What's New? Summarizing Contributions in Scientific Literature

Hiroaki Hayashi[1] **Wojciech Kryściński**[1] **Bryan McCann**[2]
Nazneen Rajani[3] **Caiming Xiong**[1]
[1]Salesforce Research [2]You.com [3]Hugging Face
hiroakihayashi@salesforce.com

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 also propose a comprehensive automatic evaluation protocol which reports the *relevance*, *novelty*, and *disentanglement* of generated outputs. Through a human study involving expert annotators, we show that in 79%, of cases our new task is considered more helpful than traditional scientific paper summarization.

1 Introduction

With the growing popularity of open-access academic article repositories, such as arXiv or bioRxiv, disseminating new research findings has become nearly effortless. Through such services, tens of thousands of scientific papers are shared by the research community every month[1]. At the same time, the unreviewed nature of mentioned repositories and the sheer volume of new publications has made it nearly impossible to identify relevant work and keep up with the latest findings.

Scientific paper summarization, a subtask within automatic text summarization, aims to assist researchers in their work by automatically condensing articles into a short, human-readable form that contains only the most essential information. In recent years, abstractive summarization, an approach where models are trained to generate fluent summaries by paraphrasing the source article, has seen impressive progress. State-of-the-art methods leverage large, pre-trained models (Lewis et al., 2020; Liu et al., 2022), define task-specific pre-training strategies (Zhang et al., 2019), and scale to long input sequences (Zhao et al., 2020; Zaheer et al., 2020). Available large-scale benchmark datasets, such as arXiv and PubMed (Cohan et al., 2018), were automatically collected from online archives and repurpose paper abstracts as reference summaries. However, the current form of scientific paper summarization where models are trained to generate paper abstracts has two caveats: 1) often, abstracts contain information which is not of primary importance, 2) the vast majority of scientific articles come with human-written abstracts, making the generated summaries superfluous.

To address these shortcomings, we introduce the task of disentangled paper summarization. The new task's goal is to generate two summaries simultaneously, one strictly focused on the summarized article's novelties and contributions, the other introducing the context of the work and previous efforts. In this form, the generated summaries can target the needs of diverse audiences: senior researchers and field-experts who can benefit from reading the summarized contributions, and newcomers who can quickly get up to speed with the intricacies of the addressed problems by reading the context summary and get a perspective of the latest findings from the contribution summary.

For this task, we introduce a new large-scale dataset by extending the S2ORC (Lo et al., 2020) corpus of scientific papers, which spans multiple

[1]https://arxiv.org/stats/monthly_submissions

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1019–1031

scientific domains and offers rich citation-related metadata. We organize and process the data, and extend it with automatically generated contribution and context reference summaries, to enable supervised model training. We also introduce three abstractive baseline approaches: 1) a unified, controllable model manipulated with descriptive control codes (Fan et al., 2018; Keskar et al., 2019; Liu et al., 2021; He et al., 2022), 2) a one-to-many sequence model with a branched decoder for multi-head generation (Luong et al., 2016; Guo et al., 2018), and 3) an information-theoretic training strategy leveraging supervision coming from the citation metadata (Peyrard, 2019). To benchmark our models, we design a comprehensive automatic evaluation protocol that measures performance across three axes: relevance, novelty, and disentanglement. We thoroughly evaluate and analyze the baselines models and investigate the effects of the additional training objective on the model's behavior. To motivate the usefulness of the newly introduced task, we conducted a human study involving human annotators in a hypothetical paper-reviewing setting. The results find disentangled summaries more helpful in 79% of cases in comparison to abstract-oriented outputs. Code and data preparation scripts introduced in this work are available at `https://github.com/salesforce/disentangled-sum`.

2 Related Work

Recent trends in abstractive text summarization show a shift of focus from designing task-specific architectures trained from scratch (See et al., 2017; Paulus et al., 2018) to leveraging large-scale Transformer-based models pre-trained on vast amounts of data (Liu and Lapata, 2019; Lewis et al., 2020), often in multi-task settings (Raffel et al., 2019). A similar shift can be seen in scientific paper summarization, where state-of-the-art approaches utilize custom pre-training strategies (Zhang et al., 2019) and tackle problems of summarizing long documents (Zhao et al., 2020; Zaheer et al., 2020). Other methods, at a smaller scale, seek to utilize the rich metadata associated with scientific articles and combine them with graph-based methods (Yasunaga et al., 2019). In this work, we combine these two lines of work and propose models that benefit from pre-training procedures, but also take advantage of task-specific metadata.

