Data Documentation for Master Thesis

A BDI-based Virtual Patient for the Education of Shared Decision Making

Introduction

This document exists as an addition to the master thesis done by T.H. Lie at the TU Delft. It provides information about each of the files stored in the 4TU data centre.

Most of the data analysis was done using R Markdown, which was used to create pdf files that go through each step of the data analysis. These pdf's contain information about raw data, data pre-processing steps and the R scripts. Therefore note that the function of this document is mainly to point out where information of the analysis can be found and what each file contains.

Files Overview

The data analysis focused on three subjects it wanted to assess:

- Usability of the Virtual Patient
- SDM Performance
- System Robustness

The Usability was assessed by acquiring raw data in the form of questionnaire data (see /final-analaysis/questionnaire-analysis/questionnaire-results-final.csv for the raw data). Both SDM performance and System robustness were assessed based on data that was acquired through the system logging conversation information in a MySQL database (see /final-analaysis/questionnaire-analysis/db-results for the raw data).

Usability

The raw questionnaire data can be found at: /final-analaysis/questionnaire-analysis/questionnaire-results-final.csv

The .rmd file that contains the explanation and R code can be found at: /final-analaysis/questionnaire-analysis/questionnaire-data.rmd

The .pdf created from the .rmd which is more readable can be found at: /final-analaysis/questionnaire-analysis/questionnaire-data.pdf

SDM Performance

The SDM performance was assessed by comparing the results of three different input types: speech, transcript and hand annotation.

The .rmd file that contains the explanation and R code can be found at: /final-analaysis/sdm-coverage-analysis/sdm-coverage-analysis.rmd

The .pdf created from the .rmd which is more readable can be found at: /final-analaysis/sdm-coverage-analysis/sdm-coverage-analysis.pdf

The folder /final-analysis/sdm-coverage-analysis contains three subfolders:

- speech-results
- transcript-results
- handannotation-results

Both speech-results and transcript-results contain the same subfolder hierarchy: 4 subfolders for the raw results of each experiment participant containing both the database results as well as the feedback results generated by the system (the latter were not used in the analysis), and 1 subfolder named formatted-results which contains 2 subfolders: item-coverage which contains pre-processed data for assessment of step 2 of the SDM protocol (see /final-analaysis/sdm-coverage-analysis/sdm-coverage-analysis.pdf under the header Subject Coverage for a deeper explanation), and sdm-coverage which contains pre-processed data for assessment of SDM performance as a whole (see /final-analaysis/sdm-coverage-analysis/sdm-coverage-analysis.pdf under the header SDM Performance for a deeper explanation).

handannotation-results contains a excel file (perfect-input.xlsx) which shows each subject found through handannotation and another excel file (sdm-coverage-hand.xlsx) in which shows how the handannotation was used to assess the SDM performance by hand. These hand counted scores were then put into a .csv file (sdm-coverage-hand.csv) to do

further analysis in R (see /final-analaysis/sdm-coverage-analysis/sdm-coverage-analysis.pdf under the header Hand Annotation Results for a deeper explanation).

System Robustness

The system robustness was assessed through three measures: SDM performance robustness, subject classification robustness and question classification robustness.

SDM Performance Robustness

The SDM performance robustness explanation can be found in:
/final-analaysis/sdm-coverage-analysis/sdm-coverage-analysis.rmd
and
/final-analaysis/sdm-coverage-analysis/sdm-coverage-analysis.pdf
under the header Feedback Robustness.

Subject Classification Robustness

The subject classification robustness explanation can be found in the folder:
/final-analaysis/system-robustness-analysis/

To do this analysis some pre-processing had to be done, which is explained in /final-analaysis/system-robustness-analysis/system-analysis.rmd
and

/final-analaysis/system-robustness-analysis/system-analysis.pdf under the header Pre-processing. It makes use of the python file parseRawData.py and the folder /final-analaysis/system-robustness-analysis/csv-results, which both are used to create the folders /final-analaysis/system-robustness-analysis/speech, /final-analaysis/system-robustness-analysis/transcript and /final-analaysis/system-robustness-analysis/handannotation which are used for the analysis in the .rmd file above.

Question Classification Robustness

The question classification is assessed by comparing to what degree a coder and the system would classify questions to a certain question subject. The .csv file at the basis of this analysis can be found in: /final-analaysis/question-analysis/question-analysis.csv

This file contains all of the questions that were asked by each participant. The column questionSubject (System) contains what the system classified and the column questionSubject (Human) contains what the coder classified (the questionType columns were not used in this analysis).

The .rmd file that contains the explanation and R code can be found at: /final-analaysis/question-analysis/question-analysis.rmd

The .pdf created from the .rmd which is more readable can be found at: /final-analaysis/question-analysis/question-analysis.pdf