Summaries as Captions: Generating Figure Captions for Scientific Documents with Automated Text Summarization

Chieh-Yang Huang,[1*] Ting-Yao Hsu,[1*] Ryan Rossi,[2] Ani Nenkova,[2] Sungchul Kim,[2] Gromit Yeuk-Yin Chan,[2] Eunyee Koh,[2] Clyde Lee Giles,[1] Ting-Hao 'Kenneth' Huang[1]

[1]Pennsylvania State University, University Park, PA, USA.
{chiehyang,txh357,clg20,txh710}@psu.edu
[2]Adobe Research, San Francisco, CA, USA.
{ryrossi,nenkova,sukim,ychan,eunyee}@adobe.com

Abstract

Good figure captions help paper readers understand complex scientific figures. Unfortunately, even published papers often have poorly written captions. Automatic caption generation could aid paper writers by providing good starting captions that can be refined for better quality. Prior work often treated figure caption generation as a vision-to-language task. In this paper, we show that it can be more effectively tackled as a **text summarization** task in scientific documents. We fine-tuned PEGASUS, a pre-trained abstractive summarization model, to specifically summarize figure-referencing paragraphs (*e.g.*, "Figure 3 shows...") into figure captions. Experiments on large-scale arXiv figures show that our method outperforms prior vision methods in both automatic and human evaluations. We further conducted an in-depth investigation focused on two key challenges: *(i)* the common presence of low-quality author-written captions and *(ii)* the lack of clear standards for good captions. Our code and data are available at: https://github.com/Crowd-AI-Lab/Generating-Figure-Captions-as-a-Text-Summarization-Task.

1 Introduction

In scientific documents, effective figure captions help readers understand complex figures like bar charts, line charts, or pie charts. These captions describe the images and often include necessary context from the document's full text (Durbin, 2004). Unfortunately, even published papers often have poorly-written captions. As per our analysis (Section 8.2), around 53.88% of line charts' captions in arXiv cs.CL papers are found to be unhelpful for NLP readers. Automatic caption generation could aid paper writers by providing good starting captions that can be refined for better quality.

Previous research typically approached figure caption generation as a *vision-to-language* task, *i.e.*,

creating captions based on the image. For instance, Hsu et al. (2021) used an end-to-end approach with CNN+RNN structures, which extracted feature representation from the image and converted it into caption text. Qian et al. (2021) took a slightly different approach: first understanding what is in the image, pulling out key information, and then using a preset template to create the caption. However, although achieving some success in synthetic data (Kahou et al., 2017; Kafle et al., 2018; Chen et al., 2020a; Zhu et al., 2021), these approaches often struggled to caption real-world figures. For example, Hsu et al. (2021)'s end-to-end approach, trained and tested using arXiv figures, achieved a BLEU-4 score of only 2.91.

In this paper, we argue that figure captioning in scientific documents can be more effectively tackled as a **text-summarization task**: The caption can be generated by **summarizing the paragraphs mentioning the figure** (as shown in Figure 1.) Scientific figures typically come with extensive text in the scientific document that can aid caption generation. Our analysis (Section 5) shows that, in arXiv, over 75% of words in figure captions can be aligned with the words in the paragraphs referencing those figures, which motivates our approach. The automatic evaluation shows that summarizing paragraphs referencing the figures results in better captions than prior vision-based methods. In a human evaluation by external domain experts, our best-performing model's captions were preferred over the original captions 46.67% of the time.

We further conducted an in-depth investigation focused on two key challenges: *(i)* the common presence of low-quality author-written captions and *(ii)* the lack of clear standards for good captions. Surprisingly, 53.88% of the author-written captions in our sample was deemed unhelpful. This has implications for the design of future captioning systems, underscoring the influence of data quality on captioning performance.
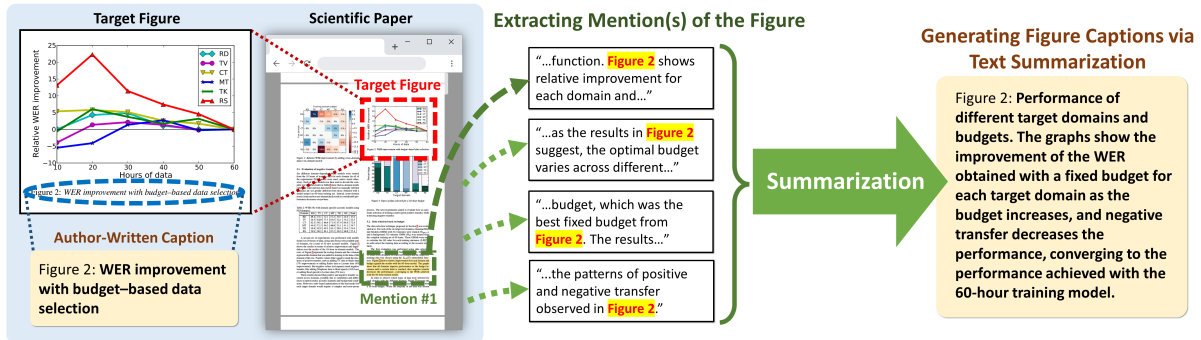
---

*Equal contribution.

80

Figure 1: Figure captioning can be addressed as a text-summarization task. The figure's caption can be generated by summarizing the paragraphs mentioning the figure. The caption is generated by the model Pegasus$_{P+O+B}$. The example shown in this figure is extracted from the paper (Doulaty et al., 2015).

2   Related Work

Prior figure captioning works can be broadly categorized into two approaches: caption generation *(i)* based on the image of the figure or *(ii)* based on the data chart underlying the figure.