Popular large-scale benchmark datasets in scientific paper summarization (Cohan et al., 2018; Cachola et al., 2020) were automatically collected from open-access paper repositories and considered article abstracts as the reference summaries. Other forms of supervision have also been investigated for the task, including author-written highlights (Collins et al., 2017) and annotations (Meng et al., 2021), citations (Yasunaga et al., 2019), and transcripts from conference presentations of the articles (Lev et al., 2019). In contrast, we introduce a large-scale automatically collected dataset with more fine-grained references than abstracts, which also offers rich citation-related metadata.

Update summarization (Dang and Owczarzak, 2008) defines a setting in a collection of documents with partially overlapping information is summarized, some of which are considered prior knowledge. The goal of the task is to focus the generated summaries on the novel information. Work in this line of research mostly focuses on novelty detection in news articles (Bysani, 2010; Delort and Alfonseca, 2012) and timeline summarization (Martschat and Markert, 2018; Chang et al., 2016) on news and social media domains. Here, we propose a novel task that is analogous to update summarization in that it also requires contrasting the source article with the content of other related articles which are considered pre-existing knowledge.

3 Task

Given a source article D, the goal of disentangled paper summarization is to simultaneously summarize the *contribution* y_{con} and *context* y_{ctx} of the source article. Here, contribution refers to the novelties introduced in the article D, such as new methods, theories, or resources, while context represents the background of the work D, such as a description of the problem or previous work on the topic. The task inherently requires a relative comparison of the article with other related papers to effectively disentangle its novelties from pre-existing knowledge. Therefore, we also consider two sets of citations: inbound citations C_I and outbound citations C_O as potential sources of useful information for contrasting the article D with its broader field. Inbound citations refer to the set of papers that cite D, *i.e.* relevant future papers, while outbound citations are the set of papers that D cites, *i.e.* relevant previous papers. With its unique set of goals, the task of disentangled paper summarization poses a novel set of challenges for automatic summarization systems to overcome: 1) identifying salient

Dataset	#Examples	Avg. #Tokens				
		Paper D	Inbound C_I	Outbound C_O	Contribution y_{con}	Context y_{ctx}
ArXiv (Train)	203037	4938	-	-	220 (Total summary)	
PubMed (Train)	119924	3016	-	-	203 (Total summary)	
Ours - Train	805152	6351	925	877	136	236
Valid	36129	6374	922	875	135	236
Test	54242	6350	927	892	136	237

Table 1: Token length statistics on the training split of our dataset compared to existing scientific paper summarization datasets. Contribution summaries tend to be shorter than context summaries.

content of D and related papers from C_I and C_O, 2) comparing the content of D with each document from the citations, and 3) summarizing the article along the two axes: contributions and context.

3.1 Dataset

Current benchmark datasets used for the task of scientific paper summarization, such as arXiv and PubMed (Cohan and Goharian, 2015), are limited in size, the number of domains, and lack of citation metadata. Thus, we construct a new dataset based on the S2ORC (Lo et al., 2020) corpus, which offers a large collection of scientific papers spanning multiple domains along with rich citation-related metadata, such as citation links between papers and annotated citation spans. Specifically, we curate the data available in the S2ORC corpus and extend it with new reference labels.

Data Curation Some papers in the S2ORC corpus[2] do not contain a complete set of information required by our summarization task: paper text, abstract, and citation metadata. We remove such instances and construct a paper summarization dataset in which each example a) has an abstract and body text, and b) has at least 5 or more inbound and outbound citations, C_I and C_O respectively. In cases where a paper has more than K incoming or outgoing citations, we first sample K citations for each of incoming and outgoing citations and sort them in descending order by the number of their respective citation from and to the target paper. K is a hyperparameter and we choose $K = 20$ in this work.[3]

Citation Span Extraction Each article in the set of inbound and outbound citations can be represented by its full text, abstract, or the span of text associated with the citation. In this study, we follow Qazvinian and Radev (2008) and Cohan and Goharian (2015) in representing citations with the sentences in which the citation occurs.[4] Thus, an outbound citation is represented by a sentence from the source paper. Usually, such sentences directly refer to the cited paper and place its content in relation to the source paper. Analogously, an inbound citation is represented by sentences from the citing paper and relates its content with the source paper.

Reference Generation Our approach relies on the availability of reference summaries for both contributions and contexts. However, such annotations are not provided or easily extractable from the S2ORC corpus, and collecting expert annotations is infeasible due to the associated costs. Therefore, we apply a data-driven approach to automatically extract contribution and context reference summaries from the available paper abstracts. First, we manually label 400 abstracts sampled from the training set. Annotations are done on a sentence-level with binary labels indicating *contribution*- and *context*-related sentences.[5] This procedure yields 3341 sentences with associated binary labels, which we refer to as golden standard references. Next, we fine-tune an automatic sentence classifier using the golden standard data. As our classifier we use SciBERT (Beltagy et al., 2019), which after fine-tuning achieves 86.3% accuracy and 0.932 for Area under ROC curve in classifying *contribution* and *context* sentences on a held-out test set. Finally, we apply the fine-tuned classifier to generate reference labels for all examples in our dataset, which we refer to as silver standard references. The statistics of the resulting dataset are shown in Table 1.

