ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization of Long and Short Summaries

Raian Rahman[†,*], Rizvi Hasan[†], Abdullah Al Farhad[†], Md Tahmid Rahman Laskar[‡,◇],
Md. Hamjajul Ashmafee[†], Abu Raihan Mostofa Kamal[†],

[†] Department of Computer Science and Engineering, Islamic University of Technology, Bangladesh
[‡] School of Information Technology, York University, Canada
[◇] Dialpad Canada Inc., Canada

Abstract

Automatic chart to text summarization is an effective tool for the visually impaired people along with providing precise insights of tabular data in natural language to the user. A large and well-structured dataset is always a key part for data driven models. In this paper, we propose ChartSumm: a large-scale benchmark dataset consisting of a total of 84,363 charts along with their metadata and descriptions covering a wide range of topics and chart types to generate short and long summaries. Extensive experiments with strong baseline models show that even though these models generate fluent and informative summaries by achieving decent scores in various automatic evaluation metrics, they often face issues like suffering from hallucination, missing out important data points, in addition to incorrect explanation of complex trends in the charts. We also investigated the potential of expanding ChartSumm to other languages using automated translation tools. These make our dataset a challenging benchmark for future research.

Keywords: Automatic Chart Summarization, Low Resource NLP, Tabular data representation

1. Introduction

Automatic chart summarization is a task where the goal is to describe important data points and trends in a chart in natural language. Chart summaries are helpful to better interpret the chart, making it useful for the visually impaired people as well as to improve the performance of different information retrieval algorithms [1–3].

Scarcity of large scale well defined datasets with chart image, metadata and well described summaries is a major challenge in automatic chart summarization. To our best knowledge, there are only four datasets [1, 4–6] available for the chart to text summarization task, making this task a low resource problem. Among these datasets, three of them [1, 6, 7] contain chart images with metadata and well defined summaries while the SciCAP dataset [5] only contains chart images and captions.

In this work, we address the scarcity of public datasets in the automatic chart summarization task. We propose "**ChartSumm**", a large scale dataset for chart to text summarization comprising of $84,363$ chart images with corresponding chart metadata and summaries. (see Figure 1 for an example). We also propose two test sets based on the summary length. In this paper, our major contributions are summarized below:

(i) Proposing a new benchmark dataset for the automatic chart summarization task. To our best knowledge, our ChartSumm dataset is currently the largest dataset proposed for this task. Meanwhile, we also introduce two different test sets to separately compare the performance on generating short and long summaries.

(ii) Conducting a series of experiments using strong baselines to demonstrate how models trained on our dataset have better generalization capability than other existing datasets. In addition, we identify the limitations of the state-of-the-art models in our proposed dataset.
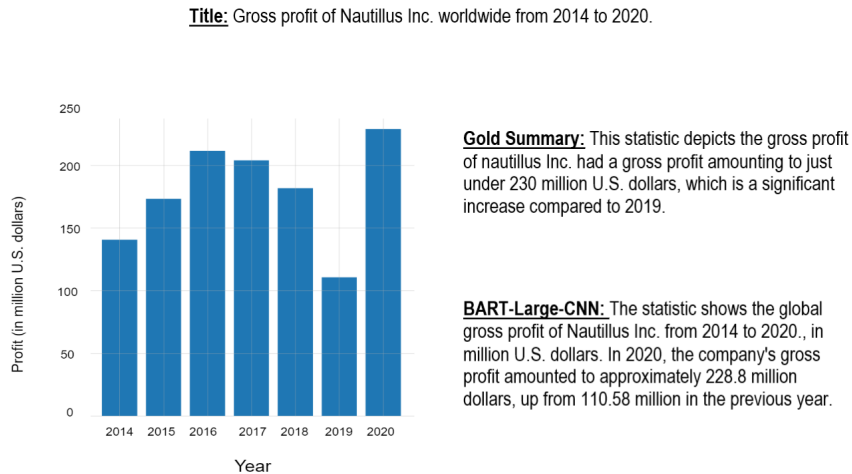
---

[*]raianrahman@iut-dhaka.edu

Title: Gross profit of Nautillus Inc. worldwide from 2014 to 2020.

Gold Summary: This statistic depicts the gross profit of nautillus Inc. had a gross profit amounting to just under 230 million U.S. dollars, which is a significant increase compared to 2019.

BART-Large-CNN: The statistic shows the global gross profit of Nautillus Inc. from 2014 to 2020., in million U.S. dollars. In 2020, the company's gross profit amounted to approximately 228.8 million dollars, up from 110.58 million in the previous year.

