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 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. 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 (Raffel et al., 2019; Lewis et al., 2020), 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 (Conan 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.

https://arxiv.org/stats/monthly_submissions
For this task, we introduce a new large-scale dataset by extending the S2ORC \cite{lo2020s2orc} corpus of scientific papers, which spans multiple 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 \cite{fan2018exploiting, keskar2019ctrl}, 2) a one-to-many sequence model with a branched decoder for multi-head generation \cite{luong2016comparing, guo2018neural}, and 3) an information-theoretic training strategy leveraging supervision coming from the citation metadata \cite{peyrard2019information}. 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, model checkpoints, and data preparation scripts introduced in this work are available at \url{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 \cite{see2017get, paulus2018iterative} to leveraging large-scale Transformer-based models pre-trained on vast amounts of data \cite{liu2019roberta, lewis2019bart}, often in multi-task settings \cite{raffel2019exploring}. A similar shift can be seen in scientific paper summarization, where state-of-the-art approaches utilize custom pre-training strategies \cite{chung2019papersum} and tackle problems of summarizing long documents \cite{zhao2020scientific, zaheer2020big}. Other methods, at a smaller scale, seek to utilize the rich metadata associated with scientific articles and combine them with graph-based methods \cite{yasunaga2019graph}. 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 \cite{cohan2018measuring} were automatically collected from open-access paper repositories and consider article abstracts as the reference summaries. Other forms of supervision have also been investigated for the task, including author-written highlights \cite{collins2017multitask}, human annotations and citations \cite{yasunaga2019graph}, and transcripts from conference presentations of the articles \cite{lev2019learning}. 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 \cite{dang2018update} 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 \cite{bysani2010novelty, delort2012novelty} and timeline summarization \cite{martschat2018timeline} 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
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.

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

$D$ 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 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 & 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 carefully curate the data available in the S2ORC corpus and extend it with new reference labels.

Data Curation Some papers in the S2ORC corpus 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.

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 & Radev (2008) and Cohan & Goharian (2015) in representing citations with the sentences in which the citation occurs. 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. 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.

---

2 Release ID: 20190928.
3 If a publication is cited multiple times within a source article we concatenate all relevant sentences.
4 Sentences not labeled as contribution are considered context, we leave finer-grained labels for future work.
Figure 1: Model diagram. Left: CONTROL CODE 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.

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 contribution 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:

CONTROL CODE (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) showed that prepending a special token or descriptive prompt 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

Peyrard (2019) proposed an information-theoretic perspective on text summarization which decomposes the criteria of a good summary into redundancy, relevance, and informativeness. Among these criteria, informativeness measures the user’s degree of surprise after reading a summary given their background knowledge, and can be formally defined as:

$$\text{Inf}(D, K) = - \sum_i P_D(\omega_i) \log P_K(\omega_i),$$

(1)

where $\omega_i$ 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.
As defined by [Peyrard, 2019], informativeness is in direct correspondence to contribution summarization. Paper contributions are novel contents introduced to the community, which causes surprisal given the general knowledge about the state of the field. Therefore, in this work we explore utilizing this measure as an auxiliary objective that is optimized during training. We define the semantic unit \( \omega_i \) as the summary itself\(^5\), which enables a simple interpretation of the corresponding probabilities. 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:

\[
\text{Inf}(D, K) = \begin{cases} 
-p(y_{\text{con}} \mid D) \log p(y_{\text{con}} \mid C_O) & \text{if generating contributions} \\
-p(y_{\text{ctx}} \mid D) \log p(y_{\text{ctx}} \mid C_I) & \text{otherwise}
\end{cases},
\]

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: \( p(y_{\text{con}} \mid C_O) \) is computed by estimating the probability of \( y_{\text{con}} \) when feeding \( C_O \) as the encoder input.

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

\[
L = L_{\text{CE}} - \lambda \text{Inf}(D, K),
\]

where \( \lambda \) is a scaling hyperparameter determined through cross-validation.

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: CONTROL
- MH, MH+INF: MULTIHEAD

5.1 Evaluation

We perform automatic evaluation of the system outputs \((s_{\text{con}}, s_{\text{ctx}})\) against the silver standard references \((y_{\text{con}}, y_{\text{ctx}})\). 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:

- **Relevance**: Generated summaries should closely correspond with the available reference summaries. We measure the lexical overlap and semantic similarity between \((s_{\text{con}}, y_{\text{con}})\) and \((s_{\text{ctx}}, y_{\text{ctx}})\) 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 \(s_{\text{con}}\) and \((y_{\text{con}}, y_{\text{ctx}})\) using NouveauROUGE\(_{\text{con}}\) \((N_{\text{con}-i})\) (Conroy et al., 2011). The metric reports an aggregate score defined as a linear combination between the two components:

\[
\text{NouveauROUGE}_{\text{con}-i} = \alpha_0 + \alpha_1 \text{ROUGE-}i(s_{\text{con}}, y_{\text{con}}) + \alpha_2 \text{ROUGE-}i(s_{\text{con}}, y_{\text{ctx}}),
\]

where weights \( \alpha_i \) were set by the original authors to favor outputs with maximal and minimal overlap with related and unrelated references, accordingly. Analogously, we calculate \( N_{\text{ctx}-i} \) in