4 Models

Our goal is to build an abstractive summarization system which has the ability to generate contribu-

[2]Release ID: 20190928.

[3]Among other values, we experimented with 20, as the trade-off between computational cost and rich citation contexts.

[4]If a publication is cited multiple times within a source article we concatenate all relevant sentences.

[5]Sentences not labeled as contribution are considered context, we leave finer-grained labels for future work.

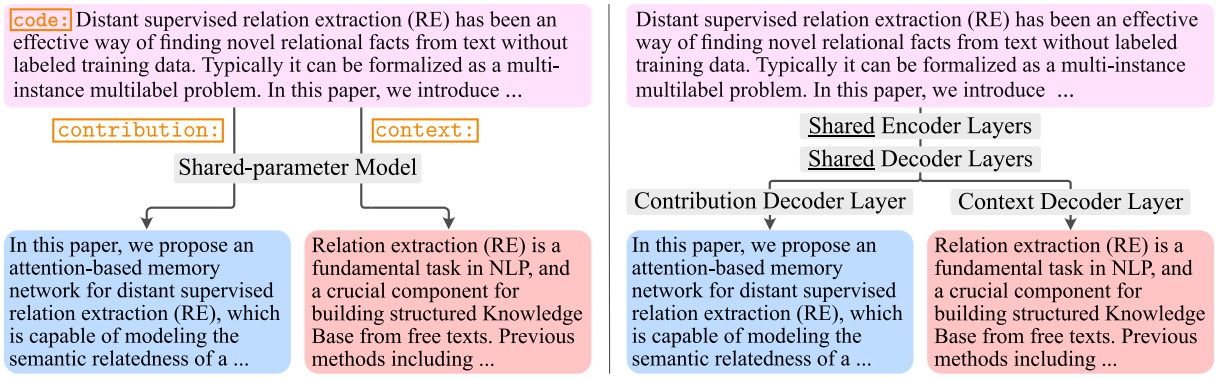

Figure 1: Model diagram. Left: CONTROLCODE model, in which inputs are prefixed with a prompt symbol and passed to a shared model to control the output mode. Right: MULTIHEAD model, which shares all of the model's parameters apart from the last decoder layer for different output modes, and chooses the final decoder layer accordingly to control the output mode.

tion and context summaries based on the source article. To achieve the necessary level of controllability, we propose two independent approaches building on encoder-decoder architectures:

CONTROLCODE (CC) A common approach to controlling model-generated text is by conditioning the generation procedure on a control code associated with the desired output. Previous work on controllable generation (Fan et al., 2018; Keskar et al., 2019; Chan et al., 2021; Luo et al., 2022) showed that prepending a special token or descriptive prompt (Liu et al., 2021) to the model's input during training and inference is sufficient to achieve fine-grained control over the generated content. Following this line of work, we modify our training instances by prepending textual control codes, `contribution:` or `context:`, to the summarized articles. During training, all model parameters are updated for each data instance and the model is expected to learn to associate the provided prompt with the correct output mode. The approach does not require changes in the architecture, making it straightforward to combine with existing large-scale, pre-trained models. The architecture is shown on the left of Figure 1.

MULTIHEAD (MH) An alternative way of controlling generation is by explicitly allocating layers within the model specifically for the desired control aspects. Prior work investigating multi-task models (Luong et al., 2016; Guo et al., 2018) showed the benefits of combining shared and task-specific layers within a single, multi-task architecture. Here, the encoder shares all parameters between the two generation modes, while the decoder shares all parameters, apart from the final layer, which splits into two generation branches. During training, each branch is individually updated with gradients from the associated mode. The model shares the softmax layer weights between the output branches under the assumption that token-level vocabulary distributions are similar in the two generation modes due to the common domain. This approach is presented on the right of Figure 1.

4.1 Informativeness-guided Training

In addition to supervising the models with gold-standard summaries, we consider guiding them to generate summaries with a focus on contributions using *informativeness* (Peyrard, 2019). Specifically, informativeness measures the user's degree of surprise after reading a summary given their background knowledge, and can be formally defined as:

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

where ω_i is a primitive semantic unit, P_K is the probability over the unit under the user's knowledge, P_D is the probability over the unit with respect to the source document, and i is an index over all semantic units within a summary.