Figure 1. An example chart-summary pair from our proposed dataset.

Furthermore, we also explore the scope of expanding our dataset to other languages through translation and evaluate the performance in a human-annotated test set in the Bengali language. To our best knowledge, this is the first work that investigated the Chart Summarization task in any languages except English. The dataset and codes are available at https://github.com/pranonrahman/ChartSumm.

2. **Related Work**

Existing Chart-To-Text summarization systems generate summaries from either the chart image [5] or the chart metadata [1, 6, 8]. Before the advent of deep learning, most early work utilized a two stage approach that applied content selection using different statistical tools in the first step followed by generating summaries using pre-defined templates [7, 9]. However, predefined template-based architectures frequently lack generality and fail to capture complex trends in data. In recent years, deep learning-based techniques have gained significant attention [1, 4, 6, 8, 10–16] due to their superior performance over the template-based approaches. Nonetheless, due to the lack of Chart-To-Text summarization datasets, not only that the models proposed for this task require improvements, the generalized effectiveness of these models is also yet to be investigated.

Among the four benchmark Chart-To-Text summarization datasets that are publicly available, the Chart2Text [1] summarization dataset is the first dataset proposed for this task that includes 8,305 samples collected from the Statista[1]. However, the size of this dataset is quite small and so effective data-driven methods cannot be trained on top of that. Later, the SciCAP [5] data was proposed for the chart captioning task from chart images. Thus, it is not suitable for methods that can only generate summary from metadata. The recently proposed AutoChart [7] dataset is based on some predetermined templates and so this dataset does not contain much variance in the chart descriptions. More recently, Kanthara et al., [6] proposed the Chart-To-Text dataset that consists of chart images, with their corresponding metadata and human written descriptions. Though this dataset is currently the largest dataset available for this task containing 44,085 charts that were collected from

---

[1]https://www.statista.com/

Dataset Name	Task	Data Source	Formulation	Summary Type	Example Count
SciCap [5]	Image → Text caption	Scientific Papers	Chart Image Captions	Short captions from scientific paper figures	290,000
Chart2Text [1]	Table → Text description	Statista	Chart Image Metadata Description	Descriptive human written summary	8,305
AutoChart [4]	Table → Text description	Different public repository	Chart Image Metadata Description	Automatic template generated summary	23,543
Chart-To-Text [6]	1. Table → Text description 2. Image → Extracted metadata using OCR → Text description	Statista Pew	Chart Image Metadata Description	Descriptive human written summary from pew and statista	44,085
ChartSumm (ours)	Table → Text description	Statista Knoema	Chart Image Metadata Description	System generated short summary from Knoema and descriptive human written summary from staista	84,363

Table 1. Comparison between existing datasets and our proposed dataset.

the Statista website and the Pew website, our proposed ChartSumm dataset is almost double in size than the Chart-To-Text dataset.

3. **The ChartSumm Dataset**

This section describes how we compile a large-scale dataset consisting of $84,363$ examples from the Knoema[2] and the Statista website for the chart to text summarization task along with their analysis.

3.1. **Dataset Construction**

Knoema: It is a statistical service-based online platform that contains the economic indicator of more than 200 countries. Knoema provides a short description for each statistic generated by its digital data assistant named Yodatai[3] that summarizes basic information about datasets. To construct our dataset, we first crawl over 1,10,000 statistics from Knoema. Then, we filter out the statistics where the source of data is not publicly available, resulting in 43,179 publicly available statistics. Afterward, we collect the chart metadata and their corresponding short descriptive captions. Since the statistics in Knoema are shown with respect to the year, we classify each chart as a simple line chart. The title and caption of the chart are then tokenized while we remove white spaces and newlines using stemming. We also normalize the numerical entities.

Statista: It is also an online platform where statistics on a wide range of topics are published along with a short human-written description of the statistics. Topics in Statista include economics, marketing, industry, and opinion research. For dataset creation, at first, we crawl over $750,000$ available pages in Statista research to collect a list of $41,184$ publicly available charts along with summaries and chart metadata. Then we classify the data into simple and complex charts depending on the number of columns in the chart. Similar to the Knoema dataset, we also apply tokenization and stemming. Since many examples in the Statista dataset did not contain the x_label, we apply the following heuristic rules to automatically classify the x_labels as Year, Month, Day, Quarter, Country, City, and Continent:

- **Year:** If all x values were integers less than 2050 and greater than 1800 we set the x label to "year".
- **Month:** If all x values were names of months, we set the x label to "month".
- **Day:** If all x values were names of days(saturday, sunday ...), we set the x label to "month".
- **Quarter:** If most x values were Q1/Q2/Q3/Q4 and optionally followed by an integer(year) we set x label as "quarter".
- **Country:** If more than 30% x values contained values from the list of all countries collected from Wikipedia, we labeled the x label as "country"

---

[2]https://knoema.com/atlas
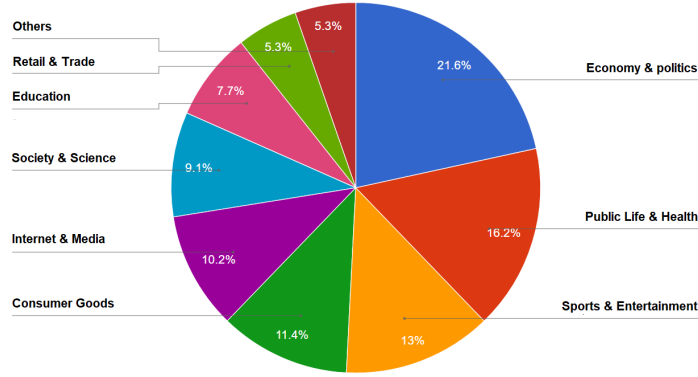[3]https://yodatai.knoema.com/

Figure 2. Topic distribution of ChartSumm

- **City:** If more than 30% x values contained values from the list of cities collected from World City Database, we labeled the x label as "city"
- **Area:** If the x labels contained the names of general areas like continent names, sub continent names, etc, we set the x label as "area".
- **NER:** We also used named entity recognition to identify some other named types such as companies, social medias etc.

For charts where the x_labels could not be automatically determined, the value for the x_label is set to "x_label".

To classify the charts into different types (bar/line/pie), we use ChartReader[4]. The charts are then divided into simple and complex categories. We need to find **missing x labels** since some of the scrapped data have missing x labels. So we manually identify them using the following methods.

3.2. Dataset Analysis

In this section, we analyze our proposed ChartSumm dataset. At first, we compare our ChartSumm dataset with some existing datasets in Table 1. We find the both Chart2Text [1] and Chart-To-Text [6] collected their data from a source called Statista in which the summaries are bit descriptive and longer in length. Whereas our dataset contains both long and short summaries.

Note that our proposed dataset contains line charts, bar charts, and pie charts. For Statista. bar chart is the most common type (64.70% in statista) followed by line chart (33.76% in statista) followed by pie chart (1.54% in statista). For knoema, all charts are line charts. In Figure 2, we show the topic distribution of our dataset. For topic modeling, we perform Latent Dirichlet allocation (LDA) [17] by creating a topic per document and a word per topic model, both of which are based on Dirichlet distributions. We find that our proposed ChartSumm dataset covers a large spectrum of topics including Economy & politics (21.60%), Society & Science (13.03%), Internet & Media (11.43%), Public life & Health (10.42%), Sports & Entertainment (9.14%), Consumer Goods (7.71%), Retail & Trade (5.35%), Education (5.32%), etc.

In Table 2, we show the average cell counts for the tables, as well as character and token counts for summaries and titles, respectively. We find that in terms of average number of tokens, the summaries of simple and complex Statista charts were about 35% and 59.9% longer than the Knoema charts, respectively. The data obtained from each source was

---

[4]https://github.com/Cvrane/ChartReader

Source	Average Cell Count	Average Summary Length		Average Title Length	
		Token	Chars	Token	Chars
Knoema	55.44	34.76	207.69	8.86	57.18
Statista - simple	13.31	46.96	288.68	9.58	63.08
Statista - complex	37.95	55.54	340.19	10.54	67.08

Table 2. Dataset analysis of ChartSumm.

	ChartSumm	Chart-To-Text	Overlaps
Simple	33067	27868	4144
Complex	8338	6943	1194
Total	41405	34811	5338

Table 3. Overlaps between Chartsumm and Chart-To-Text Statista samples.

Train-k	Train-s	Valid-k	Valid-s	Test-k	Test-s	Total
34,503	32,985	4,338	4,101	4,338	4,098	**84,363**

Table 4. Split distribution of ChartSumm.

classified into train, validation, and test sets, following an 80:10:10 split ratio. We show the number of samples in our training, validation, and test sets in table 4. Since, one source of our dataset is Statista that is similar to the Chart-To-Text [18] dataset, we measure the overlaps between the data samples from Statista in our ChartSumm dataset and the Chart-To-Text dataset. For similarity measurement, we first tokenize the captions and then calculate the percentage of matched tokens. We assume two samples are exactly similar when the similarity is greater than 90%. Table 3 shows that only 5,338 captions in our dataset overlaps with the samples from Statista in Chart-To-Text.

