SciNews: From Scholarly Complexities to Public Narratives – A Dataset for Scientific News Report Generation

Dongqi Pu, Yifan Wang, Jia Loy, Vera Demberg
Department of Computer Science
Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, Germany
{dongqipu, yifwang, jialoy, vera}@lst.uni-saarland.de

Abstract

Scientific news reports serve as a bridge, adeptly translating complex research articles into reports that resonate with the broader public. The automated generation of such narratives enhances the accessibility of scholarly insights. In this paper, we present a new corpus to facilitate this paradigm development. Our corpus comprises a parallel compilation of academic publications and their corresponding scientific news reports across nine disciplines. To demonstrate the utility and reliability of our dataset, we conduct an extensive analysis, highlighting the divergences in readability and brevity between scientific news narratives and academic manuscripts. We benchmark our dataset employing state-of-the-art text generation models. The evaluation process involves both automatic and human evaluation, which lays the groundwork for future explorations into the automated generation of scientific news reports. The dataset and code related to this work are available at https://dongqi.me/projects/SciNews.

Keywords: Scientific News Report Generation, Natural Language Generation, Text Summarization

1. Introduction

Why Studying Scientific News Report Generation is Valuable: Scientific publications capture the latest advancements and discoveries in the realm of science, but often necessitate a significant level of academic background, posing obstacles for the general public without specialized knowledge (Saikh et al., 2020; Wright and Augenstein, 2021; August et al., 2022; Wright et al., 2022a). In a bid to bridge this knowledge gap, science journalists are endeavoring to translate intricate scientific nuances and breakthroughs into concise and accessible language (Polman and Hope, 2014; Majetic and Pellegrino, 2014; Li et al., 2017; Hoque et al., 2022). This initiative seeks to promote a profound engagement between the public audience and scientific literature (Ravenscroft et al., 2018; Vadapalli et al., 2018; August et al., 2020). Figure 1 illustrates how scientific news reports/narratives may help to increase the accessibility of scientific discoveries by using simplified language, examples, and explanations for technical terms (e.g., "cybersickness"→"feeling nauseous or disorientated"). Regrettably, the pursuit of automated generation of scientific news reports faces challenges due to the insufficient availability of parallel corpora. Thus, this paper proposes (i) a new task, Automated Scientific News Report Generation (SNG), and (ii) a novel dataset, SciNews, designed for this task.

Similarities and Differences with Text Summarization and Text Simplification: Text summarization emphasizes the reduction of textual volume whilst preserving main information, without altering linguistic complexity (Liu et al., 2023b; Pu et al., 2023; Hosking et al., 2023; Cho et al., 2022; Goyal et al., 2022; Pu et al., 2022; Lai et al., 2022; See et al., 2017), while text simplification focuses on employing simplified lexicon and syntax to enhance readability (Pu and Demberg, 2023; Nisioi et al., 2017; Sulem et al., 2018; Blinova et al., 2023; Laban et al., 2023; Garimella et al., 2022; Cripwell et al., 2022). The SNG intertwines these processes, requiring both the simplification of complex concepts to more comprehensible forms and the extraction of pivotal insights from source materials (Alambo et al., 2020; August et al., 2020; Dong et al., 2021; August et al., 2022; Tan et al., 2023).

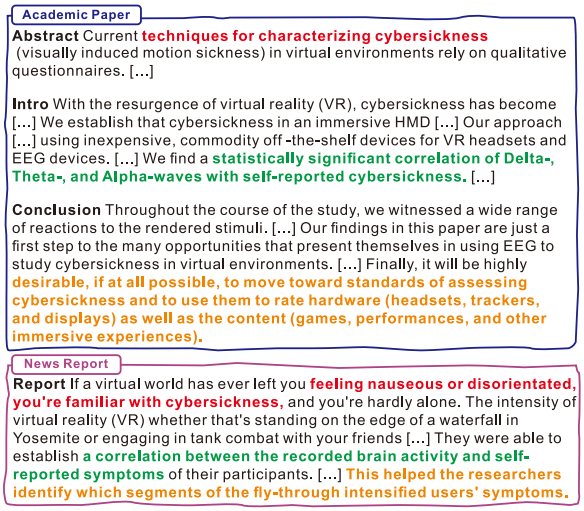

Figure 1: An example of an academic paper paired with its news report.

Unlike previous efforts mainly focusing on the

LREC-COLING 2024, pages 14429–14444

biomedical field and generating lay summaries of academic paper abstracts (Guo et al., 2021; Goldsack et al., 2022), our work across a broader range of scientific disciplines, aims for more comprehensive narrative generation. In addition, the SNG task poses a heightened challenge for text generation models, as it necessitates both a deep understanding of academic discourses and the capability to articulate coherent, long-form articles.

Our Contributions: Given the insufficient availability of benchmark datasets to gauge the potential of text generation models in the SNG task, we present SciNews[1], a novel multidisciplinary English dataset constructed for automated scientific news report generation. Our dataset leverages academic articles as source inputs, aligned with human-authored scientific news reports as target outputs. Additionally, we conduct extensive evaluations with state-of-the-art (SOTA) Natural Language Generation (NLG) models on SciNews, supplemented with assessments from human evaluators to analyze different perspectives of model outputs. Our findings suggest that the current leading models still struggle with hallucination and factual error problems. Furthermore, compared to human abilities in style-adaptive writing, SOTA NLG models exhibit an inferior capacity for converting complicated texts into understandable narratives. To summarize, our contributions are as follows:

- We introduce a task focused on the automated generation of scientific news reports, supported by the SciNews dataset, which contains 41,872 samples.
- We undertake both quantitative and qualitative analyses of the SciNews dataset, providing insights into variations in linguistic structure and readability between source articles and target reports.
- We evaluate state-of-the-art NLG models on our dataset, finding that the abstractive text generation models surpass the extractive ones on this task.
- We offer an error analysis, grounded in human evaluations, identifying primary issues in machine-generated scientific news reports.

