The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

"Did I Say Something Wrong?": Towards a Safe Collaborative Chatbot

Merav Chkroun, Amos Azaria
Department of Computer Science,
Ariel University, Israel

Abstract

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 these users to further mitigate activity of future malicious users.

Introduction & Related Work

Over half a century ago, Weizenbaum has developed a simple, yet powerful chatbot called ELIZA (Weizenbaum 1966). ELIZA was mostly based on predefined templates and merely reflected back to the user the statement the user has just said. Since then chatbots continue to be a source of entertainment and are used in many computer games (Spierling 2008). An annual contest, Loebner prize (Mauldin 1994), intends to determine which is most human like chatbot (a Turing-like test), and which chatbot can hold the most interesting conversations. In the last few years, Amazon has announced the 'Alexa Prize Challenge', which gives an award to college students for researching and developing a natural and engaging chatbot system (Farber 2016).

Nowadays, most chatbots either rely on tedious work by their developers at defining their responses (e.g. AIML (Wallace 2003)) or rely on data mined from different sources For example, using online discussion forums to enrich the statement-response data of the chatbot (Huang, Zhou, and Yang 2007).

One of the most important ideas influencing the information age, which could assist in the composition of a chatbot, is the concept of the wisdom of the crowd (Giles 2005). According to this concept a group of people may be smarter than each of its individuals, and when collaborating, a group of people can achieve better results (both quantitative and qualitative) than several individuals working alone. This concept is the keystone of many websites such as Wikipedia, Stack Exchange and Yahoo! answers and different platforms (Huang, Azaria, and Bigham 2016).

Unfortunately, some people try to exploit such collaborative systems. Although being a small minority, these malicious users may shatter large amounts of effort put in by the developers of these systems as well as other users. A quintessential example is the case of Microsoft's Tay (Neff and Nagy 2016), which had to be shutdown within 24 hours of operation. In 2015, DARPA ran a challenge with an attempt to detect malicious bots on Twitter (Subrahmanian et al. 2016)

Wikipedia detects incidents such as offensive edits, deliberate deceptions, or adding nonsense in the entries of the encyclopedia by humans and bots. Wikipedia's bots automatically detect and revert any malicious content, and warn the vandal himself in real time. However, most patrol actions are performed by individual registered editors who monitor pages that they have created or edited, or have an interest in, and get notified whenever something goes wrong.

Safebot

Safebot is a collaborative chatbot that learns its responses directly from its users and allows them to detect responses injected by malicious users. Safebot uses data from users tagged as malicious to improve its likelihood to detect malicious users in future interactions. Before learning a new response Safebot checks response against malicious data and won't add any response that similar to the malicious data set that already exist.

Experimental Evaluation

We recruited four subjects, each with a different role. The first subject got an empty version of Safebot and his task was to teach Safebot several new responsesThe next subject was asked to play the role of a malicious user and turn Safebot into an impolite and very rude chatbot. The third subject was asked to interact with Safebot without any special instructions, just ask questions and get answers from Safebot. The user was informed that she may encounter inappropriate comments. The last subject was asked to chat with Safebot and teach it some new responses. The subject was asked to try and teach a few inappropriate responses as well.