4. **Experiments**

In this section, we present the baseline models that we utilize to benchmark the performance in our proposed dataset, followed by the fine-tuning process, the evaluation metrics, and the experimental results.

4.1. **Baselines:**

We use T5-Base [19] and BART [20] as our baselines due to their effectiveness in Chart-To-Text tasks [18]. T5 is a large pre-trained language model trained on multiple sequence-to-sequence tasks. BART is a sequence-to-sequence model pre-trained for the language modeling task using the denoising autoencoder architecture. We fine-tune three variants of BART: (i) BART-Base, (ii) BART-Large-CNN, and (iii) BART-Large-XSUM. We implement all models using HuggingFace [21]. Below we describe our model fine-tuning process.

4.2. **Fine-tuning Process**

We used chart metadata (title, corresponding data table, labels) to fine-tune all four of our pre-trained baseline models. We flattened the table by rows and concatenated it with the caption of the table separated by a separator token. To mimic the pretraining process of T5, we added the prefix: "Summarize chart: " before each example. Figure 3 shows the fine-tuning stages.
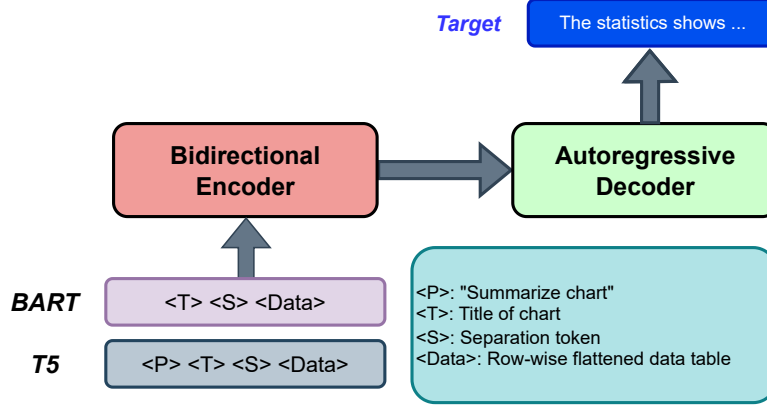
Figure 3. Fine-tuning process of baseline models

All four baseline models were fine-tuned for 3 epochs with a batch size of 8. The initial learning rate during our fine-tuning was $1e - 6$. We used AdamW [22, 23] as our optimizer and cross-entropy as the loss function. We used Google Colab[5] for our experiments.

4.3. **Evaluation Metrics:**

We use five evaluation metrics in our automated evaluation: (i) BLEU [24]: it uses n-gram overlaps between reference text and machine-generated text to determine similarity score, (ii) BLEURT [25]: it evaluates how fluent the candidate is and how well it transfers the reference's meaning (we utilize BLEURT base-128 for our evaluation), (iii) Perplexity: it is a measurement that quantifies how well a probability model predicts a sample (we utilized pre-trained GPT-2 [26] to measure perplexity), (iv) CIDEr: [27] it uses n-grams overlaps and calculates average cosine similarity between the candidate sentence and the reference sentences, to capture the grammatical qualities with richer semantics, (v): Content Selection (CS): it measures how closely the generated text matches the reference documents [28].

4.4. **Experimental Results:**

To evaluate the performance, we use three test sets: (i) *test-k*: denotes ChartSumm test set compiled from Knoema where the summaries are precise and well structured, (ii) *test-s*: denotes ChartSumm test set compiled from Statista where the summaries are descriptive and have a longer length, and (iii) *Chart-To-Text (Statista)*: the Statista version of the test set of Chart-To-Text [18]. We run experiments using the full training sets of both ChartSumm and Chart-To-Text, with additional experiments with the training subsets of Statista and Knoema from ChartSumm.