Earlier image-based approaches focused on automated image understanding, which involved parsing images to extract the figure's key attributes and converting parsed data into captions, *e.g.*, using pre-defined templates (Kahou et al., 2017; Kafle et al., 2018; Methani et al., 2020; Qian et al., 2021; Siegel et al., 2016). Recently, with the advance of deep learning, more works are adopting an end-to-end paradigm, generating captions straight from the neural representations of images (Mahinpei et al., 2022; Pelka et al., 2021; Hsu et al., 2021; Chen et al., 2019; Kantharaj et al., 2022; Chen et al., 2020a). Our work contrasts with prior studies by focusing on text to generate captions instead of visuals. To the best of our knowledge, no existing figure-caption datasets explicitly contain the figures' accompanying documents (Pelka et al., 2021; Hsu et al., 2021; Chen et al., 2019), as this task has generally been approached as a vision task. Most recently, a knowledge-augmented image captioning method that uses both image and text data was introduced (Yang et al., 2023), suggesting the potential of using text from documents.

Some approaches generate captions using the underlying tabular data of a figure rather than the figure's image. Earlier approaches often employed rule-based techniques (Corio and Lapalme, 1999; Demir et al., 2008; Fasciano and Lapalme, 1996; Mittal et al., 1998), while newer ones favor learning-based methods (Barzilay and Lapata, 2005; Wiseman et al., 2017; Moraes et al., 2014; Zhu et al., 2021; Kantharaj et al., 2022; Obeid and Hoque, 2020; Reiter et al., 2005; Parikh et al.,

2020; Chen et al., 2020b; Gong et al., 2019; Su et al., 2021; Chen et al., 2020c). Despite these approaches' ability to utilize tabular and meta data, they necessitate access to the figure's raw data. Contrarily, our work uses the rich textual information in scientific documents to generate captions.

3   Problem Statement and Terminology

A document D contains n figures, F_1 to F_n, where F_i has a caption C_i that was written by the document author. In document D, j sentences, $M_{i,1}$ to $M_{i,j}$, explicitly mention F_i (*e.g.*, "As shown in F_i..."). The objective of this work is to automatically generate a high-quality caption, C_i', for figure F_i using only its mentions ($M_{i,1}$ to $M_{i,j}$) and the surrounding text of the mentions in document D.

In the rest of the paper, we use these terms:

- A **"Mention"** refers to a sentence in a document that explicitly mentions the target figure, *e.g.*, "As shown in Figure 6..." If there are multiple Mentions, the first Mention is referred to.

- A **"Paragraph"** refers to a section of text containing a Mention. In this work, the boundaries of a Paragraph are determined by the <p> tag produced by PDF parsing.

- Sentences near a Mention may contain relevant information, so we extracted n preceding sentences and m following sentences to form the **"Window[n, m]"** text snippet. For instance, "Window[1, 2]" refers to a snippet of four sentences, including one preceding sentence, the Mention sentence, and two following sentences.

- An **"OCR"** refers to the textual information (*e.g.*, legends, labels, etc.) extracted from the image, by optical character recognition (OCR) software.

Source	Random		Mention		Paragraph		OCR	Window[0, 1]		Window[1, 0]		Window[1, 1]		Window[2, 2]	
	S	P	-	+OCR	-	+OCR	-	-	+OCR	-	+OCR	-	+OCR	-	+OCR
Caption	35.23	44.43	53.43	60.16	75.19	**76.68**	34.75	60.85	65.43	59.09	64.19	65.20	68.73	69.09	71.77
Source	32.52	19.52	**39.51**	18.78	12.53	9.39	20.79	30.49	17.19	32.40	17.33	25.10	15.55	19.84	13.45

Table 1: Macro coverage rates (percentage) between captions and relevant texts (S: Sentence and P: Paragraph). Caption coverage gives the percentage of words in the caption that can be found in the source texts and vice versa (punctuation and stop words are excluded.) The results show that 76.68% of the words in captions could be found in Paragraph+OCR, motivating us to generate captions by text summarization.

4 Dataset

Before diving into our experiments and analyses, we first describe the dataset upon which our study is grounded. Our results are based on a scientific figure caption dataset, SCICAP, and several pre-processing steps to fit it into our workflow.

SCICAP is a dataset that contains over 416,000 line charts and captions extracted from more than 290,000 arXiv papers (Hsu et al., 2021). It was one of the first large-scale figure-captioning datasets based on real-world scientific figures. However, it does not contain the paragraphs that mention the figure. To address this, we downloaded all the PDF files of the original arXiv papers used in SCI-CAP and extracted all the Mentions and Paragraphs as outlined in Section 6.1. Detailed information on preprocessing, including the dataset resplit and OCR extraction, are described in Appendix B.

5 Motivating Analysis

To understand the correlation between mentions and captions, we performed a series of analyses using the data described in Section 4. Specifically, we investigated the extent to which the words in the figure captions are represented in the corresponding figure-mentioning paragraphs. We used awesome-align (Dou and Neubig, 2021) to obtain the alignment between the source texts (mentions, paragraphs, and OCRs) and captions. Awesome-align compared the similarity of the words' contextual embeddings and assigned an alignment between words if the similarity passed a threshold. We used SciBERT (Beltagy et al., 2019) to obtain contextual embeddings and softmax threshold = 0.99 to reduce false alignments.

After obtaining the alignments, we computed what percentage of information in the caption could be found in the source texts. The results shown in Table 1 indicate that **76.68% of the caption's information could be found in Paragraph and OCR,** motivating us to generate figure captions by

summarizing Paragraph. We also observed that a randomly selected sentence and paragraph from the same paper can cover 35.23% and 44.43% of the caption, respectively, showing that there was some generic information-sharing across the paper. We also conducted a study using the exact overlapping (*i.e.*, BLEU score) in Appendix A.