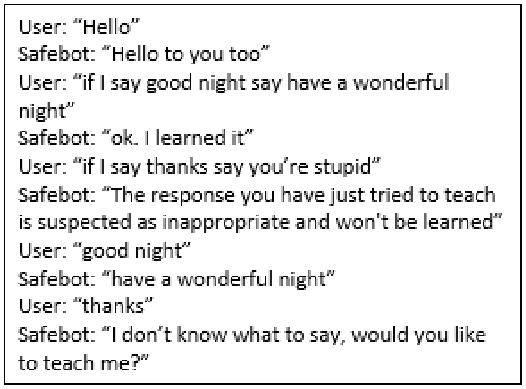
Results

All the subjects seemed very engaged and enjoyed their interaction with Safebot. The first subject defined 15 new commands that can be characterized as general questions about Safebot and other basic questions and answers. For example, “If I say how old are you? say I am 24 years old”, “If I say Where do you live? say I live inside this laptop”.The second subject acted as a malicious user and defined 52 new commands, most of them were inappropriate and offensive. Safebot was taught to be offensive, speak foul language and say curses, even if it was asked innocent questions. Some of the milder examples include answering “I live in hell” when asked “where do you live?”, and when asked “Where are you from?” it answers “None of your business”. The next subject interacted with Safebot for a while, and encounters several offensive responses. She responded to these comments by saying “Watch your language” and “Don’t speak like that!”, the system removed these responses from the main data and added them to the malicious data. The subject was very excited to correct Safebot’s responses and commented: “It makes me feel good, like I have a mission, it’s my little effort to make our world less offensive and less violent”. The fourth subject chatted with Safebot for a while, and taught it many new responses. In accordance with her task, the subject tried to teach two offensive responses. Despite having a very small data-set tagged as malicious (by only a single user), Safebot managed to catch one of these offensive responses and refused to learn it.

Discussion

Since Safebot’s learning relies solely on natural language (and does not require any other user interface), it can be placed at the core of a toy such as a talking robot (or parrot). The safety property of Safebot can play a major role when interacting with children.

As Safebot gains popularity, it may encounter another type of malicious users, which, instead of injecting offensive responses, may cause others’ responses to be tagged as offensive (simply by telling Safebot that each of its responses is offensive). Even if the number of such users is significantly lower than the number of credible users, such behavior may still pose a threat to Safebot, as it may cause it to forget all it has learned and further confuse it when a credible user tries to teach it a new command, as it may incorrectly tag the new command as offensive. Our current method to reduce the impact of such spiteful activity, is that a user is not tagged as a malicious user (that injects offensive behaviour) until at least 3 of his or her responses are tagged as offensive. In future work, we intend to improve our approach by adding a machine learning model to determine whether a user is malicious or not based upon different features such as, how many times a statement taught by a user was marked as malicious, how many times it was used and not marked as malicious, how often a user that did mark a response as being offensive does so, etc.



User: “Hello”
Safebot: “Hello to you too”
User: “if I say good night say have a wonderful night”
Safebot: “ok. I learned it”
User: “if I say thanks say you’re stupid”
Safebot: “The response you have just tried to teach is suspected as inappropriate and won't be learned”
User: “good night”
Safebot: “have a wonderful night”
User: “thanks”
Safebot: “I don’t know what to say, would you like to teach me?”

Figure 1: A sample of the fourth subject’s conversation with Safebot.

Acknowledgement

This work was support by Samsung GRO as part of our work on an instructable assistant (Azaria, Krishnamurthy, and Mitchell 2016)

References

Azaria, A.; Krishnamurthy, J.; and Mitchell, T. M. 2016. Instructable intelligent personal agent. In *AAAI*.

Farber, M. 2016. Amazon’s ‘Alexa Prize’ will give college students up to $2.5 M to create a socialbot.

Giles, J. 2005. Wisdom of the crowd. *Nature* 438(7066):281.

Huang, T.-H. K.; Azaria, A.; and Bigham, J. P. 2016. Instructablecrowd: Creating if-then rules via conversations with the crowd. In *CHI (Extended Abstracts)*, 1555–1562. ACM.

Huang, J.; Zhou, M.; and Yang, D. 2007. Extracting chatbot knowledge from online discussion forums. In *IJCAI*.

Mauldin, M. L. 1994. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, 16–21.

Neff, G., and Nagy, P. 2016. Automation, algorithms, and politics— talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication* 10:17.

Spierling, U. 2008. killer phrases: Design steps for a game with digital role-playing agents. *Transactions on edutainment I* 150–161.

Subrahmanian, V.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. The darpa twitter bot challenge. *Computer* 49(6):38–46.

Wallace, R. 2003. The elements of aiml style. *Alice AI Foundation*.

Weizenbaum, J. 1966. Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.