Datasheets for Datasets for Language-complete Abstraction and Reasoning Corpus (LARC)

Motivation

for what purpose was the dataset created?

A: the dataset was created to highlight and understand the difference between how human and machines differ when it comes to communicating and interpreting instructions.

who created the dataset?

A: the dataset was created by researchers from MIT and Autodesk.inc

who funded the creation of this dataset?

A: Yewen Pu is an employee of Autodesk Inc, all other authors are employees / students of MIT. the dataset itself was curated with the following funding object:
Title: Building Machine Common Sense the Human Way Sponsor: DARPA (a sub-award through IBM Thomas J. Watson Research) Sponsor award ID: CW3013540\PO4700205308

Composition

what does an “instance” of the dataset represent?

A: each instance represents a particular “ARC” task and how people instruct each-other on how to solve it using English.

how many instances are there total?

A: 400

does the dataset contain all possible instances or is it a sample?

A: it contains all the annotations that we curated, in that regard, this is not a sample.

what data does each instance consist of?

A: for each task/instance, there are multiple descriptions describing how to complete the task, and for each description, there are multiple attempts of attempting to solve the task using the description. for detail see : https://github.com/samacqua/LARC/tree/main/dataset#readme

is there a “target” or “label” associated with each instance?

A: no, however, for “build” it can either be successful or unsuccessful.

is there any information missing from individual instances?

A: no

are there recommended data splits?

A: no. however in the paper we used a 50-50 split, i.e. first 200 as train and rest 200 as test

is the dataset self-contained?

A: yes, it is entirely self contained and represented as vanilla json

does the dataset contain confidential information?

A: no.

does the dataset contain offensive materials?

A: no.

does the dataset identify any subpopulations (age, gender) ?

A: no.

is it possible to identify individuals from the dataset?

A: no.

does the dataset contain data that might be sensitive?
A: no.

Collection Process

how was each instance collected?
A: see appendix A.1.

what mechanism were used to collect the data?
A: using a web form with a drawing canvas and display boxes for texts

if the dataset is a sample from a larger set, what is the sampling strategy?
A: NA

who was involved in the dataset collection process?
A: Sam Acquaviva (MIT) and Yewen Pu (Autodesk)

over what timeframe was the data collected?
A: Over Spring of 2021

were there any ethical review processes?
A: Yes, both Sam Acquaviva and Yewen Pu underwent human subject training from the IRB of MIT

was the data collected directly or through a third party?
A: Through Amazon MTurk.

were the individuals notified and consent to the collection?
A: Yes

if consent was obtained, were they provided a mechanism to revoke their consent in future?
A: No, it will be impossible to identify who they were

has an analysis of the potential impact of the dataset been conducted?
A: No.

Preprocessing

the “raw” dataset has not been processed. the “summary” dataset contains a subset of the fields of the raw dataset (i.e omitting action sequences). the annotated linguistic dataset split the descriptions into phrases via markers such as period, newlines, etc before 17 linguistic tags are added

Uses

has the dataset been used for any tasks already?
A: not that the authors are aware of beside the original work.

what other tasks could the dataset be used for?
A: we believe this is a good dataset for a range of applications, from AI/ML on building a system that can understand natural programs to psychology and cogsci studies analyzing how humans communicate procedures.

is there anything about the composition of the dataset that might impact its future uses?
A: no.

are their tasks which the dataset should not be used?
A: no.

Distribution

will the dataset be distributed to 3rd parties outside of the creators?
A: yes, the dataset is public.

how will it be distributed?
A: it is hosted on github
https://github.com/samacqua/LARC/

when will the dataset be distributed?
A: it is available now

will the dataset be distributed under any IP or ToU?
A: see https://github.com/samacqua/LARC/blob/main/LICENSE

have any third party imposed IP-based or other restrictions with the instances?
A: no.

do any export controls or other regulatory restrictions apply to the dataset or individual instances?
A: no.

Maintenance

who will be maintaining the dataset?
A: Sam Acquaviva and Yewen Pu

how can the maintainers be contacted?
A: github issues

is there an erratum?
A: not at this time.

will the dataset be updated?
A: see github page for changes

if the dataset relates to people, are there applicable limits on the retention of data associated with the instances?
A: no.

will older versions of the dataset continue to be supported?
A: all versions are on github commits.

if others want to extent/augment/build on the dataset, is there a mechanism for them to do so?
A: currently no, however all source-code for the curation are publicly available on github