology/J11-1001.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*, 2020.

Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 45–54, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2706. URL https://www.aclweb.org/anthology/W18-2706.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*, 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Peter J. Liu, Mohammad Ahmad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. 2018. URL https://openreview.net/pdf?id=Hyg0vbWC-.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

Maxime Peyrard. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1059–1073, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1101. URL https://www.aclweb.org/anthology/P19-1101.

Vahed Qazvinian and Dragomir R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 689–696, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://www.aclweb.org/anthology/C08-1087.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944, 2013.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*, 2020.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of AAAI 2019*, 2019. URL https://www.aaai.org/ojs/index.php/AAAI/article/view/4727/4605.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*, 2020.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.