We Can Explain Your Research in Layman's Terms: Towards Automating Science Journalism at Scale

Rumen Dangovski,[1, *] **Michelle Shen,**[1] **Dawson Byrd,**[1] **Li Jing,**[1] **Desislava Tsvetkova,**[2]
Preslav Nakov,[3] **Marin Soljačić**[1]

[1]Massachusetts Institute of Technology, [2]Sofia University, [3]Qatar Computing Research Institute, HBKU, [*]rumenrd@mit.edu

Abstract

We propose to study *Automating Science Journalism* (ASJ), the process of producing a layman's terms summary of a research article, as a new benchmark for long neural abstractive summarization and story generation. Automating science journalism is a challenging task as it requires paraphrasing complex scientific concepts to be grasped by the general public. Thus, we create a specialized dataset that contains scientific papers and their *Science Daily* press releases. We demonstrate numerous *sequence to sequence* (seq2seq) applications using Science Daily with the aim of facilitating further research on language generation, which requires extreme paraphrasing and coping with long research articles. We further improve the quality of the press releases using co-training with scientific abstracts of sources or partitioned press releases. Finally, we apply evaluation measures beyond ROUGE, and we demonstrate improved performance over strong baselines, which we further confirm by quantitative and qualitative evaluation.

Introduction

Recent years have been characterized by rapid growth of published scientific research. Coping with this quantity is increasingly challenging, which has led to the emergence of a number of initiatives, including software applications that try to summarize and to organize research articles. For example, *Scholarcy* helps researchers and students by summarizing relevant portions of academic papers. Likewise, *Mendeley* establishes meaningful links between research papers. Furthermore, there are emerging tools, such as *Litmaps* that place scientific research in a broader perspective, thus making it accessible to layman readers.

Traditionally, this was the task of science journalism, led by media outlets such as *Science Daily*, *Scientific American*, and *Popular Science*, which establish some of the few direct connections between scientific research and the general public. As demonstrated in Table 1, this is a tremendously difficult task: it requires writing factual summaries, while also paraphrasing complex scientific concepts using a language that is accessible to the general public.

We argue that the abundance of science journalism articles enables a variety of learning approaches, most notably neural text summarization (Rush, Chopra, and Weston 2015). The latter has undergone strong evolution recently (Lin and Ng 2019): from extractive (Nallapati, Zhai, and Zhou 2017) through abstractive (Nallapati et al. 2016) to hybrid (See, Liu, and Manning 2017) models; from maximum likelihood to reinforcement learning objectives (Celikyilmaz et al. 2018; Chen and Bansal 2018); from small to large datasets (Grusky, Naaman, and Artzi 2018), which are also abstractive (Sharma, Li, and Wang 2019); from short to orders of magnitude longer sources and targets (Liu et al. 2018); from models trained from scratch to using pre-trained representations (Edunov, Baevski, and Auli 2019; Liu and Lapata 2019).

Generated: it 's no secret that women are as good as men . but when it comes to job satisfaction , a new study shows that this gender equality can affect one 's own job and make the impression that women experience higher levels of gender equity among women .
Target: male workers appear to support women becoming ceos even more than female workers do , finds new research on the proverbial glass ceiling and job satisfaction in six formerly socialist countries .
Source snippets: . . . moreover , recent data show that , in spite of significant barriers , more women reach the upper managerial ranks in the workplace . . . does gender equality in workplace promotion opportunities have consequences for job satisfaction ? we address this question by examining the link between job satisfaction and perceived prospects for women to become top manager at the firm .

Table 1: Summary from our dataset (*short Science Daily) using our model (*SciBertSumAbs*).* We see the need for extreme paraphrasing and coherent generation.

From a modelling perspective, these advances are yet to be challenged with an abstractive summarization task (*i*) from *long source* research articles into *long targets*, and (*ii*) using *extreme paraphrasing*. Here, we argue that automating science journalism is a natural testbed for this.

The task is defined as follows: *Given a scientific article, produce a layman summary of that article.*

Our contributions can be summarized as follows:

- We introduce a text summarization task: generate a layman's terms summary of a research article in the form of a press release.
- We create a specialized dataset for the task and we experiment with a number of models.
- We focus on story generation as a way to model press releases, and we propose suitable data augmentation methods, which we validate extensively.

Related Work

Summarization of Scientific Documents Abu-Jbara and Radev (2011) produced readable and coherent citation-based summaries improving upon a history of related work (Nanba, Kando, and Okumura 2000; Nakov, Schwartz, and Hearst 2004; Elkiss et al. 2008; Qazvinian and Radev 2008; Mei and Zhai 2008; Mohammad et al. 2009; Divoli, Nakov, and Hearst 2012). Collins, Augenstein, and Riedel (2017) studied extractive summarization of scientific papers to highlights, following a history of predominantly extractive summarization of scientific documents (Kupiec, Pedersen, and Chen 1995; Saggion, AbuRa'ed, and Ronzano 2016). Yasunaga et al. (2019) proposed hybrid summarization of well-annotated datasets, thus extending work by (Jaidka et al. 2016, 2017, 2018). Beltagy, Lo, and Cohan (2019) fine-tuned BERT on scientific articles and improved the baselines for some downstream scientific tasks. Subramanian et al. (2020) performed summarization of very long documents, but did not address the task of extreme paraphrasing, nor did they use a seq2seq architecture. Luu et al. (2020) explained the relationship between two scientific documents via citations. Finally, recent advances in efficient Transformers (Beltagy, Peters, and Cohan 2020) made it possible to process long scientific documents efficiently, and thus scale ASJ. Recent proof-of-concept work has approached automating scientific reviewing (Wang et al. 2020; Yuan, Liu, and Neubig 2021). Furthermore, workshops, such as CL-SciSumm/ CL-LaySumm (Chandrasekaran et al. 2019) and LongSumm (Chandrasekaran et al. 2020) have offered opportunities for developing summarization of scientific documents.

