Ethical Challenges in Data-Driven Dialogue Systems

Peter Henderson, Koustuv Sinha,
Nicolas Angelard-Gontier, Nan Rosemary Ke,
Genevieve Fried, Ryan Lowe, Joelle Pineau

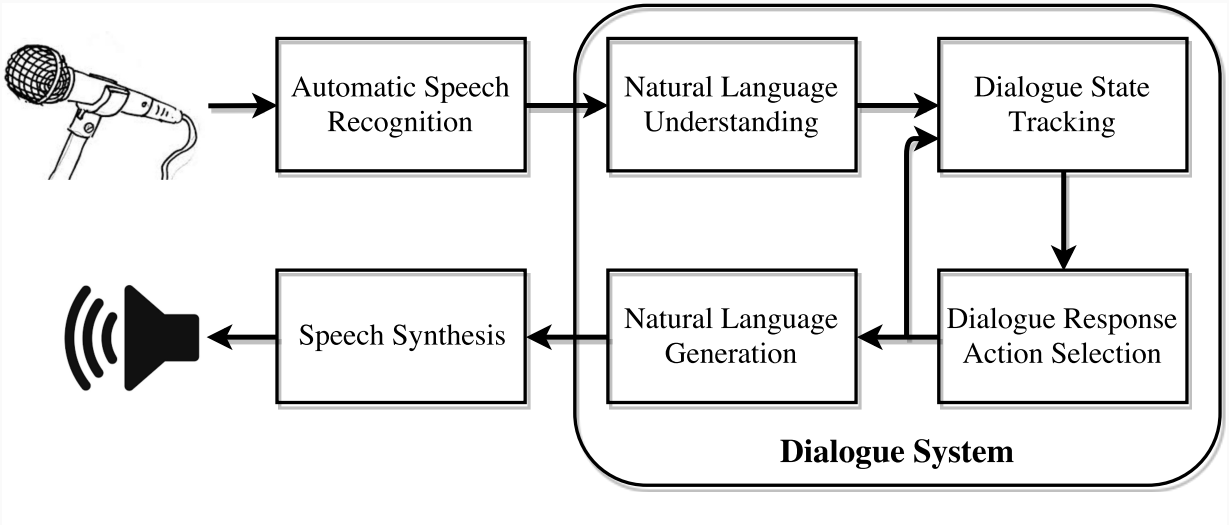
Presented by Peter Henderson*

@ The AAAI/ACM Conference on AI, Ethics, and Society

* This work is done solely at McGill University and does not reflect presenter's current position or research at Amazon.

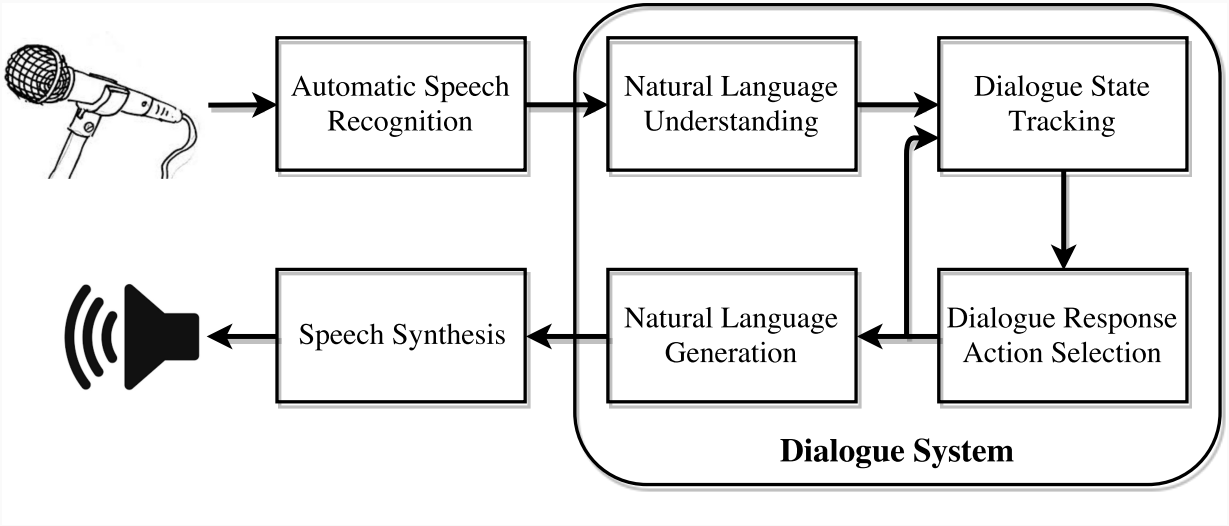
What is a dialogue system?