\(^5\)For simplicity in modeling, we chose the entire summary. However, this goes against the requirement set by Peyrard, 2019 that \( \omega_i \) is a primitive semantic unit, because a paragraph’s meaning can be decomposed into higher granular units.
reverse direction between $s_{ctx}$ and $(y_{ctx}, y_{con})$. Purity $P_i$ is defined as the average novelty in both directions:

$$Purity\cdot i = \frac{N_{con\cdot i} + N_{ctx\cdot i}}{2}; \quad (P_i).$$

**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:

- DisROUGE-$i = 100 - \text{ROUGE-}i; \quad (D_1)$
- DisBERTScore = $100 - \text{BERTScore}; \quad (DBS)$

### 5.2 Implementation Details

Our models build upon distilBART\(^6\) (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.\(^7\) 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 with a fixed learning rate of $3.0 \times 10^{-5}$ and choose the best checkpoints in terms of ROUGE-1 scores on the validation set. The loss scaling hyparameter $\lambda$ (Eq. 3) was set to 0.05 and 0.01 for the CONTROL CODE 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).

### 5.3 Results

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

**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 D. 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 CONTROL CODE-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

\(^6\)We did not observe a substantial difference in performance between distilBART and BART.

\(^7\)Model weights are available at [https://huggingface.co/sshleifer/student_cnn_6_6](https://huggingface.co/sshleifer/student_cnn_6_6)
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 measured on the pairs of contribution and context summaries.

<table>
<thead>
<tr>
<th>Model</th>
<th>Relevance</th>
<th></th>
<th>Purity</th>
<th></th>
<th>Disentanglement</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-2</td>
<td>R-L</td>
<td>BS</td>
<td>D-1</td>
<td>D-2</td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Con</td>
<td>39.16</td>
<td>15.96</td>
<td>24.65</td>
<td>63.22</td>
<td>2.77</td>
<td>3.69</td>
</tr>
<tr>
<td>Ctx</td>
<td>41.84</td>
<td>17.24</td>
<td>24.55</td>
<td>63.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC+INF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Con</td>
<td>38.92</td>
<td>15.95</td>
<td>24.65</td>
<td>62.94</td>
<td>2.75</td>
<td>3.68</td>
</tr>
<tr>
<td>Ctx</td>
<td>41.49</td>
<td>17.03</td>
<td>24.50</td>
<td>63.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Con</td>
<td>39.20</td>
<td>15.98</td>
<td>24.72</td>
<td>63.04</td>
<td>2.73</td>
<td>3.68</td>
</tr>
<tr>
<td>Ctx</td>
<td>41.67</td>
<td>17.23</td>
<td>24.65</td>
<td>63.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH+INF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Con</td>
<td>38.74</td>
<td>15.90</td>
<td>24.59</td>
<td>62.70</td>
<td>2.68</td>
<td>3.60</td>
</tr>
<tr>
<td>Ctx</td>
<td>40.39</td>
<td>16.31</td>
<td>23.83</td>
<td>62.85</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

is not strong enough to focus on generating the appropriate content (first term in Eq 2). It is also clear that the MULTHEAD 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. However, the factual correctness of summaries could not be assessed due to the highly specialized content and language of the summarized articles. An artifact noticed in a few instances of the inspected outputs was leakage of contribution information into context summaries. Other examples of generated summaries are included in the Appendix.

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 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

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

The remaining models exhibit the same pattern.
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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2ORC</td>
<td>82%</td>
<td>78%</td>
<td>70%</td>
<td>77%</td>
</tr>
<tr>
<td>CORD</td>
<td>88%</td>
<td>76%</td>
<td>78%</td>
<td>81%</td>
</tr>
</tbody>
</table>

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. We show a screen shot of the annotation user interface in Fig. Abstract-based summaries 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 was 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

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

less familiar with the topics described in Covid-related publications and would require more help to review such articles.

6.4 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 outperforming 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.

Table 6: Automatic evaluation results on 100 samples from the test set manually annotated for contributions. For all metrics, higher values indicate better results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Relevance</th>
<th>Purity</th>
<th>Disentanglement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-2</td>
<td>R-L</td>
</tr>
<tr>
<td>CC</td>
<td>39.37</td>
<td>15.86</td>
<td>24.73</td>
</tr>
<tr>
<td>Cx</td>
<td>30.59</td>
<td>11.22</td>
<td>19.08</td>
</tr>
<tr>
<td>CC+INF</td>
<td>38.38</td>
<td>15.21</td>
<td>23.47</td>
</tr>
<tr>
<td>Cx</td>
<td>30.14</td>
<td>11.10</td>
<td>19.00</td>
</tr>
<tr>
<td>MH</td>
<td>38.63</td>
<td>15.53</td>
<td>24.68</td>
</tr>
<tr>
<td>Cx</td>
<td>29.82</td>
<td>10.61</td>
<td>18.51</td>
</tr>
<tr>
<td>MH+INF</td>
<td>39.43</td>
<td>15.75</td>
<td>24.77</td>
</tr>
<tr>
<td>Cx</td>
<td>29.14</td>
<td>10.25</td>
<td>18.48</td>
</tr>
</tbody>
</table>

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.
REFERENCES


Under review as a conference paper at ICLR 2021