2. Related Work

In tasks of news-related generation, some investigations have been conducted into the automated generation of general news articles (Sigita et al., 2013; Nesterenko, 2016; Mosallanezhad et al., 2020; Horvitz et al., 2020; Shu et al., 2021), headlines (Gusev, 2019; Bukhtiyarov and Gusev, 2020; Liu et al., 2020; Ao et al., 2021; Panthaplackel et al., 2022; Cai et al., 2023), comments (Yang et al., 2019), and summaries (Nallapati et al., 2016). Similarly, science-related generation efforts have focused on producing academic summaries (Cohan et al., 2018; Cachola et al., 2020; Lu et al., 2020), contributions (Hayashi et al., 2023), related work (Hu and Wan, 2014), definitions (August et al., 2022), paraphrases (Dong et al., 2021), and claims (Wright et al., 2022b; Hayashi et al., 2023; Tan et al., 2023). However, thus far, attempts to study the automated generation of scientific news reports from academic papers across various fields have been less than comprehensive, with a concurrent dataset containing just over 2,400 samples that align with source research papers (Cardenas et al., 2023). In addition, the dataset from Cardenas et al. (2023) primarily focuses on generating press releases from news articles. In contrast, our initiative aims at using academic papers as a foundation to produce news articles. This complementary dataset underscores our shared goal of enhancing public engagement with science, albeit through different lenses of scientific communication. We next explore related areas of scientific lay summarization and text simplification to contextualize our approach within the broader landscape of making science accessible.

2.1. Scientific Lay Summarization

Scientific Lay Summarization (SLS) strives to produce accessible summaries that enable researchers in the field to quickly grasp the main content of current papers. For example, at EMNLP 2020, Chandrasekaran et al. (2020a) released a small-scale corpus and organized a shared task. However, this small corpus has proven challenging for training neural NLG models(Chandrasekaran et al., 2020b). To alleviate this problem, subsequent studies (Guo et al., 2021, 2022) introduced two larger-scale datasets, demonstrating the effectiveness of neural architectures in SLS. Further contributions (Goldsack et al., 2022, 2023) have added two expansive lay summary datasets focused on bio-medicine, enriching the domain. Additionally, the introduction of RSTformer (Pu et al., 2023) explored the role of discourse structure in improving SLS. It is essential to note that while current efforts predominantly convert academic paper abstracts into lay summaries, generating scientific news articles requires adopting a narrative-driven approach. This storytelling style poses challenges related to text length and content, typically including aspects such as research background, findings, and impacts.

2.2. Scientific Text Simplification

Scientific Text Simplification (STS) seeks to make complicated texts more readable through text style transfer. Previous attempts, such as the one by

[1]SciNews can only be used for academic purposes.

Coster and Kauchak (2011), introduced a sentence-level parallel simplification dataset sourced from Wikipedia. Building upon this, Kim et al. (2016) established an additional corpus focusing on the lexical simplification of scientific articles, and Grabar and Cardon (2018) curated a simplification corpus tailored to Medical French. Laban et al. (2021) devised a reinforcement learning-based system for simplifying multi-sentence structures, while Devaraj et al. (2021) applied the Transformer model for paragraph-level simplification of medical texts. Most recently, Ermakova et al. (2022, 2023) initiated a scientific simplification task at CLEF2022/3. Furthermore, Blinova et al. (2023) proposed SIMSUM, a strategy for document-level text simplification via simultaneous summarization. However, the majority of the current studies center on simplification at the lexical, sentence, paragraph/short-document levels, leaving substantial unexplored room in long-document simplification (Devaraj et al., 2021; Garimella et al., 2022; Laban et al., 2023; Cripwell et al., 2023a; Fatima and Strube, 2023). In contrast to STS, which simplifies language while preserving academic integrity and depth, ensuring no critical information is lost (Cripwell et al., 2023b), scientific news narratives, although precise, may alter the depth of discussion and incorporate additional explanatory information for enhanced clarity and reader engagement.

3. The SciNews Dataset

3.1. Task Formulation

The task of SNG can be formalized as follows: Given a scientific paper x_i and its corresponding news article y_i, we have a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and $(x_i, y_i) \in D$. Our objective is to train NLG models N, such that the model learns a conditional probability distribution $P(Y|X)$, where $Y = \{y_1, y_2, \ldots, y_n\}$ and $X = \{x_1, x_2, \ldots, x_n\}$. When a new $x_j \notin X$ is fed into N, it should generate the corresponding y_j.

3.2. Data Acquisition

The SciNews is derived from the Science X platform, an important open-access hub featuring news on science, technology, and medical research. It is noteworthy that the news articles on this platform are contributed by authors or their affiliated institutions and are carefully revised by skilled editors to ensure narrative consistency and mitigate potential ethical issues. In compliance with Science X's terms, which permit data collection for academic research without prior written consent, we collect data such as news titles, news content, associated URLs, relevant DOIs, and domain tags for our study.

Our work focuses on the generation of one-to-one news reports, thus we exclude samples derived from multiple research papers. Leveraging the DOI information, we identify and select articles that are open access and published under the "Creative-Commons" CC-BY-4.0 license[2]. For data extraction, we employ web scraping tools Selenium and BeautifulSoup, which facilitate the retrieval of article content and citation details, ensuring compliance with copyright licenses.

3.3. Data Cleaning

We follow the steps of Cohan et al. (2018); Cachola et al. (2020) to clean the acquired data. In the first phase, we apply PySBD rule-based parser (Sadvilkar and Neumann, 2020) and spaCy to remove line breaks, emoticons, and web links from news articles and scientific papers. Next, we delete news reports (and their corresponding papers) associated with multiple disciplines, identified by their domain tags. For academic papers, we limit the extraction to the text between the abstract and the references section. Finally, we exclude documents exceeding 30,000 words, likely dissertations or monographs, and those under 2000 words, typically tutorials or research proposals[3].

3.4. Quality Control

Documents from the Science X platform are high-quality, sourced from reputable academic origins, and authored by both original researchers and professional journalists. Our dataset creation process bypasses the need for further human annotations but incorporates a dual-phase quality control method, including both automated and human assessments.