Chatbots, virtual personal assistants, natural language interfaces for various systems.



WHAT IS A *DATA-DRIVEN* DIALOGUE SYSTEM?

These components are learned from datasets or end-to-end from actual conversations.



Why should you care?

- Dialogue systems are convenient interfaces and are becoming increasingly prevalent in society.
- Applications contexts include automation in healthcare, in-the-home assistive devices, hands-free interaction in automobiles, and more.

Why should you care?

Adversarial example, using VHRED model from (Serban et al., 2017) with an intentional single-character edit.

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.

So what are the ethical challenges?

We outline several main aspects:

- Bias
- Privacy
- Adversarial Examples
- Safety
- Special Considerations for Reinforcement Learning
- Reproducibility

So what are the ethical challenges?

We outline several main aspects:

- Bias
- Safety
- Reproducibility
- Adversarial Examples
- Privacy
- Special Considerations for Reinforcement Learning

Bias in Dialogue Datasets

- Bias can be defined as *prejudice for or against a person, group, idea, or thing particularly expressed in an unfair way.*
- **Rule-based dialogue systems:** Bias introduced by the rule designer
- **Data-driven dialogue systems:** Bias introduced by the data, including choice of dataset, collection procedure

Data-driven models encode underlying biases

Commonly used datasets for training end-to-end dialogue models in the literature contain bias and the state of the art models learn it!

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: Bias using *Hutto et al., 2015.* bias model. Hate speech and offensive content classified via *Davidson et al., 2017.*

Performance of predicting bias models can be variable

Reddit Dataset:

- **Max Bias (3.93)** : "fresh off apology nugent compares obama administration to nazis"
- **Min Bias (-1.44)**: "american hostage <...> held by isis confirmed dead nbcnews."

Movie Dataset:

- **Max Bias (8.31)** : "him. him..."
- **Min Bias (-3.45)**: "no. pray. we never find out."

Pre-Trained Word Embeddings Contain Bias

(Bolukbasi et al. 2016)

We examine if a language model trained with debiased word embeddings still contains the same sorts of biases (it does).

	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

Bias in Dialogue Systems

- Need better ways to detect biases in natural language, account for contextual information (may require reasoning)
- Can try to prune datasets, but difficult at large scale without better bias detection mechanisms.
- Need more methods of preventing generative models from exhibiting natural language biases even if underlying data contains it

Safety

Goal : Avoid unintended harmful consequences from dialogue systems

Safety

Aim to provide:

- Performance guarantees: stability and predictability of output.
- Proper objective specification: adapt to preferences and tolerability.
- Model interpretability: Understand behavior in case where it deviates from objective.

Safety

Safety critical settings:

- Medical domains (incl. mental illness): diagnostic and intervention
- Transportation (e.g. can't distract the driver)
- Contextual awareness for any chat agent (e.g. if dialogue system realistic, may have subtle effects on mental health)

Reproducibility

> *Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. (...) Reproducibility is a minimum necessary condition for a finding to be believable and informative.*

K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, J. L. Olds, Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science, National Science Foundation, 2015.

Thank you!

Adversarial Examples

If an adversary can augment you input signal to direct the agent's output, compromise:

- Safety
- Performance
- Neutrality (freedom from biases)

Adversarial Examples

Adversarial examples in text?

- Adding distracting sentences to paragraph (Jia and Liang 2017)
- Misspelled words (remove/replace/insert characters).
- Paraphrased sentences (similar meaning, different words).

Adversarial Examples

Use VHRED model from Serbal et al., 2017, causing intentional single-character edit.

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.

Privacy

Information leakage:

- Device in “listening” mode, records side conversation with private information
- Information uploaded to train shared model.

