Plan Overview

A Data Management Plan created using DMPonline

Title: A Machine With Human-Like Memory Systems

Creator: Taewoon Kim

Principal Investigator: Taewoon Kim, Mark Neerincx, Michael Cochez, Piek Vossen, Vincent François-Lavet

Data Manager: Taewoon Kim

Project Administrator: Taewoon Kim, Mark Neerincx, Michael Cochez, Piek Vossen, Vincent François-Lavet

Affiliation: Vrije Universiteit Amsterdam

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: VU DMP template 2021 (NWO & ZonMW certified)

Project abstract:

Although modern machines are very good at answering factual questions (semantic memory), they aren't good at answering their personal questions (episodic memory). In this project, we explicitly train an agent that has both semantic and episodic memory systems, as we humans do. Our experiments show that such an agent is better at locating objects than otherwise.

ID: 89053

Start date: 22-09-2020

End date: 21-09-2024

Last modified: 15-12-2021

Grant number / URL: HI-project-12

Copyright information:

A Machine With Human-Like Memory Systems

0. General information

Document version & date

Version 2.0
Date: 21-Nov-2021

Project title

A Machine With Human-Like Memory Systems

Project summary

Although modern machines are very good at answering factual questions (semantic memory), they aren't good at answering their personal questions (episodic memory). In this project, we explicitly train an agent that has both semantic and episodic memory systems, as we humans do. Our experiments show that such an agent is better at locating objects than otherwise.

Your contact details

Name: Taewoon Kim
phone: 0613777556
website: https://taewoonkim.com/
ORCID: 0000-0003-2892-0194
University: Vrije Universiteit Amsterdam
Faculty / Institute: The faculty of Humanities
Department / Research Group: Computational Linguistics, Department of Linguistics

Please list the other people involved in this project

Mark Neerincx
m.a.neerincx@tudelft.nl
Human-Centered Computing, Delft University of Technology
Michael Cochez
m.cochez@vu.nl
Departement of Computer Science, Vrije Universiteit Amsterdam
Piek Vossen
p.t.j.m.vossen@vu.nl
Vrije Universiteit Amsterdam
Computational Linguistics, Department of Linguistics, Vrije Universiteit Amsterdam
Vincent François-lavet
vincent.francoislavet@vu.nl
Vrije Universiteit Amsterdam
Departement of Computer Science, Vrije Universiteit Amsterdam

The contacts mentioned are my supervisors who have nothing to do with my data. Their job is to supervise me with methodologies and experiments.

Funding organisation & grant number (if applicable)

NWO, HI-project-12

Project code (if applicable)

N/A

Consulted data management expert(s)

N/A

1. Data description

Will you collect and/or process personal data in this project?

- No

The data I collect is mock data. Since it's costly to collect real human data, often times in computer science we use simulators and probability distributions to create mock data. There are no personal data involved here.

Will you use existing data? If yes, what is their source?

For example, the probability distributions that I use to sample instances from the distribution can be considered the "source" of data. Another example is the source of common sense human knowledge. More specifically, I use ConceptNet (https://conceptnet.io/) to generate common sense knowledge. This kind of database is normally open-source and open-data.

Will you collect or produce new data? If yes, please describe how.

Collection and production of the data is the same thing in my case. As stated above, I'll collect the data using the mentioned tools.

What kinds of outputs will you produce in this project? Please describe these data assets.

The data is mock data that an AI will observe while it's interacting with its environment. The simplest case is quadruples (e.g., <Tae's laptop,AtLocation,Tae's desk, now>). Throughout the PhD research, I'll make this environment more complicated by adding additional modalities (e.g., natural language, image, etc.)
The data is processed by the agent itself, using its artificial neural network.

How much digital data storage will your project require?

- 0 - 50 GB

The data itself is very small. The core of the data is mathematics and the API, which I can always query with HTTP requests. This means that the size is nearly negligible.

Will you collect physical data? If yes, please describe these.

No, I don't collect physical data. My data are all ones and zeros.

Will you take measures to ensure data quality? Please describe these, if applicable.

I'll try to make the mock data reflect the real world as much as possible. In order to do so, I have to have a deeper understanding of what part of data can come from what kind of distributions. For example, if I randomly sample an adult Dutch male and measure his height, this is sampling from a Gaussian distribution whose mean value is 183 cm.

2. Legal and ethical requirements, codes of conduct

What legislation applies to your research project? Please tick the relevant boxes for your project.

- General Data Protection Regulation (GDPR)/ Algemene Verordening Gegevensbescherming (AVG)

I had to tick a box so I just chose GDPR.
But in the end, no legislation applies to my data and its collection, as its mathematics and derived from open-source and open-data.

