Datasheet for Annotated Durative Action Dataset

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

The dataset was created for the purposes of building a model that can extract temporal PDDL code from natural language. Due to the specific nature of this task, no dataset of this exact kind was found online. This is why it was necessary to create this novel dataset by annotating existing temporal PDDL code obtained from Artificial Intelligence and Machine Learning Group.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

The first use of this dataset was in an undergraduate thesis at Queen's University, focused on the extraction of temporal PDDL code from natural language. The code created for this project is freely available on Github, @QuMuLab/temporal-nl-to-pddl.

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

The dataset can be used for any tasks that require either natural language descriptions of a select set of temporal actions, the PDDL code for this select set of actions, or both. A wide range of NLP tasks could potentially make use of this dataset, although if the task is not specifically related to temporal PDDL, than this dataset may be less useful than other datasets due to its small size.

D. Who funded the creation dataset?

The dataset was created by members of the MuLab at Queen's University.

E. Any other comment?

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Each instance corresponds to a single durative action in PDDL.

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

There are 8 instances in total.

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance consists of 5 annotations for a single durative action.

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

The target is the PDDL code for the durative action.

E. 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.

N/A

F. 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.

N/A

G. 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 durative actions are sampled from domains obtained from Artificial Intelligence and Machine Learning Group [1].

H. 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.

N/A

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

Some annotations are incomplete (i.e. do not describe all aspects of the corresponding action).

J. 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.

III. Collection Process

A. 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?

Members of the MuLab created natural language annotations for durative actions from the PDDL code for these actions.

B. 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.

Each instance consists of 5 natural language descriptions of the same durative action. Each of the descriptions for a single action was created by a different member of the MuLab at Queen's University.

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

Each instance is a durative action from a different temporal PDDL domain. Actions were selected manually based on their perceived comprehensibility and complexity. Actions that appeared straightforward to annotate were selected.

D. 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 actions were selected by an undergraduate student at Queen's University responsible, and the annotations were completed by members of the MuLab at Queen's University.

E. 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.

The dataset was created over the period from January 2022 to February 2022.

IV. Data Preprocessing

A. 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 preprocessing was performed.

B. 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.

N/A

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

N/A

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

The procedure of having several members of the MuLab annotate temporal PDDL code achieves the motivations for creating the dataset. The dataset that resulted from this procedure is sufficient for the purposes stated in the first section of this datasheet.

E. Any other comments

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The dataset shall be made publicly available on GitHub, @QuMuLab/temporal-nl-to-pddl.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset shall be released in April 2022 under an MIT licence.

C. Are there any copyrights on the data?

There are no copyrights on the data.

D. Are there any fees or access/export restrictions?

There are no fees or access/export restrictions.

E. Any other comments?

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The MuLab at Queen's University.

B. Will the dataset be updated? If so, how often and by whom?

There are currently no plans to update the dataset, but this could change.

C. How will updates be communicated? (e.g., mailing list, GitHub)

Updates will be communicated through GitHub.

D. If the dataset becomes obsolete how will this be communicated?

N/A

E. Is there a repository to link to any/all papers/systems that use this dataset?

A list of papers/systems that use this dataset shall be maintained in the README file that accompanies the dataset.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Everyone is free to copy and modify the dataset for their own purposes. If anyone would like to modify or extend the original dataset, they can do so through a pull request on GitHub. These changes will be reviewed by members of the MuLab at Queen's University.

VII. Legal and Ethical Considerations

A. 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 formal ethical review processes were conducted.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

C. 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.

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

The dataset does not relate to people.

E. 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.

N/A

F. 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.

N/A

G. 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.

N/A

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

N/A

I. 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.

N/A

J. 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.

N/A

K. 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).

N/A

L. 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.

N/A

M. Any other comments?

REFERENCES