1 Datasheets for AIT-QA

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.

Transformers that are frequently pre-trained on open-domain content such as Wikipedia, effectively encode questions and tables from Wikipedia as seen in Table QA datasets such as WikiTableQuestions and WikiSQL to achieve state-of-the-art performance. However, web tables in Wikipedia are notably *flat* in their layout, with the first row as the sole column header. The layout lends to a relational view of tables where each row is a tuple. However, tables in domain-specific business or scientific documents often have a much more *complex* layout, including hierarchical row and column headers, in addition to having specialized vocabulary terms from that domain.

Who created this 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 IBM Research with collaborators from Rensselaer Polytechnic Institute and IIT Bombay.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

No funding was used.

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.

The dataset consists of 515 questions authored by human annotators on 116 tables extracted from public U.S. SEC filings (SEC Filings publicly available at: `https://www.sec.gov/edgar.shtml`) of major airline companies for the fiscal years 2017-2019. The tables and questions are stored in JSONL format.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 515 questions authored by human annotators on 116 tables.

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 questions are generated by humans and not taken from any other larger set.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

JSONL format files with tables and questions in natural language (English).

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each question is binary labeled as "requiring KPI", "using table hierarchy" or "paraphrased".

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.

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.

No, no such relationship or mapping is made available or stored.

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 provide a dev and test set as the dataset is meant to be a zero-shot evaluation.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

All annotation are checked atleast by twice, we hope there are no significant errors. We will accept PR to the github repo and log any changes to the dataset in the CHANGELOG.md.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self contained in the same format as popularly used for WikiSQL and WikiTableQuestions datasets.

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.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how

these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

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 dataset is based on 10-K forms; comprehensive annual reports that publicly traded companies file with the U.S. Securities and Exchange Commission (SEC). For this dataset, we focused on the airline industry and retrieved recent 10-K forms of all 5 airlines included in the Standard & Poor's 500 (S&P 500) stock market index. The covered airlines include: Alaska Air Group (ALK), American Airlines Group (AAL), Delta Air Lines Inc. (DAL), Southwest Airlines (LUV), and United Airlines Holdings (UAL). The 10-K forms were downloaded through the SEC EDGAR online system. Tables were taken from these files and questions were annotated.

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?

Downloaded through SEC EDGAR online system.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

All the tables in 10K reports where provided to the annotators and they were allowed to choose tables to create questions on.

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)?

IBM employees.

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 timeframe in which the data associated with the instances was created.

It was collected in 2020. These are data from 2017-2019 SEC filings. These information are constant.

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 remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

NA

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

NA.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

NA.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

NA.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

NA.

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.

No.

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.

Yes. Available on request.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No, screenshot is provided in the paper appendix for reference.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

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.

No.

What (other) tasks could the dataset be used for?

other relevant tasks to Table QA, such as question generation, question based table retrieval etc.

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.

Are there tasks for which the dataset should not be used? If so, please provide a description.
No.

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.
The dataset will be published for anyone to use under the license: CDLA-Sharing-1.0. See LICENSE.md in Github here: https://github.com/IBM/AITQA

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
It is shared on Github at https://github.com/IBM/AITQA

When will the dataset be distributed?
The Dev set is released. The Test set will be released upon acceptance of the paper.

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.
The dataset will be published for anyone to use under the license: CDLA-Sharing-1.0. See LICENSE.md in Github here: https://github.com/IBM/AITQA

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.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The MAINTAINERS.md lists the maintainers of the dataset and Github repo here: https://github.com/IBM/AITQA

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The MAINTAINERS.md lists the maintainers of the dataset and Github repo here: https://github.com/IBM/AITQA

Is there an erratum? If so, please provide a link or other access point.

Any erratum can be found in CHANGELOG.md of Github repo here: https://github.com/IBM/AITQA

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)?

All annotation are checked atleast by twice, we hope there are no significant errors. We will accept PR to the github repo and log any changes to the dataset in the CHANGELOG.md.

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)? If so, please describe these limits and explain how they will be enforced.

NA

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.

Yes.

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.

Yes, refer to the repo: https://github.com/IBM/AITQA