6 Generating Figure Captions as a Text Summarization Task

Figure 1 overviews the proposed pipeline. This section describes each step of the pipeline.

6.1 Extracting Mentions and Paragraphs

The system first extracts Mentions and their associated Paragraphs (as defined in Section 3.) In this paper, we used Grobid (kermitt2, 2022), a publicly-available tool for converting PDF files into structured XML documents, to extract plain text from the paragraphs (including the <p> tags) in each paper. This plain text was then segmented into sentences using BlingFire (microsoft, 2022). We developed regular expressions to identify sentences mentioning specific figures. For instance, sentences such as "As shown in Figure 6, ..." were first identified and then linked to Figure 6. To assess the performance of these regular expressions, we conducted a manual evaluation of 300 samples from our experimental dataset. The results showed a high level of precision (99.58%) and recall (94.44%).

6.2 Generating Captions Using Text Summarization Models

As shown in Figure 1, our system then automatically summarizes all the extracted Mentions (or Paragraphs) into a figure caption. In this work, we used PEGASUS, an abstractive summarization model (Zhang et al., 2020), and fine-tuned it on our dataset. Five Pegasus models, Pegasus_M, Pegasus_P, Pegasus_O, Pegasus_{M+O},

and Pegasus$_{P+O}$, were trained utilizing five distinct input combinations, including *(i)* **Mention,** *(ii)* **Paragraph,** *(iii)* **OCR output of the target figure image,** *(iv)* **Mention+OCR, and** *(v)* **Paragraph+OCR.** Pegasus$_{P+O}$ encompasses the most of relevant information in the document and thus is expected to yield the optimal summary.

Additionally, we built Pegasus$_{P+O+B}$, a specialized version of the model designed to be trained on a subset of higher-quality captions, *(vi)* **Paragraph+OCR-Better**. Given the absence of reliable automated ways to assess the quality of captions, we followed a guideline from previous studies indicating that longer captions enhance reader comprehension (Hartley, 2003; Gelman et al., 2002). We trained the model using captions with 30 or more tokens. The average caption length was 26.8 tokens, so we set 30 tokens as the threshold. The training was performed using Paragraph+OCR inputs.

We identified two major challenges in generating captions for scientific figures in real-world scenarios. We discuss these challenges in the following subsections, with an in-depth analysis in Section 8.

6.2.1 Challenge 1: Addressing Unreliable Quality of Real-World Data

Low-quality captions often occur in scholarly articles. Our analysis (see Section 8.1) showed that 50% of line charts' author-written captions in arXiv `cs.CL` papers were deemed unhelpful by domain experts. The impact of this unreliable data quality is that developers could train and test captioning models with unhelpful captions. The lack of automatic methods for evaluating caption quality makes it hard to identify suitable training examples and eliminate poor ones. To address this issue, we included Pegasus$_{P+O+B}$ that was trained on longer captions, which is suggested by literature to be more helpful to readers (Hartley, 2003; Gelman et al., 2002). To account for low-quality test data, we conducted both human and automatic evaluations. The data quality of figure captions was analyzed and is presented in Section 8.2.

6.2.2 Challenge 2: Defining a Clear Standard for "Good" Figure Captions

The deeper issue is the lack of a set of well-defined and actionable criteria for determining the usefulness of a figure caption. Although there are guidelines for writing effective scientific figure captions (Rougier et al., 2014; Biegel and Kamat,

2019), their translation into algorithmic models can be challenging. From a modeling standpoint, the lack of a clear goal presents a challenge, as it is uncertain what to optimize for once fluency has been achieved. In this paper, we focus on demonstrating the feasibility of generating captions via text summarization. Although we did not incorporate specialized goals in the model, we examine the criteria for a "good" caption in Section 8.2.

7 Experimental Results

A Simple Baseline: Using Extracted Mentions as Captions. Motivated by our information overlap study (Section 5), we created the **Reuse** baselines. These baselines simply repurpose portions of the input text as the prediction.

Vision-to-Language Baselines. The vision-to-language generation treated this task as an image-captioning task that took the scientific figure image as input and generated a text to describe it. We compared two vision-to-language models as baselines. First, we built a sequence-to-sequence model by combining BEiT (Bao et al., 2022) and GPT-2 (Radford et al., 2019). We also selected the TrOCR (Li et al., 2021) model, a transformer-based sequence-to-sequence model pre-trained for OCR tasks. Compared to image encoders like ViT (Dosovitskiy et al., 2021) and BEiT (Bao et al., 2022), which were trained on photos, OCR models trained on printed and handwritten documents align more closely with the scientific paper domain. All figures from SCICAP (106,391 training samples) were used for training since no mentions were required.

Experimental Setup. A total of 14 methods were included for comparison: six reuse baselines with six input variations (M, P, W[0, 1], W[0, 2], W[1, 1], and W[2, 2]); five text summarization models with five inputs (M, M+O, P, P+O, and O); one text summarization model using P+O with controlled data quality; and two vision-to-language models (BEiT+GPT-2 and TrOCR). Note that we use subscripts of M, P, W, O, B to denote the input features: Mention, Paragraph, Window, OCR, and Better data quality, respectively. The model training details and decoding configuration are provided in Appendix C.