Since paper contributions are novel contents introduced to the community, we argue that contributions cause surprisal given the general knowledge about the state of the field. Quantified such surprisal as informativeness, we explore utilizing this measure as an auxiliary objective that is optimized during training. We define the semantic unit ω_i as the summary itself,[6] which enables a simple interpretation of the corresponding probabil-

[6]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.

ities. We estimate P_D as the likelihood of the summary given the paper content, $P_D(\omega_i) = p(y \mid D)$. Since each paper is associated with a unique context and background knowledge, we treat the background knowledge as all relevant papers published before the source paper, *i.e.*, outbound citations C_O. Therefore, P_K is estimated as the likelihood of the summary given the previous work, $P_K(\omega_i) = p(y \mid C_O)$. We formulate the informativeness function as:

$$Inf(D, K) = \begin{cases} -p(y_{con} \mid D) \log p(y_{con} \mid C_O) \\ -p(y_{ctx} \mid D) \log p(y_{ctx} \mid C_I) \end{cases}, \quad (2)$$

where the conditioning depends on the generation mode of the model, and aim to maximize it during the training procedure. The estimation of each term is done by a forward pass on the model with corresponding input and output pairs: $p(y_{con} \mid C_O)$ is computed by estimating the probability of y_{con} when feeding C_O as the encoder input.

Combined with a cross entropy loss L_{CE}, we obtain the final objective which we aim to the minimize during training:

$$L = L_{CE} - \lambda \, Inf(D, K), \quad (3)$$

where λ is a scaling hyperparameter determined through cross-validation. Note that C_I, C_O are only used during training. Models trained with this objective is applicable to papers without citation information at inference time.

5 Experiments and Results

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

- **CC**, **CC+INF**: CONTROLCODE model without and with the informativeness objective,
- **MH**, **MH+INF**: MULTIHEAD model without and with the informativeness objective.

5.1 Evaluation

We perform automatic evaluation of the system outputs (s_{con}, s_{ctx}) against the silver standard references (y_{con}, y_{ctx}). For this purpose, we have designed a comprehensive evaluation protocol, shown in Figure 2, based on existing metrics that evaluates the performance of models across 3 dimensions:

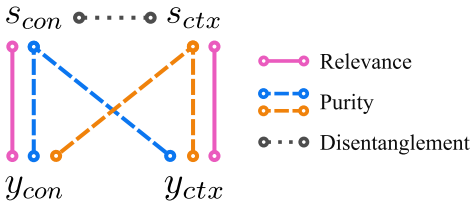

Figure 2: Diagram illustrating the evaluation protocol assessing summaries along 3 axes: relevance, purity, and disentanglement.

Relevance Generated summaries should closely correspond with the available reference summaries. We measure the lexical overlap and semantic similarity between (s_{con}, y_{con}) and (s_{ctx}, y_{ctx}) using ROUGE (R-i) (Lin, 2004) and BERTScore (Zhang et al. 2020; BS), respectively.

Purity Generated contribution summary should closely correspond with its respective reference summary, but should not overlap with the context reference summary. We measure the lexical overlap between s_{con} and (y_{con}, y_{ctx}) using NouveauROUGE $_{con}$ (N$_{con}$-i) (Conroy et al., 2011). The metric reports an aggregate score defined as a linear combination between the two components:

$$\begin{aligned} \text{NouveauROUGE}_{con}\text{-}i = \alpha_0^i \\ + \alpha_1^i \text{ROUGE-}i(s_{con}, y_{con}) \\ + \alpha_2^i \text{ROUGE-}i(s_{con}, y_{ctx}), \end{aligned}$$

where weights α_j^i were set by the original authors to favor outputs with maximal and minimal overlap with related and unrelated references, accordingly. Analogously, we calculate N$_{ctx}$-i in reverse direction between s_{ctx} and (y_{ctx}, y_{con}). Purity P-i is defined as the average novelty in both directions:

$$\text{Purity-}i = (\text{N}_{con}\text{-}i + \text{N}_{ctx}\text{-}i)/2; \quad \text{(P-}i\text{)}.$$

Disentanglement Generated contribution and context summaries should have minimal overlap. We measure the degree of lexical overlap and semantic similarity between (s_{con}, s_{ctx}) using ROUGE and BERTScore, respectively. To maintain consistency across metrics (higher is better) we report disentanglement scores as complements of the associated metrics:

$$\begin{aligned} \text{DisROUGE-}i = 100 - \text{ROUGE-}i; \quad \text{(D-}i\text{)}, \\ \text{DisBERTScore} = 100 - \text{BERTScore}; \quad \text{(DBS)}. \end{aligned}$$

5.2 Results

In Table 2 we report results from the automatic evaluation protocol described in Subsection 5.1.

Model		Relevance				Purity		Disentanglement			
		R-1	R-2	R-L	BS	P-1	P-2	D-1	D-2	D-L	DBS
CC	Con	39.16	15.96	24.65	63.22	2.77	3.69	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	2.75	3.68	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	2.73	3.68	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	2.68	3.60	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. Con and Ctx refer to contribution summary and context summary, respectively. Purity and Disentanglement are measaured on the pairs of contribution and context summaries.

