# Hierarchical Reinforcement Learning for Open-Domain Dialog (AAAI ‘20)

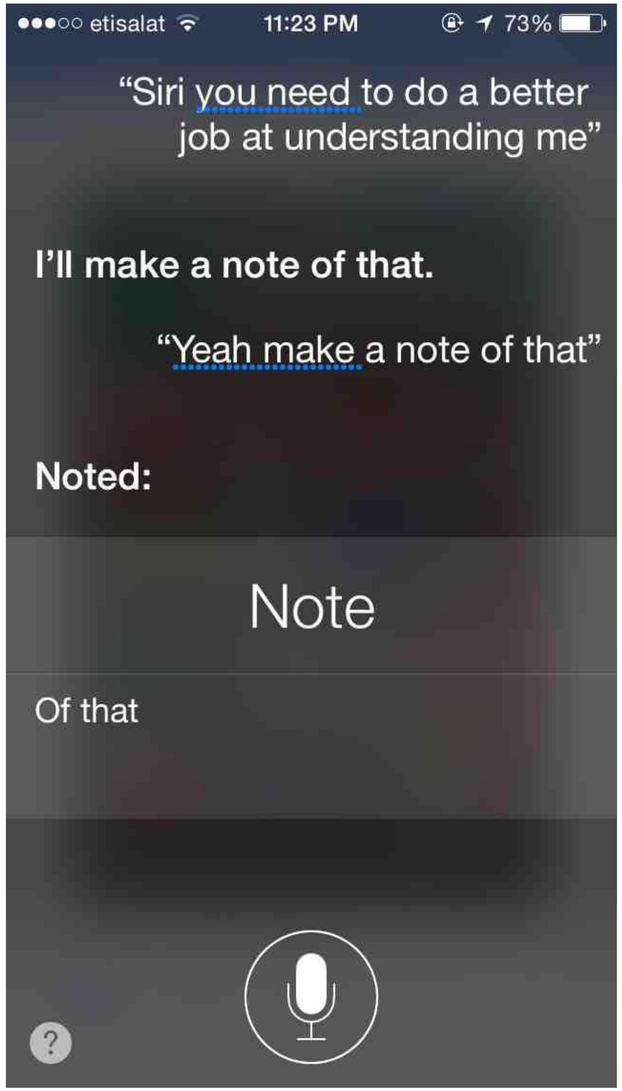
**Abdul Saleh***, Natasha Jaques*, Asma Ghandeharioun, Judy Hanwen Shen,

Rosalind Picard

# What is Open-Domain Dialog?

## Goal-oriented systems:

- Do predefined tasks.
- Scripted responses common.



# What is Open-Domain Dialog?

## Goal-oriented systems:

- Do predefined tasks.
- Scripted responses common.

## Open-domain systems:

- Mimic human conversations.
- Here we generate language.

| Open-Domain Dialog |
| --- |
| [Usr]: hello! how are you? |
| [Bot]: I'm doing well, how about yourself? |
| [Usr]: great! I'm at a conference now. |
| [Bot]: what are you doing there? |
| [Usr]: presenting and meeting people! |

