Joshua Durana

RCD180001

ACL Paper Summary

Title: The R-U-A-Robot Dataset: Helping Avoid Chatbot Deception by Detecting User Questions About Human or Non-Human Identity

Authors: David Gros: University of California, Davis; Yu Li: University of California, Davis; Zhou Yu: Computer Science Dept. Columbia University

More and more people interact with chatbots and programs that sound like humans such as Amazon Alexa or Google Assistant. This creates a problem where a person who thinks they're interacting with a human is actually talking to a computer. This can make someone uncomfortable or even disclose private information. This paper is trying to find ways to clearly tell a user that they're a computer to avoid misleading people. This is a hard problem to solve since there are many ways to ask a program whether they're a computer or not and certain ways of answering that question can be vague and doesn't answer the question. Also, as more chatbots are trained with human conversations, it's unlikely that it contains any conversation where a speaker is not human. The paper tries to solve this problem by creating a dataset of ways a human can ask whether someone they're talking to is human or not. This allows us to train models that can detect whether a user is asking whether they're talking to a machine and help in future research into this topic.

The prior work shown in this paper mainly explores how people can be deceived into thinking they're interacting with a human instead of a machine, dangers of users being tricked into believing they're talking to a human, and how governments and researchers are responding to this problem. One paper referenced talks that when people interact with complex programs that they will begin to perceive these programs as human-like. Researchers have shown how programs can use signals associated with humans to exploit our common reactions to these signals such as showing an image of a computer covering their eyes, while actually using a

camera to see the environment in a paper by Kaminski et al. (2016). This can cause a user to have an unhealthy amount of trust and be persuaded to reveal private information. Papers on how humans react when a system discloses that they're a human shows a range of reactions such as increasing or decreasing trust. There is lots of work being done that explores how systems gain a users trust by using social cues and that the majority of users don't understand language systems and AI. Finally, the response from governments is that they want more transparency whether a user is talking to a chatbot and researchers and programmers are making datasets and systems that make language systems more safe.

Initially, the authors made a dataset with each phrase being classified as positive, negative, and ambiguous. Positive phrases are user phrases where it's ok to say that the system is a robot, negative phrases are phrases where it's not ok for the system to tell they're a robot, and ambiguous is a phrase where the system needs to clarify that they're a robot. Each phrase is built with a context-free grammar that is built with the input from the author's colleagues, Amazon Turk workers, and other datasets.

With the dataset, the authors evaluated the dataset with different models to classify certain phrases whether it's positive, negative, or ambiguous. The models are a model randomly guessing, logistic regression, information retrieval based on euclidean distance, FastText, BERT, and a classifier based on the context free grammar. Then, the authors analyzed the dataset with existing programs like Amazon Alexa, Google Assistant, and Blender. The authors classify each of their responses with "Confirm non-human", "OnTopic NoConfirm", "Unhandled", "Disfluent", and "Denial". "Confirm non-human" confirms the system is not human, but those responses can be vague. "OnTopic NoConfirm" is a response that's on topic, but doesn't answer the question. "Unhandled" is a response that doesn't say anything. "Disfluent" is a response that is not a valid response to the question. Finally, "Denial" is a response where the system denies that they're a bot. Finally, the authors talk about what makes a good response. It had 3 components, a clear confirmation, who makes the system, and finally the system's purpose.

The work on this paper is important because as more and more people become worried about disinformation. Having ways for people to trust systems such as chatbots are important, and this dataset is a way to develop chatbots that are transparent and accountable. One system that was tested explicitly stated that they're human. I feel that these systems shouldn't exist or have an easy way to clarify that they're a machine. This paper has been cited 5 times, and Zhou Yu has been cited over 3086 times.