Unlike the above work, we use orders of magnitude larger datasets with diverse content domains, and we generate meaningful abstractive summaries in layman's terms. To our knowledge, we are the first to explore scaling automating science journalism through summarization of long sources, which require extreme paraphrasing and long generation.

Scientific Datasets Dangovski et al. (2019) presented pioneering results on the *Science Daily* dataset using a seq2seq model with novel RNN units, based on rotation. However, their work was limited to short source and target pairs. Moreover, they performed summarization from a *journalistic* article in Science Daily article to the highlight of that article, again in Science Daily.

In contrast, we perform summarization from a *research journal* article to Science Daily highlights. This is an important distinction, as research articles use very different style, language, and terminology compared to journalistic articles.

Other work preserved the style of the source (Teufel and Moens 2002; Nikolov, Pfeiffer, and Hahnloser 2018; Cohan et al. 2018) or generated very short targets taking the form of blog titles (Vadapalli et al. 2018). Sharma, Li, and Wang (2019) introduced BigPatent as a new challenge for abstractive summarization, which is a good parallel to our task, as it still summarizes scientific content in an abstractive manner. Lev et al. (2019) proposed a dataset, TalkSumm, for generating summaries using conference talks. Recently, Cachola et al. (2020) introduced SciTldr for extreme summarization of scientific papers in Computer Science. Gidiotis and Tsoumakas (2020) used the RNN units from (Dangovski et al. 2019) and a divide-and-conquer approach to improve summarization of ArXiv and PubMed (Cohan et al. 2018) articles to abstracts. However, none of the above work addressed our task of producing a press release for a research article in layman's terms.

Data-Augmentation and Multitask Learning for Language Generation Our task and the corresponding datasets made it possible to use recent advances in transfer learning for NLP (Raffel et al. 2020; Ruder 2019). Namely, we combine datasets sharing a source domain, i.e., scientific articles, with different target domains, i.e., abstracts and press releases. We take inspiration from recent work on automatically generating news articles (Zellers et al. 2019), trained on multiple variations of the same dataset, e.g., in some instances, the headline might be used to generate the body, while in other, the body can be used to generate the headline. Similarly, via a special tag, we can signal to the decoder to generate either an abstract or a press release, or to generate the target in several steps by conditioning on intermediate outputs. Other ways to signal to the decoder were proposed in the context of summarization with user preferences (Fan, Grangier, and Auli 2018), neural machine translation (Lample and Conneau 2019; Aharoni, Johnson, and Firat 2019), and controllable text generation (Keskar et al. 2019) that contain tags, similarly to pre-training contextual word embeddings (Peters et al. 2018; Delvin et al. 2019). Finally, we should mention multitask learning (Raffel et al. 2020; Lewis et al. 2019; Cachola et al. 2020) for improving summarization.

The *Science Daily* Dataset

Dataset

We introduce two versions of *Science Daily*: (*i*) for long summarization, consisting of pairs of full-text scientific papers and their corresponding *Science Daily* press releases, and (*ii*) for short summarization, made of pairs of scientific papers cut after the first 400 words and corresponding short highlights in the press releases. Even though in this paper we put emphasis on long summarization, the *short Science Daily* is a task that is closer, in terms of length of sources and targets, to the one considered in (Dangovski et al. 2019).

Moreover, they both contain another challenging aspect, that is the difference in the style of language between the source and the target. See Table 2 for some statistics.

Science Daily	short	long
# pairs	50,308	50,134
# source words	400 ± 0	$5{,}975 \pm 2{,}731$
# target words	45 ± 19	488 ± 219
train/ dev/ test	90%/5%/5%	80%/10%/10%

Table 2: Statistics about the *Science Daily* datasets.

Rank	Journal	# dataset entries
1	PNAS	5,482
2	Science	4,006
14	Nature Geoscience	472
15	Nature Medicine	425
16	Nature Neuroscience	397
17	Nature Climate Change	396

Table 3: *Science Daily* covers diverse journals.

Note that the number of pairs in these datasets do not match, as not all *Science Daily* articles had highlights. The training split for *long Science Daily* is lower by 10% since its pairs contain more tokens than their counterparts in the short dataset. Below, we explain how we created our datasets.

Note also that our *Science Daily* dataset differs from existing datasets for summarization of scientific content as it is extremely diverse and covers a wide range of scientific fields, as shown in Table 3, and as it features a drastic change in style between the source and the target.

Press Releases. *Science Daily*[1] is a website that aggregates and publishes lightly edited press releases about science. We were granted access to download about 100,000 HTML pages from their website, each containing a public story about a recent research paper. From each HTML page, we extracted the main content, a short highlight, and a title.

Scientific Articles. We further parsed each HTML page of the press releases to obtain information about the target scientific article: title, short description, main content and DOI. Then, we sent the DOI to the *Crossref API*[2] to obtain the meta information about the target paper. We downloaded the papers as PDF files, and we then converted them to raw text. These papers span a large range of publishers including *American Association for the Advancement of Science* (AAAS), *Elsevier*, *Public Library of Science* (PLOS), *Proceedings of the National Academy of Sciences* (PNAS), *Springer* and *Wiley*. We ignored publishers with fewer than 100 papers.

There are many such publishers and the style of their PDFs is not consistent; hence, we opted to convert to text and to use scientific papers from the most prevalent publishers only.

[1]http://www.sciencedaily.com/

[2]https://www.crossref.org/

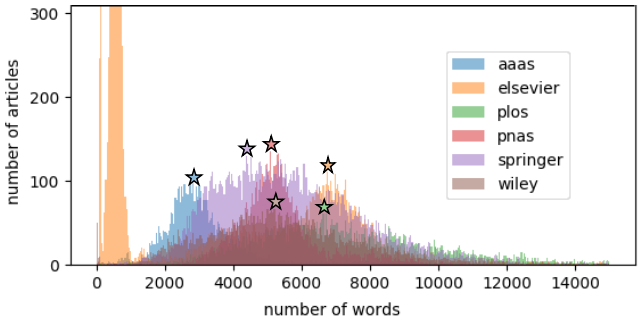

Figure 1: Histogram for number of articles vs. number of words for the selected publishers in *Science Daily*. Stars indicate modes of the histograms (excluding the outliers with fewer than 1,000 words for Elsevier).