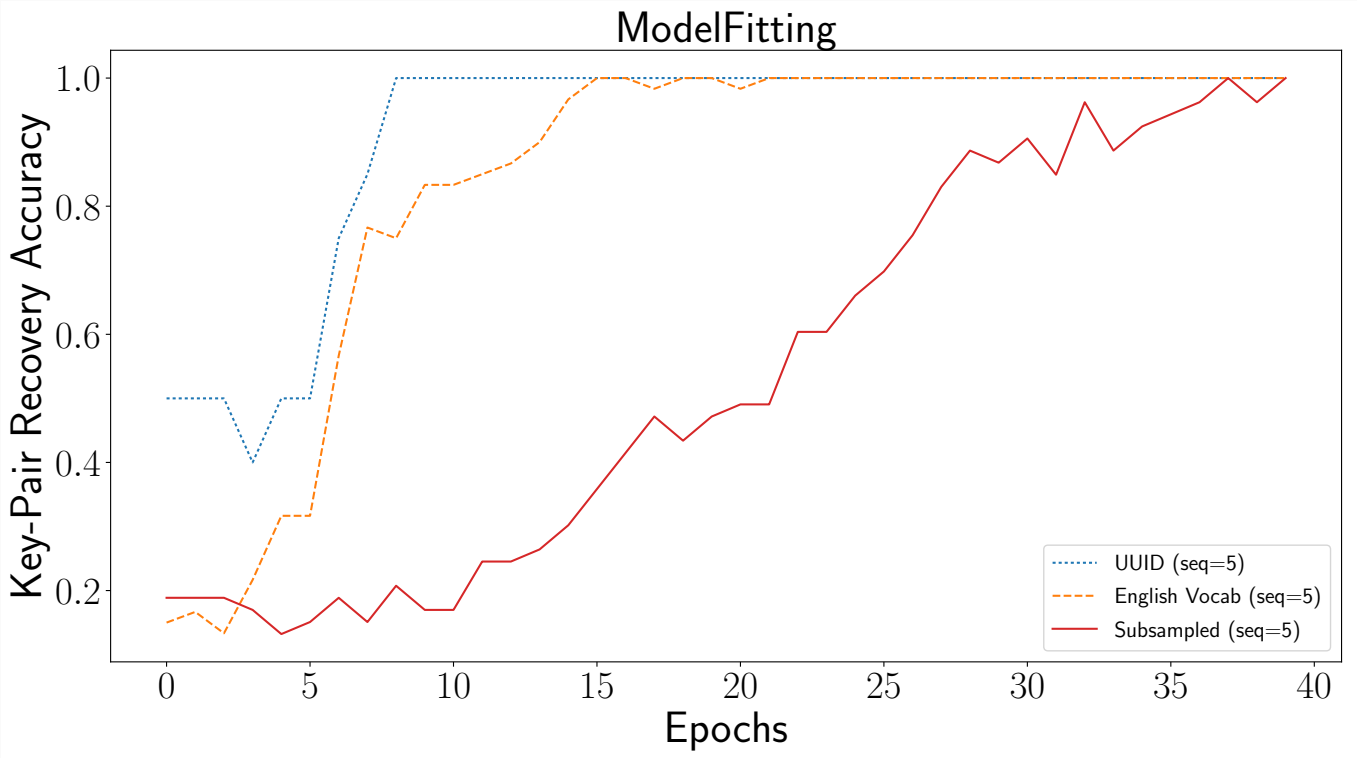
Simple experiment:

- Introduce 10 private input-output keypairs in data.
- Train simple seq2seq language model.
- When seeing on the input, does the model generate the matching output (and vice versa)?

Privacy

Types of keypairs:

- unique keypairs that do not exist in any vocabulary (UUID)
- words from the English natural language vocabulary (NL)
- words sub-sampled from the 10k dialogue pairs.



Special Considerations for RL Agents

- During live exploration, need guarantees not to enter dangerous state spaces
- Need performance and stability guarantees
- *However!* Evaluating dialogue is hard, how can you place guarantees on performance if your reward is variable?
- Need to improve automated methods for detection and evaluation of dialogues, biases, etc.