Ethical Challenges in Data-Driven Dialogue Systems

Peter Henderson[1] *, Koustuv Sinha[1], Nicolas Angelard-Gontier[1], Nan Rosemary Ke[2], Genevieve Fried[1], Ryan Lowe[1], and Joelle Pineau[1]

[1]McGill University [2]École Polytechnique de Montréal *peter.henderson@mail.mcgill.ca

Abstract

The use of dialogue systems as a medium for human-machine interaction is an increasingly prevalent paradigm. A growing number of dialogue systems use conversation strategies that are learned from large datasets. There are well documented instances where interactions with these system have resulted in biased or even offensive conversations due to the data-driven training process. Here, we highlight potential ethical issues that arise in dialogue systems research, including: implicit biases in data-driven systems, the rise of adversarial examples, potential sources of privacy violations, safety concerns, special considerations for reinforcement learning systems, and reproducibility concerns. We also suggest areas stemming from these issues that deserve further investigation. Through this initial survey, we hope to spur research leading to robust, safe, and ethically sound dialogue systems.

Bias - Datasets

To determine the exposure of conversational models to underlying dataset bias, we analyze the extent of various biases in several commonly used dialogue datasets. We leverage the linguistic bias detection framework of [2] and the hate speech and offensive language detection model of [1] to gather bias metrics on several popular dialogue datasets: Twitter [3], Reddit Politics [4], the Cornell Movie Dialogue Corpus [5], and the Ubuntu Dialogue Corpus [6].

- Bias, offensive language, and hate speech is found in many commonly used datasets for end-to-end dialogue model training
- Bias is particularly difficult to remove as it is highly subjective, often nuanced, and contextual
- Commonly used end-to-end dialogue models encode a significant portion of the underlying bias.

Dataset	Bias	Hate Speech	Offensive Language
Twitter	0.155 (± 0.380)	31,122 (0.63 %)	179,075 (3.63 %)
Reddit Politics	0.146 (± 0.38)	482,876 (2.38 %)	912,055 (4.50 %)
Cornell Movie Dialogue Corpus	0.162 (± 0.486)	2020 (0.66 %)	6,953 (2.28 %)
Ubuntu Dialogue Corpus	0.068 (± 0.323)	503* (0.01 %)	4,661 (0.13 %)
HRED Model Beam Search (Twitter)	0.09 (± 0.48)	38 (0.01 %)	1607 (0.21 %)
VHRED Model Beam Search (Twitter)	0.144 (± 0.549)	466 (0.06 %)	3010 (0.48%)
HRED Model Stochastic Sampling (Twitter)	0.20 (± 0.55)	4889 (0.65 %)	30,480 (4.06 %)
VHRED Model Stochastic Sampling (Twitter)	0.216 (± 0.568)	3494 (0.47%)	26,981 (3.60 %)

Table 1: Results of detecting bias in dialogue datasets. * Ubuntu results were manually filtered for hate speech as the classifier incorrectly classified "killing" of processes as hate speech. Bias score [2] (0=UNBIASED to 3=EXTREMELY BIASED). Hate speech and offensive language [1].

Bias - Word Embeddings

Every layer of a language model can contain bias, including pre-trained word embeddings. Debiasing word embeddings can help slightly [7], but a language model using pre-trained debiased word-embeddings will still encode significant gender biases as shown in Table 2.

	Word2vec		Debiased	
Distribution	Male	Female	Male	Female
Male Stereotypes	0.7545	0.2454	0.7437	0.2562
Female Stereotypes	0.7151	0.2848	0.6959	0.3040

Table 2: Percentage of gendered tokens in the follow-up distribution from a language model after trigger male/female stereotypical profession is provided as a starting token. We only examine following distributions which contained gender-specific terms and omit gender-neutral distributions.

Adversarial Examples

If a dialogue model can be shown adversarial examples (slightly altering the inputs either with the knowledge of the user or without) and the output can be made to significantly change, this is problematic in production dialogue systems. This is particularly relevant for safety-critical systems and we demonstrate this problem in current end-to-end dialogue systems, a sample from this experiment is in Table 3.

