Towards an Explainer-agnostic Conversational XAI

Navid Nobani[1,4*], **Fabio Mercorio**[2,3], **Mario Mezzanzanica**[2,3]
[1]Dept. of Informatics, Systems & Communication, Univ. of Milan-Bicocca, Milan, Italy
[2]Dept. of Statistics and Quantitative Methods, Univ. of Milan-Bicocca, Milan, Italy
[3]CRISP Research Centre, Univ. of Milan-Bicocca, Milan, Italy
[4]Digital Attitude, Milan, Italy
{navid.nobani, fabio.mercorio, mario.mezzanzanica}@unimib.it

Abstract

Explainable Artificial Intelligence (XAI) is gaining interests in both academia and industry, mainly thanks to the proliferation of darker more complex black-box solutions which are replacing their more transparent ancestors. Believing that the overall performance of an XAI system can be augmented by considering the end-user as a human being, we are studying the ways we can improve the explanations by making them more informative and easier to use from one hand, and interactive and customisable from the other hand.

1 Introduction

Despite the superiority of Natural Language (NL) explanations in efficiency and their ease of use for non-technical users, the majority of the XAI systems either make use of non-language presentation forms (e.g. list of features or simple graphics) or utilise basic Natural Language Generation (NLG) techniques like template-filling which while effective, 1) fail to reinforce the sense of reliability humans associate with explanations coming from machines and 2) are rigid and work only with predefined templates. To address these concerns, we have designed our research activities to first create a system that is agnostic towards black-box and explainer models and provides a conversational interface regardless of the chosen data, black-box or explainer. Secondly, we are studying methods that enable us to compose interactive systems which allows end-users to directly interact with them while considering the true needs of the user, based on her prior knowledge of the system and how she utilises it.

2 An XAI Framework Focusing on Humans

Surveying 70 recent XAI papers, we came up with a roadmap that can map the whole process of an XAI system from generating context and explanations and conveying them to the final user (Figure 1). Given that our focus is on the content, context of explanations and the interaction between the system and the human, we deliberately ignore technical aspects which directly play a role in constructing the explanations like black-box and data types.

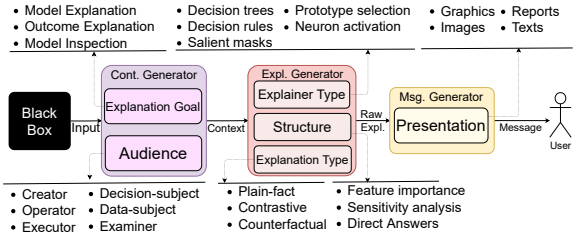

Figure 1: A graphical overview of framework

Motivation: Why XAI Should Be Interrogable? Although XAI techniques are the best hope to open black-box models and to discover their underlying decision-making process unless designed with humans in mind, XAI systems often fail to provide comprehensible explanations to laypersons, are static or provide explanations only for a fixed and presumed set of problems and are limited to producing simplistic plain-fact answers which often ignore the root problem [Miller, 2019]. To clarify the matter, let us consider an imaginary conversation among a user (👤) and the system (⚙) that has been trained to classify online job ads:

> 👤: Why have you classified this job ad as *"Data engineer"*?
> ⚙: I saw *"Data"* and *"Architect"* words.
> 👤: Why you ignored the word "senior"?
> ⚙: I consider this word as "too general".
> 👤: Answer me what "too general" means by considering me as a data scientist.
> ⚙: The TF-IDF score of the word "senior" is distant from the score of the rest of the terms.
> 👤: What if I tell you that I'm your developer?
> ⚙: In that case, I would say that I took that decision because the word senior is on the list of stopwords.

Such a tailored conversation can directly contribute to making a decision, for instance, in the example above, this decision could be adding the word *senior* to the list of stopwords.