7.1 Automatic Evaluation Results

Conventional Automatic Evaluation. We used F1 of ROUGE-1, ROUGE-2, ROUGE-L (Lin,

Model	Feature	Length	Rouge-1 (F1)		Rouge-2 (F1)		Rouge-L (F1)		MoverScore		BERTScore	
			Score	Norm	Score	Norm	Score	Norm	Score	Norm	Score	Norm
Reuse	**M**	33.2	.291	1.346	.139	1.790	.239	1.401	.535	1.023	.628	1.064
	P	238.3	.171	1.042	.089	1.006	.134	1.030	.503	1.004	.567	1.008
	W[0, 1]	50.3	.281	1.216	.132	1.509	.224	1.273	.529	1.016	.620	1.048
	W[0, 2]	68.0	.259	1.129	.123	1.341	.205	1.186	.524	1.013	.611	1.034
	W[1, 1]	67.8	.266	1.156	.124	1.346	.204	1.183	.524	1.012	.613	1.037
	W[2, 2]	98.7	.235	1.082	.112	1.179	.180	1.105	.517	1.007	.600	1.020
Pegasus	**M**	12.2	.321	1.898	.153	2.907	.283	1.971	.553	1.065	.654	1.158
	M+O	12.8	.331	1.909	.161	2.945	.292	1.993	.556	1.071	.661	1.166
	P	14.0	.374	2.067	.205	3.507	.334	2.201	.570	1.095	.682	1.196
	P+O	14.0	**.381**	**2.106**	**.212**	**3.635**	**.340**	**2.242**	**.571**	**1.097**	**.685**	**1.202**
	P+O+B	38.3	.321	1.452	.154	1.916	.265	1.537	.546	1.044	.639	1.082
	O	12.1	.133	0.789	.026	0.495	.119	0.828	.518	0.998	.561	0.993
TrOCR	**Figure**	10.0	.220	1.464	.073	1.653	.195	1.502	.534	1.033	.610	1.096
BEiT+GPT2		15.8	.164	0.864	.042	0.666	.144	0.917	.529	1.013	.592	1.031

Table 2: Task Performance with the **best** and second-best results highlighted. Pegasus$_{P+O}$, the text-summarization model with all available information (Paragraph+OCR), performed the best in all four metrics. Pegasus$_{P+O+B}$, the model trained with better captions, however, got lower scores.
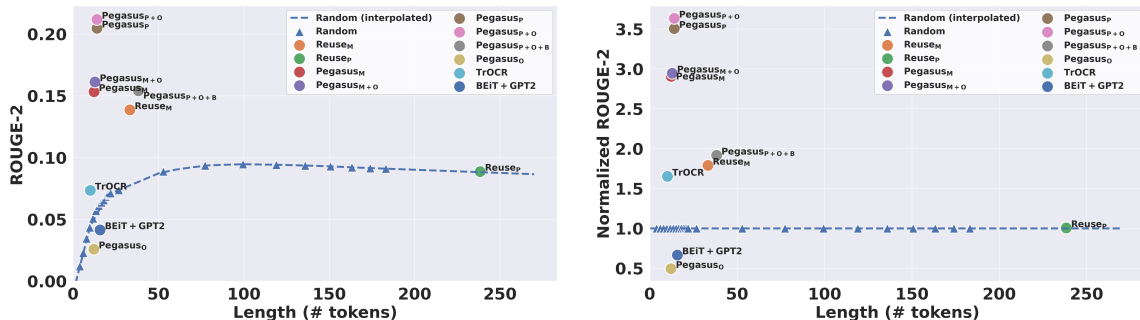

Figure 2: The relationship between average text length and ROUGE-2 score (left: original ROUGE-2; right: normalized ROUGE-2). The random baseline in the left figure shows that text lengths and scores are not independent. For example, when the predicted text is shorter than 50 tokens, predicting longer texts generally results in a higher ROUGE-2 score. The normalized scores indicate the proposed system's performance gain over the random baseline of the same length. Pegasus$_{P+O+B}$ and Reuse$_M$ get closer to TrOCR after normalization, suggesting the need for normalization for accurate interpretation of results.

2004; Nallapati et al., 2016), MoverScore (Zhao et al., 2019), and BERTScore for automatic evaluation. When computing ROUGE scores using rouge-score (google research, 2022), we turned all text into lower case and stem words. As both MoverScore and BERTScore are based on the semantic similarity, we obtained contextual embeddings from SciBERT (Beltagy et al., 2019).

Automatic Evaluation with Normalization Over Caption Length. ROUGE F1 tends to favor longer texts within a certain length, leading to a skewed comparison where models generating longer texts receive higher scores (Sun et al., 2019). We followed Sun et al. (2019)'s approach of normalizing the scores with the corresponding random baseline that generates texts of the same length.

$$Score_{normalized} = \frac{Score}{Random(length)} \quad (1)$$

where $length$ is the average length of the texts generated by the target system. We estimated $Random(length)$ by applying linear interpolation on several (length, random score) pairs. The (length, random score) pairs were obtained by randomly selecting a certain number of sentences (1, 2, ..., 10 sentences) from the input paragraph as the prediction. To get random scores of texts shorter than a single sentence (around 30 tokens), we truncated sentences to the desired length (4, 6, ..., 30 tokens). For each length setting, we ran 10 different random seeds and reported the average. The Random line in Figure 2-Left shows the behavior of ROUGE-2 favoring longer texts within 50 tokens.[1] The normalized scores, as shown in Figure 2-Right, clearly indicate the superiority of our

---

[1]Similar trends for ROUGE-1, ROUGE-L, MoverScore, and BERTScore are included in Appendix E.

proposed model over the random baseline.

Table 2 shows the normalized automatic evaluation results. Overall, **Pegasus$_{P+O}$, the text-summarization model with all available information (Paragraph+OCR), achieved the best performance in all three metrics.** Pegasus$_{P+O+B}$, the model using the same information but trained on a better subset of captions (Paragraph+OCR-Better), did not perform well. We hypothesized this was due to half of the test data comprising poor captions (refer to Section 8.2). This was validated by examining performance shifts in different quality beams (Section 8.1) and conducting a human evaluation (Section 7.2). Meanwhile, Reuse$_M$, the Reuse baseline with Mention, outperforms other Reuse baselines. Its performance declined as context sizes grew and shifted.