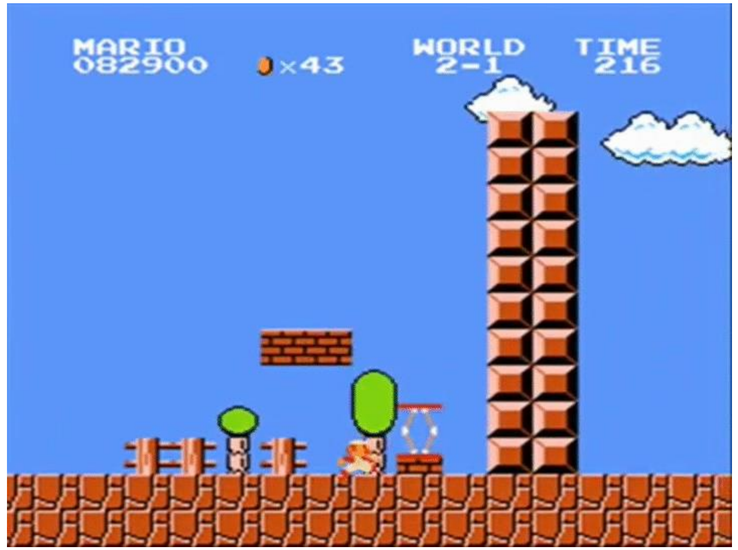
# Limitations of Open-Domain Systems

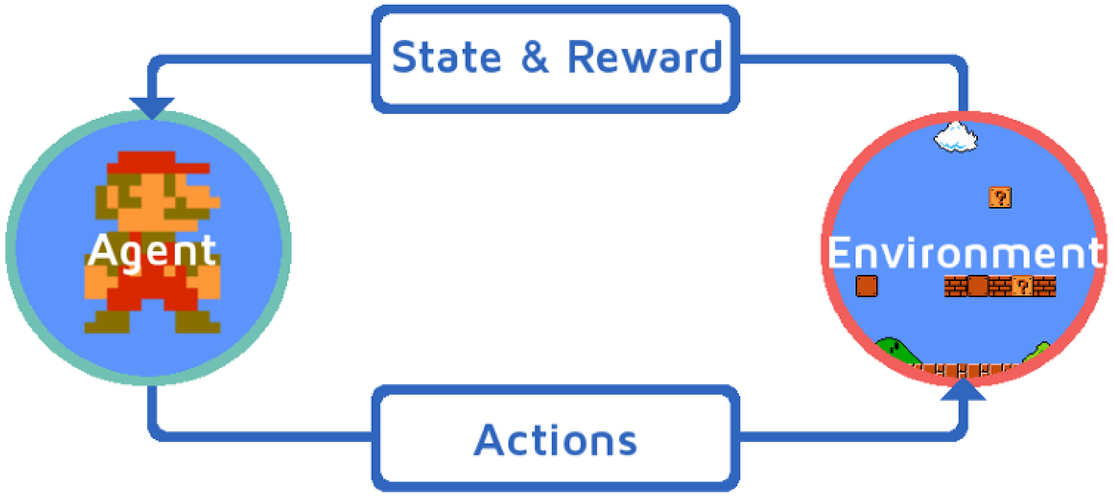
- **Repetitive and dull responses**
- **No long-term coherence**
- **Inappropriate and toxic responses**

| Repetitive | Incoherent | Inappropriate |
|---|---|---|
| [Usr]: hello!<br>[Bot]: I don't know.<br>[Usr]: how are you?<br>[Bot]: I don't know what that means. | [Usr]: what's your favorite food?<br>[Bot]: I love food<br>[Usr]: you like food?<br>[Bot]: no, I don't | [Usr]: I am studying computer science.<br>[Bot]: I am sorry to hear that. 😅 |

# Solutions?

- **Reinforcement Learning!**
  - Jaques et al. 2019, Shin et al. 2019, Li et al. 2017, Li et al. 2016, …

