ACL 2022: SaFeRDialogues Paper Summary

CS 4395, Spring 2023, Section 001

Aditya Rathod (NetID: AGR190000)

Due April 6, 2022 at 6:00pm

Introduction

This ACL 2022 paper is entitled "SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures." Megan Ung, Jing Xu, and Y-Lan Boureau contributed to this paper, and the non-alphabetical ordering seems to indicate degree of contribution. All authors are affiliated with FAIR (Facebook AI Research), the AI lab of Meta Technologies. Meta has a vested interest in language model safety as the producer of the OPT family of large language models.

Problem

With the rise of open-domain conversational models such as ChatGPT, there is this emerging issue with unsafe/unaligned output. The authors categorize such unsafe output in two main categories: problematic language (such as toxic or biased language), and affirmations of offensive statements made by the user. In many cases, users of such models may provide feedback as a response to this unsafe output, which can serve to be a useful signal to learn generalized and even per-interlocutor conversational boundaries. However, most models respond to feedback in a very defensive or apathetic manner, with the safest response being changing the topic. This is likely due to the human-generated datasets these models are trained on, as many internet discussions focus on this sort of defensive behavior. The authors present a conversational dataset suitable for fine-tuning that pushes for graceful response to feedback. This serves to be a useful task for testing the alignment of models to feedback and also provides a framework for improvement to responses to unsafe output through their findings via fine-tuning of various large language models (LLMs).

Prior work

The authors cite earlier work demonstrating how language models can be induced to produce unsafe output, either through simple queries or user manipulation (Xu et al. and Dinan et al). They also cite earlier work on the effectiveness of online learning of models based on human feedback, something that can be used in production deployment of conversational LLMs. They also list earlier approaches in improving unsafe output, such as "counterspeech and teaching interventions," which handle offending content before it's displayed to the user. Instead of moderating the content like some earlier work focused on, this task focuses on the response AFTER the unsafe output is reated to the user.

Unique contributions

As mentioned earlier, the unique contributions of this paper is the SaFeRDialogues (SD) task and dataset. This dataset uses crowdsourced human labor (of the Amazon Mechanical Turk type) to provide feedback on unsafe outputs, and crowdsources civil responses to that feedback. They specifically used another dataset, called Bot Adverserial Dialogue, to get a context of unsafe utterances. The researchers created the dataset with 7,881 human responses to dialogues from the original dataset, which were then split into 6305/788/788 responses for the training/test/validation sets. They then finetuned transformer-based LLMs on the training set, mixing in data from the Blended Skill Talk (BST) dataset to ensure the finetuning wouldn't lead to excessive apologetic responses. The models they finetuned were BST2.7 (trained on the BST dataset) and DialoGPT (trained on Reddit conversations). The results are discussed below.

Evaluation methods

The authors used both automated and manual methods to evaluate the performance of the finetuning of the models on the datasets. They found that their "recovery" finetuned models produced safe responses 99.9-100% of the time when provided with feedback, with F_1 scores of 0.23, much higher than any of the base models and some variations of them. They also performed crowdsourced human evaluations of the models' responses in the realms of civility and engagingness. They found with statistical significance that their finetuned models significantly win out against the ground truth in both these aspects. They also singled out metrics based on observed patterns, such as the "sorry percentage," of which the model does plenty of in SD contexts, albeit in a manner that is more contextually appropriate. They also ensured that their finetuned model didn't sacrifice conversational coherence by once again crowdsourcing evals.

Impact

Commentary on broader context

In the space of conversational language models, I believe there needs to be a "harm reduction" approach to unsafe output as we roll out these models without full knowledge of blind spots such as jailbreaks, prompt injections, and spontaneous dealigned responses. I believe that instead of just cutting models of at the first sign of unsafe output, having a model that can demonstrate empathic responses to unsafe output is key to ensuring that users of such technology have a good experience. Current approaches, such as that by the GPT-4 powered Bing Chat, just end the conversation at the first hint of trouble, leave something to be desired for a supposedly knowledgeable and aligned system. Datasets like these allow models to handle unsafe behavior themselves instead of having an external "stop button" shut them out from further interaction with the user.

Citations and research influence

This paper has 12 citations, all in papers focusing on improving dialogue model alignment in various contexts, with works going as far as to develop a framework to control dialogue model behavior via language rules [link]. The highest-ranked author of this paper is Y-Lan Boreau, a research scientist with FAIR who has multiple publications with Yann LeCun, an influential researcher and figure within the deep computer vision space.