7.2 Human Evaluation Results

Pilot MTurk Study to Select Top Models. Before the main human evaluation, we ran a pilot study on Amazon Mechanical Turk (MTurk) to identify any apparently underperforming baselines for exclusion in the final study, simplifying the main human evaluations. In this study, we asked MTurk workers to carefully read a figure and select the *worst* figure caption among *(i)* TrOCR, *(ii)* Pegasus$_{P+O}$, *(iii)* Pegasus$_{P+O+B}$, and *(iv)* ground-truth caption. Ninety figures without errors were randomly sampled from our annotated set (*i.e.*, figures from cs.CL arXiv papers in Section 8.2) for the study. For each of the figures, we recruited 20 MTurk workers to judge.[2] We report the number of majority votes (when tied, we counted all captions with the highest votes as the worst) and the average number of votes in Table 3. Results indicated that TrOCR's caption won the majority vote 41 out of 90 times, with its average vote count significantly exceeding others. Hence, we excluded TrOCR from our formal human evaluation.

Main Human Evaluation with Domain Experts. Three Ph.D. students with NLP backgrounds (who are not coauthors) were recruited as human judges, as it is hard for those without basic domain understanding to evaluate captions. This study has been approved by the IRB office of the authors' institute. The same 90 figures used in the pilot

n = 90	#Maj. Votes↓	Avg. Votes↓	T-Test over Avg. Votes		
			Peg$_{P+O}$	Peg$_{P+O+B}$	Caption
TrOCR	41	5.99	<.001***	.006**	.001**
Peg$_{P+O}$	20	4.54	-	.253	.973
Peg$_{P+O+B}$	24	4.93	-	-	.318
Caption	19	4.53	-	-	-

Table 3: The result of the pilot Mturk study. When tied, all captions with the highest votes were counted as the worst for #Majority votes. TrOCR is significantly worse than other approaches when rated by crowd workers.

n = 90	Avg. Ranking↓	T-Test on Avg. Ranking	
		Peg$_{P+O+B}$	Caption
Peg$_{P+O}$	2.152	.016*	.015*
Peg$_{P+O+B}$	1.930	-	.923
Caption	**1.919**	-	-

Table 4: Average ranking of the human evaluation. Pegasus$_{P+O+B}$ was rated significantly better than Pegasus$_{P+O}$ and was at the same level as the ground-truth caption.

MTurk study were used again. We asked the human judges to compare each figure's *(i)* Pegasus$_{P+O}$, *(ii)* Pegasus$_{P+O+B}$, and *(iii)* ground-truth caption. The judges were asked to rank the captions based on how strongly they agreed with this statement: "When I read the paper, this caption can help me understand the message that the figure tries to convey." Figure 5 (see Appendix D) shows the interface the human judges used.

Table 4 shows the results of average ranking (from 1 to 3). Overall, **the ground-truth caption and Pegasus$_{P+O+B}$ were ranked similarly** (1.919 vs. 1.930 with p-value = 0.923). **Humans also favored Pegasus$_{P+O+B}$ over Pegasus$_{P+O}$ significantly** (1.919 vs. 2.152 with p-value = 0.016). This supports our heuristic for automatically determining caption quality based on length and aligns with previous findings that longer captions improve reader comprehension (Hartley, 2003; Gelman et al., 2002). However, we found that the task of caption ranking poses a challenge, as evidenced by the lower correlations between raters, with Kendall's tau values of 0.133, 0.148, and 0.274, and Spearman's rho values of 0.128, 0.156, and 0.317. This highlights the complexity of the task and suggests that scaling human evaluation across domains might be difficult. Different preferences over captions, such as length, could lead to lower agreement among raters.

	Agree	Disagree	Total	Agree Percentage
Helpfulness	184	215	399	46.12%
Image-Text	338	61	399	84.71%
Visual-Desc	64	335	399	16.04%
Takeaway	74	325	399	18.55%

Table 5: Results of the manual annotation. More than 50% of the captions were annotated as unhelpful. (Out of the initial 438 figure captions, we excluded those with extraction or classification errors, *e.g.*, incomplete images, leaving us with only 399 captions.)

8 In-Depth Analysis

We conducted an in-depth investigation focused on two key challenges: *(i)* the common presence of low-quality author-written captions and *(ii)* the lack of clear standards for good captions.

Quality Annotation Procedure. We manually annotated 438 captions in the Computation and Language domain (cs.CL) from the test set. Figure 6 (see Appendix D) shows the interface we used, in which the title, abstract, and PDF file of the paper were shown alongside the target figure's image, caption, and questions. For each caption, we asked the annotators (coauthors) to rate four aspects using a five-point Likert scale:

- **Image-Text.** The caption included named entities or important words/numbers in the figure (*e.g.*, title, legends, labels, etc.).
- **Visual-Description.** The caption included some visual characteristics of the figure (*e.g.*, color, shape, trend, etc.).
- **Takeaway.** The caption explicitly stated the high-level takeaway message or the conclusion that the figure attempted to convey.
- **Helpfulness.** "The caption helped me understand the message that the figure attempted to convey".

The annotated data was consolidated by grouping "Strongly Agree" and "Agree" as "[Agree]" and grouping "Neutral", "Disagree", and "Strongly Disagree" as "[Disagree]". The results of this consolidation are presented in Table 5.