Original 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. The control schemes (ideal thermal comfort, a traditional control and the RL control) are implemented in MATLAB. Using the Building Control Virtual Test Bed (BCVTB), the control of the thermostat schedule during each sample time is implemented for the office in EnergyPlus alongside local weather data. Results from training and validation indicate that the RL controller impoves thermal comfort by an average of 15% and energy efficiency by an average of 2.5% as compared to other strategies mentioned.

Generated Abstract: 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 (PID) controllers. These classical methods do not perform optimally. The high thermal inertia of buildings induces large time delays in the building dynamics, which cannot be handled efficiently by the simple on/off controllers. However, due to the high non-linearity in building dynamics coupled with uncertainties such as weather, energy pricing, etc., these PID controllers require extensive retuning or auto-tuning capabilities, which increases the difficulty and complexity of the control problem. In this work, we introduce novel control algorithms from a branch of machine learning called reinforcement learning. From a controls perspective, reinforcement learning algorithms can be considered as direct adaptive optimal control. Like optimal control, reinforcement training algorithms minimize the cumulative sum of costs over a time horizon. Unlike traditional optimization algorithms can learn optimal control actions

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. However, due to the high nonlinearity in building dynamics coupled with uncertainties such as weather, energy pricing, etc., these PID controllers require extensive retuning or auto-tuning capabilities, which increases the difficulty and complexity of the control problem. The high thermal inertia of buildings induces large time delays in the building dynamics, which cannot be handled efficiently by the simple on/off controllers.

Table 3: Generated samples compared with the original and generated abstracts of the associated paper. The second rows shows the output decoded from DistilBART fine-tuned on our dataset, the third rows shows the outputs from CONTROLCODE model. Our model successfully generates disentangled content, thus making it easier to follow than the abstract.

Relevance Across most models and metrics, relevance scores for context generation are higher than those for contribution summarization. Manual inspection revealed that in some cases generated context summaries also include article contribution information, while this effect was not observed in the reverse situation. Considering that silver standard annotations may contain noisy examples with incorrectly separated references, we suspect that higher ROUGE scores for context summaries may be caused by noisy predictions coinciding with noisy references. Examples of such summaries are shown in the Appendix C. We also observe that informativeness-guided models (+INF) perform on par with their respective base versions, and the additional training objective does not affect the performance on the relevance metric. This insight corroborates with Peyrard (2019) who defines informativeness and relevance as orthogonal criteria.

Purity While the informativeness objective was designed to improve the novelty of generated summaries, results show an opposite effect, where informativeness-guided models slightly underperform their base counterparts. The true reason for such behavior is unknown, however, it might be an indicator that the outbound citations C_O are not a good approximation of reference context summaries y_{ctx}, or the relationship between the two is weak. This effect is more evident in the Medical and Biology domains, which are the two most frequent domains in the dataset.

Disentanglement Results indicate that CONTROLCODE-based models perform better than MULTIHEAD approaches in terms of generating disentangled outputs. This comes as a surprise

given that the CC models share all parameters between the two generation modes, but might indicate that the two tasks contain complementary training signals. We also noticed that, both informativeness-guided models performed better in terms of D-1.

Based on both purity and disentanglement evaluations, we suspect that the informativeness objective does guide the models to output more disentangled summaries (second term in Eq 2), but the signal is not strong enough to focus on generating the appropriate content (first term in Eq 2). It is also clear that the MULTIHEAD model benefits more from the additional training objective.

6 Analysis

6.1 Qualitative Analysis

To better understand the strengths and shortcomings of our models, we performed a qualitative study of model outputs. Table 3 shows an example of generated summaries compared with the original abstract of the summarized article. Our model successfully separates the two generation modes and outputs coherent and easy to follow summaries. The contribution summary clearly lists the novelties of the work, while the context summary introduces the task at hand and explains its importance. In comparison, the original abstract briefly touches on many aspects: the context, methods used, and contributions, but also offers details that are not of primary importance, such as the detailed about the simulation environment.

More generally, the described trends hold across summaries generated by our models. The model outputs are fluent, abstractive, offer good separation between modes, and are on topic. An artifact noticed in a few instances of the inspected outputs was leakage of contribution information into context summaries. Factual correctness of summaries is discussed in Section 6.4. Other examples of generated summaries are included in the Appendix C.