Our experimental results are shown in Table 5. In terms of BLEURT, CIDER, and CS metrics, we observe that BART-Large models (Bart-Large-CNN and Bart-Large-XSUM) outperform other models in all three test sets, while T5 performs the best in terms of PPL in all test sets. Furthermore, we observe that models fine-tuned on ChartSumm-K always perform better than other models in test sets containing data from Knoema only (except the PPL metric). We again observe similar trends in terms of models fine-tuned on ChartSumm-S. More importantly, all models fine-tuned on our ChartSumm-S dataset outperform the baselines that are fine-tuned on Chart-To-Text even in the Chart-To-Text test set, showing the effectiveness of our proposed dataset. Meanwhile, models fine-tuned

---

[5]https://colab.research.google.com/

Fine-tuned on	Model	BLEU (↑)			BLEURT (↑)			CIDER (↑)			CS (↑)			PPL (↓)		
		Test-S	Test-K	Chart-To-Text	Test-S	Test-K	Chart-To-Text	Test-S	Test-K	Chart-To-Text	Test-S	Test-K	Chart-To-Text	Test-S	Test-K	Chart-To-Text
Chart to Text [6]	T5-Base	18.06	8.89	14.98	-0.2466	-0.9940	-0.0969	2.735	1.143	3.239	68.86	55.2	67.2	10.8742	8.2657	11.3778
	BART-Base	13.3	6.79	12.99	-0.0483	-0.4208	0.0080	3.028	1.304	3.283	63.03	55.43	66.1	11.0813	9.4579	9.9658
	BART-Large-CNN	22.79	6.34	22.98	0.0685	-0.5161	0.0462	2.923	1.051	3.111	83.76	66.49	85.41	9.9658	6.7668	8.248
	BART-Large-XSUM	22.18	6.58	21.35	0.0496	-0.4798	0.0259	2.866	1.067	2.973	83.23	65.61	85.15	9.2262	6.996	8.1973
ChartSumm	T5-Base	26.86	48.84	19.15	0.1942	0.2400	0.0867	3.82	4.149	3.136	77	77.25	71.23	8.1932	4.9536	8.9733
	BART-Base	11.69	21.23	11.58	-0.0265	-0.0259	-0.0305	3.479	3.073	3.134	65.16	70.04	66.61	10.0825	10.5109	11.7486
	BART-Large-CNN	29.91	24.37	22.87	0.1869	-0.0276	0.0910	3.401	2.614	2.992	85.33	76.59	**84.52**	8.7723	11.6479	9.3914
	BART-Large-XSUM	31.73	26.15	24.94	0.1963	0.0030	0.1107	3.551	2.718	3.092	**85.58**	77.2	84.02	9.1612	11.262	9.1076
ChartSumm-k	T5-Base	10.1	**50.74**	7.6	-0.2283	0.2624	-0.2753	2.047	4.31	1.964	61.36	77.63	55.38	14.1075	5.1099	12.8642
	BART-Base	7.29	37.57	6.79	-0.1128	0.2255	-0.1471	1.443	3.865	1.431	54.62	77.17	53.13	10.6306	9.0209	8.7061
	BART-Large-CNN	10.31	29.83	10.21	-0.2537	0.0085	-0.2546	1.594	3.036	1.667	69.62	**81.24**	72.22	24.3287	12.2263	15.1434
	BART-Large-XSUM	7.12	43.27	6.12	-0.1569	**0.3416**	-0.2128	1.672	**4.422**	1.638	58.72	80.06	56.45	18.7177	8.5596	14.015
ChartSumm-s	T5-Base	26.84	14.39	21.11	0.1970	-0.1449	0.1328	3.811	1.656	3.435	76.98	62.95	73.8	8.2032	5.6382	**7.9724**
	BART-Base	21.93	10.23	17.25	0.1582	-0.1622	0.0978	3.349	1.341	3.219	74.78	62.93	71.7	9.4447	6.2375	9.7854
	BART-Large-CNN	**32.27**	7.6	**28.71**	0.1998	-0.3784	0.1628	3.563	1.101	3.235	85.33	66.97	83.6	9.4005	6.9521	9.5459
	BART-Large-XSUM	31.67	10.67	25.42	**0.2832**	-0.1178	**0.2377**	**3.929**	1.271	**3.563**	81.26	65.2	78.97	9.315	5.9528	9.572

Table 5. Experimental results on different test sets. Here, '↑' means 'Higher' and '↓' means 'Lower' value is better.

on Chart-To-Text performs very poorly in ChartSumm-K (about 82.48% lower in terms of best performing T5 model), indicating that the Chart-To-Text dataset is not generalizable to generate summaries that are shorter and precise. To further investigate the performance, we do some error analyses in the following section.

4.5. **Error Analysis and Challenges:**

For error analysis, we randomly sampled 100 instances with their summaries generated by different baseline models.

