SciGen: Dataset Documentation

Nafise Sadat Moosavi[1], **Andreas Rücklé**[1], **Dan Roth**[2], **Iryna Gurevych**[1]

[1]UKP Lab, Technische Universität Darmstadt
[2]Department of Computer and Information Science, UPenn
[1]https://www.ukp.tu-darmstadt.de
[2] https://www.seas.upenn.edu/~danroth

We use the dataset documentation framework proposed by Gebru et al. (2018).

1 Dataset Documentation and Intended Uses

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. This dataset is created to enable research on reasoning-aware table-to-text generation based on scientific tables. Existing table-to-text generation datasets mostly contain descriptions that are surface-level summaries of the data (Chen and Mooney, 2008; Belz et al., 2011; Lebret et al., 2016; Gardent et al., 2017; Dušek et al., 2018; Koncel-Kedziorski et al., 2019; Radev et al., 2020; Parikh et al., 2020). The generation of descriptions in SciGen, on the other hand, requires performing arithmetic reasoning. As our results show, despite the impressive progress of large pretrained language models on various NLP tasks including generation tasks, their performance is severely limited when the task requires performing arithmetic reasoning. The availability of SciGen opens new research directions in reasoning-aware text generation and its evaluation.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This dataset is created by researchers from UKP Lab from the Computer Science Department at Technische Universität Darmstadt and the Department of Computer and Information Science at UPenn.

Who funded the creation of the dataset? The expert annotations are done voluntarily by one of the authors of the included scientific articles. The selection step—i.e., selecting table-descriptions that contain arithmetic reasoning and removing table-description pairs that contain errors—, and human evaluation were funded by the third parties which we acknowledged in the paper.

Any other comments? N/A

2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. Each instance in SciGen presents a table-description pair. Both tables and descriptions are taken from the articles of the computer science field from arXiv.org. Figure 1 shows a sample table-description pair.

How many instances are there in total (of each type, if appropriate)? SciGen is released in three different settings. The few-shot setting contains 1338 instances. The medium setting contains 18097 instances. The large setting contains 53136 instances.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). The instances in SciGen are mainly from the "Computation and Language" and "Machine Learning" fields of "Computer Science" English articles from arXiv.org. We consider the table-description pairs from "Computation and Language" as in-domain training and evaluation data and those from "Machine Learning" as out-of-domain. Our automatic annotation pipeline can be used to extend the data with any other English articles that are accompanied by their corresponding LaTeX sources. However, the limitation is that our pipeline is developed based on the English language.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. Each instance—i.e., an entry in a json file—represents a table-description and consist of the following fields:

- paper: the title of the corresponding article
- paper-id: the arXiv identifier of the corresponding article
- table-caption: the caption of the table
- table-column-names: the first row of the table, i.e., the header row
- table-content-values: the list of rows, excluding the header row, in the table
- text: the description of the table

Is there a label or target associated with each instance?If so, please provide a description. The target of each instance is the description of the table—i.e., the "text" field described for the above question.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. No, the dataset does not contain missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. Yes. The table-description pairs the belong to the same paper can be identified by both of "paper-id" and "paper" data fields. There is no explicit relation between table-description pairs that belong to different papers.

Are there recommended data splits (e.g., training, development/validation,testing)? If so, please provide a description of these splits, explaining the rationale behind them. Yes. We have provided a fixed training, development, and test splits regarding two different configurations: (1) based on the amount of available training data, and (2) the domain of training vs. test data. The provided splits enable evaluating examined models in two different scenarios: (1) when the amount of training data increases, and (2) when evaluated on in-domain vs. out-of-domain instances.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. A subset of the training data in the medium and large settings that are extracted automatically may contain additional information—e.g., describing the intuitions and reasons for the results of the tables—that cannot be generated by only using the table content and its caption. In addition, the table-description pairs are extracted automatically from LaTeX files using the AxCell tool (Kardas et al., 2020). While extracting tables and texts from LaTeX files is straightforward and accurate, there may be some cases in which the extracted content is noisy. The test set, on the other hand, is annotated and validated by experts and we are not aware of any errors or noise in the test set.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? SciGen is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?If so, please provide a description. SciGen does not contain any confidential data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. SciGen contains table-description pairs with tables that mostly contain numerical values and their corresponding descriptions. There are few tables, which are related to articles about toxicity or emotion detection, in which some of the header cells in the table may contain swear words.

Does the dataset relate to people? If not, you may skip the remaining questions in this section. No, SciGen does not relate to people.

3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. The table-description pairs are all directly selected (observed) from the content of scientific articles.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? The data is collected in two different ways: (1) manual annotations by experts, and (2) automatic annotation extraction. Manual annotations are done in two different steps (a) annotating all the table-descriptions pairs in the articles by one of the authors of the articles, and (b) selecting annotated table-description pairs that require arithmetic reasoning and discarding noisy pairs. Both steps of the manual annotation are done by experts.