6.2 Per-domain Performance

Taking advantage of the rich metadata associated with the S2ORC corpus, we analyze the performance of models across the 10 most frequent scientific domains. Table 4 shows the results of contribution summarization using the CONTROLCODE[7] model. While ROUGE-1 scores oscillate around 40 points for most academic fields, the results indicate that summarizing documents from the Medical domain is particularly difficult, with models scoring about 7 points below average. Manual inspection of instances with low scores (R-1 < 20), exposed that contribution summaries in the Medical domain are highly quantitative (*e.g.* "Among these treated ... retinopathy was noted in X%"). While other domains such as Biology also suffer from the same phenomenon, low-scoring quantitative summaries were 1.9 times more frequent in Medicine than in Biology. An investigation into the domain distribution in our dataset (Appendix) revealed that Biology and Medicine are the two best represented fields in the corpus, with Biology having over twice as many examples. We hypothesize that the poor performance of models stems from the fact that generating such quantitative summaries requires a deeper, domain-specific understanding of the source document and the available in-domain training data is insufficient to achieve that goal.

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: Relevance evaluation of contribution summaries for the top 10 domains generated using the CONTROLCODE model. Performance on Medicine domain is paricularly low.

Dataset	A1	A2	A3	AVG.
S2ORC	82%	78%	70%	77%
CORD	88%	76%	78%	81%

Table 5: Usefulness of disentangled summaries in percentage, *e.g.*, Annotator 1 (A1) chose the disentangled summaries 82% out of all the samples from S2ORC.

6.3 Human Evaluation of Usefulness

To assess the usefulness of the newly introduced task to the research community, we conducted a human study involving expert annotators. The study aimed to compare disentangled papers summaries with traditional, abstract-based summaries in a hypothetical paper reviewing setting. Judges were shown both types of summaries side by side and asked to pick one which would be more helpful for conducting the paper review.[8] Abstract-based sum-

[7] The remaining models exhibit the same pattern.

[8] We include a screen shot of the annotation user interface in Appendix B.

maries were generated by a model with a configuration identical to the models previously introduced in this work, trained to generate full abstracts using the same training corpus. Annotators that participated in this study hold graduate degrees in technical fields and are active in the research community, however, they were not involved or familiar with this work prior to this experiment. The study used 100 examples, out of which 50 were decoded on the test split of the adapted S2ORC dataset, while the other 50 were generated in a zero-shot fashion from articles in the CORD dataset (Wang et al., 2020), a recently introduced collection of papers related to COVID-19. The inter-annotator agreement measured by Fleiss' Kappa were 0.41 and 0.33 for the S2ORC and CORD datasets, respectively. Results in Table 5 show the proportion of all examples where the annotators preferred the disentangled summaries over the generated abstracts. The numbers indicate a strong preference from the judges for disentangled summaries, in the case of both S2ORC and CORD examples. The values on CORD samples are slightly higher than those on S2ORC; we suspect this being due to the fact that the annotators were less familiar with the topics described in Covid-related publications and would require more help to review such articles.

6.4 Factuality of Generated Summaries

In the scientific literature domain, the truthfulness of generated summaries with the input articles is a crucial aspect. Thus, we measured the factual consistency of 10 pairs of summaries sampled from Computer Science domain by assessing the validity of each *sentence* in the generated summaries against the input article and representing truthfulness as the proportion of sentences deemed consistent. We compared summaries from CC+INF (contribution, context) and DistilBART, resulting in 57, 71, 67 sentences to evaluate, respectively. As shown in Table 7, we found that most sentences copy segments from various positions in the input and lightly paraphrased or fused them together, which led to high percentage of factually consistent sentences.

6.5 Evaluation against Gold Annotations

As discussed in Section 3.1, contribution and context labels are assigned automatically using a data-driven classifier, which could introduce errors in the process. Therefore, we created a gold standard evaluation set by manually annotating 100 samples from the test set and report the evaluation results in Table 6. A sharp drop in ROUGE scores for the context summaries is caused by some examples receiving zero scores for generating context summaries when the manual annotators judged that there are not existent. The overall trend of CONTROLCODE model outpeforming MULTIHEAD model is still observed in the evaluation. More importantly, a reverse tendency is noticeable when the two models are trained with the informativeness objective. Specifically, the MULTIHEAD model showed significant improvement in terms of novelty and disentanglement.

7 Conclusions

In this paper, we propose *disentangled paper summarization*, a new task in scientific paper summarizing where models simultaneously generate contribution and context summaries. With the task in mind, we introduced a large-scale dataset with fine-grained reference summaries and rich metadata. Along with the data, we introduced three abstractive baseline approaches to solving the new task and thoroughly assessed them using a comprehensive evaluation protocol design for the task at hand. Through human studies involving expert annotators with motivated the usefulness of the task in comparison to the current scientific paper summarization setting. Together with this paper, we release the code, trained model checkpoints, and data preprocessing scripts to support future work in this direction. We hope this work will positively contribute to creating AI-based tools for assisting scientists in the research process.