Figure 1 shows statistics about the publishers. The figure gives a peek into the differences in style among the publishers. For example, AAAS publishes short letters, while PNAS publishes longer articles. We treat articles with fewer then 1,000 words as outliers, and we do not include them in the dataset.

Analysis: Comparison to Related Datasets

Compared to other datasets, *Science Daily* summaries are significantly more *abstractive*. To see this, we compare to the *ArXiv* dataset, which summarizes scientific articles to their abstracts. We use two statistics from (Grusky, Naaman, and Artzi 2018):

$$\text{coverage}(A, S) = (1/|S|) \sum_{f \in \mathcal{F}(A,S)} |f|$$

and

$$\text{density}(A, S) = (1/|S|) \sum_{f \in \mathcal{F}(A,S)} |f|^2$$

where $\mathcal{F}(A, S)$ is the set of extractive fragments, a sequence of words that is shared between the source and the target for a set of articles $\{A\}$ and a corresponding set of summaries $\{S\}$, $|f|$ is the number of words in fragment f, and $|S|$ is the number of words in summary S.

The *coverage* represents the fraction of words in an extractive fragment, and the *density* is the average length of these fragments. Figure 2 compares *Science Daily* to established datasets. We can see that the coverage is around 0.4 for *Science Daily* vs. 0.8 for *ArXiv*. Moreover, while the density for *Science Daily* is on the order of a few absolute density points, it is in the hundreds for *ArXiv*.

Another important characteristic of our *Science Daily* dataset is that both the source and the target are relatively long, with source articles and target press releases containing about 6,000 and 500 word tokens, respectively. For comparison, the *CNN/ Daily Mail* dataset is much shorter, with sources of 800 word tokens and targets of just 50 word tokens, and even the *ArXiv* dataset has substantially shorter targets of around 200 word tokens.

We further computed standard measures of language complexity such as SMOG, CLI, and LIX, as implemented in the NELA toolkit (Horne et al. 2018). The results are shown

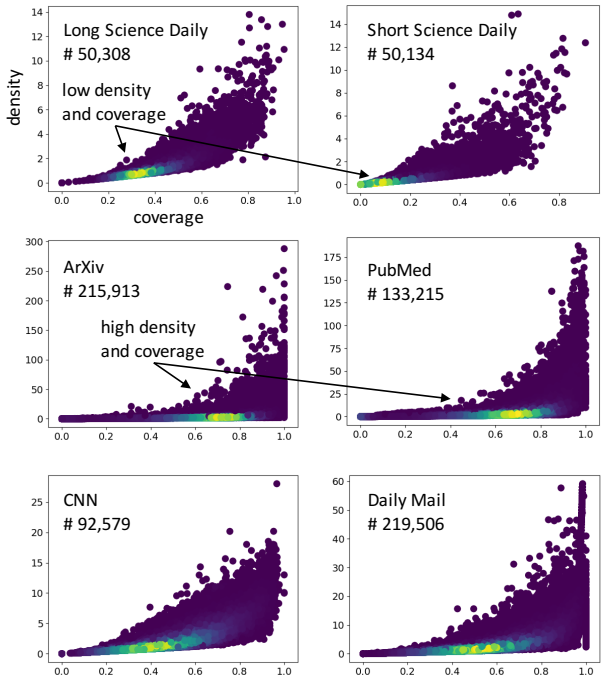

Figure 2: Density vs. coverage of source-target pairs for *Science Daily*, *ArXiv*, *PubMed*, and *CNN and Daily Mail*. Warmer colors show more data entries, and # is number of pairs. Outliers with extreme densities are omitted. Arrows indicate the modes of the datasets.

in Table 4, where we can see that the texts from scientific sources use more complex language.

We further used natural language inference (NLI) to explore which parts of the source text contain the most relevant information for summarizing *Science Daily* research articles. For each sentence from the target summary, we found a corresponding one in the source text that entailed it with the highest probability, and we marked the relative position of that sentence in the source text.

We repeated the procedure for all summaries, and we generated aggregated statistics about the relative positions of these source sentences (in bins), as shown in Figure 3. We can see on the left side of the figure that the gold journalistic summaries use information not only from the introduction and from the conclusion of the input research articles, but also from the entire input text. On the right side of the figure, we show a similar analysis for summaries generated by our model: we can see a similar pattern, (albeit different from the left histogram, the right histogram spreads throughout its entire range too), which means that the model learns to look at the entire input when generating a summary.

Evaluation

ROUGE. We use the standard ROUGE 1/2/L scores (Lin and Hovy 2003).

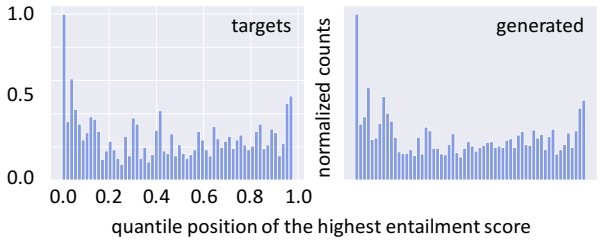

Figure 3: Positions of the source sentences that maximize the NLI entailment of the summary sentences for *Science Daily*. On the left are gold summaries, and on the right are summaries by our model (*Story+Parts*). The counts are normalized, so that the bin with the highest counts is at 1.0.

Dataset	SMOG	CLI	LIX
Science Daily	15.23 ± 1.51	14.34 ± 1.21	55.60 ± 4.66
PubMed	16.98 ± 1.65	14.21 ± 1.67	59.00 ± 6.73
ArXiv	13.74 ± 1.53	12.09 ± 1.64	50.26 ± 6.19
CNN	12.01 ± 1.67	10.66 ± 1.87	45.31 ± 8.20
Daily Mail	12.29 ± 1.61	10.35 ± 1.50	49.01 ± 7.80

Table 4: Complexity of related datasets' sources based on readability scores such as SMOG, CLI, and LIX. The datasets from scientific sources (the top half) use more complex language (bigger numbers indicate higher complexity).