For the automatic annotation, we have used the AxCell tool (Kardas et al., 2020) and a set of heuristic rules for extracting table-description pairs from LaTeX sources of scientific articles. As discussed in Section 3.3, we have validated our pipeline by comparing them with expert annotations. To do so, we extract the table-description pairs from those articles for which we have the expert annotations. We then compare the automatically extracted pairs with those that are manually annotated and have reported the results.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The candidate arXiv articles are selected from the period 2010-2020 and mainly from the "Computation and Language" and "Machine Learning" fields of "Computer Science" English articles from arXiv.org.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g.,how much were crowdworkers paid)? The expert annotations of table-description pairs are done voluntarily by one of the authors of the included scientific articles. We paid 10 per hour for the table-description pair selection step, and human evaluation.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time-frame in which the data associated with the instances was created. The manual annotations were collected from January 2020 to June 2020. The automatic annotations are extracted in July 2020. SciGen is self-contained and its content will not change over time.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section. No.

4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section. Yes. As described in Section 3.2, for the manual annotations, 970 table-description pairs were discarded from the manual annotations that were collected in the first step. The purpose of this selection step is to only include pairs that contain arithmetic reasoning and to discard pairs that contain errors. Also, as described in Section 3.3, our annotation extraction pipeline includes a post-processing step that discards a subset of initially extracted pairs based on a set of rules. For instance, we can see from the analysis of Table 3 that the post-processing step reduces the extracted table-description pairs from 950 to 380.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. No. The discarded pairs are those that do not require arithmetic reasoning or contain an error. Therefore, they are not related to the SciGen dataset, i.e., a reasoning-aware data-to-text generation dataset.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point. The postprocessing.py for the post-processing step of automatic annotations is available in the extraction-pipeline directory of the supplementary materials. The data cleaning for the manual annotations was done manually and by experts.

5 Uses

Has the dataset been used for any tasks already? If so, please provide a description We have used ScGen for the experiments of Section 5 for the task of reasoning-aware data-to-text generation.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. `https://paperswithcode.com/dataset/scigen`

What (other) tasks could the dataset be used for? If we also include the whole text of the articles to SciGen, the annotations can be used for the task of relevant content selection. The input will be a table and the text of its corresponding article, and the task is to select all the text spans from the article that describe the table.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms? No, there is minimal risk for undesirable harm from arXiv papers.

Are there tasks for which the dataset should not be used? If so, please provide a description. The dataset should not be used for generating fake scientific articles.

6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?If so, please provide a description. Yes, the dataset is publicly available at `https://github.com/UKPLab/SciGen`. We will update

the repository with human evaluation, the extraction pipeline, and baseline codes, which are already available in the supplementary materials.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? The dataset is distributed on GitHub. We will make the DOI available after finalizing the GitHub repository.

When will the dataset be distributed? The dataset itself is already distributed. The repository will be updated with the human evaluation, the extraction pipeline, and baseline codes after the NeurIPS 2021 camera-ready deadline.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. SciGen is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

7 Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset will be supported, hosted, and maintained by the UKP Lab at Technische Universität Darmstadt.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? All comments and questions about SciGen can be sent to Nafise Sadat Moosavi: moosavi@ukp.informatik.tu-darmstadt.de Other contacts can be found at: `https://www.informatik.tu-darmstadt.de/ukp/ukp_home/staff_ukp/index.en.jsp`.

Is there an erratum? If so, please provide a link or other access point. All changes to the dataset will be announced on `https://github.com/UKPLab/SciGen`.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list,GitHub)? Yes. Any updates to the dataset will be communicated via the GitHub repository, and the UKP Lab website. Potential updates are creating new training settings that contain larger automatically extracted training data.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users All versions of SciGen will be continue to be supported and maintained on `https://github.com/UKPLab/SciGen`. We will post the updates on the GitHub repository.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Filenames	dataset/train/few-shot/train.json dataset/train/medium/train.json dataset/train/large/train.json dataset/development/few-shot/dev.json dataset/development/medium/dev.json dataset/development/large/dev.json dataset/test/test-CL.json dataset/test/test-Other.json
Format	json
URL	`https://github.com/UKPLab/SciGen`
Domain	Scientific articles
Keywords	data-to-text generation, arithmetic reasoning, scientific articles
License	CC BY-NC-SA 4.0

Table 1: Metadata of SciGen

Yes. Others can contact the authors of this paper describing their proposed extension or contribution. We would discuss their proposed contribution to confirm its validity, and if confirmed, we will release a new version of the dataset on GitHub and will announce it accordingly.

8 Metadata

Table 1 presents the metadata of the SciGen dataset.

9 Responsibility

The authors bear all responsibility in case of violation of rights, etc. We confirm that the dataset is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

References

Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 217–226, Nancy, France. Association for Computational Linguistics.

David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. Axcell: Automatic extraction of results from machine learning papers.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Ankur P 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)*. Association for Computational Linguistics.

Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Nazneen Fatema Rajani, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. 2020. Dart: Open-domain structured data record to text generation. *arXiv preprint arXiv:2007.02871*.