

SWP 2023/24 Research Plan

Chris Pracht, Finn Hillengass, Lea Kyveli Chrysanthopoulou
12.11.2023

Objective

Our project aims to leverage different prompting architectures to extract Task Oriented Dialogue (TOD) datasets from Large Language Models (LLMs) that are available free of charge, e.g., Llama-2 (7B, 13B or 32B). This has the aim of building high-quality TOD datasets at low cost, which could then be used to train smaller chatbots on, enhancing their performance. To achieve this aim, we intend to deploy a variety of automated metrics, such as GRUEN, DEAM, GRADE and FactScore. Based on the performance of the prompting architectures on the metrics, we then intend to optimize and refine them as needed.

Methods

We generate our TODs via variations of the prompt architecture introduced Labruna et al. (2022) and inspired the insertion of commonsense knowledge through knowledge triplets as detailed by Kim et al. (2023).

We initially perform Dialogue Generation with the **One Shot Approach**, i.e., asking one LLM to generate both the conversational turns of the user and the system. In the initial prompt the model will be given a triplet of dialogue states the to-be generated dialogue is to include, e.g. (hotel, Italian, expensive). Then, the dialogue is generated as described above.

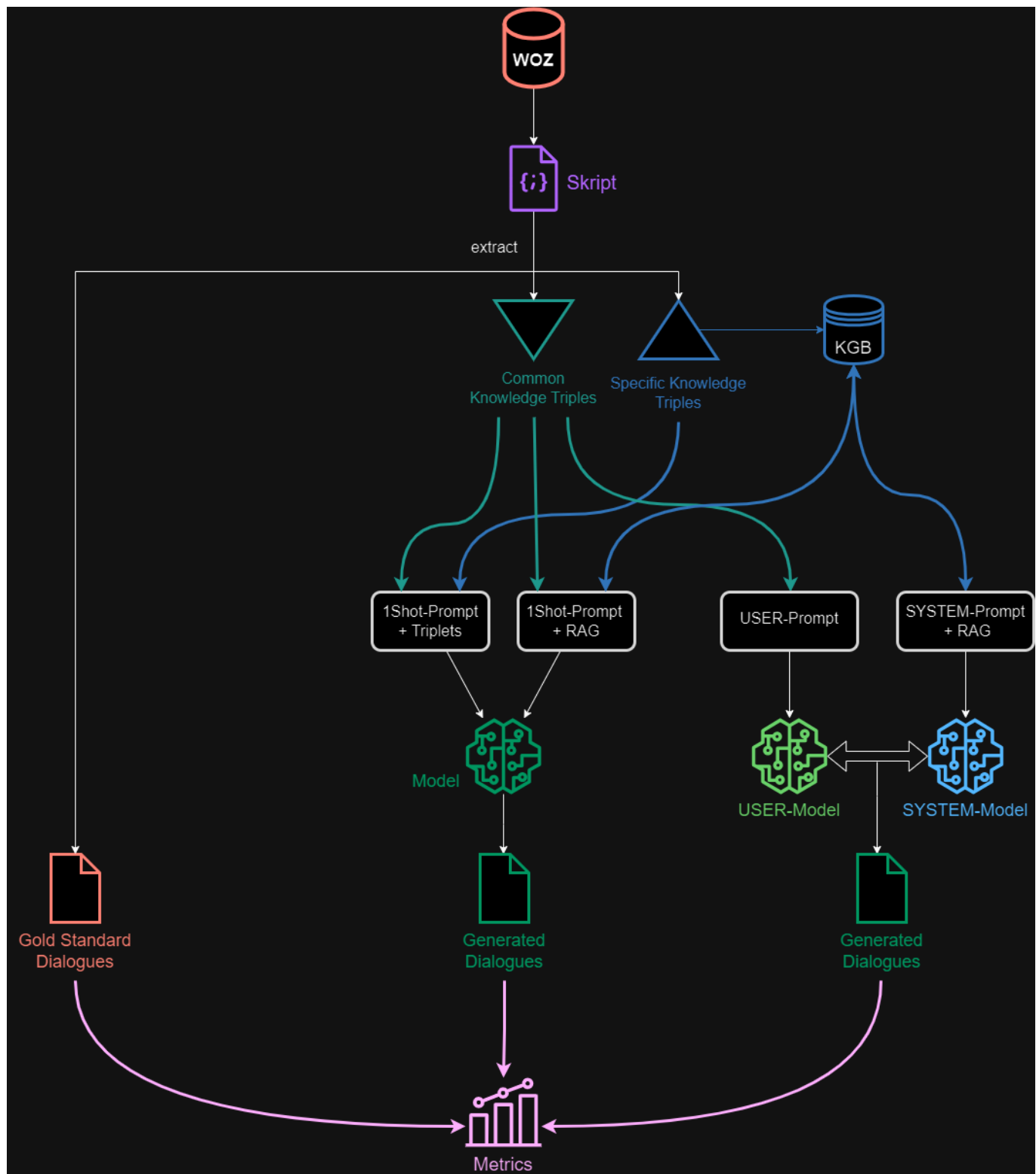
The **labelling of the dialogue states** on the generated dialogue can happen in two manners, the more suitable of which we will decide on during the course of our project:

1. The labels are generated together with the dialogue
2. The dialogue is generated first and then annotated by the model

We furthermore want to implement **Retrieval Augmented Generation** (RAG), i.e., based on the triplet of requirements described above (hotel, Italian, expensive), the model should fetch an option fulfilling these parameters from a knowledge base.