Natural Language Inference (NLI). Ideally, each summary should be fully entailed from the source text. With this in mind, Falke et al. (2019) proposed an evaluation measure for text summarization that uses NLI and tries to find for each sentence in the summary the maximal probability of it being entailed from some sentence in the source text. The final score is calculated as the average of these probabilities:

$$\sigma(S) = \frac{1}{n}\sum_{j=1}^{n} \max_{d \in D} N(d, s_j) \quad (1)$$

where $N(d, s_j)$ is the probability that sentence s_j from the summary S is entailed from sentence d in the source document D, and n is the number of sentences in the summary.

This approach resembles the NLI analysis method we used above, but here the focus is on the score, while above we were interested in the relative position of the best-matching source sentence.

Prompt Ranking (PR). For *long Science Daily*, we further used an evaluation measure, inspired by the *prompt ranking measure* from (Fan, Lewis, and Dauphin 2018). For a target in the dataset, it takes the source and nine additional sources of different targets. Then, it tests whether the generator assigns higher probability to the target when conditioned on the correct source (by feeding the source into the encoder) compared to conditioning on the incorrect sources,

and measures the success rate of that test on a selected number of targets from the dataset.

Here, we follow the same procedure, but with the important modification that instead of taking the full source, we select a random substring of 100 words to feed into the encoder. To provide results in the context of existing work on prompt ranking, our aim is to mirror the original prompt ranking measure, which was used to rank the prompts (short prompts, such as a title of a movie) based on the probability that the true story (long generation) has, when conditioned on the prompts. In the *long Science Daily*, the press releases are similar to the stories in (Fan, Lewis, and Dauphin 2018), but the sources (the scientific papers) are not similar to the prompts. Hence, we take 100-word random substrings to form prompts for the press releases. We calculate the prompt ranking score on a hold-out set of 1,000 long *Science Daily* pairs, and we report its value in percentage points.

Experiments

Given that the size of the Science Daily dataset is not that large compared to existing summarization corpora, our task should benefit from using pre-trained models or from augmenting the data. Below, we present experiments that demonstrate techniques in both directions, which lay the foundations for our task.

Summarization with Pre-trained BERT

We begin by exploring familiar ground: short summarization using the *short Science Daily* (Table 2) à la CNN/ Daily Mail (See, Liu, and Manning 2017), i.e., our sources are up to 512 tokens long, and the targets are up to 140 tokens long. We choose an abstractive seq2seq model, following a strong neural summarization baseline with pre-trained BERT (Liu and Lapata 2019). In particular, we experiment with their BertSumAbs, which uses a pre-trained BERT model as an encoder and a Transformer (Vaswani et al. 2017) trained from scratch as a decoder. We denote this experiment with *BertSumAbs* as well.

Scientific Pre-training Since we are in the scientific domain, we replace the BERT (Delvin et al. 2019) encoder with SciBERT (Beltagy, Lo, and Cohan 2019), which is fine-tuned on scientific papers, and we dub the resulting model *SciBertSumAbs*. We train the model for 200K steps. The hyper-parameter values coincide with those for *BertSumAbs*.

In Table 5, we show how *BertSumAbs* and *SciBertSumAbs* compare in terms of ROUGE 1/2/L scores using beam search decoding and trigram blocking (Paulus, Xiong, and Socher 2018), thereby following the decoding setup in (Liu and Lapata 2019), but limiting the generation to 50–200 tokens. We observe sizable gains from using SciBERT. This result is expected since *Science Daily* focuses on the scientific articles (Table 3).

In the following experiments, we focus on the *long Science Daily* (Table 2) dataset.

Model	1	2	L
LEAD	19.7	3.7	13.1
BertSumAbs	27.16	4.54	21.45
SciBertSumAbs	**30.30**	**6.24**	**24.00**

Table 5: *Short Science Daily*: SciBERT pre-training improves over vanilla BERT (ROUGE scores in %). LEAD takes the first 45 words from the input.

Efficiency with CNN seq2seq

For the *long Science Daily*, we use CNN-based seq2seq architectures, which can handle long input. We start with a small vanilla convolutional seq2seq model (Gehring et al. 2017), corresponding to fairseq's ISWLT German–English (de-en) model (Ott et al. 2019), which we take directly from the library.

We train the model until convergence on the dev set with a learning rate of 0.25, Nesterov accelerated gradient (NAG) descent, 0.2 dropout, and a 0.1 gradient threshold. We name this experiment *Fconv*.

ASJ as Story Generation

We can frame ASJ as story generation, since a press release can be viewed as a story shaped around a scientific paper. The scientific paper itself can be viewed as a "writing prompt" for the story. Hence, our second model is a modification of a state-of-the-art model for neural story generation (Fan, Lewis, and Dauphin 2018, 2019). It introduces attention (Bahdanau, Cho, and Bengio 2015) between the output of the encoder and the decoder layers, as well as multi-head self-attention on the decoder layers (Vaswani et al. 2017) that is gated (Dauphin et al. 2017) and equipped with a multi-scale mechanism for down-sampling (Fan, Lewis, and Dauphin 2018). Since our sources are three orders of magnitude larger than the writing prompt sources for which the original story model has been used, we *additionally equip the encoders* with gated multi-scale multi-head self-attention. Thus, we extend the fairseq implementation with additional four-gated self-attention heads both on the encoders and on the decoders with projected inputs and down-sampling. We train the model until convergence on Dev with a learning rate of 0.25, NAG, dropout of 0.2, and a gradient threshold of 1.0. We call this experiment *Story*.

Training for all our fairseq models takes about 20-30 epochs depending on the batch size, which is around 30-40. In preprocessing, we only keep words that appear at least ten times in the source, or at least ten times in the target. Moreover, for these models we converted all textual data to *byte pair encoding* (BPE) (Sennrich, Haddow, and Birch 2016) with 32,000 BPE tokens both on the source and on the target side following the guidelines for fairseq.