Limitations

The importance of factuality in scientific literature summarization is crucial, as the summaries will serve as evidence for scientific discussion and citations. While our human annotation showed over 90% of the summary sentences were truthful to the input documents, the gap needs to be filled towards zero non-truthful sentences. In addition, failing to construct discourse-level coherence is another factor non-truthful summaries, which our models do not take into account explicitly.

Our models are developed upon DistilBART, which can process up to 1024 tokens for input. However, scientific documents (such as an 8-page paper like this submission) tend to exceed 1024 tokens. While there might be tendencies as to

Model		Relevance				Purity		Disentanglement			
		R-1	R-2	R-L	BS	P-1	P-2	D-1	D-2	D-L	DBS
CC	Con	39.37	15.86	24.73	63.28	2.30	3.22	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	2.17	3.10	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	2.21	3.13	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	2.26	3.13	51.56	68.57	64.97	32.35
	Ctx	29.14	10.25	18.48	54.92						

Table 6: Automatic evaluation results on 100 samples from the test set with **manual** annotation on contributions. For all metrics, higher values indicate better results.

Model		Truthfulness [%]
CC+INF	Con	93.0
	Ctx	90.1
DistilBART	-	91.0

Table 7: Proportion of factually consistent sentences in the summaries.

which sections likely discuss contributions or backgrounds, it is important to consider all the documents whenever possible, not to mention those of cited documents. Efficient incorporation of relevant documents at scale (*e.g.*, efficient attention, retrieval-augmented generation) is an active research area and should be considered for future work.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Praveen Bysani. 2010. Detecting novelty in the context of progressive summarization. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2, 2010, Los Angeles, California, USA - Proceedings of the Student Research Workshop*, pages 13–18. The Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained Markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232.

Yi Chang, Jiliang Tang, Dawei Yin, Makoto Yamada, and Yan Liu. 2016. Timeline summarization from social media with life cycle models. In *IJCAI*, pages 3698–3704.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. 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)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400. Association for Computational Linguistics.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada. Association for Computational Linguistics.

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. 2011. Nouveau-ROUGE: A novelty metric for update summarization. *Computational Linguistics*, 37(1):1–8.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task.

Jean-Yves Delort and Enrique Alfonseca. 2012. Dualsum: a topic-model based approach for update summarization. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27,*

2012, pages 214–223. The Association for Computer Linguistics.

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

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 687–697. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. TalkSumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2125–2131, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Sebastian Martschat and Katja Markert. 2018. A temporally sensitive submodularity framework for timeline summarization. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 230–240. Association for Computational Linguistics.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long 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 2: Short Papers)*, pages 1080–1089, Online. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696, Manchester, UK. Coling 2008 Organizing Committee.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. 2019. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of AAAI 2019*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. Seal: Segment-wise extractive-abstractive long-form text summarization. *arXiv preprint arXiv:2006.10213*.

A Implementation Details

Our models build upon DistilBART[9] (Sanh et al., 2019; Wolf et al., 2019), a Transformer-based (Vaswani et al., 2017), pre-trained sequence-to-sequence architecture distilled from BART (Lewis et al., 2020). Specifically, we used a model with 6 self-attention layers in both the encoder and Decoder. Weights were initialized from a model fine-tuned on a news summarization task that are available at https://huggingface.co/sshleifer/student_cnn_6_6. For the MULTIHEAD model, the final layer of the decoder was duplicated and initialized with identical weights. We fine-tuned on the training set for 80000 gradient steps on one NVIDIA V100 GPU with a fixed learning rate of 3.0×10^{-5} and choose the best checkpoints in terms of ROUGE-1 scores on the validation set. The loss scaling hyperparameter λ (Eq. 3) was set to 0.05 and 0.01 for the CONTROLCODE and MULTIHEAD models, accordingly. Input and output lengths were set to 1024 and 200, respectively. At inference time, we decoded using beam search with beam size 5. The evaluation was performed using SummEval toolkit (Fabbri et al., 2020).

B Human Evaluation of Disentanglement

In addition to various automatic evaluation, we perform human evaluation on disentanglement to understand which models human annotators prefer. We use Best-Worst scaling (Kiritchenko and Mohammad, 2017) over the 4-tuples of summaries on 50 random samples from the test set and have 3 annotators pick the best and the worst contribution and context summary pairs in terms of disentanglement. The rating in Table 8 shows the percentage a model is chosen as the best minus the percentage a model is chosen as the worst; the rating ranges from -1 to 1. A similar trend to automatic disentanglement evaluation is observed here as well in that introducing the informativeness objective significantly improves MULTIHEAD model, while it conversely affects CONTROLCODE model.

Table 8: Best-Worst scaling scores for Disentanglement.

Model	Rating
CC	**0.027**
CC+INF	0.020
MH	-0.073
MH+INF	**0.027**

We show the annotation user interface in Fig 3.