Automated Quality Control: Adapting methods from Mao et al. (2022), we calculate pairwise BERT similarity score (Zhang* et al., 2020) between sentences in the news report and their corresponding academic paper. We remove pairs where over half of the news report sentences have BERT similarity scores below 0.5, indicating significant dissimilarity. This procedure is also applied to named entities within the texts, excluding pairs failing to meet this benchmark. Through this vetting process, we remove 612 pairs from our initial set of 42,484 samples.

Human Quality Control: Inspired by Sun et al. (2021), we randomly select 100 article pairs for a manual quality check to evaluate their overall simplicity without sacrificing quality. We utilize a binary judgment to determine if the news narrative

[2] https://creativecommons.org/licenses/by/4.0/

[3] We use spaCy to count the number of words.

is simpler than the academic paper while maintaining its quality. We recruit two evaluators, each having a Master's degree in either Computer Science or Computational Linguistics. Among the 100 samples, only one sample receives divergent assessments – being labeled as 'accepted' by one evaluator and 'rejected' by another. The reason given for being 'rejected' is that the scientific news report is longer and less concise compared to other test samples, but there are no complaints about other factors, such as simplicity, faithfulness, etc. This sample is retained after a second review confirming its validity. No sample is unanimously rated as 'rejected'.

3.5. Data Splits

After quality control, our dataset comprises 41,872 samples spanning nine scientific domains, as illustrated in Figure 2 on topic distribution. We divide the data into training (80%), validation (10%), and test set (10%) by randomly sampling from the entire dataset while keeping the proportion of papers from the different domains constant. The detailed distribution of samples across these subsets is provided in Table 1. All of our experiments described in Sections 5 and 6 use this split.

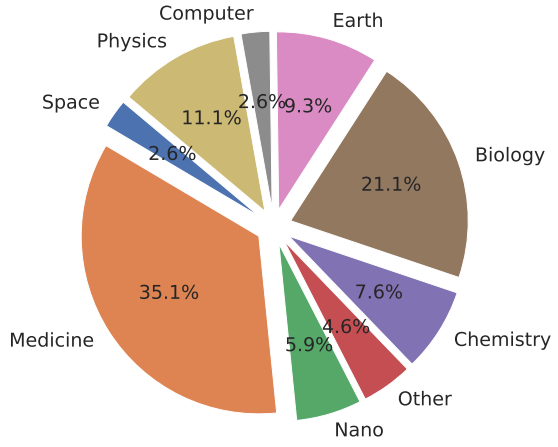

Figure 2: Topic distribution of our dataset

4. Dataset Analysis

4.1. Dataset Comparison

Table 2 presents a comparison between our SciNews dataset and datasets for scientific lay summarization and scientific text simplification (as discussed in Section 2). Two document-level corpora have a similar size to SciNews (41,872 samples): CSJ has 50,132 samples and PLOS contains 27,525 samples. SciNews stands out due to its multidisciplinary coverage and its provision of category labels for each field. Additionally, the SciNews scientific news reports are longer (averaging 695 tokens), in comparison to PLOS summaries (176 tokens on average), and the simplified texts from CSJ (average length 361 tokens). It is also important to highlight that CSJ derives its data from Wikipedia for multidisciplinary data (without domain labels), in contrast to scholarly articles. Furthermore, CSJ is a paragraph/short-document level simplification dataset, setting it apart from SciNews.

Property	Value
# Training Set	33497
# Validation Set	4187
# Test Set	4188
Avg. # Tokens (Papers)	7760.90
Avg. # Tokens (News)	694.80
Avg. # Sents. (Papers)	290.52
Avg. # Sents. (News)	25.17
Compression Ratio	12.71
Coverage	0.74
Density	0.94
1-gram Novelty	0.52
2-gram Novelty	0.91
3-gram Novelty	0.98
4-gram Novelty	0.99

Table 1: Dataset statistics

4.2. Dataset Statistics

We apply metrics from prior studies (Grusky et al., 2018; Bommasani and Cardie, 2020; Hu et al., 2023) for corpus-level analysis. As Table 1 shows, on average, scientific papers consist of 7760.90 tokens and 290.52 sentences, whereas news reports contain an average of 694.80 tokens and 25.17 sentences; the *Compression Ratio* in our dataset is thus 12.71. The *Coverage* metric measures the percentage of tokens in the news report that originate from the original article. A value of 0.74 in *Coverage* indicates substantial inclusion of core information or content from the source in the news articles. The *Density* score assesses the extent to which news reports can be characterized as a set of extractive fragments. The value of 0.94 implies that academic news reports contain only short contiguous text fragments extracted from source papers, indicating a highly abstractive rewriting process.

To measure the textual overlap between news reports and the original papers, we use the methodology from Narayan et al. (2018) and Sharma et al. (2019) to calculate the proportion of 1/2/3/4-grams in news reports that are not present in the original reference texts. The high n-grams novelty scores indicate significant reformation of the material by human authors, suggesting that the news narratives are not just simplified versions of the source texts but involve the creation of novel n-grams through

Dataset	Task	Language	Data Scope	Data Source	Scale	Input Level	Output Level	Multi-disciplinary?
LaySumm (Chandrasekaran et al., 2020c)	SLS	English	Archaeology, Hepatology, etc.	Research Papers	572	Document	Paragraph	✓
CDSR (Guo et al., 2021)	SLS	English	Healthcare	Research Papers	7805	Document	Paragraph	✗
CELLS (Guo et al., 2022)	SLS	English	Biomedicine	Research Papers	47157	Sentence	Sentence	✗
eLife (Goldsack et al., 2022)	SLS	English	Biomedicine	Research Papers	4828	Document	Paragraph	✗
PLOS (Goldsack et al., 2022)	SLS	English	Biomedicine	Research Papers	27525	Document	Paragraph	✗
SimpleScience (Kim et al., 2016)	STS	English	Biomedicine	Research Papers	293	Sentence	Vocabulary	✗
CLEAR (Grabar and Cardon, 2018)	STS	French	Biomedicine	Research Papers	663	Sentence	Sentence	✗
PLS (Devaraj et al., 2021)	STS	English	Medicine	Research Papers	4459	Paragraph	Paragraph	✗
SimpleText (Ermakova et al., 2022, 2023)	STS	English	Medicine & Computer Science	Research Papers	648	Sentence	Sentence	✓
CSJ (Fatima and Strube, 2023)	STS	English & German	Astronomy, Biology, etc.	Wikipedia	50132	Document	Paragraph	✓
SciNews (ours)	SNG	English	Science & Technology & Medicine	Research Papers	41872	**Document**	**Document**	✓

Table 2: Dataset comparison

combination, rearrangement, or interpretation of information from the source scientific papers.