Table 6 shows a comparison between *Fconv* and *Story*. Surprisingly, the simple *Fconv* baseline outperforms the *Story* model both on ROUGE scores and Prompt Ranking. We speculate that this might be due to *Fconv* being more extractive, which might influence the scores marginally.

Model	1	2	L	NLI	PR
LEAD	39.6	10.1	16.1	N/A	N/A
Fconv	39.2	9.5	36.9	0.23	38.0
FconvTopK	39.2	**10.8**	**37.0**	0.23	38.0
Story	38.9	7.8	36.4	0.12	22.7
StoryTopK	38.2	**8.5**	36.0	**0.14**	22.7

Table 6: *Long Science Daily*: baselines. *Fconv* outperforms *Story* in ROUGE 1/2/L and Prompt Ranking (PR); top-k sampling generally helps for *Fconv*. PR does not depend on the decoding scheme. LEAD takes the first 488 input words.

Model	1	2	L	NLI	PR
LEAD	39.6	10.1	16.1	N/A	N/A
Fconv	39.2	9.5	36.9	0.23	38.0
Fconv+ArXiv	**41.2**	**10.2**	**38.6**	**0.28**	**77.8**
FconvTopK	39.2	10.8	37.0	0.23	38.0
FconvTopK+ArXiv	**41.8**	**11.6**	**38.6**	**0.25**	**77.8**
Story	38.9	7.8	36.4	0.12	22.7
Story+ArXiv	**41.0**	**9.2**	**38.6**	**0.15**	**64.1**
StoryTopK	38.2	8.5	36.0	0.13	22.7
StoryTopK+ArXiv	**41.4**	**10.6**	**38.8**	**0.14**	**64.1**

Table 7: *Long Science Daily*: Training with *ArXiv*. We can observe sizeable and consistent improvements.

I.e., the model might optimize for generating words that overlap between the source paper and the target, e.g., by copying scientific terms. Thus, high ROUGE scores do not necessarily imply a good story (Fan, Lewis, and Dauphin 2018), and we will proceed with both models as baselines.

Moreover, sampling from the top-k candidates ($k = 10$) has been shown useful for story generation (Fan, Lewis, and Dauphin 2018), and we try it here as well. We label such experiments by appending *TopK*; Table 6 shows that top-k decoding yields sizable improvements for ROUGE-2.

Data Augmentation with ArXiv

As summarization in *Arxiv* to generate abstracts and our ASJ task share similar domains for their sources, namely scientific papers, it is natural to try to augment our *Science Daily* dataset with the *ArXiv* dataset. We do so using specially designed tags: (*i*) we prepend the tag **[begin-paper]** and we append the tags **[end-paper] [begin-press]** for *Science Daily*.

For *ArXiv* examples, we do the same, but we replace **press** with **abstract**. (*ii*) We also append the target with **[end-press]** or **[end-abstract]**, respectively. These tags indicate the source domain (*ArXiv* or *Science Daily*) and the target domain (*abstract* or *press release*). To ensure equal balance between the two datasets, we take 40,000 examples from their training sets, 5,000 from their test, and 5,000 from their dev set, for a final train/dev/test split of 80,000/10,000/10,000.

Model	1	2	L	NLI	PR
LEAD	39.6	10.1	16.1	N/A	N/A
Fconv	**39.2**	**9.5**	**36.9**	0.23	38.0
Fconv+Parts	32.8	7.8	31.2	**0.25**	**77.1**
FconvTopK	**39.2**	**10.8**	**37.0**	0.23	38.0
FconvTopK+Parts	31.1	9.0	29.6	**0.27**	**77.1**
Story	38.9	7.8	36.4	0.12	22.7
Story+Parts	**42.8**	**10.6**	**40.2**	**0.17**	**73.8**
StoryTopK	38.2	8.5	36.0	0.14	22.7
StoryTopK+Parts	**41.4**	**11.0**	**39.1**	**0.16**	**73.8**

Table 8: Training in parts yields improvements: sizable for Prompt Ranking, but partial for ROUGE 1/2/L.

We hypothesize that the encoder layers and the decoder attention mechanism will focus on these tags while processing the source and while generating the output, respectively. Table 7 shows that using *ArXiv* yields sizable improvements both for ROUGE 1/2/L and for our Prompt Ranking score. Note that we did not use ArXiv source-target paris for generation and calculation of ROUGE, NLI and PR. We only used the originally designated *Science Daily* test source-target pairs (even though the model has been trained using *ArXiv* source-target pairs too). We believe that the ability to co-train with other datasets offers important flexibility in our experimental setup.

Data Augmentation with Targets in Parts

In order to increase the total number of training examples and to focus the summarization on particular parts, we experimented with augmenting *Science Daily* with partitioned targets as follows:

1. For each source–target pair in *Science Daily*, we preserve the source **body**, and we divide the target into three equal parts: **part-1**, **part-2**, and **part-3**.
2. We construct the source-target pairs as follows: for all bodies **body**, for indices **i** equal to 2 or 3, the source is
   [begin-body]body[end-body][begin-part-(i-1)]
   part-(i-1)[end-part-(i-1)][begin-part-i]
   and for **i** equal to 1, the source is
   [begin-body]body[end-body][begin-part-i],
   where the corresponding target to the source is **part-i [end-part-i]**.
3. At inference, we generate the parts **part-i** autoregressively from **part-1** to **part-3**.

Instead of training the model to generate the full press release, we train it to generate specific sections only. Thus, we increase the data split threefold, which yields a train/dev/test split of size 120,741/15,087/15,087. Recently, similar divide-and-conquer approaches have improved the state of the art on scientific summarization (Gidiotis and Tsoumakas 2020). Table 8 shows results when using this partition.

Note that to compute ROUGE, NLI, and PR, we generate each designated part, concatenate the generations, and then we calculate the scores. We can see in Table 8 sizable improvements over the baselines for the in-parts training method, both for ROUGE 1/2/L and for PR, which confirms that this data augmentation scheme is indeed helpful.