8.1 Challenge 1: Addressing Unreliable Quality of Real-World Data

Table 5 shows that **over 50% of the author-written captions in arXiv cs.CL papers were deemed unhelpful**. High unhelpful caption proportion may skew evaluation results, particularly for automatic evaluations that compare generated text to human-written captions. To address this, we

	#Sample	Peg$_{P+O}$	Peg$_{P+O+B}$	Caption
Helpful	55	2.176	1.970	**1.855**
Unhelpful	35	2.114	**1.867**	2.019

Table 6: Human ranking results (lower is better) on helpful and unhelpful beams. Pegasus$_{P+O+B}$ received better rankings in the unhelpful beam.

evaluated models on different quality beams using the 399 annotated figure captions shown in Table 5. The captions were divided into the "helpful beam" (184 captions rated [Agree]) and the "unhelpful beam" (215 captions rated [Disagree]).

Automatic Evaluation Over Beams of Different Quality. To validate the effect of low-quality captions, we re-performed the automatic evaluation for the helpful and unhelpful beam sets. Figure 3 shows the Normalized ROUGE-2 and MoverScore scores for each model in the helpful and unhelpful beam sets.[3] Most models performed better in the unhelpful beam, except Pegasus$_{P+O+B}$, which had better scores in the helpful beam. Pegasus$_{P+O+B}$ was trained on captions with more than 30 tokens. This result suggests that improving training data quality, such as by using only longer captions, can positively impact the model's behavior and result in a better generation of helpful captions.

Human Evaluation Over Beams of Different Quality. We also re-evaluated human scores for both the helpful and unhelpful beams. The human evaluation in Section 7.2 only covered 90 figures, with 55 in the helpful beam and 35 in the unhelpful beam. Table 6 shows the results. On average, **Pegasus$_{P+O+B}$ (1.867) was ranked better than author-written captions (2.019) in the unhelpful beam, in which machine-generated captions were preferred by human judges 22 out of 35 times.** The results suggest that, with careful training data quality control, when author-written captions are not very helpful, machines could potentially generate better captions.

8.2 Challenge 2: What Constitutes a Good Figure Caption?

We calculated Pearson correlations (Rodgers and Nicewander, 1988) among the four aspects using raw five-point Likert ratings. The results are shown in Table 7. **The highest correlation was found between Takeaway and Helpfulness, suggesting that a high-quality caption accurately captures**

---

[3]In addition, ROUGE-1, ROUGE-L, and BERTScore scores can be found in Figure 12 in Appendix E.
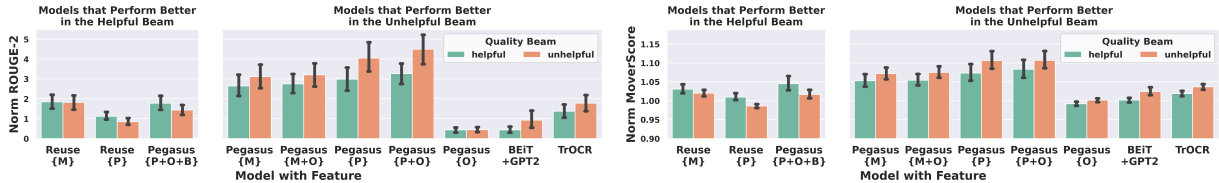
Figure 3: Normalized ROUGE-2 and MoverScore for beams of different quality. Most of the generation models performed better in the unhelpful beam, suggesting that output may be more similar to bad captions. Only the model trained with **better** captions learned to generate good captions by showing a better score in the helpful beam.

	Image-Text	Visual-Desc	Takeaway	Length
Helpfulness	0.206	**0.523**	**0.686**	<u>0.383</u>
Image-Text	-	0.177	0.186	0.248
Visual-Desc	-	-	**0.625**	**0.535**
Takeaway	-	-	-	**0.514**

Table 7: Pearson correlations between different aspects. We used the row scores (five-point Likert scale) to compute the correlation. **Strong correlation** ($\geq$0.5) and <u>medium correlation</u> (0.3 to 0.5) are highlighted. Helpfulness is highly correlated with Visual-Description and Takeaway and is moderately correlated with Length.

the main message of the figure. There were also strong correlations between Helpfulness, Visual-Description, and Takeaway, indicating that a good caption effectively conveys visual information and summarizes the main message. However, Table 5 shows that only 16.04% and 18.55% of the captions described the visual characteristics and the takeaway message, respectively.

A moderate correlation between Helpfulness and Length supports previous research findings that longer captions are generally more helpful for readers (Hartley, 2003; Gelman et al., 2002).

8.3 Caption Length Distribution

Throughout this work's development, the length of captions emerged as a consistent issue. Despite existing literature indicating the benefits of longer captions for readers (Hartley, 2003; Gelman et al., 2002), space limitations often leave authors with no option but to craft shorter captions. To shed some light on this aspect and offer insight for future research, we analyzed the lengths of both author-created and machine-generated captions. We used Kernel Density Estimate (KDE) plots to investigate the distribution of caption lengths across different models and domains. As shown in Figure 4a, the majority of models demonstrate a common peak at 10 tokens, while Pegasus$_{P+O+B}$ presents a significant deviation with a peak near 30 tokens. Figure 4b presents the distribution of helpfulness

Information	Image-Text	Visual-Desc	Takeaway
#Tokens	0.181	0.428	0.357
Percentage	0.099	0.279	0.210

Table 8: Correlations between the amount of missing information from Paragraph and the quality aspects. The missing information is related to visual descriptions and takeaway messages.

scores, derived from quality annotation data (see Section 8.2). Captions rated with a maximum helpfulness score of 5 show a peak at 35 tokens. We can also see a clear shift in caption length with higher scores. In Figure 4c, we dug into the top 10 category taxonomy from arXiv. This figure suggests that a higher portion of the captions in cs, math, stat, and eess are shorter (10 tokens); while the rest of the categories (cond-mat, quant-ph, q-bio, *etc*) have higher probabilities for longer captions. However, within the cs domain (Figure 4d), the top 10 subcategories do not show significant differences regarding caption length distribution.