4.3. Papers vs. News

Academic papers typically employ a first-person perspective, in contrast to the third-person narrative found in scientific news articles (as shown in Figure 1). Beyond the differences in writing tone, we analyze the disparities between these mediums at the lexical (vocabulary), syntactic (sentence)[4], discourse (intersentential)[5] and readability (document)[6] levels. As shown in Table 3, we find that news articles exhibit a higher type-token ratio, indicating greater lexical diversity. Both mediums maintain substantial lexical density, but the news articles contain fewer difficult words.

News articles also use simpler syntactic structures, with fewer modifiers per noun phrase and a reduced average depth of the dependency trees. Moreover, an examination of readability shows a more reader-friendly profile for news texts, corroborated by lower scores in the Flesch-Kincaid Grade Level (FKGL) and the Automated Readability Index (ARI). The statistical significance observed in all metrics of Table 3, as verified by the Wilcoxon signed-rank test[7] ($p<0.05$), suggests that scientific news narratives function as a more accessible medium with respect to lexical, syntactic and readability features compared to original research papers.

Property	Papers	News
Type-Token Ratio↑	0.20	0.44
Lexical Density↑	0.42	0.46
Avg. # Difficult Words↓	773.08	134.84
Avg. # Modifiers per Noun Phrase↓	0.58	0.51
Avg. Depth of Dep Tree↓	6.94	6.25
FKGL↓	14.57	13.31
ARI↓	17.94	16.32

Table 3: Papers and News comparison

Figure 3A provides additional details on the distribution of part-of-speech tags between the two text types: news reports contain a higher proportion of verbs and adjectives, while original articles feature more proper nouns, numbers, and punctuation. Regarding rhetorical structure (discourse relations), as shown in Figure 3B, news reports tend to utilize more 'example', 'contrast', and 'cause & effect' relations, which may enhance their appeal and accessibility. In contrast, academic texts often favor 'temporal', 'coordinating', and 'progressive' relations to convey research trajectories and findings.

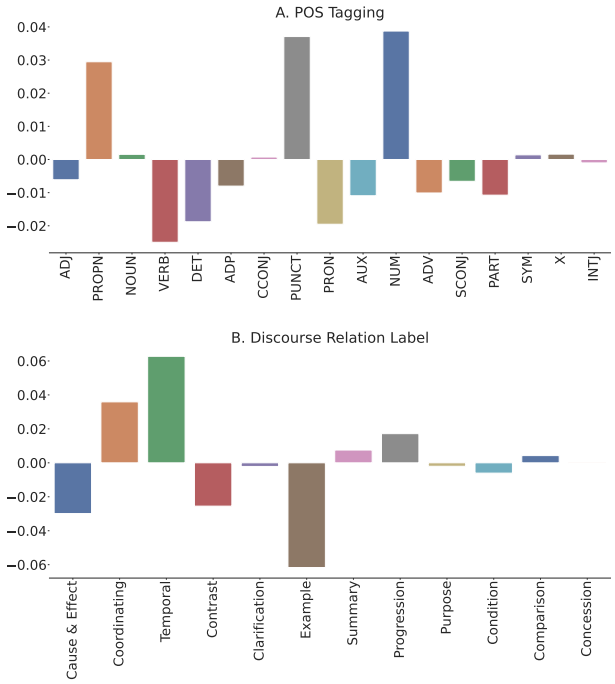

Figure 3: Absolute differences of proportion in linguistic structures (academic papers−news articles).

5. Experiments

5.1. Baseline Models

To promote future work, we benchmark our datasets using two types of baselines: extractive methods and abstractive approaches. Extractive methods involve directly retrieving sentences or phrases from the source text, while abstractive ap-

[4] https://spacy.io/
[5] https://github.com/seq-to-mind/DMRST_Parser
[6] https://github.com/textstat/textstat
[7] https://scipy.org/

proaches generate outputs by comprehending and paraphrasing the content.

5.1.1. Extractive Methods

We select several prevalent algorithms, including heuristic methods like Lead-3/K, Tail-3/K, and Random-3/K (with K denoting the average number of sentences in news reports of the training set, K=25), and non-heuristic algorithms like Latent Semantic Analysis (LSA) (Steinberger et al., 2004), LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), Ext-oracle (Narayan et al., 2018), and PacSum (Zheng and Lapata, 2019). We also include the original papers as a trivial output to establish the performance of the lower bound for the task.

5.1.2. Abstractive Methods

For abstractive methods, we utilize SOTA models based on Seq2Seq like Longformer (Beltagy et al., 2020), RSTformer (Pu et al., 2023) and SIMSUM (Blinova et al., 2023), and the Generative Pre-trained Transformer (GPT) architecture like Vicuna7B-16k (Zheng et al., 2023) and GPT-4 (OpenAI, 2023). Notably, RSTformer (current SOTA model in SLS task) enhances Longformer's attention mechanism by incorporating discourse knowledge. SIMSUM (current SOTA model in STS task) simplifies documents by aggregating information from several sentences into a single one, omitting some content and breaking down complex sentences. Vicuna7B-16k model is a competitive open-source large language model based on LLaMA 2 (Touvron et al., 2023). We also conduct a comparison using GPT-4 in a zero-shot setting (ZS).

