Datasets

Descriptions of datasets and how to work with them.

- ASIST Study 3 Dataset
- MultiCAT

ASIST Study 3 Dataset

The ASIST Study 3 dataset was collected in 2022 as part of DARPA's Artificial Social Intelligence for Successful Teams (ASIST) program.

Experiment description

In this experiment, teams of 3 participants conduct urban search-and-rescue missions in Minecraft, with some teams being advised by AI agent advisors.

Each team participates in three 'missions' (or 'trials' -- we will use those terms interchangeably here): one 'training' mission, followed by two 'real' missions.

The preregistration for the experiment describes the motivation for the experiment as well as details on the data collection.

How to access the dataset

Option 1: ASU Dataverse

The dataset is publicly available via the ASU Dataverse Research Data Repository.

However, this option is less than ideal if you need to frequently access different subsets of the data, or work with the data programmatically. This option is recommended only if you do not have SSH access to the copy of this dataset on the lab servers.

Option 2: Access via the lab servers

This dataset is currently hosted on one of the lab's network storage volumes, `mule`. Specifically, the raw ASIST Study 3 data is located in the following directory:

```
/media/mule/projects/tomcat/protected/study-3_2022
```

The `mule` volume is mounted on the `orca`, `kraken`, and `leviathan` VMs (see Compute and Storage), so you should be able to access the data if you have SSH access to any of these VMs and permission to access the data (if you have SSH access to the VM and cannot access the data, please contact Adarsh).

File naming conventions

The general naming convention for the files is as follows:

```
<Completeness>_<Data type>[-Part 1 | _Trial-<Trial ID>_Team-<Team ID>_Member-<Participant ID>_CondBtwn-<Advisor>_CondWin-Vers-<Version number>.<extension>
```

EBNF Grammar

```
filename = [validity]
           "_HSRData_"
           remainder

validity = "Missing" | "Terminated"

remainder =
  "ClientAudio" [part] trial team member condbtwn condwin "Vers-1" "wav"
| "DockerLogs" [part] "_Trial-na_" team "_Member-na" condbtwn condwin "Vers-1" ".tar.gz"
| "MetaData_Study-3.csv"
| "OBVideo" [part] trial team member condbtwn condwin "Vers-1" ".mp4"
| "QCTrialMessages" trial team "_Member-na_" condbtwn condwin version "txt"
| "Surveys" survey_part [ "Fulltext" | "Numeric" ] "Trial-na_Team-na_Member-na_CondBtwn-na_CondWin-na_Vers-07272022." ("csv" | "sav")
| "TrialMessages" [part] trial team "Member-na" condbtwn condwin version ".metadata"
| "ZoomAudio" "_Trial_na" team "_Member-na" condbtwn condwin "Vers-1.m4a"
```

```
| "ZoomAudioTranscript" "_Trial_na" team "_Member-na" condbtwn condwin "Vers-1.vtt"
| "ZoomVideo" "_Trial_na" team "_Member-na" condbtwn condwin "Vers-1.mp4"

survey_part = "0" | "1" | "2" | "3"
part = "-Part" part_number
part_number = "1" | "2"

trial = "_Trial-" trial_id
trial_id = "Training" | trial_number
trial_number = "T000XXX" | ...

condwin = "_CondWin-na"
version = "_Vers-" version_number
version_number = "1" | "2" | "3" ...

condbtwn = "CondBtwn-" condition
condition = "none"  | "Human-01" | "ASI-" performer "-TA1"

performer = "UAZ" | "DOLL" | "CRA" | "USC" | "SIFT" | "CMURI"

team = "_Team-" team_id
team_id = "TM00323" | ...

member = "_Member-" member_id
member_id = "na" | "HumanAdvisor" | "E000726" | ...
```

Completeness

The `Completeness` part of the filename above can take on the following three values:

- `HSRData`: The data is valid. In general, you will want to only use files from this dataset with names starting with `HSRData`.
- `Missing`: Missing data
- `Terminated`: Data from a trial that was terminated early.

File descriptions

There are multiple types of files in the dataset. They are described below.

Metadata

- `HSRData_MetaData_Study-3.csv`: Metadata about the experiment, filenames, etc.

Message bus data

- `*.metadata`: Messages sent on the message bus, one file for each trial. Each line of this file is a JSON object. The messages contain information about participant positions, actions, etc. This also includes automated transcriptions of the participants' dialog done in real time via Google Cloud Speech.

Documentation of the message formats can be found here.

Video recordings

- `*.mp4`: Individual and team video recordings (of the missions).

Example rsync invocation for downloading all the videos (assumes you are able to SSH into `kraken` using the invocation `ssh kraken`):

```
rsync \
  -avP \
  kraken:/media/mule/projects/tomcat/protected/study-3_2022 \
  --include "*/" \
  --include "HSRData_*.mp4" \
  --exclude "*"
```

Audio recordings

- `*.m4a`: Zoom audio recordings, one per team
- `*.wav`: Individual participant audio recordings. These audio recordings are captured via the participants' browser rather than Zoom, in order to have real-time, source-separated audio streams for automated speech recognition.

Survey data

- `HSRData_Surveys*.csv`: Data from Qualtrics surveys filled out by participants.
- `*.sav`: Alternate file format for Qualtrics survey exports

Other data

- `*.tar.gz`: Docker logs from the different testbed components. One `.tar.gz` compressed archive per team.
- `*.txt`: Quality control reports for the `.metadata` files.
- `*.vtt`: Automated transcriptions of the experimental sessions generated by Zoom (one per team).

MultiCAT

MultiCAT webpage