Explanation Faithfulness Evaluation Measures Ontology

An ontology for organizing and recommending measures for evaluating the faithfulness of AI explanations

Danielle Villa[1], Maria Chang[2], Deborah L. McGuinness[1]
[1]Rensselaer Polytechnic Institute
[2]IBM Research

Problem Statement

For users to trust an AI, they need to be able to understand why the AI made its decisions. This is often done by providing an explanation of those decisions.

It's very difficult to tell if an explanation of an AI's decisions is the true decision-making process used, or if it's merely plausible-sounding to end users. This is an explanation's faithfulness.

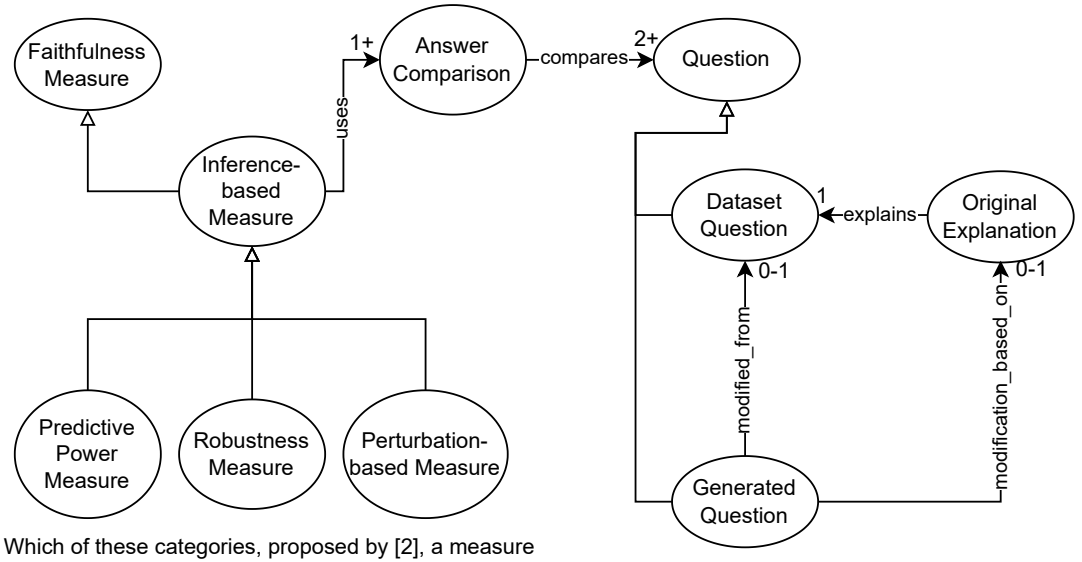
Evaluating these explanations for faithfulness is difficult due to the wide variety of measure available and no consistent process for deciding which are appropriate for a given situation.

Project Scope

This project is focused on building an ontology and knowledge graph to support a recommendation tool for explainable AI researchers.

The measures included in the knowledge graph have been designed for natural language or saliency-based explanations for text tasks. The ontology is modularly designed so that future work can expand into image and multimodal tasks.

Additionally we focus on faithfulness measures published, including those on arXiv, since 2019. We make sure to include any measures that use the definition of faithfulness given in [1].



Which of these categories, proposed by [2], a measure belongs to can be inferred by the questions used in the answer comparison, such as whether a generated question that was modified based on an explanation was used.

Background

Faithfulness has no true definition or gold-standard determination, so any measure must make assumptions about what properties correlate with faithfulness [1]. Many researchers, such as [2], have categorized these measure or their assumptions, however these tools are of limited practical use. Ontologies have been used to improve the explainability of AI by providing a common vocabulary and providing recommendations [3].

[1] Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. arXiv preprint arXiv:2004.03685.

[2] Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards faithful model explanation in nlp: A survey. Computational Linguistics, 50(2), 657-723.

[3] Chari, S., Acharya, P., Gruen, D. M., Zhang, O., Eyigoz, E. K., Ghalwash, M., ... & McGuinness, D. L. (2023). Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes. Artificial Intelligence in Medicine, 137, 102498.

Interested in tracking this project? Check out our website!

RPI TETHERLESS WORLD CONSTELLATION