"Did I Say Something Wrong?": A Safe Collaborative Chatbot

Chatbots have been a core measure of AI since Turing has presented his test for intelligence, and are also widely used for entertainment purposes. In this paper we present a platform that enables users to collaboratively teach a chatbot responses, using natural language. We present a method of collectively detecting malicious users and using the commands taught by the malicious users to further mitigate activity of future malicious users. We ran an experiment with 192 subjects and show the effectiveness of our chatbot.