NLI Scores. We computed the NLI scores using RoBERTa-large (Liu et al. 2019), fine-tuned for natural language inference on the MNLI dataset. We noted an increase in the scores when training with *ArXiv* (+ArXiv) compared to the baseline models. Although the *TopK* strategy also improves the scores for the baseline models, the *ArXiv* (+ArXiv) models performed better on their own. Training parts (+Parts) also yielded a higher score for both the *Story* and the *Fconv* models. However, we should note that there is a significant difference between the scores of the *Story* and of the *Fconv* models due to the more extractive nature of the *Fconv* model, which ultimately yields higher NLI scores.

PR Scores. For *Fconv* models, training with *ArXiv* (+ArXiv) and in parts (+Parts) outperforms the baseline Fconv/ FconvTopK significantly by 39.8 and 39.1 absolute percentage points, respectively. For *Story* models, both training with *ArXiv* (+ArXiv) and in parts (+Parts) outperforms the *Story* baseline significantly by 42.6 and 51.1 percentage points absolute, respectively. Tables 6, 7 and 8 further show that, in general, Prompt Ranking is in agreement with the ROUGE scores, but it is more sensitive to training using data augmentation with *ArXiv* or using parts.

Discussion

Short Science Daily For *short Science Daily*, we observe that the results are particularly coherent and fluent, given the short sources. For example, in Table 1, in contrast to the gold summary, the source does not mention *male* or *socialist countries*. Yet, *SciBertSumAbs* gets it correctly even though *SciBERT* and *Science Daily* are biased towards biomedical topics, which is not the case here.

Long Science Daily Table 9 shows a summary from *long Science Daily*, which is fluent and logical, and focuses on specific information relevant to the source paper. It demonstrates structured and concise writing with sections that are both relevant and conceptually accurate. For example, it mentions that *x-ray crystallography* was used to determine the three-dimensional structure of the proteins. The target article says that this was done by the study's authors in previous work, but this technique is not mentioned in the source, which is all the model sees. This demonstrates a very important and promising phenomenon: similarly to (Tshitoya, Dagdelen, and Weston 2019), where unsupervised word embeddings captured information about materials, the model learns representations of key concepts such as *x-ray crystallography*, and applies this knowledge at generation time. In contrast, the baseline *Fconv* generates fragments like *in the new study , the scientists used a technique called " dna, " the researchers say*, which misreads the meaning of DNA.

Generated: . . . histone proteins are the building blocks of proteins , and they are involved in a variety of biological functions , marino said . histones are the amino acids that make up the cell 's dna . when dna is copied , the proteins are copied to form proteins , which are called histones . histones have been studied for more than a decade , but until now it has been difficult to determine how these histone proteins assemble and how histones are assembled in the cell . . . . the researchers used a technique called **x-ray crystallography** , which allows scientists to determine the atomic structure of proteins . . . other co-authors of the paper are postdoctoral researcher zachary sandman , a former ohio state graduate student in biochemistry and molecular biology and a member of the marino lab.
Target: the colorado state university researcher studies how these hardy microbes – which constitute one of three surviving domains of life – express their genes , produce their energy , and thrive in hot , lightless environments . . . . in 1997 , luger and colleagues first reported the exact structure of eukaryotic nucleosomes via x-ray crystallography . . .
Source snippets: . . . small basic proteins present in most archaea share a common ancestor with the eukaryotic core histones . we report the crystal structure of an archaeal histone-dna complex . . . our data establish that most features of eukaryotic dna compaction into nucleosomes are conserved in archaeal histone-based chromatin . . . .

Table 9: Summary from *long Science Daily.* Shown are some snippets (generated, gold, and original) when using the *Story* model with top-k sampling and data augmentation using *ArXiv* (*StoryTopK+Arxiv*).

Overall, the advantages of our transfer learning experiments include (*i*) topical and factual generation, (*ii*) memorization and utilization of scientific concepts beyond the current source, and (*iii*) clear semantic and syntactic structure.

Limitations We found that in some cases, the output of *Fconv+ArXiv*, *Story+ArXiv*, and *Fconv+Parts* is repetitive, unable to match named entities (e.g., *Zachary Sandman* in Table 9 is not a real person), diverging from the topic, and limited in the sense that it only has access to a single scientific paper. Moreover, the *Story* model sometimes overfits to a set of concepts, and then creates a story around those concepts rather than based on the input sequence. For example, a source paper about the structural similarities of DNA in archaea and eukaryotes might not be accurately summarized by story-based experiments: they might elaborate on related topics, even though still focusing on DNA.

Human Evaluation on IEEE Articles Using our *SciBertSumAbs* model on *short Science Daily*, we generated summaries for five IEEE articles, randomly selected by an IEEE expert. The summaries were manually evaluated by that expert using the following criteria, which he selected independently from us:

#	Rel.	Read.	Compr.	As-is	Cons.
1	Y	Y	Y	Y	ML
2	Y	P	Y	N	NMT
3	Y	P	Y	N	NMT
4	Y	P	Y	N	NMT
5	Y	P	Y	N	NFT

Table 10: Manual expert analysis of the utility of models trained with *SciBertSumAbs* on *short Science Daily*. See the text for a definition of the criteria and their abbreviations. Legend: Y=Yes, N=No, P=Probably, ML=Most Likely, NMT=Needs Minor Tweaks, NFT=Needs Few Tweaks.

- (Rel.) *Is the generated summary relevant to the article in context?*
- (Read.) *Is the generated summary readable by the market of interest?*
- (Compr.) *Can the summary be comprehended by the market of interest?*
- (As-is) *Is the summary acceptable As-Is?*
- (Cons.) *Can the summary be consumed by the market of interest as is (leads to effort level required from IEEE to polish the summaries before they are market-ready)?*

We present the evaluation results in Table 10. Overall, our summaries appear deployable after some polishing by IEEE experts. Note that, in general, human evaluation is hard, as it requires a domain expert, as opposed to evaluating topics that are common sense (Chang et al. 2009). Evaluating even a small number of articles properly is a difficult task.