5.2. Experimental Settings

For unbiased comparison, we operate models based on the open-source codes provided by the authors and adhere to the original implementations' default settings, such as model size, batch size, optimizer configuration, and learning rate. During the decoding process, abstractive algorithms are set to a uniform beam search with beam size=3 and trigram blocking, we also set temperature and top-p parameters to 1. For Vicuna model, the initial learning rate is set to 5e-5, with a cosine learning rate schedule, batch size of 16, and fully fine-tuned for 30 epochs. The optimizer used is Adam, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-9}$, weight decay = 0.1, and a warm-up ratio of 0.2. To prevent overfitting, we apply early stopping and L2 regularization techniques. Unless stated otherwise, all other parameters align with those in the original publications.

5.3. Automatic Metrics

In alignment with other text-generation work (Narayan et al., 2018; Pu and Sima'an, 2022; Liu et al., 2023c; Pu et al., 2023; Blinova et al., 2023), we examine model performance against the human reference news articles using the following metrics:

- ROUGE (Lin, 2004) measures the overlap of n-grams between machine-generated output and human-crafted reference. We apply F1 scores of Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and Rouge-Lsum (RLsum) in our analysis.
- BERTScore (Zhang* et al., 2020) examines word overlap between texts, using contextual BERT embedding for semantic similarity analysis.
- METEOR (Banerjee and Lavie, 2005) calculates the harmonic mean of uni-gram precision and recall with an enhanced emphasis on recall for balanced evaluation.
- sacreBLEU (Post, 2018) gauges linguistic congruence and translation fluidity between generated and reference texts for comparative analysis of text generation systems.
- NIST (Lin and Hovy, 2003) evaluates the informativeness of n-grams, assigning weights based on corpus frequency-derived information content.
- SARI (Xu et al., 2016) assesses text simplification competency across three dimensions: retention, deletion, and integration of pertinent n-grams for the streamlined rendition of the original text.

Additionally, we also use reference-free automatic evaluation metrics from Section 4.3 to evaluate the differences between the top-performing models in their respective categories and human performance on the same test subset.

6. Results and Analysis

6.1. General Results

Table 4 depicts the performance of benchmark models on the same test split. Heuristic models such as Lead-3/K, Tail-3/K, and Random-3/K serve as baseline comparison models. Furthermore, we also adopt several popular extractive and abstractive algorithms to explore which algorithm paradigm is more suitable for our dataset.

Overall, abstractive models significantly outperform both heuristic and extractive models. Specifically, the RSTformer demonstrates superior performance in terms of ROUGE metrics, indicating its enhanced lexical selection capability. Meanwhile, Vicuna surpasses the RSTformer in the SARI metric, highlighting its strengths in simplification and paraphrasing. When it comes to BERTScore, METEOR, sacreBLEU, and NIST metrics, RSTformer and Vicuna exhibit comparable performance.

Model	R1$_{f1}$↑	R2$_{f1}$↑	RL$_{f1}$↑	RLsum$_{f1}$↑	BERTscore$_{f1}$↑	Meteor↑	sacreBLEU↑	NIST↑	SARI↑
Full article (lower bound)	14.42	5.21	6.90	13.94	58.55	0.21	1.49	0.55	34.83
Lead-3	14.65	4.47	8.93	13.47	54.69	0.06	0.12	0.00	35.79
Lead-K	41.99	10.96	16.13	39.68	58.55	0.27	5.25	2.34	37.21
Tail-3	8.43	1.46	5.41	7.77	43.61	0.03	0.05	0.01	33.94
Tail-K	32.16	5.58	13.37	30.49	51.83	0.20	2.16	1.76	35.50
Random-3	10.20	1.84	6.43	9.30	47.68	0.04	0.05	0.01	34.23
Random-K	35.91	6.90	14.10	33.83	54.41	0.22	2.68	1.97	35.94
LSA (Steinberger et al., 2004)	39.75	8.45	15.10	37.40	56.43	0.25	3.42	2.19	36.13
LexRank (Erkan and Radev, 2004)	35.59	7.98	14.97	33.62	54.49	0.24	3.22	1.92	36.16
TextRank (Mihalcea and Tarau, 2004)	35.64	7.85	14.77	33.52	53.80	0.23	3.17	1.94	36.13
PacSum (Zheng and Lapata, 2019)	41.03	10.53	15.47	38.75	57.64	0.27	4.82	2.28	36.85
Ext-oracle (Narayan et al., 2018)	42.58	11.92	16.16	40.38	56.60	**0.30**	5.90	2.43	37.28
GPT-4$_{ZS}$ (OpenAI, 2023)	41.38	9.03	15.25	39.01	58.33	0.19	4.64	1.12	37.52
SIMSUM (Blinova et al., 2023)	44.38	12.20	18.13	41.46	60.09	0.27	6.31	2.38	40.54
Longformer (Beltagy et al., 2020)	47.60	14.74	19.09	44.83	62.84	0.28	7.64	2.47	41.52
RSTformer (Pu et al., 2023)	**48.21**‡	**14.92**	**20.12**‡	**45.19**‡	62.80	0.28	**7.70**	**2.55**	41.56
Vicuna7B-16k (Zheng et al., 2023)	47.75	14.88	19.92	45.01	**62.88**	**0.30**	7.69	2.53	**41.71**‡

Table 4: Model performance. The bold numbers represent the best results with respect to the given test set. ‡ denotes that the value is significantly superior to those of all other models according to the Wilcoxon signed-rank test in the corresponding indicator ($p<0.05$).

6.2. Comparison with Human-authored News Articles

Table 5 contrasts the lexical diversity, syntactic complexity, and readability of the best models for extractive and abstractive methods, as listed in Table 4, against human ability.