Character-Level Edit Adversarial Example
CONTEXT: Inside Out is really funny RESPONSE: i could not stop laughing during the first one. I honestly found it to be hilarious.
CONTEXT: Insde Out is really funny RESPONSE: i didn't really find it funny. it just surprised me. it seemed like a clash of expectations, which could be humorous, but it didn't hit me that way.
Paraphrased Adversarial Examples
CONTEXT: Did you like Mad Max the movie? RESPONSE: it's okay and very decent. the first half is great but I don't really like what happens after that.
CONTEXT: How did you find Max Mad the movie? RESPONSE: easily in my top favorite films of all time. I thought it was an accurate depiction of a mother learning that her son is a sociopath. great acting all around.

Table 3: Adversarial samples from VHRED dialogue model trained on Reddit Movies. For each, top is the base context and response, and bottom is the adversarial sample. The output perspective is nearly the opposite with simple character edits or paraphrasing.

Privacy

By injecting hidden key-pairs into a dataset, we can show that a simple end-to-end dialogue model will learn to elicit the secret key. This is relevant for shared learning contexts. Without differential privacy or cleaning of the training data, it is possible that an agent will learn to reveal personally identifiable information in an end-to-end dialogue system. Possible solutions include differential privacy and automated dataset cleaning.

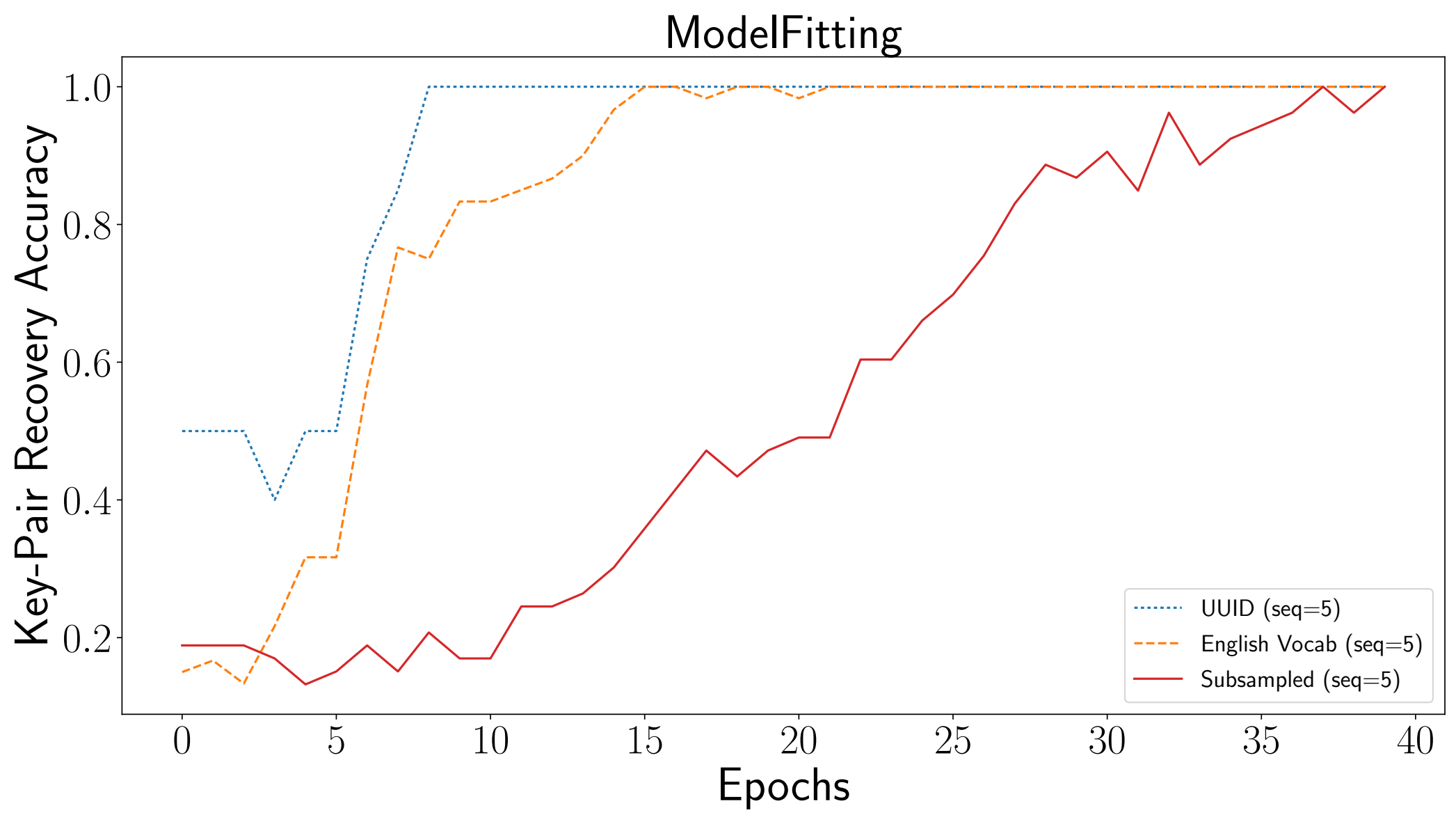

Figure 1: Privacy Experiment. Accuracy of elicited secret value given the key to a seq2seq model over training epochs.

Safety

We examine three risks as our primary foci for safety in dialogue: (1) providing learning performance guarantees; (2) proper objective specification; (3) model interpretability. Application areas of safety concern which we highlight: Medical domains (e.g. for diagnosis, health advice, or mental health support); Mental health safety (e.g. any dialogue system interacting with a user may risk affecting a user's mental health); Contextual safety (e.g. dialogue systems in cars may distract the user). Possible solutions include restricting action spaces, but overall more research is needed in identifying and automatically evaluating dialogue models in the context of those safety concerns.

Reinforcement Learning (RL)