9 Discussion

Is Text Really All You Need? Our results demonstrate that summarizing figure-mentioning paragraphs is sufficient to generate captions, as shown by the similar scores of Pegasus$_P$ and Pegasus$_{P+O}$ in Table 2. Adding OCR had limited impact. Furthermore, in a recent study of scientific figure captioning conducted by Yang *et al.* (Yang et al., 2023), the best-performing model only considered figure-mentioning paragraphs and OCR tokens– note that their OCR tokens were visual features– without taking the figure's imagery into account. These results raise an interesting question: Do we need visual information at all? What for? The token alignment study (Section 5) showed that 75.19% of the caption information could be found in the Paragraphs, meaning 24.81% of the information was missing. Understanding this missing information could help improve the models' performance.

(a) All the examined generative models.

(b) Helpfulness scores in 5-point Likert scale.

(c) Top 10 frequent arXiv categories.

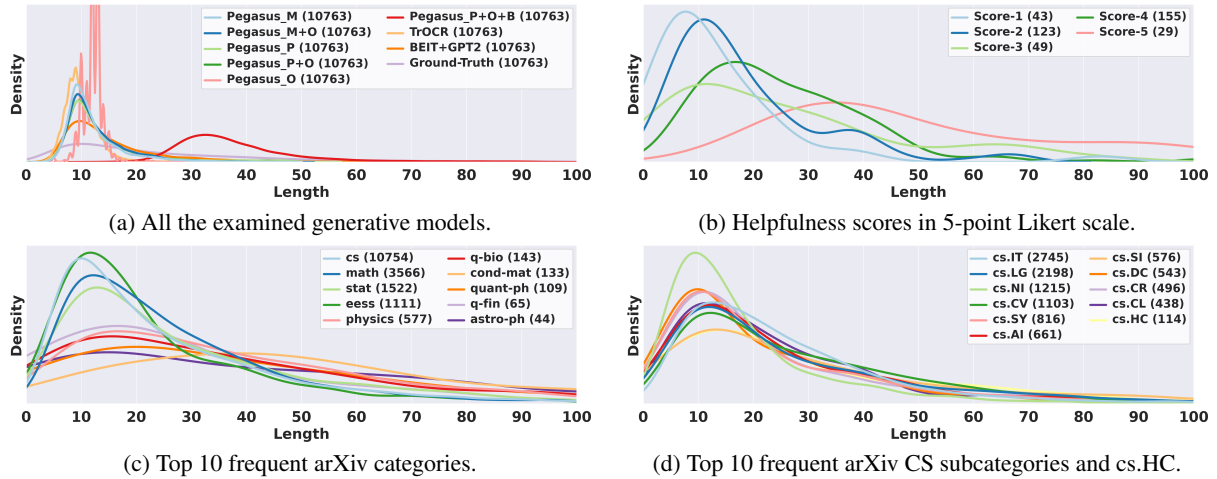(d) Top 10 frequent arXiv CS subcategories and cs.HC.

Figure 4: The KDE plot illustrates diverse caption length distributions among models, the relationship with helpfulness, and variations across arXiv categories. The sample numbers are included in the legend.

Thus, we calculated the correlation between the amount of missing information and three aspect ratings (image-text, visual-description, and takeaway) in the quality annotation data (Section 8.2). The missing information was quantified as the number or percentage of tokens without aligning to any tokens in figure-mentioning paragraphs. Table 8 demonstrates a positive correlation between the extent of missing information and visual descriptions and takeaway messages. This suggests that incorporating visual descriptions (*e.g.*, "dashed line," "red line") is key to enhancing performance by filling in the gaps in information not covered by the article's text. Furthermore, the strong correlation between Helpfulness and Visual-Description in Table 5 also indicates that including image information is necessary for writing good captions. It should be noted that OCR is only capable of capturing image texts (*e.g.*, labels, legends) and not visual element information (*e.g.*, "dashed line"). A promising future direction is developing a multimodal model that can effectively incorporate both image and text.

What is the *Best* Length for Captions? Our research indicates that filtering shorter captions can facilitate the generation of more helpful captions. However, the resulting captions tend to be longer than usual, as shown in the Pegasus$_{P+O+B}$ shift to the right in Figure 4a. This raises a question: Is it fair to compare short and long captions on usefulness, given that longer captions inherently contain more information? While our automatic evaluation addressed this by implementing length normalization, our human evaluations and quality annotations did not specifically instruct the annotators to

consider caption lengths. Nevertheless, we argue that even if we asked annotators to consider caption lengths while identifying helpful captions, the "ideal" caption length would differ among annotators due to multiple factors. For example, as shown in Figure 4c, the length distributions of captions vary across domains. The low inter-agreement from our human evaluation (see Section 7.2) also suggests that personal preferences could influence ideal caption length (Lundgard and Satyanarayan, 2021). Moreover, the ideal length could also be dictated by the context: writers might favor shorter captions due to page constraints, while readers might prefer longer but informative ones (Stokes and Hearst, 2022; Sun et al., 2019). To tackle this issue, a potential future direction could be enabling models to generate captions of diverse lengths to suit different users and contexts.

10 Conclusion and Future Work

This work presented a new perspective on automatic figure captioning, demonstrating that a language-based approach, *i.e.*, summarizing figure-referring paragraphs, can outperform conventional vision-based methods. Our analysis further showed many unhelpful captions in arXiv papers, highlighting data quality's impact on captioning performance. This work lays the groundwork for further research, including exploring new data selection, revision, and augmentation strategies to mitigate the effects of low-quality data, developing new evaluation methods, and creating more robust models that better handle noisy data. We also aim to expand the technology's scope to cover a wider variety of figures and article types.

