Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D11C83

Title: A Multi-level Knowledge-driven Customized Dialogue System for General Use

Creator: Daivd Ho - **ORCID:** 0000-0001-8097-4910

Affiliation: The Chinese University of Hong Kong

Principal Investigator: Prof. WONG Kam Fai

Funder: The Chinese University of Hong Kong

Funding opportunity number: n/a

Grant: n/a

Template: CUHK Data Management Plan Template

Project abstract:

Dialogue systems have been widely applied in the applications such as Apple's Siri and Google Voice. It helps us to control our smart devices and to complete tasks, such as making an order in a restaurant or booking a flight ticket with our mobile phones. Unfortunately, there is no dialogue system available for the Small and Middle Enterprises (SMEs) in Hong Kong due to the limited Cantonese dialogue dataset information. There are around 320,000 SMEs in Hong Kong, which constitute over 98% of all business establishments. With the limited resources, there is a high demand for more automation strategies to improve manufacturing and service quality, especially in the restaurant industry. The dialogue system could be a waitress's assistant to help the customer to accomplish their order. Different from the electrical order, the dialogue system can provide the customer with more individual advice and personalized recommendation. But in Hong Kong, limited to the Cantonese dialogue dataset and the newest dialogue system technology, there are no mature technical solutions to make it.

For this reason, we need to publish the first Cantonese knowledge-driven Dialogue Dataset for General Use (CK-DDD) in Hong Kong. In this case, it collects the information in multi-turn conversations with restaurants. The dataset comes from Reddit, and we will crawl the dialogues from representative restaurants concerning data collection policy. The dialogue will be annotated by the labeler and generate dialogue states and dialogue actions. With the labeled data, we will design the dialogue system model to support human-computer

interaction. The dialogue system composes advanced natural language processing techniques, such as pre-trained language models and few-shot learning settings. It will efficiently integrate the data into the natural language understanding and response generation in the conversation between the customer and the dialogue system.

After the above steps, the data will be accessible to the online repository, and we believe the publication of CK-DDD will be a necessary supplement to current dialogue datasets and more suitable and valuable for SMEs of society, such as building a customized dialogue system for each application. Finally, the corpus and benchmark models will be publicly available.

Start date: 10-01-2022

End date: 09-30-2023

Last modified: 01-18-2024

Copyright information:

A Multi-level Knowledge-driven Customized Dialogue System for General Use

- Yes

- Experiment
- Survey

We first choose 10 representative restaurants with different styles in Hong Kong from the Web to build the database. To better imitate the real scenario, we designed 7 primary slots and 14 secondary slots and extracted those slots from the restaurants' information. Then we collected 832 dialogues by developing a website specifically to enable two workers to play the role of customers and the restaurant to talk and exchange information with each other. Meanwhile, the slots and corresponding values were explicitly chosen and filled by the workers during the dialogue collection stage.

- Text

A general summary of the types and the estimated amount of scientific data to be generated and/or used in the research. Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

- Text - .xml
- Others

Data and metadata formats

- Microsoft EXCEL format is used for listing slots information, and data statistics.
- JSON format is used for dialogue data, slots annotations, and goal
- Python format (UTF8-encoded text files with .py extensions) is used for
- The dialogue models will be stored and distributed in binary (PyTorch) format, and their associated processing files (vocabulary, tokenizer,) will be stored in JSON format.
- The corresponding documentation will be stored in Markdown

All the original data described above under "used data" will be stored and managed by our research group. And we will make it publicly accessible once the project is finished. Also, the codes and models of the dialogue system will be preserved and shared openly in a data repository after our product launch.

We will make accessible all the data, codes, and models, as well as a detailed instruction document about how we collect and annotated the data, a technical document about how to use our codes and models, and a research paper to introduce the whole research project.

Users do not need any specialized tools to use our collected dialogue data. But the codes and dialogue models to reproduce our dialogue system will require the Python environment and Pytorch library. Python (https://www.python.org/) and PyTorch (https://pytorch.org/) are all freely accessible for most modern computers and at the moment there are no plans to discontinue their support. More information is available at their respective websites.

The codes, dialogue models, and documentation will be made publicly available through GitHub and the HuggingFace website (widely used within the machine-learning community) under an Apache 2.0 License. The research paper will be put on arXiv.

The collected data, trained dialogue models, and associated code will be findable on Github at https://github.com/.

The PIs for this project will ensure that the data management plan is followed by auditing the project personnel on a monthly basis and monitoring the project through an online project management tool. The RA is in charge of data collection, annotation, and annotated file storage. We hire some workers (number = 6) for annotation, which is required for training supervised machine learning models. The project coordinator is responsible for updating the repository and version record keeping.

- Cloud storage
- Data recovery software
- By value of data
- 5 years
- CUHK Research Data Repository
- No
- Yes
- Upon project completion
- Public
- Via data repository
- No
- readme.txt
- Data Documentation Initiative (DDI)
- No
- Others

No confidential information in the data.

- Others

No security issue involved.

- CUHK

- CC BY-NC

- No

- Principal investigator
- Data creator

- No

- No

Planned Research Outputs

Dataset - "A Multi-level Knowledge-driven Customized Dialogue System for General Use"

Ref.*	Description	Number**	Scale**	Data type	Preserved	Shared
A)	Dialogue data	832	200MB	Json	Permanently	Yes
B)	Meta data	2	2KB	Excel	Permanently	Yes
C)	code	2	20KB	Python file	Permanently	Yes
D)	Dialog model	1	334MB	Binary file	Permanently	Yes
E)	Annotation Instruction documentation	1	4MB	Markdown file	Permanently	Yes
F)	Technical documentation	1	5MB	Markdown file	Permanently	Yes
G)	Research paper	1	4MB	PDF	Permanently	Yes

All the original data described above under "used data" will be stored and managed by our research group. And we will make it publicly accessible once the project is finished. Also, the codes and models of the dialogue system will be preserved and shared openly in a data repository after our product launch.

We will make accessible all the data, codes, and models, as well as a detailed instruction document about how we collect and annotate the data, a technical document about how to use our codes and models, and a research paper to introduce the whole research project.

Users do not need any specialized tools to use our collected dialogue data. However, the codes and dialogue models to reproduce our dialogue system will require the Python environment and Pytorch library.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
A Multi-level Knowledge-driven Customized Dialogue ...	Dataset	2023-08-31	Open	CUHK Research Data Repository GitHub	1 GB	Creative Commons Attribution Non Commercial 4.0 International	Dublin Core	No	No