Metric	Human	Ext-oracle	RSTformer	Vicuna7B
Avg. # Tokens	696.19	1274.54	653.37	782.21
Avg. # Sents.	25.29	44.51	22.85	25.03
Type-Token Ratio↑	0.45	0.40	**0.47**	0.37
Lexical Density↑	**0.46**	0.44	**0.46**	0.42
Avg. # Difficult Words↓	**134.65**‡	217.37	141.75	164.5
Avg. # Modifiers per NP↓	**0.50**	0.61	0.57	0.62
Avg. Depth of Dep Tree↓	**6.24**‡	6.68	7.62	6.72
FKGL↓	**13.27**‡	15.80	14.95	14.12
ARI↓	**16.26**‡	19.20	18.22	16.90

Table 5: Models vs. Humans; ‡ indicates that the value significantly differs from those of all other candidates in the same test set, according to the Wilcoxon signed-rank test for the corresponding indicator ($p<0.05$).

We find that texts generated by the RSTformer model most closely resemble human-written news articles in both length and lexical diversity, while Vicuna-generated texts tend to include slightly longer and more complex words. Additionally, human-written texts are classified as significantly more readable than any model-generated texts, based on FKGL and ARI metrics. Texts generated by Ext-oracle are notable for being much longer and containing more difficult words compared to those written by humans.

6.3. Automatic Inconsistency Detection

Figure 4 shows the outcomes of the automated consistency evaluation for different models on the same test set. We observe that the SummaC consistency scores (Laban et al., 2022) for news reports generated by abstractive models fall below those generated by humans in scientific news articles. On the other hand, extractive models, which directly extract text segments from the source, achieve the highest consistency scores without introducing or reorganizing content.

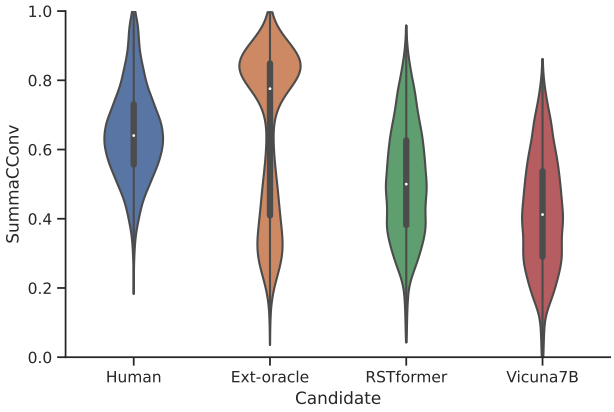

Figure 4: Consistency check

6.4. Human Evaluation

In order to gain more insight into the quality of the generated news articles compared to human-authored news articles, we randomly choose 10 samples and present them to human evaluators. The evaluators are asked to read the corresponding original academic article, as well as four candidate news reports (from Ext-oracle, RSTformer, Vicuna, and the original human-authored text). The human evaluators are blind to the condition, i.e., they do not know which article comes from which system (or a human author). Each of the 10 samples is assessed by three different judges, resulting in a total

of 30 evaluated samples. The recruited evaluators are all Master's or Doctoral students with Computer Science or Computational Linguistics backgrounds, with high proficiency in English. All annotators are compensated at the prevailing hourly rates set by the university.

The annotators assess the texts based on the following criteria:

- **Relevant:** How well the news article reflects the source text.
- **Simple:** How understandable the text is for the general public.
- **Concise:** The extent to which the text omits less important information from the source article.
- **Faithful:** The extent to which the text contradicts the information from the source text.

Evaluators should assign scores to candidate texts on a scale of 1 to 3 for each criterion, with higher scores indicating better generation quality. Annotators are also required to use different colors to highlight any errors in generated news articles and link them to the corresponding section in the source text. After scoring all candidates, evaluators are asked to identify the best and worst news texts.

Table 6 displays human evaluation results. For each metric, we calculate the average value to assess the candidate system's performance. The terms 'Best' and 'Worst' denote the frequency with which a model's output is ranked highest or lowest among the four candidates, respectively. Additionally, we count issues flagged by evaluators in the generated news texts, with identical issues highlighted in the same color considered a single instance.

Ext-oracle performs poorly in terms of 'simple' and 'concise', as an extractive method, it shows strength in 'relevant' and 'faithful'. However, it is never chosen as the best candidate by the annotators. RSTformer outperforms Vicuna in 'relevant' and 'faithful', whereas Vicuna bests RSTformer in 'simple' and 'concise'. Notably, both abstractive models face challenges with maintaining faithfulness across all generated news texts, a critical issue for practical deployment. Overall, our findings suggest that NLG models have yet to match the proficiency of human writing. This underscores a significant opportunity for future research in enhancing model reliability.

6.5. GPT-4 Evaluation

We also employ the same guidelines used for human evaluation to ask GPT-4 via API queries (OpenAI, 2023) to assess our benchmark models. For consistency, all experiments adhere to OpenAI's default hyper-parameter settings. To ensure no influence from previous interactions, we reset the conversation history before each GPT-4 query. Initially, we sought to confirm whether GPT-4's evaluations align with human judgments on a subset of 10 samples as discussed in Section 6.4 (maintaining the same ranking of news report scores), and indeed, we find consistent results across all four criteria. Subsequently, we randomly pick an additional 100 samples from the test set, with the results displayed in Table 7.

Candidate	Relevant	Simple	Concise	Faithful	Best | Worst
Human	**2.67**$_{0.23}$	**2.83**‡$_{0.33}$	**2.43**‡$_{0.33}$	**2.73**‡$_{0.10}$	70.00% | 3.33%
Ext-oracle	2.63$_{0.33}$	1.30$_{1.00}$	1.20$_{1.00}$	2.63$_{0.17}$	0.00% | 80.00%
RSTformer	2.63$_{0.40}$	2.27$_{0.67}$	2.03$_{0.73}$	2.17$_{1.00}$	20.00% | 3.33%
Vicuna7B	2.47$_{0.60}$	2.47$_{0.67}$	2.17$_{0.60}$	1.96$_{1.00}$	10.00% | 13.33%

Table 6: Human evaluation results: average ratings (on a scale from 1 to 3). The number following the slash represents the percentage of evaluation samples in which an issue identified by evaluators occurs at least once.

