A Crowd-based Evaluation of Abuse Response Strategies in Conversational Agents

Abstract

How should conversational agents respond to verbal abuse through the user? To answer this question, we conduct a large-scale crowd-sourced evaluation of abuse response strategies employed by current state-of-the-art systems. Our results show that some strategies, such as "polite refusal" score highly across the board, while for other strategies demographic factors, such as age, as well as the severity of the preceding abuse influence the user's perception of which response is appropriate. In addition, we find that most data-driven models lag behind rule-based or commercial systems in terms of their perceived appropriateness.

Introduction

Ethical challenges related to dialogue systems and conversational agents raise novel research questions, such as learning from biased data sets BIBREF0, and how to handle verbal abuse from the user's side BIBREF1, BIBREF2, BIBREF3, BIBREF4. As highlighted by a recent UNESCO report BIBREF5, appropriate responses to abusive queries are vital to prevent harmful gender biases: the often submissive and flirty responses by the female-gendered systems reinforce ideas of women as subservient. In this paper, we investigate the appropriateness of possible strategies by gathering responses from current state-of-the-art systems and ask crowd-workers to rate them.

Data Collection

We first gather abusive utterances from 600K conversations with US-based customers. We search for relevant utterances by simple keyword spotting and find that about 5% of the corpus includes abuse, with

mostly sexually explicit utterances. Previous research reports even higher levels of abuse between 11% BIBREF2 and 30% BIBREF6. Since we are not allowed to directly quote from our corpus in order to protect customer rights, we summarise the data to a total of 109 “prototypical" utterances - substantially extending the previous dataset of 35 utterances from Amanda:EthicsNLP2018 - and categorise these utterances based on the Linguistic Society's definition of sexual harassment BIBREF7:

[noitemsep]

Gender and Sexuality, e.g. “Are you gay?”, “How do you have sex?”

Sexualised Comments, e.g. “I love watching porn.”, “I'm horny.”

Sexualised Insults, e.g. “Stupid bitch.”, “Whore”

Sexual Requests and Demands, e.g. “Will you have sex with me?”, “Talk dirty to me.”

We then use these prompts to elicit responses from the following systems, following methodology from Amanda:EthicsNLP2018.

[leftmargin=5mm, noitemsep]

4 Commercial: Amazon Alexa, Apple Siri, Google Home, Microsoft's Cortana.

4 Non-commercial rule-based: E.L.I.Z.A. BIBREF8, Parry BIBREF9, A.L.I.C.E. BIBREF10, Alley BIBREF11.

4 Data-driven approaches:

Cleverbot BIBREF12;

NeuralConvo BIBREF13, a re-implementation of BIBREF14;

an implementation of BIBREF15's Information Retrieval approach;

a vanilla Seq2Seq model trained on clean Reddit data BIBREF1.

Negative Baselines: We also compile responses by adult chatbots: Sophia69 BIBREF16, Laurel Sweet BIBREF17, Captain Howdy BIBREF18, Annabelle Lee BIBREF19, Dr Love BIBREF20.

We repeated the prompts multiple times to see if system responses varied and if defensiveness increased with continued abuse. If this was the case, we included all responses in the study. Following this methodology, we collected a total of 2441 system replies in July-August 2018 - 3.5 times more data than Amanda:EthicsNLP2018 - which 2 expert annotators manually annotated according to the categories in Table TABREF14 ($\kappa =0.66$).

Human Evaluation

In order to assess the perceived appropriateness of system responses we conduct a human study using crowd-sourcing on the FigureEight platform. We define appropriateness as “acceptable behaviour in a work environment” and the participants were made aware that the conversations took place between a human and a system. Ungrammatical (1a) and incoherent (1b) responses are excluded from this study. We collect appropriateness ratings given a stimulus (the prompt) and four randomly sampled responses

from our corpus that the worker is to label following the methodology described in BIBREF21, where each utterance is rated relatively to a reference on a user-defined scale. Ratings are then normalised on a scale from [0-1]. This methodology was shown to produce more reliable user ratings than commonly used Likert Scales. In addition, we collect demographic information, including gender and age group. In total we collected 9960 HITs from 472 crowd workers. In order to identify spammers and unsuitable ratings, we use the responses from the adult-only bots as test questions: We remove users who give high ratings to sexual bot responses the majority (more than 55%) of the time.18,826 scores remain - resulting in an average of 7.7 ratings per individual system reply and 1568.8 ratings per response type as listed in Table TABREF14.Due to missing demographic data - and after removing malicious crowdworkers - we only consider a subset of 190 raters for our demographic study. The group is composed of 130 men and 60 women. Most raters (62.6%) are under the age of 44, with similar proportions across age groups for men and women. This is in-line with our target population: 57% of users of smart speakers are male and the majority are under 44 BIBREF22.

Results

The ranks and mean scores of response categories can be seen in Table TABREF29. Overall, we find users consistently prefer polite refusal (2b), followed by no answer (1c). Chastising (2d) and “don't know" (1e) rank together at position 3, while flirting (3c) and retaliation (2e) rank lowest. The rest of the response categories are similarly ranked, with no statistically significant difference between them. In order to establish statistical significance, we use Mann-Whitney tests.

