WikiDBs: Dataset Documentation

Liane Vogel[1], Jan-Micha Bodensohn[2,1], Carsten Binnig[1,2]
[1]Technical University of Darmstadt, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Darmstadt, Germany

We use the *datasheet for datasets* dataset documentation framework proposed by [1]. The questions are taken from version v8 of the paper.

1 Motivation

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. WikiDBs is created to foster the development of foundation models for relational databases. Currently, there is a lack in large-scale collections of relational databases, most existing datasets contain only individual tables that are not connected by foreign key relationships.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset is created by researchers from the Systems Group of the Computer Science Department at the Technical University of Darmstadt.

Who funded the creation of the dataset? This work has been supported by the BMBF and the state of Hesse as part of the NHR Program and the HMWK cluster project 3AI. It was also partially funded by the LOEWE Spitzenprofessur of the state of Hesse. We also thank DFKI Darmstadt and hessian.AI for their support.

Any other comments? N/A

2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Each instance of the dataset represents one relational database, consisting of the tables of the database in CSV format, a schema description in JSON format, and a visualization in the form of an ERD-Diagram as a PDF file.

How many instances are there in total (of each type, if appropriate)? In total, WikiDBs has 100,000 instances of individual relational databases. In total, the dataset contains around 1,6 million tables.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? It would be possible to create even more than 100,000 databases from Wikidata using our method. As our method is based on property relationships in Wikidata, we uniformly traversed all suitable (i.e. resulting in tables with more than 10 rows) relationships until 100,000 instances were reached. We make our code openly available so that databases tailored to specific use-cases can be created if necessary.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

What data does each instance consist of? Each instance consists of:

- tables: a folder with CSV files, one CSV file per table in the database
- tables_with_item_ids: a folder with CSV files, one CSV file per table in the database, each cell value is a Wikidata Q-ID
- schema.json: a JSON file containing information on every table's column names (our rephrased names and the original names from Wikidata) and datatypes, as well as foreign key connections to other tables in the database
- schema_diagram.pdf: an ERD diagram visualizing the database tables and foreign key connections

Is there a label or target associated with each instance? No, there are no designated labels or targets for each database.

Is any information missing from individual instances? No, each of the 100,000 databases contains the contents mentioned above.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? There are no relationships between different databases in the dataset. Relationships between individual tables per database are made explicit in each schema.json file describing the schema of each database.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes, we provide the training/validation/testing split that we used for our experiments along with the dataset for download. The splits were created by splitting the 20,000 files of the preliminary version submitted for review into 14,000 for training, 2,000 for validation, and 4,000 for testing. We kept the 3,000 databases that were not paraphrased yet in the training set and otherwise split the rest randomly for the three splits.

Are there any errors, sources of noise, or redundancies in the dataset? As our corpus is grounded in Wikidata, the underlying data may potentially be noisy, untruthful, and hard to attribute to individual authors. Since our automatic paraphrasing procedure is based on large language models, it is vulnerable to their well-known weaknesses, including hallucinations and social biases.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? WikiDBs is self-contained. We make the Q-IDs of items from Wikidata available to open up the opportunity to adapt our corpus for a variety of table-based end tasks such as schema matching, entity matching, and deduplication.

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)? All included data is already openly available in the Wikidata knowledge base. Wikidata is actively monitored and moderated, adhering to strict guidelines and policies regarding personal information (`https://www.wikidata.org/wiki/Wikidata:Living_people`). Our used paraphrasing method is unlikely to add additional information beyond the provided inputs, therefore the publication of our dataset does not expose any new personally identifiable information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? Our dataset is grounded in Wikidata, which is actively monitored and moderated. However, there might still be content that does not fulfill the Wikidata guidelines and might be harmful.

Does the dataset relate to people? If not, you may skip the remaining questions in this section. The dataset contains data on individuals that is already openly available in Wikidata. Wikidata adheres to strict guidelines and policies regarding personal information (`https://www.wikidata.org/wiki/Wikidata:Living_people`). We do not add any additional data related to people to the dataset apart from the data from Wikidata.

