Supplemental Materials

1 Statement

We license PubTables1M under a CDLAv2 License. The authors, Brandon Smock, Rohith Pesala, and Robin Abraham, bear all responsibility in case of any violation of rights.

2 Data URL

The README with dataset access and reading instructions can be found here: https://pubtables1m.blob.core.windows.net/pubtables1m/README

3 Data hosting plan

Upon acceptance for publication, we will upload the data to the Microsoft Research Open Data repository https://msropendata.com/ for long-term hosting.

4 Datasheet

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.

This dataset was created for the problem of table extraction from unstructured documents. It addresses all three subtasks of: table detection, table structure recognition, and functional analysis. It attempts to address the lack of a single large, high-quality dataset that can be used for training and evaluation of deep learning models across the entire range of different architectures proposed for this task.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Brandon Smock and Rohith Pesala, AI & Research at Microsoft.

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.

There was no additional funding for the creation of this dataset.

Any other comments?

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 main entity type in the dataset is presentation tables from scientific articles. The dataset has a second entity type which are pages from scientific articles that have tables in them.

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

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

There are 947,642 fully-annotated tables and there are 460,589 fully-annotated pages in the dataset.

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 dataset does not contain all possible instances of tables from the set of scientific articles they were extracted from. It is missing a small percentage of the full set of tables and missing a somewhat larger percentage of the full set of pages.

The dataset was assembled using automated processes, particularly a sequence alignment step between the same text in a PDF and an XML version of the document. Quality checks were run at different stages to try to filter out any errors in the data. When the automated process failed, there could have been numerous causes. Some failures would be caused by errors in or differences between the source PDF or XML documents. One main cause of failure attributable to our automated process is that the text in the PDF document had to be extracted as a string, and this string may not always be in the same order as the text in the XML document. The goal for the data is high precision, at the expense of recall.

In terms of representation, we only differentiated between two kinds of tables: simple and complex. The final set has roughly equal representation from both (52.7% complex, 47.3 % simple).

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

The dataset has the same object annotations on both PDF documents, which can be considered raw data, and images, which can be considered a processed version of the raw PDF data.

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

The tables are annotated with object instances that appear within them: rows, columns, cells (grid cells, spanning cells), column headers, projected row headers, and words. These entities have hierarchical relationships. Grid cells lie at the intersection of each row and column. Spanning cells contain grid cells. Projected row headers contain rows. Column headers contain rows. The cells of a table are the spanning cells in it and the grid cells in it that are not contained inside spanning cells. Cells contain words.

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.

The tables are missing row header annotations, which were not provided in the source data that this dataset is derived from. The page images in the dataset are not missing any data for table detection, but there are many other entities that appear within the pages that are not annotated.

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.

It is noted in the dataset which document and which page each table is located on.

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.

We randomly split the source documents into 80/10/10 training/validation/test sets. We then split the pages and tables into training/validation/test sets based on which set their source document is in. We split based on the source documents to ensure no table from the same document appeared in both the training set and either the validation set or the test set. This is for two reasons: 1. Tables in the same document could be very similar, so it could be harmful to generalization to split these into different sets. 2. Different models, one trained on pages, and one trained on tables, might end up being used together. To evaluate the performance of these models when used together, we need to ensure no test data was used for training for either model. This is accomplished by splitting the source data for both models together into training/validation/test, rather than splitting the tables and the pages independently.

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

Yes, there are errors in the data. The data is sourced from annotations from thousands of authors, without which it would be much harder to get so many diverse, realistic samples. But this source data contains both errors as well as annotation patterns that are not consistent with each other.

One of the main purposes for us in creating this data was to introduce methodology to eliminate as many of these errors as possible. We document these efforts in our paper. One way we demonstrate that the data contains high quality annotations is by training deep learning models on it, which we measured to have very high performance on the test set.

Despite our best efforts, there will still be some amount of errors and inconsistencies in the data.

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 processed image data is self-contained, but the original source PDF is linked to. This data is part of the PubMed Central Open Access database (https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/). All of the source data is licensed as either CC BY and CC0, with no restrictions. This is a crucial source of data for many institutions and is expected to remain available indefinitely.

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.

The dataset contains images of pages from scientific articles and it's possible that some of these pages could contain graphic medical-related images.

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.

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.

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.

Any other comments?

It is worth noting that the tables in this dataset are those produced by authors of scientific articles, and while diverse, these may not be representative of tables from all document types.

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.

When we acquired the source data, it was already annotated. That annotation process is a requirement from PubMed for each scientific article to include an XML description of any tables that appear within the article. The papers are peer reviewed and the XML annotations are checked to be present but we are not aware of the review process these annotations go through, if any, before they are accepted.

We originally downloaded several hundred thousand scientific articles along with their with XML annotations in August 2019.

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 original XML annotations are authored and manually verified, but we do not know what the procedure is for determining quality, if any. We describe in our paper several types of checks we put the data through to verify and improve upon their quality.

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

We randomly sampled blocks of thousands of scientific articles at a time, with blocks ordered by PMC ID, to process from the full pool of scientific articles. Some blocks, containing potentially related articles, were not sampled during the random process.

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 original annotators were not compensated as part of the scientific publishing process. Brandon Smock and Rohith Pesala were employed by Microsoft at the time they assembled the derived dataset.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data asso-

ciated 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.

The scientific articles are primarily recent articles, and were downloaded August 2019.

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)?

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.

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.

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).

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.

Any other comments?

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, this is a crucial part of the dataset creation process. Before and after the automated alignment process to create the dataset from the source annotations, we checked for several kinds of errors and filtered out samples that did not pass our quality checks. We describe these in our paper. Some of the removed instances failed the automated alignment process and cannot be included. Others were thrown out after alignment due to errors, including some that were considered outliers (less than 0.1% instances were considered outliers and removed). All of the removed instances could be recovered from the original source data if necessary.

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, the raw XML data is included in the dataset, and the raw PDF data can be downloaded from: https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/.

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

This software is not released as of this writing, but will be officially released with the final paper.

Any other comments?

Uses

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

The dataset is not currently used in any published projects.

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?

The dataset could be used for unsupervised training or transfer learning for other tasks in document layout understanding or document object detection. The images could be used for supervised document layout understanding or document object detection tasks if additional annotations were added to them.

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.

Any other comments?

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 will be publicly released.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The plan is to host the data on the Microsoft Research Open Data repository. It does not have a DOI yet but we will ask that the repository provide one.

When will the dataset be distributed?

Immediately following acceptance for publication.

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.

It will be licensed under the Community Data License Agreement (CDLA) Permissive 2.0: https://github.com/Community-Data-License-Agreements/Working-Drafts/blob/main/CDLA-Permissive-2.0.md.

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.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Microsoft Research.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Email either Brandon Smock (brsmock@microsoft.com) or Rohith Pesala (rohith.pesala@microsoft.com).

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

No.

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)?

Likely yes, if we receive or identify enough corrections we will plan to release a version 2.0 of the data and announce on GitHub.

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.

No.

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, we will maintain prior versions of the dataset on the data 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.

We will accept any corrections for review on a case-by-case basis, through GitHub, but changes to the dataset

will only be officially distributed through major releases and not on an ongoing basis. Others are free to extend the dataset and publish their own versions.

Any other comments?