In our setting, it took more than an hour per paper by an expert. Naturally, such settings are very difficult to scale, and they take up a sizable portion of the expert's time and effort. The challenge becomes even more acute when we recognize that outsourcing such evaluations would be harder than for domains closer to a layman.

Conclusion and Future Work

We have proposed to study *Automating Science Journalism* (ASJ), which is the process of producing a layman's terms summary of a research article, as a new benchmark for long neural abstractive summarization and story generation. We further created a specialized dataset that contains scientific papers and their *Science Daily* press releases: short and long versions. We demonstrated numerous *sequence to sequence* (seq2seq) applications using Science Daily with the aim of facilitating further research on language generation, which requires extreme paraphrasing and coping with long research articles. We further improved the quality of the press releases using co-training with scientific abstracts of sources or partitioned press releases. Finally, we further confirmed our results using quantitative and qualitative evaluation, including manual evaluation and analysis by a domain expert. The results suggested that our model is potentially usable in practice, possibly after post-editing.

There are many exciting directions that we plan to explore in future work. One possibility is to use more efficient linear Transformers that can model long sequences better. Another option is to encourage factuality more explicitly during training and inference, e.g., by combining variants of the NLI score and Prompt Ranking measures with the maximum likelihood objective at training time, and with the generation method at inference time. More explicit text simplification and style transfer methods could also improve the performace. Finally, we could apply our models directly to many practical problems, which would truly test generalization, and could serve as the basis of fruitful applications of automating science journalism.

Acknowledgements

We would like to express deep gratitude to Mićo Tatalović and to Dan Hogan for their help with the Science Daily dataset, Lavanya Sayam for her help with evaluation of the summaries of IEEE papers, Nicholas Gibbons, Henning Schoenenberger, Christian Chiarcos, Niko Schenk and Chris Watkins for fruitful discussions, and Daniel Dardani and Matthew Fucci for their advice.

The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center (Reuther et al. 2018) for providing HPC and consultation resources that have contributed to the research results reported within this paper/report.

This research was sponsored by the United States Air Force Research Laboratory and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and the conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and to distribute reprints for Government purposes notwithstanding any copyright notation herein. This work was also supported in part by the Army Research Office under Cooperative Agreement W911NF-18-2- 0048.

Ethics Statement

On the positive side, automating science journalism could be helpful both to journalists, who would be able use such tools to create press releases, and also to readers, who could learn about scientific discoveries in layman's terms. We further believe that research on Science Daily and similar corpora of scientific text and their summaries in layman's terms could benefit the overall field of text summarization by offering an interesting and challenging reformulation of the general problem.

On the negative side, models trying to solve the problem could produce misleading summaries, which could result in false reporting. Thus, such models should be used with care as they would typically need some polishing and double-checking by experts.

References

Abu-Jbara, A.; and Radev, D. 2011. Coherent Citation-Based Summarization of Scientific Paper. In *ACL*.

Aharoni, R.; Johnson, M.; and Firat, O. 2019. Massively Multilingual Neural Machine Translation. In *NAACL-HLT*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP-IJCNLP*.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150* .

Cachola, I.; Lo, K.; Cohan, A.; and Weld, D. S. 2020. TLDR: Extreme Summarization of Scientific Documents. *arXiv preprint arXiv:2004.15011* .

Celikyilmaz, A.; Bosselut, A.; He, X.; and Choi, Y. 2018. Deep Communicating Agents for Abstractive Summarization. In *NAACL-HLT*.

Chandrasekaran, M. K.; Feigenblat, G.; Hovy, E.; Ravichander, A.; Shmueli-Scheuer, M.; and de Waard, A. 2020. Overview and Insights from the Shared Tasks at Scholarly Document Processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*.

Chandrasekaran, M. K.; Yasunaga, M.; Radev, D.; Freitag, D.; and Kan, M.-Y. 2019. Overview and Results: CL-SciSumm Shared Task 2019. In *BIRNDL 2019*.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-graber, J. L.; and Blei, D. M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *NeurIPS*.

Chen, Y.-C.; and Bansal, M. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *ACL*.

Cohan, A.; Dernoncourt, F.; Kim, D. S.; Bui, T.; Kim, S.; Chang, W.; and Goharian, N. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *NAACL-HLT*.

Collins, E.; Augenstein, I.; and Riedel, S. 2017. A Supervised Approach to Extractive Summarization of Scientific Papers. In *CoNLL*.

Dangovski, R.; Jing, L.; Nakov, P.; Tatalović, M.; and Soljačić, M. 2019. Rotational Unit of Memory: a Novel Representation Unit for RNNs with Scalable Applications. *TACL* 7: 121–138.

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language Modeling with Gated Convolutional Networks. In *ICML*.

Delvin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Divoli, A.; Nakov, P.; and Hearst, M. 2012. Do Peers See More in a Paper Than Its Authors? *Adv. Bioinformatics* 2012: 750214:1–750214:15.

Edunov, S.; Baevski, A.; and Auli, M. 2019. Pre-trained Language Model Representations for Language Generation. In *NAACL-HLT*.

Elkiss, A.; Shen, S.; Fader, A.; Erkan, G.; States, D.; and Radev, D. 2008. Blind Men and Elephants: What Do Citation Summaries Tell Us About a Research Article? *J. Am. Soc. Inf. Sci. Technol.* 59: 51–62.

Falke, T.; Ribeiro, L. F. R.; Utama, P. A.; Dagan, I.; and Gurevych, I. 2019. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *ACL*.

Fan, A.; Grangier, D.; and Auli, M. 2018. Controllable Abstractive Summarization. *NMT at ACL* .

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In *ACL*.

Fan, A.; Lewis, M.; and Dauphin, Y. 2019. Strategies for Structuring Story Generation. In *ACL*.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional Sequence to Sequence Learning. In *ICML*.