We notice that in many cases, even though the generated summary is fluent and readable, it contains factually incorrect information and predicted wrong trend in data (see the first example in Table 6. We also notice that models sometimes fail to generate informative summaries while also failing to predict anything about data (see the second example in Table 6). In both Bart-Large models, we also find some cases where the models generate information that is fully irrelevant to the chart (i.e., the Hallucination effect [1, 8, 29]) (see the third example in Table 6).

4.6. **Evaluation of ChartSumm in Other Languages**

To assess the potential for expanding the use of ChartSumm to other languages, we undertook a study where we translated ChartSumm into Bengali and fine-tuned a pre-trained mT5 [30] model to evaluate its performance. To our best knowledge, this is the first study that has explored the task of summarizing charts in languages other than English.

Translation: To translate the training and validation sets of ChartSumm into Bengali, we utilized NLLB [31], which is a state-of-the-art neural machine translation model. To ensure proper evaluation, the test data was translated into Bengali with the assistance of human annotators who are undergraduate students and have proficiency in both English and Bengali.

Baseline: In our study, we employed mT5 [30] as a baseline model due to its efficacy in text summarization tasks in Bengali. We fine-tuned a variant of mT5 which was pre-trained on multilingual XL-SUM [32]. Our baseline model was fine-tuned for 4 epochs using a batch size of 8, with an initial learning rate of 0.000001. The AdamW optimizer was used for the fine-tuning process and cross-entropy was utilized as the loss function.

Evaluation: In our study, we conducted automatic evaluation of the text summarization models using the BLEU, which measures the degree of similarity between the reference and the machine-generated text using n-gram overlaps. However, we were unable to employ other evaluation metrics such as BLEURT and CIDEr for the Bengali language, as they require language-specific models that are not available for Bengali. Therefore, we only used BLEU as an evaluation metric for our Bengali text summarization experiment. We present the outcomes of our experiments in Table 7. It is evident from the results that models fine-tuned on
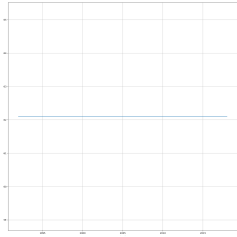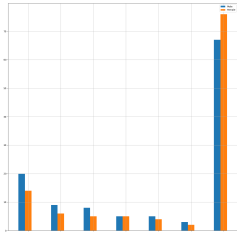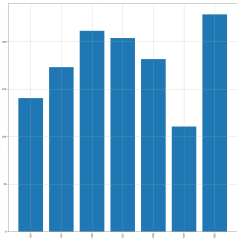
	Gold: Between 1992 and 2018, Georgia renewable surface water remained stable at around 62.1 billion cubic meters per year.	**T5-Base:** In 2018 , renewable surface water for Georgia was 62.1 million cubic meters per year. Though Georgia renewable surface water fluctuated substantially in recent years , it tended to increase through 1997 - 2018 period ending at 62.1 million cubic meters per year in 2018 .	**Comments:** Model predicts factually incorrect information and captures incorrect trend.
	Gold: A survey from April 2020 revealed that during the coronavirus (COVID-19) lockdown about three out of ten Italian adults subscribed to paid digital services, with slightly higher penetration rates among male respondents. 20 percent of male adults subscribed to video on demand platforms, while just 14 percent of female respondents did the same.	**BART-Base:** This statistic shows the number of paid digital subscriptions during the coronavirus outbreak in Italy in 2020, by user gender.	**Comments:** Model fails to generate informative summary.
	Gold: This statistic depicts the gross profit of Nautilus Inc. worldwide from 2014 to 2020 . In 2020 , Nautilus Inc. had a gross profit amounting to just under 230 million U.S. dollars , which is a significant increase compared to 2019 .	**BART-Large-CNN:** The statistic shows the gross profit of Nautilus Inc. worldwide from 2014 to 2020 . In 2020 , the company 's global gross profit amounted to approximately 228.8 million U.S. dollars . The company is headquartered in Seattle , Washington , United States .	**Comments:** Model hallucinated about its headquarter.

Table 6. Examples of generated summaries containing error. Red indicates factually incorrect output, orange indicates failure to capture trend and violet indicates hallucination error

Fine-tuned On	Test-k			Test-s		
	BLEU-1	**BLEU-2**	**BLEU-N**	**BLEU-1**	**BLEU-2**	**BLEU-N**
Full dataset	17.07	6.41	3.22	28.45	12.28	7.69
Knoema	**17.80**	**6.62**	**3.79**	12.02	3.60	1.25
Statista	15.44	5.02	1.73	**28.93**	**12.63**	**7.97**

Table 7. Experimental results on different test sets of ChartSumm translated to Bengali.