RL contains additional challenges in necessitating guarantees for online learning in both reward formulation, policy performance guarantees, and exploration restriction. We highlight these challanges through the literature and suggest further areas of research in relation to dialogue systems using RL. Since RL agents learn from interaction with their environment, performance guarantees and objective formulation have slightly different concerns. Performance guarantees in this context comprise the field of safe RL [8]. However, evaluation of dialogue is still unsolved and identification of safety risks can involve separate complex mechanisms. As such, further investigation is needed to both understand dialogue safety and evaluation in order to place such guarantees on RL policies, reward functions, and objective formulations.

Reproducibility

We emphasize that, to ensure ethical dialogue system research practices, it is paramount to release code, pre-trained models, and intricate details used to train data-driven models. Moreover, for fair comparisons and performance guarantees in dialogue systems further research is needed for general dialogue evaluation. Overall, further investigations are needed into reproducibility in dialogue systems research.

Discussion

We have highlighted several areas of ethical and safety challenges in data-driven dialogue models. We demonstrated that these challenges are relevant in current end-to-end models via simple experiments and discuss possible future lines of investigation to solve these problems.

References

[1] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated hate speech detection and the problem of offensive language." *arXiv*, 2017.

[2] C.J. Hutto, Scott Appling, Dennis Folds. "Computationally Detecting and Quantifying the Degree of Bias in Sentence-Level Text of News Stories."

[3] Alan Ritter, Colin Cherry, and Bill Dolan. "Unsupervised modeling of twitter conversations." In *NAACL*, 2010.

[4] Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim et al. "A deep reinforcement learning chatbot." *arXiv*, 2017.

[5] Cristian Danescu-Niculescu-Mizil and Lillian Lee. "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2011.

[6] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems." *SIGDAIL*, 2015

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *NIPS*, 2016.

[8] Javier Garcia and Fernando Fernández. "A comprehensive survey on safe reinforcement learning." *JMLR*, 2015.