As it can be seen in Figure 2, our proposed system is composed of four components, namely Data, black-box, Explainer and Dialogue System, with the latter is further built from NLU, Dialogue state tracker, Dialogue policy and NLG

*Contact Author

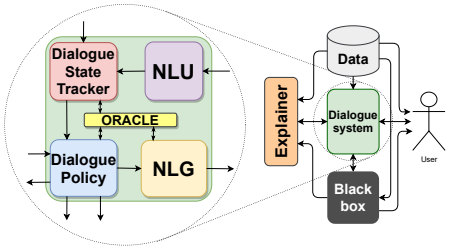

Figure 2: An overall schema of the interrogable XAI system

units. This system allows the modification of the first three components while the dialogue system (e.g. NLU and NLG units) is retrained based on the attached data, black-box and explainer. Similar to [Sokol and Flach, 2020], the NLU unit is made using RASA[1] while unlike them, we generate explanations using SOTA models which covers both plain-fact and contrastive explanations, instead of generating them from scratch. Such a flexible design allows going beyond textual explanations for instance, by offering PDP and ICE plots.

State of the art. Despite the ever-growing interest in XAI, only a small portion of XAI research provide solutions or discuss systems that are able to directly interact with the final user, customise the results based on the user's profile or provide explanations in a natural language. This could lead to solutions that, while potent and effective from a technical point of view, cannot be directly utilised by non-expert, non-technical users, defying the principal objective of an XAI system (see [Miller, 2019]). The following works partially address the specified characteristics mentioned above: [Hohman *et al.*, 2019] present a prototype that combines visualisations and verbalisations to support interactive exploration of ML models. [Amarasinghe and Manic, 2019] present a methodology for linguistically explaining a deep learning classifier based on fuzzy logic, capable of customising the results and interact with users. Using contrastive explanations [Sokol and Flach, 2018] propose a system that can generate interactive natural language explanations for ML models. [Chang *et al.*, 2016] propose a process that generates personalised natural language explanations for recommendation systems.

3 Explainer-agnostic Conversational XAI

Currently, the system has three main components: 1) a graph which represents connections between system elements, e.g. entities, intents, data[set] types and explainers, 2) a chatbot with units depicted in Figure 2 and 3) a web interface which enables the user to interrogate the system and in the same time records the conversation and the user evaluation of the conversation. While the last two parts are used in a few other XAI methods, all of them have a fixed structure which results in limited explanation and data types (see Figure. 1). On the other hand, by considering the building blocks of the system as individual elements and mapping them through a graph, not only we are able to facilitate the process of adoption to new cases (e.g. dataset) but also we can achieve a solution that is explainer agnostic and is not bounded to a specific data type or black-box. A brief description of the aforementioned components follows. In order to reduce the effort and time to adopt the system for new applications (e.g. new dataset/black-box), we define generic entities and intents, which are later mapped to different types of input data and explainers through a matrix. Our system utilises both ML and rule-based techniques inside the chatbot block, and while the dialogue state tracker and dialogue policy both use rules to manage the user inputs, they perform these tasks through the components matrix, instead of having rules which handle singular cases. Such a design reduces the time and effort a classic rule-based system demands to be adapted to a new environment. In order to achieve this, by arranging a focus group of fifteen ML practitioners, we identified thirteen intents from which five are related to model-related queries while the rest cover the cases regarding the natural situations which happened in human conversation like mind-changing, frustration and misunderstanding. We developed the dialogue system using Python for analysing the user queries and generating explanations in textual and graphical forms while React and Flask were used to create a web-based user interface.

Currently, we are working on the following elements: 1) Multi-turn open-domain dialogue management through Cohesion, Coherence and Ellipsis resolution (see [McTear, 2020]) 2) performing a user study to evaluate the efficacy of the system towards facilitating the explanation process 3) using Amazon Mechanical Turk to enhance the NLU training data by increasing the possible user queries.

[1]https://github.com/RasaHQ/rasa

References

[Amarasinghe and Manic, 2019] Kasun Amarasinghe and Milos Manic. Explaining what a neural network has learned: Toward transparent classification. In *FUZZ-IEEE*. IEEE, 2019.

[Chang *et al.*, 2016] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. Crowd-based personalized natural language explanations for recommendations. In *ACM Conference on Recommender Systems*, 2016.

[Hohman *et al.*, 2019] Fred Hohman, Arjun Srinivasan, and Steven M Drucker. Telegam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 2019.

[McTear, 2020] Michael McTear. Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 2020.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2019.

[Sokol and Flach, 2018] Kacper Sokol and Peter A Flach. Conversational explanations of machine learning predictions through class-contrastive counterfactual statements. In *IJCAI*, 2018.

[Sokol and Flach, 2020] Kacper Sokol and Peter Flach. One explanation does not fit all. *KI-Künstliche Intelligenz*, pages 1–16, 2020.