ChartSumm-s and ChartSumm-k performed better on their corresponding test sets in comparison to the scenario when evaluated on the combined test set. From this, we can see that the model can perform better in the respective test sets even with machine-generated translations. This opens up the possibility to investigate the performance of ChartSumm in other languages through automatic machine translations. Meanwhile, performance evaluation via human-annotated training data in other languages is also something worth investigating in the future.

5. **Conclusion**

In this work, we present a new large scale benchmark dataset for the automatic chart summarization task to address the low resource problem in such tasks. Our proposed dataset is almost double in size than the existing largest dataset available for this task. Thus, the proposed ChartSumm dataset will serve as a strong benchmark for researchers in this relatively new area of natural language generation. We utilize three BART models and one T5 model as baselines and conduct extensive experiments using various evaluation metrics to identify the challenges in this task. Experimental results showed that models fine-tuned with our proposed ChartSumm dataset could achieve better domain generalization than other existing benchmark datasets. We also explored the possibility of extending ChartSumm to other languages through automatic machine translation. In the future, we would like to extend ChartSumm to a multilingual dataset to address the scarcity of well-formatted datasets in other low-resource languages. We will also study how to incorporate query relevance [33–35], question-answering [36–39], and entity recognition [40–42] capabilities in this task.

References

[1] J. Obeid and E. Hoque. "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*. Dublin, Ireland: Association for Computational Linguistics, Dec. 2020, pp. 138–147. URL: https://aclanthology.org/2020.inlg-1.20.

[2] G. Carenini, C. Conati, E. Hoque, and B. Steichen. "User Task Adaptation in Multimedia Presentations." In: *UMAP Workshops*. Citeseer. 2013.

[3] Z. Li, M. Stagitis, S. Carberry, and K. F. McCoy. "Towards retrieving relevant information graphics". In: *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval*. 2013, pp. 789–792.

[4] J. Zhu, J. Ran, R. K.-w. Lee, K. Choo, and Z. Li. "AutoChart: A Dataset for Chart-to-Text Generation Task". In: *arXiv preprint arXiv:2108.06897* (2021).

[5] T.-Y. Hsu, C. L. Giles, and T.-H. Huang. "SciCap: Generating Captions for Scientific Figures". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3258–3264. DOI: 10.18653/v1/2021.findings-emnlp.277. URL: https://aclanthology.org/2021.findings-emnlp.277.

[6] S. Kanthara, R. T. K. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty. "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization". In: *arXiv preprint arXiv:2203.06486* (2022).

[7] J. Zhu, J. Ran, R. K.-W. Lee, Z. Li, and K. Choo. "AutoChart: A Dataset for Chart-to-Text Generation Task". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., Sept. 2021, pp. 1636–1644. URL: https://aclanthology.org/2021.ranlp-1.183.

[8] L. Gong, J. M. Crego, and J. Senellart. "Enhanced transformer model for data-to-text generation". In: *Proceedings of the 3rd Workshop on Neural Generation and Translation*. 2019, pp. 148–156.

[9] E. Reiter. "An Architecture for Data-to-Text Systems". In: *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*. Saarbrücken, Germany: DFKI GmbH, June 2007, pp. 97–104. URL: https://aclanthology.org/W07-2315.

[10] S. Gajbhiye and M. Lopes. "Template-based NLG for tabular data using BERT". In: Feb. 2021, pp. 1–5. DOI: 10.1109/GHCI50508.2021.9514032.

[11] Y.-T. Hsu, S. Garg, Y.-H. Liao, and I. Chatsviorkin. "Efficient Inference For Neural Machine Translation". In: *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Online: Association for Computational Linguistics, Nov. 2020, pp. 48–53. DOI: 10.18653/v1/2020.sustainlp-1.7. URL: https://aclanthology.org/2020.sustainlp-1.7.

[12] F. Zhou, Y. Zhao, W. Chen, Y. Tan, Y. Xu, Y. Chen, C. Liu, and Y. Zhao. "Reverse-engineering bar charts using neural networks". In: *Journal of Visualization* 24.2 (2021).

[13] K. Dadhich, S. C. Daggubati, and J. Sreevalsan-Nair. "BarChartAnalyzer: Digitizing Images of Bar Charts." In: *BarChartAnalyzer*. 2021.

[14] J. Sreevalsan-Nair, K. Dadhich, and S. C. Daggubati. "Tensor fields for data extraction from chart images: bar charts and scatter plots". In: *Topological Methods in Data Analysis and Visualization VI*. Springer, 2021.