Do you require approval of an ethical committee for this project? If yes, please indicate which ethical committee and whether you have obtained approval for this project.

- No

No ethics play a role in my project.

Will you work with data for which intellectual property and/ or confidentiality are an issue? If yes, please describe.

- No

My data is created by myself, and it will be open-sourced. If I have to publish my data, I'll stamp the MIT license, a permissive free software license.

Do you plan on generating a marketable product from your research project? if yes, please describe

- No

No, my data and project are not for commercial usage.

3. Storage and back-up during the research process

What measures will you take to secure and protect data during the research process? Please describe, for each separate data asset you described for question 1.5, how you will ensure data security, where the data assets are stored & backed up, and who has authorization to access the asset.

The data and the code to generate it will be hosted on Github. It will live there forever. Github has a project to save it data indefinitely. Below is the summary of its project from Wikipedia:
In July 2020, GitHub stored a February archive of the site[54] in an abandoned mountain mine in Svalbard, Norway, part of the Arctic World Archive and not far from the Svalbard Global Seed Vault. The archive contained the code of all active public repositories, as well as that of dormant, but significant public repositories. The 21TB of data was stored on piqlFilm archival film reels as matrix (2D) barcode (Boxing barcode), and is expected to last 500–1,000 years.[95][96][97][98]
The GitHub Archive Program is also working with partners on Project Silica, in an attempt to store all public repositories for 10,000 years. It aims to write archives into the molecular structure of quartz glass platters, using a high-precision laser that pulses a quadrillion (1,000,000,000,000,000) times per second.

Is it necessary to transfer the (physical or digital) data assets to other locations or research partners? If yes, please describe how you secure the file transfer.

- No

It's not necessary. I've been sending my Github repository addresses to my collaborators and this is enough.

4. Data archiving and publishing

Which data assets will be archived and which will be published?

They will be archived and published on Github, with the MIT license.

Where will you archive your data assets?

On Github servers. Perhaps it will be buried in Svalbard, Norway, if I join the GitHub Archive Program.

For how long will the data be available in the archive?

At least about 1,000 years, according to the Github official page:
https://archiveprogram.github.com/

Where will you publish your data assets?

On Github

How will you ensure your data assets get a persistent identifier (e.g. a DOI-code)?

The identifier is the Github repository address.

Will you register your datasets in an online registry other than PURE? If yes, where?

At the moment, I am only considering Github, since it's the best place to do so. But perhaps if there are better alternatives, I'm always up for it.

Are there restrictions to data publishing? If yes, please specify the reasons and list the data assets you do not wish to share publicly.

There are no restrictions.

When will you share the data? If not immediately after completion of the project, please specify the reasons.

At the moment, my Github repository is private, which means no one except me and my collaborators can see. I want to make this public as soon as possible when I submit a paper to a conference.

Please indicate the license and/ or terms of use under which you share your data.

The MIT License:
Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:
The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.
THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

5. Documentation

How will you document your project?

I document most of the thing using markdown language (e.g., README.md) on Github, as it is the most standard practice in my field.

What metadata and documentation will accompany the data assets?

I'll specify what probability distributions I have used, and what other open-source tools and data I used.

What methods, software or hardware are needed to access and use your data?

I've only used x86-64 Ubuntu machine to create and access my data. But any other modern devices can also access it easily.

6. Data management responsibilities and resources

Who will be responsible for management of the data assets during the project? Please specify their name, position, role in the project, and faculty/ institution/ group.

Name: Taewoon Kim
phone: 0613777556
website: https://taewoonkim.com/
ORCID: 0000-0003-2892-0194
University: Vrije Universiteit Amsterdam
Faculty / Institute: The faculty of Humanities
Department / Research Group: Computational Linguistics, Department of Linguistics

Who will be responsible for management of the data assets after completion of the project (e.g. the project lead/ dedicated data manager/ department head)? Please specify their name, position, role in the project, and faculty/ institution/ group.

Name: Taewoon Kim
phone: 0613777556
website: https://taewoonkim.com/
ORCID: 0000-0003-2892-0194

University: Vrije Universiteit Amsterdam
Faculty / Institute: The faculty of Humanities
Department / Research Group: Computational Linguistics, Department of Linguistics

For data that are only available upon request, what methods will be used to handle requests for access and how will data be made available to those requesting access?

The person who's interested in my data will have first seen it on my Github repo. There he/she can just access it or send me a message on Github.

What resources (for example financial and time) will be dedicated to research data management? Please estimate their cost.

It only costs my labor force and electricity for computation.
So far it took me about 3 months. But it's not done yet, since I'll update the data / code in the next few years.