[9]We did not observe a substantial difference in performance between DistilBART and BART.

C Generated Full Samples from Different Models

We show additional samples generated from CONTROLCODE model in Table 9. A common failure case of all our models is the lack of disentanglement between the two summaries. While samples start generating on-topic, the model for context summary generation starts generating contributions at one point.

Table 9: Generated Sample. In this example, the red highlighted segments describe the contribution.

Original Abstract: Domain adaption (DA) allows machine learning methods trained on data sampled from one distribution to be applied to data sampled from another. It is thus of great practical importance to the application of such methods. Despite the fact that tensor representations are widely used in Computer Vision to capture multi-linear relationships that affect the data, most existing DA methods are applicable to vectors only. This renders them incapable of reflecting and preserving important structure in many problems. We thus propose here a learning-based method to adapt the source and target tensor representations directly, without vectorization. . . .

Generated Abstract: The difficulty of securing an appropriate and exhaustive set of training data, and the tendency for the domain of application to drift over time, often lead to variations between the distributions of the training (source) and test (target) data. In Machine Learning this problem is labeled domain mismatch. Failing to model such a distribution shift may cause significant performance degradation. Domain adaptation (DA) techniques capable of addressing this problem of distribution shift have thus received significant attention recently [24]. . . .

Contribution: To address these issues, we propose a novel approach termed Tensor-Aligned Invariant Subspace Learning (TAISL) to learn an invariant tensor subspace that is able to adapt the tensor representations directly. By introducing a set of alignment matrices, the tensors from the source domain are aligned to an underlying tensor space shared by the target domain. Instead of executing a holistic adaptation (where all feature dimensions would be taken into account), our approach performs mode-wise partial adaptation where each mode is adapted separately to avoid the curse of dimensionality. . . .
Context: Deep convolutional neural networks (CNNs) represent the state-of-the-art method for a substantial number of visual tasks. The activations of such CNNs, and the interactions between them, are naturally represented as tensors, meaning that DA should also be applied using this representation. . . . The proposed direct tensor method uses much lower dimensional entities, thus avoiding these estimation problems. To address these issues we propose to learn an invariant tensor subspace that is able to adapt the tensor representations directly. We show in Section 5 that the proposed method outperforms all comparators in DA of the Convolutional Activation of CNNs. Higher-order tensor modeling offers us an opportunity to investigate multiple interactions and couplings that capture the commonality and differences between domains. Following this idea, a novel approach

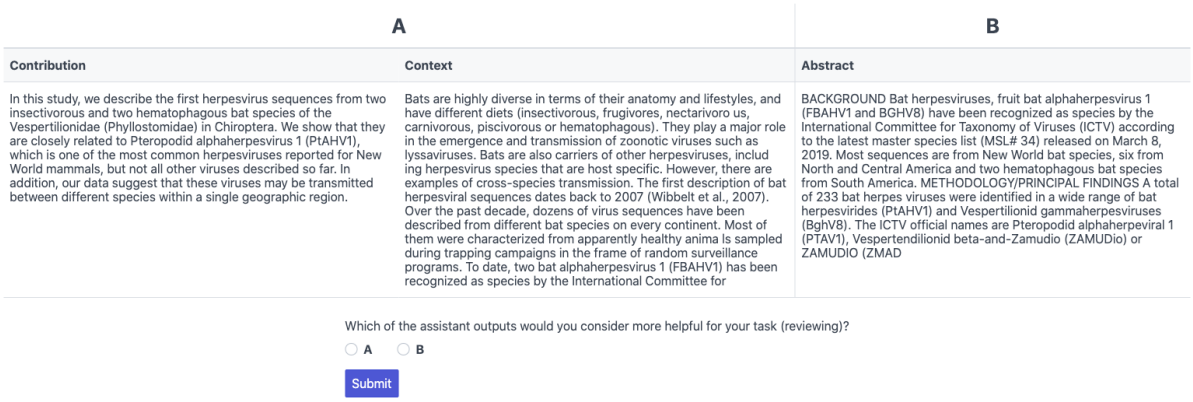

Figure 3: The annotation interface. Summaries indicated with A and B are disentangled summaries and a generated abstract, respectively.

D Ethical Considerations

While we achieve reasonable automatic evaluation results using the proposed models, we note that these models pose ethical risks in two ways. From readers' perspective, entirely trusting machine-generated summaries would lead to a wrong understanding of the articles, thus potentially harming the progress of the research community. Even though we show that more than 90% of sentences from the annotated summaries were truthful to the input articles, the remaining sentences that were not truthful are impactful enough to misunderstand the contributions.

From writers' perspective, our proposed models could be used maliciously to appear valuable. In a hypothetical situation where our model outputs are regarded trustworthy enough for people to assess articles, "hacking" our summarization model to output over-claming contributions could be possible.