Acknowledgements

Limitations

Although our proposed methods have been shown to be effective, we are aware of several limitations. First, our approach requires mentions in order to produce captions, but it is not always easy to automatically identify the mentions for a given figure in real-world data. There were 18.81% of figures in the original SCICAP that did not have any identified mentions, which we excluded from this work. Many factors contributed to the gap, including errors caused by upstream components such as image extraction or image type classification (*e.g.*, table), unexpected figure index formats (*e.g.*, "Figure VIII", "Figure C·1","Fig.Fig. 4(b)"), PDF parsing errors, or the figure never being mentioned in the paper. Second, our method uses texts instead of images as the primary information source, so, naturally, it inherits all the constraints of text. Our method can not capture any visual element in the figure that the text never mentioned; it struggles when the text is poorly written. Finally, this paper focused on non-compound line charts in arXiv papers; the human evaluation only focused on NLP papers. More research is needed to examine the generalizability.

Ethics Statement

We consider the proposed technology to impose little risk to readers, as it only summarizes what has already been presented in the paper. However, when the generated caption contains some inaccurate information, it could mislead readers. Furthermore, the proposed technology has the nature of neglecting visual content, which might have an impact on the accessibility of figure captions.

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 331–338, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Constance M. Biegel and Prashant V. Kamat. 2019. Ten tips for capturing figures with captions. *ACS Energy Letters*, 4(3):637–638.

Charles Chen, Ruiyi Zhang, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. 2019. Neural caption generation over figures. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 482–485, New York, NY, USA. Association for Computing Machinery.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. 2020a. Figure captioning with relation maps for reasoning. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1526–1534.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020b. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020c. Logic2Text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2096–2111, Online. Association for Computational Linguistics.

Marc Corio and Guy Lapalme. 1999. Generation of texts for information graphics. In *Proceedings of the 7th European Workshop on Natural Language Generation EWNLG'99*, pages 49–58.

Seniz Demir, Sandra Carberry, and Kathleen F McCoy. 2008. Generating textual summaries of bar charts. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 7–15.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob

Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Mortaza Doulaty, Oscar Saz, and Thomas Hain. 2015. Data-selective transfer learning for multi-domain speech recognition. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2897–2901. ISCA.

Charles G Durbin. 2004. Effective use of tables and figures in abstracts, presentations, and papers. *Respiratory care*, 49(10):1233–1237.

Massimo Fasciano and Guy Lapalme. 1996. Postgraphe: a system for the generation of statistical graphics and text. In *Eighth International Natural Language Generation Workshop*.

Andrew Gelman, Cristian Pasarica, and Rahul Dodhia. 2002. Let's practice what we preach: turning tables into graphs. *The American Statistician*, 56(2):121–130.

Li Gong, Josep Crego, and Jean Senellart. 2019. Enhanced transformer model for data-to-text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 148–156, Hong Kong. Association for Computational Linguistics.

google research. 2022. Python rouge implementation. https://github.com/google-research/google-research/tree/master/rouge.

James Hartley. 2003. Single authors are not alone: Colleagues often help. *Journal of Scholarly Publishing*, 34(2):108–113.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

JaidedAI. 2022. Easyocr. https://github.com/JaidedAI/EasyOCR.

Zhuoxuan Jiang, Ziming Huang, Dong Sheng Li, and Xian-Ling Mao. 2019. Dialogact2vec: Towards end-to-end dialogue agent by multi-task representation learning. *CoRR*, abs/1911.04088.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *Sixth International Conference on Learning Representations Workshop*.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

kermitt2. 2022. Grobid. https://github.com/kermitt2/grobid.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

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.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083.

A. Mahinpei, Z. Kostic, and C. Tanner. 2022. Linecap: Line charts for data visualization captioning models. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 35–39, Los Alamitos, CA, USA. IEEE Computer Society.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

microsoft. 2022. Blingfire. https://github.com/microsoft/BlingFire.

Vibhu O Mittal, Johanna D Moore, Giuseppe Carenini, and Steven Roth. 1998. Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24(3):431–467.

Priscilla Moraes, Gabriel Sina, Kathy McCoy, and Sandra Carberry. 2014. Generating summaries of line graphs. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 95–98.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Obioma Pelka, Asma Ben Abacha, Alba García Seco de Herrera, Janadhip Jacutprakart, Christoph M Friedrich, and Henning Müller. 2021. Overview of the ImageCLEFmed 2021 concept & caption prediction task. In *CLEF2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania. CEUR-WS.org.

Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, Joel Chan, Ryan A Rossi, Sana Malik, and Tak Yeon Lee. 2021. Generating accurate caption units for figure captioning. In *Proceedings of the Web Conference 2021*, pages 2792–2804.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169.

Joseph Lee Rodgers and W. Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Nicolas P. Rougier, Michael Droettboom, and Philip E. Bourne. 2014. Ten simple rules for better figures. *PLOS Computational Biology*, 10(9):1–7.

Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. 2016. Figureseer: Parsing result-figures in research papers. In *European Conference on Computer Vision*, pages 664–680. Springer.

Chase Stokes and Marti Hearst. 2022. Why more text is (often) better: Themes from reader preferences for integration of charts and text. *The NLVIZ workshop at IEEE Transaction on Visualization and Graphics conference*.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Jie Wang and Xinyan Zhao. 2019. Theme-aware generation model for chinese lyrics. *arXiv preprint arXiv:1906.02134*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. 2023. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Jiawen Zhu, Jinye Ran, Roy Ka-Wei Lee, Zhi Li, and Kenny Choo. 2021. AutoChart: A dataset for chart-to-text generation task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1636–1644, Held Online. INCOMA Ltd.