WikiExtract

Varun Nagpal

April 2024

Motivation

For what purpose was the dataset created?

The dataset was created to facilitate research and development in the field of information extraction and text generation. When talking about text generation, it aims to support tasks in the domain of Table-to-Text or Text-to-Table generation. It also aims to support tasks such as automated infobox generation and updation from Wikipedia page content. The motivation behind this is to enable more efficient and accurate extraction of structured information from unstructured (and structured) text, particularly from Wikipedia articles, which are rich in structured and unstructured data.

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

This dataset has been created by Varun Nagpal, a student of Mehta Family School of Data Science and Artificial Intelligence, IIT Guwahati as part of a course project for the course DA323.

Who funded the creation of the dataset?

No funding involved.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The instances in the dataset represent Wikipedia articles, specifically those related to various people in various sports categories. Each instance consists of two parts: the infobox section and the non-infobox page content. The infobox section contains structured information about the entity, while the non-infobox page content includes unstructured text describing the entity in detail.

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

The dataset consists of 29,051 entities across 11 sports categories. The categories and their entity counts are:

1. Alpine Sking - 340
2. Badminton - 3157
3. Baseball - 1419
4. Basketball - 4768
5. Cricketer - 4945

6. Cyclist - 4945
7. Equesterian - 188
8. Football - 4607
9. Squash - 774
10. Tennis - 3077
11. Table Tennis - 831

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset is a sample of instances extracted from Wikipedia articles related to sports categories. It is not exhaustive and does not cover all possible instances in this domain. However, efforts were made to ensure diversity in terms of the types of sports categories included.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

Each instance consists of text data. The infobox section contains structured data in the form of key-value pairs (possible nested), while the non-infobox page content consists of unstructured text describing the sportsperson in detail.

Is there a label or target associated with each instance?

There is no specific label or target associated with each instance in this dataset because this dataset can be utilized for multiple tasks. The label depends on the task being done, for example if the task is text to table generation, the label is the set of infoboxes.

Are there recommended data splits (e.g., training, development/validation, testing)?

The data splits depend on the type of evaluation we want to do, for example if we want to evaluate OOD generalization, we can make the categories across splits mutually exclusive.

Are there any errors, sources of noise, or redundancies in the dataset?

Some of the page content or infobox content might be missing or the formatting of the structured data might not be correct.

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

The dataset has been completely extracted from Wikipedia. While the dataset in it's current iteration is self-contained, any updates rely on the status of Wikipedia.

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)?

Depends on the wikipedia article.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

Depends on the wikipedia article.

Does the dataset relate to people?

Yes, it is about sportspersons.

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?
Yes

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)?
Depends on the wikipedia article.

Collection Process

How was the data associated with each instance acquired?
The data was acquired from Wikipedia articles.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
The dataset was scraped from the wikipedia articles using Selenium. First, for each category/template the most relevant entities were extracted from the wikipedia lists and then the non infobox page content was extracted directly from the page HTML while the raw infobox content was extracted from the index.php page associated with the article followed by heavy parsing and processing to extract the structure and content.

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

The dataset consists of some of the most visited entities associated with each template and the templates are selected from the larger set of all sports templates on wikipedia.

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)?
The data has the most recent versions associated with the entity pages as of 27th April 2024.

Does the dataset relate to people?

Yes, the dataset relates to sportspersons across multiple sports.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data was obtained from Wikipedia, so the data is about the people but doesnt belong to those people and is free to use.

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.
The extracted infoboxes underwent alot of processing to ensure proper structure while the wikipedia page content are kept raw.

Was the “raw” data saved in addition to the prepro-

cessed/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 raw infobox data was not saved.

Uses

Has the dataset been used for any tasks already?

No

What tasks could the dataset be used for?

The dataset could be used for tasks in the information extraction (tasks like automated infobox generation and updation from page content) and text generation (tasks like text to table or table to text generation) domains.

Distribution

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)

The dataset is available on GitHub (https://github.com/SpyzzVVarun/wiki-dataset)

When will the dataset be distributed?

The dataset is public as of April 27, 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 is distributed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

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

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset will be updated by the creator.

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

Email Address: n.varun@iitg.ac.in

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

The dataset will not be updated on a fixed schedule but might be updated as and when necessary

Will older versions of the dataset continue to be supported/hosted/maintained?

The multiple versions of the dataset will be available on GitHub.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Anyone who wishes to contribute to the data in anyway can contact us at n.varun@iitg.ac.in