[15] J. Luo, Z. Li, J. Wang, and C.-Y. Lin. "ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021.

[16] C. Chai, G. Li, J. Fan, and Y. Luo. "CrowdChart: Crowdsourced Data Extraction From Visualization Charts". In: *IEEE Transactions on Knowledge and Data Engineering* 33.11 (2021), pp. 3537–3549. DOI: 10.1109/TKDE.2020.2972543.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3.null (2003), 993–1022. ISSN: 1532-4435.

[18] S. Kantharaj, R. T. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty. "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)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4005–4023. DOI: 10.18653/v1/2022.acl-long.277. URL: https://aclanthology.org/2022.acl-long.277.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *arXiv preprint arXiv:1910.10683* (2019).

[20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703.

[21] T. Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

[22] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[23] I. Loshchilov and F. Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. 2018.

[24] M. Post. "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. DOI: 10.18653/v1/W18-6319. URL: https://aclanthology.org/W18-6319.

[25] T. Sellam, D. Das, and A. Parikh. "BLEURT: Learning Robust Metrics for Text Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. URL: https://aclanthology.org/2020.acl-main.704.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[27] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. "CIDEr: Consensus-Based Image Description Evaluation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[28] S. Wiseman, S. M. Shieber, and A. M. Rush. "Challenges in data-to-document generation". In: *arXiv preprint arXiv:1707.08052* (2017).

[29] S. Wiseman, A. Backurs, and K. Stratos. "Data-to-text Generation by Splicing Together Nearest Neighbors". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4283–4299. DOI: 10.18653/v1/2021.emnlp-main.352. URL: https://aclanthology.org/2021.emnlp-main.352.

[30]   L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41.

[31]   M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. "No language left behind: Scaling human-centered machine translation". In: *arXiv preprint arXiv:2207.04672* (2022).

[32]   T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubassir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4693–4703. DOI: 10.18653/v1/2021.findings-acl.413. URL: https://aclanthology.org/2021.findings-acl.413.

[33]   M. T. R. Laskar, E. Hoque, and J. X. Huang. "Domain Adaptation with Pre-trained Transformers for Query-Focused Abstractive Text Summarization". In: *Computational Linguistics* 48.2 (June 2022), pp. 279–320. DOI: 10.1162/coli_a_00434. URL: https://aclanthology.org/2022.cl-2.2.

[34]   M. T. R. Laskar, E. Hoque, and J. Huang. "Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models". In: *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13–15, 2020, Proceedings 33*. Springer. 2020, pp. 342–348.

[35]   M. T. R. Laskar, E. Hoque, and J. X. Huang. "WSL-DS: Weakly Supervised Learning with Distant Supervision for Query Focused Multi-Document Abstractive Summarization". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5647–5654. DOI: 10.18653/v1/2020.coling-main.495. URL: https://aclanthology.org/2020.coling-main.495.

[36]   A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. "Chartqa: A benchmark for question answering about charts with visual and logical reasoning". In: *arXiv preprint arXiv:2203.10244* (2022).

[37]   S. Kantharaj, X. L. Do, R. T. K. Leong, J. Q. Tan, E. Hoque, and S. Joty. "OpenCQA: Open-ended Question Answering with Charts". In: *arXiv preprint arXiv:2210.06628* (2022).

[38]   E. Hoque, P. Kavehzadeh, and A. Masry. "Chart question answering: State of the art and future directions". In: *Computer Graphics Forum*. Vol. 41. 3. Wiley Online Library. 2022, pp. 555–572.

[39]   M. T. R. Laskar, X. Huang, and E. Hoque. "Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 5505–5514.

[40]   M. T. R. Laskar, C. Chen, A. Martsinovich, J. Johnston, X.-Y. Fu, S. B. Tn, and S. Corston-Oliver. "BLINK with Elasticsearch for Efficient Entity Linking in Business Conversations". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July 2022, pp. 344–352. DOI: 10.18653/v1/2022.naacl-industry.38. URL: https://aclanthology.org/2022.naacl-industry.38.

[41]   M. T. R. Laskar, C. Chen, J. Johnston, X.-Y. Fu, S. Bhushan TN, and S. Corston-Oliver. "An auto encoder-based dimensionality reduction technique for efficient entity linking in business phone conversations". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 3363–3367.

[42]   M. T. R. Laskar, C. Chen, X.-y. Fu, and S. Bhushan Tn. "Improving Named Entity Recognition in Telephone Conversations via Effective Active Learning with Human in the Loop". In: *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 88–93. URL: https://aclanthology.org/2022.dash-1.12.