Gidiotis, A.; and Tsoumakas, G. 2020. A Divide-and-Conquer Approach to the Summarization of Academic Articles. *IEEE/ACM TASLP* 28: 3029–3050.

Grusky, M.; Naaman, M.; and Artzi, Y. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *NAACL-HLT*.

Horne, B. D.; Dron, W.; Khedr, S.; and Adali, S. 2018. Assessing the News Landscape: A Multi-module Toolkit for Evaluating the Credibility of News. In *WWW*.

Jaidka, K.; Chandrasekaran, M. K.; Jain, D.; and Kan, M.-Y. 2017. The CL-SciSumm Shared Task 2017: Results and Key Insights. In *BIRNDL*.

Jaidka, K.; Chandrasekaran, M. K.; Rustagi, S.; and Kan, M.-Y. 2016. Overview of the CL-SciSumm 2016 Shared Task. In *BIRNDL*.

Jaidka, K.; Yasunaga, M.; Chandrasekaran, M. K.; Radev, D.; and Kan, M.-Y. 2018. The CL-SciSumm Shared Task 2018: Results and Key Insights. In *BIRNDL at SIGIR*.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858* .

Kupiec, J.; Pedersen, J.; and Chen, F. 1995. A Trainable Document Summarizer. In *SIGIR*.

Lample, G.; and Conneau, A. 2019. Cross-lingual Language Model Pretraining. *arXiv preprint arXiv:1901.07291* .

Lev, G.; Shmueli-Scheuer, M.; Herzig, J.; Jerbi, A.; and Konopnicki, D. 2019. TalkSumm: A Dataset and Scalable Annotation Method for Scientific Paper Summarization Based on Conference Talks. In *ACL*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

Lin, C.-Y.; and Hovy, E. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *HLT-NAACL*.

Lin, H.; and Ng, V. 2019. Abstractive Summarization: A Survey of the State of the Art. In *AAAI*.

Liu, P. J.; Saleh, M.; Pot, E.; Goodrich, B.; Sepassi, R.; Kaiser, L.; and Shazeer, N. 2018. Generating Wikipedia by Summarizing Long Sequences. In *ICLR*.

Liu, Y.; and Lapata, M. 2019. Text Summarization with Pretrained Encoders. In *EMNLP-IJCNLP*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* .

Luu, K.; Koncel-Kedziorski, R.; Lo, K.; Cachola, I.; and Smith, N. A. 2020. Citation Text Generation. *arXiv preprint arXiv:2002.00317* .

Mei, Q.; and Zhai, C. 2008. Generating Impact-based Summaries for Scientific Literature. In *COLING*.

Mohammad, S.; Dorr, B.; Egan, M.; Hassan, A.; Muthukrishan, P.; Qazvinian, V.; and Radev, D. 2009. Using Citations to Generate Surveys of Scientific Paradigms. In *ACL-HLT*.

Nakov, P.; Schwartz, A.; and Hearst, M. 2004. Citances: Citation Sentences for Semantic Analysis of Bioscience Text. In *SIGIR workshop: Search and Discovery in Bioinformatics*.

Nallapati, R.; Zhai, F.; and Zhou, B. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *AAAI*.

Nallapati, R.; Zhou, B.; dos Santos, C. N.; Gülçehre, Ç.; and Xiang, B. 2016. Abstractive Text Summarization Using Sequence-to-sequence RNNs and Beyond. In *CoNLL*.

Nanba, H.; Kando, N.; and Okumura, M. 2000. Classification of Research Papers Using Citation Links and Citation Types: Towards Automatic Review Article Generation. In *ASIS SIG/CR*.

Nikolov, N.; Pfeiffer, M.; and Hahnloser, R. 2018. Data-driven Summarization of Scientific Articles. In *LREC*.

Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; and Ng, N. 2019. FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling. In *NAACL-HLT*.

Paulus, R.; Xiong, C.; and Socher, R. 2018. A Deep Reinforced Model for Abstractive Summarization. In *ICLR*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*.

Qazvinian, V.; and Radev, D. R. 2008. A Supervised Approach to Extractive Summarization of Scientific Papers. In *COLING*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21(140): 1–67.

Reuther, A.; Kepner, J.; Byun, C.; Samsi, S.; Arcand, W.; Bestor, D.; Bergeron, B.; Gadepally, V.; Houle, M.; Hubbell, M.; et al. 2018. Interactive Supercomputing on 40,000 Cores for Machine Learning and Data Analysis. In *IEEE HPEC*.

Ruder, S. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Rush, A. M.; Chopra, S.; and Weston, J. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP*.

Saggion, H.; AbuRa'ed, A.; and Ronzano, F. 2016. Trainable Citation-enhanced Summarization of Scientific Articles. In *BIRNDL*.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the Point: Summarization with Pointer-Generator Networks. In *ACL*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Sharma, E.; Li, C.; and Wang, L. 2019. BIGPATENT: A Large-Scale Datasetfor Abstractive and Coherent Summarization. In *ACL*.

Subramanian, S.; Li, R.; Pilault, J.; and Pal, C. 2020. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. In *EMNLP*.

Teufel, S.; and Moens, M. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Comput. Linguist.* 28(4): 409–445.

Tshitoya, V.; Dagdelen, J.; and Weston, L. 2019. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* 571: 95–98.

Vadapalli, R.; Syed, B.; Prabhu, N.; Vasan Srinivasan, B.; and Varma, V. 2018. When Science Journalism Meets Artificial Intelligence: An Interactive Demonstration. In *EMNLP*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*.

Wang, Q.; Zeng, Q.; Huang, L.; Knight, K.; Ji, H.; and Rajani, N. F. 2020. ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. In *INLG*.

Yasunaga, M.; Kasai, J.; Zhang, R.; Fabbri, A. R.; Li, I.; Friedman, D.; and Radev, D. R. 2019. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. In *AAAI*.

Yuan, W.; Liu, P.; and Neubig, G. 2021. Can We Automate Scientific Reviewing? *arXiv preprint arXiv:2102.00176* .

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending Against Neural Fake News. In *NeurIPS*.