Candidate	Relevant	Simple	Concise	Faithful	Best | Worst
Human	**2.86**‡	**2.77**‡	**2.83**‡	**2.91**‡	92.00% | 0.00%
Ext-oracle	2.73	1.73	1.55	2.70	0.00% | 93.00%
RSTformer	2.69	2.41	2.42	2.47	6.00% | 2.00%
Vicuna7B	2.56	2.59	2.53	2.32	2.00% | 5.00%

Table 7: GPT-4 evaluation results on 100 samples

According to Table 7, GPT-4's evaluations mirror those of human evaluations. All models underperform compared to human answers. The scores of GPT-4 for the two SOTA abstractive models are comparable to each other. Across all test samples, GPT-4 prefers the human answer as the best answer, while the extractive method is frequently rated as the worst.

6.6. Model Errors

In conjunction with the above-mentioned human evaluation, we conduct a qualitative analysis to identify the prevalent challenges in current models:

1. Hallucinations: Models may produce ungrounded information. For instance, a model might suggest future research areas for chatbots, even if such discussions are absent from the source document.

2. Factual Errors: Models often misstate facts, especially numerical values. For example, in a cancer identification paper, the original mentions sensitivity at 96.7% and specificity at 97.5%, but the model reports them as 88.2% and 98.3% respectively.

3. Generalization: While models generally grasp the primary subject, they sometimes diverge into irrelevant specifics. A case in point is a paper on cybersickness, where the model drifts from the main topic into unrelated areas, unlike a focused human-written article.

7. Conclusion

We introduce the scientific news report generation task, and present a novel dataset "SciNews". This dataset comprises over 40,000 scientific papers spanning nine distinct domains, each paired with a corresponding news report. We conduct an exploratory analysis of the SciNews dataset and provide benchmark results highlighting the challenges faced by current state-of-the-art models. The SciNews dataset not only offers some research prospects, such as fostering the advancement of improved models for scientific news report generation which are faithful to the facts in the original papers but also suggests the potential enhancement of news reports through the integration of relatable explanations. Additionally, beyond its primary purpose, the SciNews dataset can also serve as a valuable resource for other natural language processing tasks, including topic classification and news headline generation.

Ethical Considerations

All data in our dataset are sourced from publicly accessible resources, adhering to the respective copyright and web crawling regulations. Each data sample explicitly displays the relevant source URL and author attributions. Moreover, every data sample has undergone rigorous examination and penning by journalists on the Science X website to mitigate ethical or moral apprehensions. Our methodology discerns no privacy infringements during the data processing, experimental analysis, and model training/evaluation phases. Regarding human evaluation, all contributors participate voluntarily and are fairly compensated. We provide a secure and comfortable environment for evaluations, strictly following the ACM Code of Ethics throughout this study's experiments and analyses.

Limitations

Data: The SciNews dataset comprises academic papers in English along with their corresponding news reports. Despite the high-quality sourcing of the data, which involves contributions from domain experts, it is possible that biases specific to certain fields persist. Moreover, we only explore scientific news reports in nine research fields, and these data are only a small part of the real-world data and do not contain all of the academic fields, such as Humanities and Social Sciences. The exclusivity of English within our dataset can be perceived as a limitation, as it presently does not incorporate data in other languages.

Model: In our utilization of the SciNews dataset, we have employed several state-of-the-art models, which may carry biases embedded during their pretraining. However, we have not conducted rigorous assessments regarding the magnitude of these biases within the models as it is beyond the scope of this study. Moreover, the data we gathered all originates from online publicly available resources, so we cannot ascertain whether ChatGPT/GPT-4 has been exposed to or trained on our data during their development (data contamination risks). We acknowledge this limitation and earmark this as a potential space for exploration in our future studies. In addition, we also leave the discussion of differences in the performance of NLG models across different disciplines as part of future work.

Automated Evaluation: Despite employing nine popular automated evaluation algorithms systematically assessing the baseline models from various angles on the test set with human gold answers, and contrasting the baseline models with human performance through multiple reference-free metrics, we recognize that all automated metrics have inherent limitations. Consequently, they might not furnish a comprehensive evaluation of the model's performance.

Human Assessment: The size of the data samples used for human evaluation is constrained by the nature of long document generation and the extensive length of the original texts, often spanning multiple pages. Consequently, expanding the evaluation process through means such as crowdsourcing becomes challenging. As a result, we can only assess a limited set of 10 documents, which may not offer a fully representative view of the entire dataset. While all of our recruited human evaluators are Master's or Ph.D. students, not all of them are domain experts/lay readers, nor can they be experts/lay readers across multiple scientific fields. Therefore, their judgments cannot be solely relied upon.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). We are grateful to the reviewers and area chairs for their exceptionally detailed and helpful feedback.

8. Bibliographical References

Amanuel Alambo, Cori Lohstroh, Erik Madaus, Swati Padhee, Brandy Foster, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Michael Raymer. 2020. Topic-centric unsupervised multi-document summarization of scientific and news articles. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 591–596.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5327–5344, Online. Association for Computational Linguistics.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Alexey Bukhtiyarov and Ilya Gusev. 2020. Advances of transformer-based models for news headline generation. In *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9*, pages 54–61. Springer.

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.

Pengshan Cai, Kaiqiang Song, Sangwoo Cho, Hongwei Wang, Xiaoyang Wang, Hong Yu, Fei Liu, and Dong Yu. 2023. Generating user-engaging news headlines. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3265–3280, Toronto, Canada. Association for Computational Linguistics.

Ronald Cardenas, Bingsheng Yao, Dakuo Wang, and Yufang Hou. 2023. 'don't get too technical with me': A discourse structure-based framework for automatic science journalism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1186–1202, Singapore. Association for Computational Linguistics.