3 Collection Process

How was the data associated with each instance acquired? All data from the dataset is openly available in Wikidata. We use the inherent triple structure of Wikidata (subject, predicate, object) to build relational tables and foreign key connections between them.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? The dataset is based on the Wikidata dump 'latest-all.json.gz' of May 15, 2024, downloaded from `https://dumps.wikimedia.org/wikidatawiki/entities`. To process the data, it was loaded into a MongoDB database, the database creation is written in Python. For rephrasing, the OpenAI API was used to prompt GPT-4o (*gpt-4o-2024-08-06*).

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? It would be possible to create even more than 100,000 databases from Wikidata using our method. As our method is based on property relationships in Wikidata, we uniformly traversed all suitable (i.e. resulting in tables with more than 20 rows) relationships until 100,000 instances were reached. We make our code openly available so that databases tailored to specific use-cases can be created if necessary.

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)? Only the authors were involved in the data collection process.

Over what timeframe was the data collected? The dataset is based on the Wikidata dump 'latest-all.json.gz' of May 15, 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section. The dataset contains data on individuals that is already openly available in Wikidata. Wikidata adheres to strict guidelines and policies regarding personal information (`https://www.wikidata.org/wiki/Wikidata:Living_people`). As Wikidata is made available under the Creative Commons Public Domain License, no individuals needed to be notified about the data collection.

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)? Yes, the data from the Wikidata dump 'latest-all.json.gz' of May 15, 2024 was reformatted into a format suitable to process for the approach of creating relational databases and loaded into a MongoDB instance.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes, we provide the pre-processed data as a MongoDB export along with our code to do the pre-processing.

Is the software that was used to preprocess/clean/label the data available? Yes, we openly release our code to preprocess the data.

Any other comments? N/A

5 Uses

Has the dataset been used for any tasks already? Initial experiments have been conducted for the tasks of predicting missing table names, column names, and cell values using the dataset within the paper of this submission.

Is there a repository that links to any or all papers or systems that use the dataset? Apart from this submission, the dataset has not yet been used. It is planned to link to papers or systems that use the dataset on a Website of the dataset.

What (other) tasks could the dataset be used for? The dataset can be used to train a foundation model on relational databases. The included Q-IDs, which link every cell value to the corresponding Wikidata item, open up the opportunity to adapt our corpus for a variety of table-based end tasks such as schema matching, entity matching, and deduplication. Additionally, it can be a useful resource for data discovery tasks, like table union and joinability search or table retrieval.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? As our corpus is grounded in Wikidata, the underlying data may potentially be noisy, untruthful, and hard to attribute to individual authors. Since our automatic paraphrasing procedure is based on large language models, it is vulnerable to their well-known weaknesses, including hallucinations and social biases.

Are there tasks for which the dataset should not be used? The dataset should not be used for any tasks that result in unfair treatment of individuals or groups.

Any other comments?

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? The dataset is openly available under CC-BY 4.0 license.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset is uploaded to Zenodo (`https://zenodo.org/records/11559814`). Additionally, we plan to upload it to the HuggingfaceDataset Hub.

When will the dataset be distributed? The dataset will be published when the submitted paper gets published to incorporate reviewer feedback into the final version. Update: It has been published on October 30, 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The dataset will be openly available under CC-BY 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

Any other comments? N/A

7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The dataset will be hosted on Zenodo and the HuggingfaceDatasets Hub.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Comments and Questions can be sent to Liane Vogel (liane.vogel@cs.tu-darmstadt.de).

Is there an erratum? We plan to use Zenodo, and Zenodo allows to upload further versions if error corrections are necessary along with the documentation of what was changed.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If necessary, the authors will fix errors, update the dataset, and upload updated versions to Zenodo.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the 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? Zenodo hosts all versions of the dataset that are ever uploaded.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, we make our source code openly available to make extensions/augmentations, etc., and customized versions of the dataset possible.

Any other comments? N/A

References

[1] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.