Synthetic Dialogue Dataset Generation using LLM Agents

Yelaman Abdullin
Macquarie University
yelaman.abdullin@hdr.mq.edu.au
Presenter

Diego Molla-Aliod
Macquarie University
diego.molla-aliod@mq.edu.au

Bahadorreza Ofoghi
Deakin University
b.ofoghi@deakin.edu.au

John Yearwood
Deakin University
john.yearwood@deakin.edu.au

Qingyang Li
The University of Melbourne
ql5@student.unimelb.edu.au

Abstract

Linear programming (LP) problems are pervasive in real-life applications. However, despite their apparent simplicity, an untrained user may find it difficult to determine the linear model of their specific problem. We envisage the creation of a goal-oriented conversational agent that will query the user for information related to the problem so that the agent can generate the linear model. In this work, we present an approach for the generation of sample dialogues that can be used to develop such a conversational agent. Using prompt engineering, we develop two agents, one acting as the conversational agent the other acting as the user. Using a set of text descriptions of linear problems from NLP4Opt(Ramamonjison et al., 2022, 2023) available to the user only, the agent and the user engage in conversation until the agent has retrieved all key information from the original problem description. We also propose an extrinsic evaluation of the dialogues by assessing how well the summaries generated by the dialogues match the original problem descriptions. We conduct human and automatic evaluations, including an evaluation approach that uses GPT-4 (OpenAI, 2023) to mimic the human evaluation metrics. The evaluation results show an overall good quality of the dialogues, though research is still needed to improve the quality of the GPT-4 evaluation metrics. The resulting dialogues are available to the research community, and the conversational agent used for the generation of the dialogues can be used as a baseline.

In this presentation, we will show the results and conclusions of our paper that was submitted to the GEM Workshop (Abdullin et al., 2023). We aim to present to the ALTA community a dataset for the task of eliciting information from the user through a dialogue with a conversation agent. The specific use of the information elicited is for the automatic modeling of linear optimization problems. This is itself a very useful task with broad potential applications, but the methods for data generation and evaluation proposed here can be adopted easily for other possible tasks. The results indicate a reasonable correlation between ROUGE-L (Lin, 2004), BERTScore Precision (Zhang et al., 2020), and the average human information precision scores and this is slightly better than the correlation between the GPT4 agent and the human IP scores. As further work, we intend to refine the prompts used for the evaluation approach with GPT-4. In addition, we will conduct more exhaustive types of evaluation on the data set that might be more suitable to the specific domain of linear programming modelling.

References

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. Synthetic dialogue dataset generation using llm agents. Submitted to the Third Generation, Evaluation Metrics (GEM) Workshop at EMNLP 2023.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Ramamonjison et al. 2023. Nl4opt competition: Formulating optimization problems based on their natural language descriptions.

Rindra Ramamonjison, Haley Li, Timothy Yu, Shiqi He, Vishnu Rengan, Amin Banitalebi-dehkordi, Zirui Zhou, and Yong Zhang. 2022. Augmenting operations research with auto-formulation of optimization models from problem descriptions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 29–62, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.