Finally, if there is time, we want to try a **MultiAgent Approach**, in which two models simulate a dialogue between each other, wherein the one takes the part of the system and the other the part of the user in the Task Oriented Dialogue (TOD).

Visualisation of Project Plan



Models

While selecting the LLM for our project, one of our concerns was for it to be open-source, i.e., for it to be available to download and for us to host it ourselves, so we would not have to pay for an API or interact with it manually over its web-interface, as would be the case with a Chat-GPT model.

Of the open-source models available, we decided on Llama-2 (Touvron et al. 2023) in its fine-tuned chat versions. We were able to run the 7B version on the CLuster and locally. By utilizing both GPU-nodes of the CLuster, we should be able to run Llama-2-13B. Should we get access to the bwForCluster, even the 70B model would be possible to use for inference.

We chose Llama-2, as it performs very well in comparison to other open-source models on number of benchmarks (see [figure 1](#)).

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

Figure 1: Llama-2: overall performance on grouped academic benchmarks compared to open-source base models (from the paper by [Touvron et al. 2023](#))

Additionally, a lot of emphasis has been placed on respectful and non-discriminatory language use in the fine-tuning of the Llama-2 models to the chat-versions through **Reinforcement learning from Human Feedback** (RLHF). We consider this to be of great importance in generating data points as potential training data for other chatbots (our [objective](#)). This is why we preferred Llama-2 over Mistral-7B, even though **Mistral-7B** ([Jiang et al. 2023](#)) in its base version outperforms the Llama-2-7B base version and sometimes even the Llama-2-13B base version on a number of benchmarks (see [figure 2](#)).

Model	Modality	MMLU	HellaSwag	WinoG	PIQA	Arc-e	Arc-c	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	Pretrained	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	Pretrained	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	29.0%	69.6%	18.9%	35.4%	6.0%	34.3%
Code-Llama 7B	Finetuned	36.9%	62.9%	62.3%	72.8%	59.4%	34.5%	11.0%	34.9%	31.1%	52.5%	5.2%	20.8%
Mistral 7B	Pretrained	60.1%	81.3%	75.3%	83.0%	80.0%	55.5%	28.8%	69.9%	30.5%	47.5%	13.1%	52.2%

Figure 2: Comparison of Mistral 7B with Llama (from the paper by [Jiang et al. 2023](#))

If we have time, we might attempt a comparison of the performance on our metrics with Mistral-7B as well and explore the adaptability of each model to the dialogue generation tasks.

Dataset

As far as datasets go, we will use the **MultiWOZ 2.2** ([Zang et al. 2020](#)) dataset as a foundational resource. This entails both using the provided dialogues as gold-standard reference-dialogues, as well as extracting the Knowledge-Graph-Triplets with a script for the prompts and to build the Knowledge-Base. We chose MultiWOZ 2.2, as it is a widely used TOD dataset with over 10,000 dialogues annotated for dialogue states, as well as the values the slots take. Additionally, the version 2.2 has significantly reduced the noise present in the earlier versions by correcting erroneous annotations and user utterances.

We will also leverage the prompting architecture described in the **data of Labruna et al. (2023)**, as well as use the dialogues they generated (if available) as preliminary data for the first implementation of our metrics or potentially to compare to the dialogues generated by us at the end of our project.

Evaluation Metrics

An essential part of our project are the **evaluation metrics** we intend to deploy. They are divided in two main categories: **automated metrics** and **manual assessment**.

Our **automated metrics** are the following: We will utilise **GRUEN** (Zhu and Suma 2020) to assess the linguistic quality of the generated dialogues. Specifically, this metric is designed to provide separate scores for the **grammaticality, non-redundancy, focus**, as well as **structure and coherence** of the generated utterances.

Further information regarding **coherence** is provided by **GRADE** (Huang et al. 2020) and **DEAM** (Ghazarian et al. 2022). **GRADE** specifically evaluates topic coherence, i.e., whether the transition between topics are sufficiently coherent and natural, not too abrupt. **DEAM** tests for subtler forms of incoherence through AMR-based semantic manipulations. As both of these metrics were initially developed with Open Dialogues in mind, we will attempt their implementation and then evaluate their relevance to our task throughout the course of our project.