Muthu Kumar Chandrasekaran, Anita de Waard, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Petr Knoth, David Konopnicki, Philipp Mayr, Robert M. Patton, and Michal Shmueli-Scheuer, editors. 2020a. *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Online.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020b. Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020c. 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*, pages 214–224, Online. Association for Computational Linguistics.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. SciXGen: A scientific paper dataset for context-aware text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1483–1492, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu, and Dong Yu. 2022. Toward unifying text segmentation and long document summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 106–118, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2022. Controllable sentence simplification via operation classification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2091–2103, Seattle, United States. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. ParaSCI: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Liana Ermakova, Patrice Bellot, Jaap Kamps, Diana Nurbakova, Irina Ovchinnikova, Eric SanJuan, Elise Mathurin, Sílvia Araújo, Radia Hannachi, Stéphane Huet, et al. 2022. Automatic simplification of scientific texts: Simpletext lab at clef-2022. In *European Conference on Information Retrieval*, pages 364–373. Springer.

Liana Ermakova, Eric SanJuan, Stéphane Huet, Olivier Augereau, Hosein Azarbonyad, and Jaap Kamps. 2023. Clef 2023 simpletext track: What happens if general users search scientific texts? In *European Conference on Information Retrieval*, pages 536–545. Springer.

Mehwish Fatima and Michael Strube. 2023. Cross-lingual science journalism: Select, simplify and rewrite summaries for non-expert readers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1843–1861, Toronto, Canada. Association for Computational Linguistics.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. 2022. Text simplification for legal domain: Insights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. BioLaySumm 2023 shared task: Lay summarisation of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and*

BioNLP Shared Tasks, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *arXiv preprint arXiv:2211.03818*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Ilya Gusev. 2019. Importance of copying mechanism for news headline generation. *arXiv preprint arXiv:1904.11475*.

Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2023. What's new? summarizing contributions in scientific literature. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1019–1031, Dubrovnik, Croatia. Association for Computational Linguistics.

Pengcheng He, Baolin Peng, Song Wang, Yang Liu, Ruochen Xu, Hany Hassan, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. 2023. Z-code++: A pre-trained language model optimized for abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5095–5112, Toronto, Canada. Association for Computational Linguistics.

Md Reshad Ul Hoque, Jiang Li, and Jian Wu. 2022. Sciev: Finding scientific evidence papers for scientific news. *arXiv preprint arXiv:2205.00126*.

Zachary Horvitz, Nam Do, and Michael L. Littman. 2020. Context-driven satirical news generation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 40–50, Online. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633, Doha, Qatar. Association for Computational Linguistics.

Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. SimpleScience: Lexical simplification of scientific terminology. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1071, Austin, Texas. Association for Computational Linguistics.

Connie A Korpan, Gay L Bisanz, Jeffrey Bisanz, and John M Henderson. 1997. Assessing literacy in science: Evaluation of scientific news briefs. *Science Education*, 81(5):515–532.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes-Larrarte. 2022. An exploration of post-editing effectiveness in text summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 475–493, Seattle, United States. Association for Computational Linguistics.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197, Santiago de Compostela, Spain. Association for Computational Linguistics.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. 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.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Dayiheng Liu, Yeyun Gong, Yu Yan, Jie Fu, Bo Shao, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. Diverse, controllable, and keyphrase-aware: A corpus and method for news multi-headline generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6241–6250, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Zechun Liu, Barlas Oguz, Aasish Pappu, Yangyang Shi, and Raghuraman Krishnamoorthi. 2023c. Binary and ternary natural language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 65–77, Toronto, Canada. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Cassie Majetic and Catherine Pellegrino. 2014. When science and information literacy meet: An approach to exploring the sources of science

news with non-science majors. *College teaching*, 62(3):107–112.

Yuning Mao, Ming Zhong, and Jiawei Han. 2022. CiteSum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10922–10935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ahmadreza Mosallanezhad, Kai Shu, and Huan Liu. 2020. Topic-preserving synthetic news generation: An adversarial deep reinforcement learning approach. *arXiv preprint arXiv:2010.16324*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Liubov Nesterenko. 2016. Building a system for stock news generation in Russian. In *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, pages 37–40, Edinburgh, Scotland. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Sheena Panthaplackel, Adrian Benton, and Mark Dredze. 2022. Updated headline generation: Creating updated summaries for evolving news stories. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6438–6461, Dublin, Ireland. Association for Computational Linguistics.

Daraksha Parveen, Mohsen Mesgar, and Michael Strube. 2016. Generating coherent summaries of scientific articles using coherence patterns. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 772–783, Austin, Texas. Association for Computational Linguistics.

Joseph L Polman and Jennifer MG Hope. 2014. Science news stories as boundary objects affecting engagement with science. *Journal of Research in Science Teaching*, 51(3):315–341.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Dongqi Pu, Xudong Hong, Pin-Jie Lin, Ernie Chang, and Vera Demberg. 2022. Two-stage movie script summarization: An efficient method for low-resource long document summarization. In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 57–66, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Dongqi Pu and Khalil Sima'an. 2022. Passing parser uncertainty to the transformer: Labeled dependency distributions for neural machine translation. In *Proceedings of the 23rd Annual*

Conference of the European Association for Machine Translation, pages 41–50, Ghent, Belgium. European Association for Machine Translation.

Dongqi Pu, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.

James Ravenscroft, Amanda Clare, and Maria Liakata. 2018. HarriGT: A tool for linking news to science. In *Proceedings of ACL 2018, System Demonstrations*, pages 19–24, Melbourne, Australia. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Tanik Saikh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. ScholarlyRead: A new dataset for scientific article reading comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5498–5504, Marseille, France. European Language Resources Association.

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.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2021. Fact-enhanced synthetic news generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13825–13833.

Marlisa Sigita, Ali Ridho Barakbah, Entin Martiana Kusumaningtyas, and Idris Winarno. 2013. Automatic representative news generation using online clustering. *EMITTER International Journal of Engineering Technology*.

Josef Steinberger et al. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *7th International Conference ISIM.*, page 93–100.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 2023. Multi2Claim: Generating scientific claims from multi-choice questions for scientific fact-checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2652–2664, Dubrovnik, Croatia. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. 2018. When science journalism meets artificial intelligence : An interactive demonstration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 163–168, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, pages 10824–10836, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022a. Modeling information change in science communication with semantically matched paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022b. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation. 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 5077–5089, Hong Kong, China. Association for Computational Linguistics.

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*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.