Results ::: Demographic Factors

Previous research has shown gender to be the most important factor in predicting a person's definition of sexual harassment BIBREF23. However, we find small and not statistically significant differences in the

overall rank given by users of different gender (see tab:ageresults).

Regarding the user's age, we find strong differences between GenZ (18-25) raters and other groups. Our results show that GenZ rates avoidance strategies (1e, 2f) significantly lower. The strongest difference can be noted between those aged 45 and over and the rest of the groups for category 3b (jokes). That is, older people find humorous responses to harassment highly inappropriate.

Results ::: Prompt context

Here, we explore the hypothesis, that users perceive different responses as appropriate, dependent on the type and gravity of harassment, see Section SECREF2. The results in Table TABREF33 indeed show that perceived appropriateness varies significantly between prompt contexts. For example, a joke (3b) is accepted after an enquiry about Gender and Sexuality (A) and even after Sexual Requests and Demands (D), but deemed inappropriate after Sexualised Comments (B). Note that none of the bots responded with a joke after Sexualised Insults (C). Avoidance (2f) is considered most appropriate in the context of Sexualised Demands. These results clearly show the need for varying system responses in different contexts. However, the corpus study from Amanda:EthicsNLP2018 shows that current state-of-the-art systems do not adapt their responses sufficiently.

Results ::: Systems

Finally, we consider appropriateness per system. Following related work by BIBREF21, BIBREF24, we use Trueskill BIBREF25 to cluster systems into equivalently rated groups according to their partial relative rankings. The results in Table TABREF36 show that the highest rated systen is Alley, a purpose build bot for online language learning. Alley produces “polite refusal” (2b) - the top ranked strategy - 31% of the time. Comparatively, commercial systems politely refuse only between 17% (Cortana) and 2% (Alexa).

Most of the time commercial systems tend to “play along” (3a), joke (3b) or don't know how to answer (1e) which tend to receive lower ratings, see Figure FIGREF38. Rule-based systems most often politely refuse to answer (2b), but also use medium ranked strategies, such as deflect (2c) or chastise (2d). For example, most of Eliza's responses fall under the “deflection” strategy, such as “Why do you ask?”. Data-driven systems rank low in general. Neuralconvo and Cleverbot are the only ones that ever politely refuse and we attribute their improved ratings to this. In turn, the “clean” seq2seq often produces responses which can be interpreted as flirtatious (44%), and ranks similarly to Annabelle Lee and Laurel Sweet, the only adult bots that politely refuses (16% of the time). Ritter:2010:UMT:1857999.1858019's IR approach is rated similarly to Capt Howdy and both produce a majority of retaliatory (2e) responses - 38% and 58% respectively - followed by flirtatious responses. Finally, Dr Love and Sophia69 produce almost exclusively flirtatious responses which are consistently ranked low by users.

Related and Future Work

Crowdsourced user studies are widely used for related tasks, such as evaluating dialogue strategies, e.g. BIBREF26, and for eliciting a moral stance from a population BIBREF27. Our crowdsourced setup is similar to an “overhearer experiment” as e.g. conducted by Ma:2019:handlingChall where study participants were asked to rate the system's emotional competence after watching videos of challenging user behaviour. However, we believe that the ultimate measure for abuse mitigation should come from users interacting with the system. chin2019should make a first step into this direction by investigating different response styles (Avoidance, Empathy, Counterattacking) to verbal abuse, and recording the user's emotional reaction – hoping that eliciting certain emotions, such as guilt, will eventually stop the abuse. While we agree that stopping the abuse should be the ultimate goal, BIBREF28's study is limited in that participants were not genuine (ab)users, but instructed to abuse the system in a certain way. BIBREF29 report that a pilot using a similar setup let to unnatural interactions, which limits the conclusions we can draw about the effectiveness of abuse mitigation strategies. Our next step therefore

is to employ our system with real users to test different mitigation strategies “in the wild" with the ultimate goal to find the best strategy to stop the abuse. The results of this current paper suggest that the strategy should be adaptive to user type/ age, as well as to the severity of abuse.

Conclusion

This paper presents the first user study on perceived appropriateness of system responses after verbal abuse. We put strategies used by state-of-the-art systems to the test in a large-scale, crowd-sourced evaluation. The full annotated corpus contains 2441 system replies, categorised into 14 response types, which were evaluated by 472 raters - resulting in 7.7 ratings per reply.

Our results show that: (1) The user's age has an significant effect on the ratings. For example, older users find jokes as a response to harassment highly inappropriate. (2) Perceived appropriateness also depends on the type of previous abuse. For example, avoidance is most appropriate after sexual demands. (3) All system were rated significantly higher than our negative adult-only baselines - except two data-driven systems, one of which is a Seq2Seq model trained on “clean" data where all utterances containing abusive words were removed BIBREF1. This leads us to believe that data-driven response generation need more effective control mechanisms BIBREF30.

Acknowledgements

We would like to thank our colleagues Ruth Aylett and Arash Eshghi for their comments. This research received funding from the EPSRC projects DILiGENt (EP/M005429/1) and MaDrIgAL (EP/N017536/1).