In the section on our [methods](#) we outlined how we intend to label our generated dialogues with dialogue states. In order to evaluate the accuracy of these labels we will employ **Slot Accuracy** and **Joint-Goal Accuracy**.

For the overall evaluation of **semantic similarity** of our generated dialogues to their reference dialogues from the MultiWOZ dataset, we will use **S3BERT** (Opitz and Frank 2022).

Our final automated metric is **FactScore** (Min et al. 2023) with which we will evaluate our **Retrieval Augmented Generation**, i.e., the extent to which the model is able to retrieve the correct information to match with the requests posed by the (modelled) user. This metric is designed to evaluate the factuality of statements generated by LLMs against a knowledge-base.

Our **Manual Assessment** will be predominantly designed to evaluate the same linguistic categories as in **GRUEN** to provide good comparability. It will be divided into two categories:

1. Prompting an LLM (perhaps the same as we used for generating the data or perhaps another, such as GPT-3.5 or GPT-4) to evaluate our generated data for the **GRUEN** metrics, imitating an annotator. If the results for a manual evaluation via an LLM are comparable to that of a human annotator that would be of great interest, as such annotation is very costly to be performed by human annotators.
2. Conducting a regular manual assessment by human annotators, for the **GRUEN** metrics as well.

We intend to measure inter-annotator agreement between the human annotators, the manual LLM annotation, and the automated GRUEN metric by calculating the Spearmann and Pearson Correlation between the different annotators. Thus we can evaluate both the efficacy of the automated annotation, as well as the manual LLM annotation, by operating under the assumption that the human annotation is the gold-standard.

Tools

The tools we will be using for our project are the following:

- **GitLab** for code management and collaboration.
- The **bwForCluster/Helix** and **CoLi-Cluster** as computational resources.
- **ChatGPT** potentially for manual LLM feedback.
- Potentially the **Redis** Database system for our knowledge base (this will be decided during the course of our project).
- **Docker** for the deployment of the database and the processing pipelines.
- **LangChain** as a LLM framework for RAG and Multi-Agent pipeline.

Project Timeline

1. Preparation and Setup (Target Date: 12.12): Parallel Processes
 - Set up knowledge database and API (Chris)
 - Extract knowledge from WOZ slots into Common Knowledge (CK) for Prompt Injection (PI) and Specific Knowledge (SK) for Knowledge Graph Retrieval (KGR)
 - Import into \$DB
 - Embed datapoints with SBERT
 - Implement LangChain Retrieval Augmented Generation (RAG) adapter
 - Select and prepare models and access to computing power (CLuster & bwForCluster) (Finn)
 - Evaluate memory requirements and hardware availability
 - Select model
 - Test workflows
 - Implement automated metrics on preliminary data (Lea)
2. Join Parallel Processes (Target Date: 02.01)
 - Generate dialogues using varied methods (Chris & Finn)
 - One Shot approach
 - RAG approach
 - Apply evaluation metrics for quality assessment (Finn & Lea)
 - Develop processing pipeline for automated evaluation
 - Plotting of results
3. Comparison and Optimization (Target Date: 23.01) (Chris & Finn & Lea)
 - Analyze results from different methods based on the metrics
 - Potentially adjust prompting architecture based on metric insights
4. Given Spare Time:
 - Multi-Agent Approach

References

1. Ghazarian, Sarik, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. 'DEAM: Dialogue Coherence Evaluation Using AMR-Based Semantic Manipulations'. arXiv. <https://doi.org/10.48550/arXiv.2203.09711>.
2. Huang, Lishan, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. 'GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems'. arXiv. <https://doi.org/10.48550/arXiv.2010.03994>.
3. Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. 'Mistral 7B'. arXiv. <https://doi.org/10.48550/arXiv.2310.06825>.

4. Labruna, Tiziano, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. 'Unraveling ChatGPT: A Critical Analysis of AI-Generated Goal-Oriented Dialogues and Annotations'. arXiv. <http://arxiv.org/abs/2305.14556>.
5. Kim, Hyunwoo, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, et al. 2023. 'SODA: Million-Scale Dialogue Distillation with Social Commonsense Contextualization'. arXiv. <http://arxiv.org/abs/2212.10465>.
6. Min, Sewon, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 'FactScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation'. arXiv. <https://doi.org/10.48550/arXiv.2305.14251>.
7. Opitz, Juri, and Anette Frank. 2022. 'SBERT Studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features'. arXiv. <https://doi.org/10.48550/arXiv.2206.07023>.
8. Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. 'Llama 2: Open Foundation and Fine-Tuned Chat Models'. arXiv. <http://arxiv.org/abs/2307.09288>.
9. Zang, Xiaoxue, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. 'MultiWOZ 2.2: A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines'. In Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, edited by Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah, 109–17. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.nlp4convai-1.13>.
10. Zhu, Wanzheng, and Suma Bhat. 2020. 'GRUEN for Evaluating Linguistic Quality of Generated Text'. arXiv. <http://arxiv.org/abs/2010.02498>.