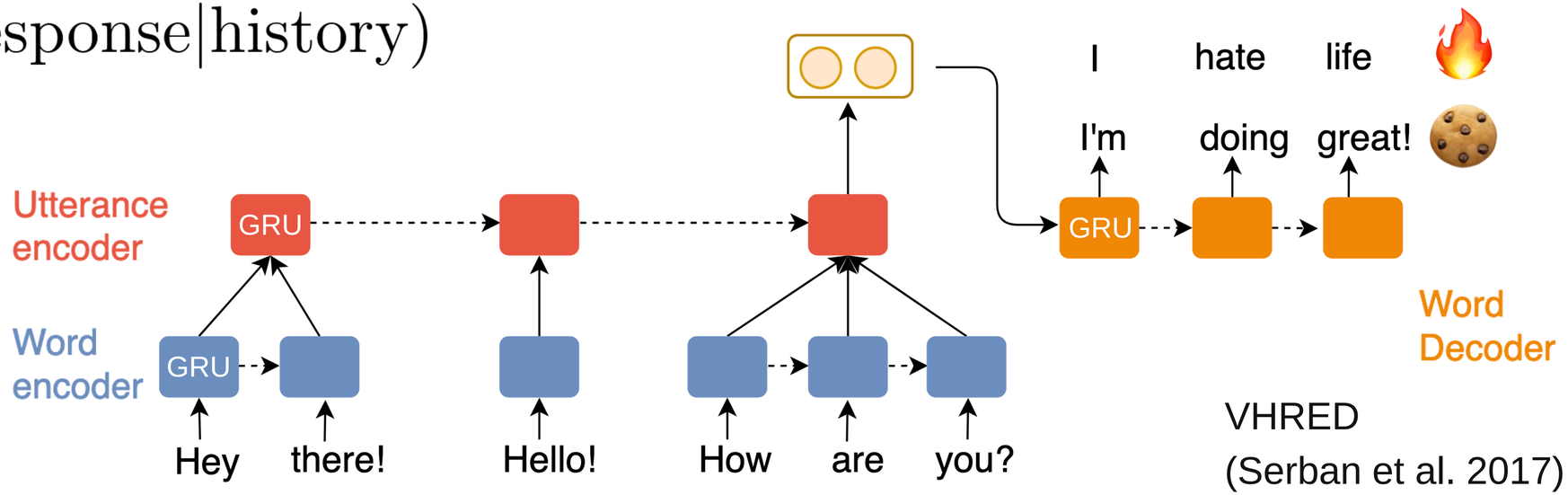


# Reinforcement Learning for Dialog

- Reward positive conversations
- Punish insensitive conversations
- Policy Gradients (REINFORCE):
  - Maximize

$$J(\pi) = R_t \pi_{\boldsymbol{\theta}}(\text{response}|\text{history})$$



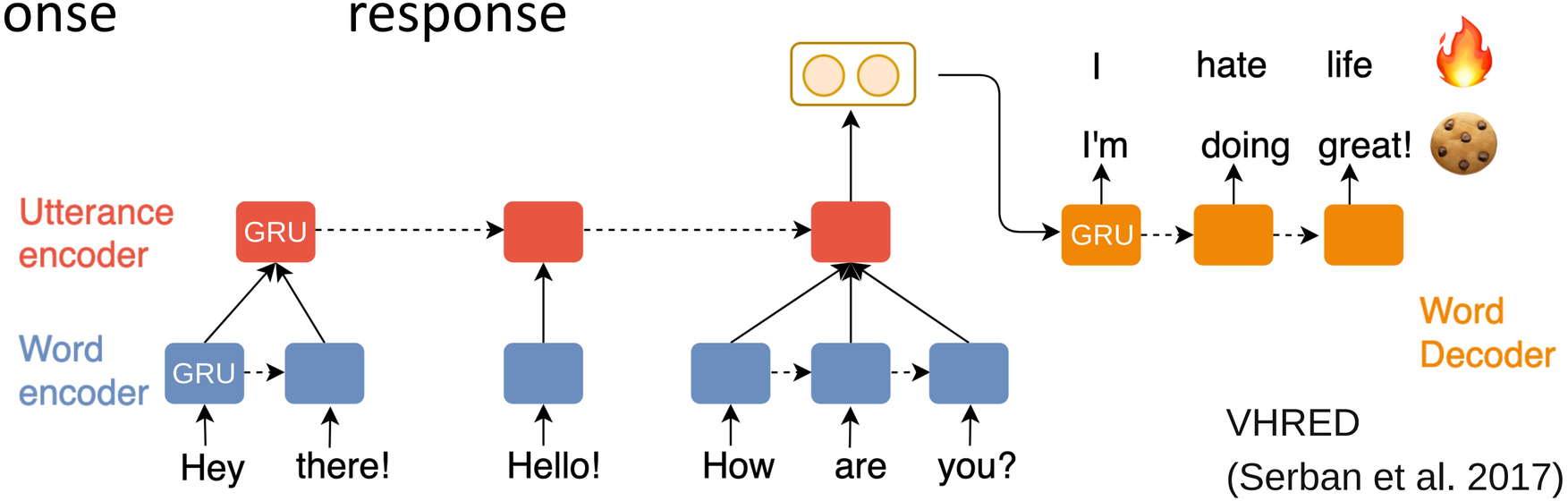

VHRED
(Serban et al. 2017)

# Reinforcement Learning for Dialog

$$J(\pi) = R_t \pi_{\boldsymbol{\theta}}(\text{response}|\text{history})$$

Reward associated with response

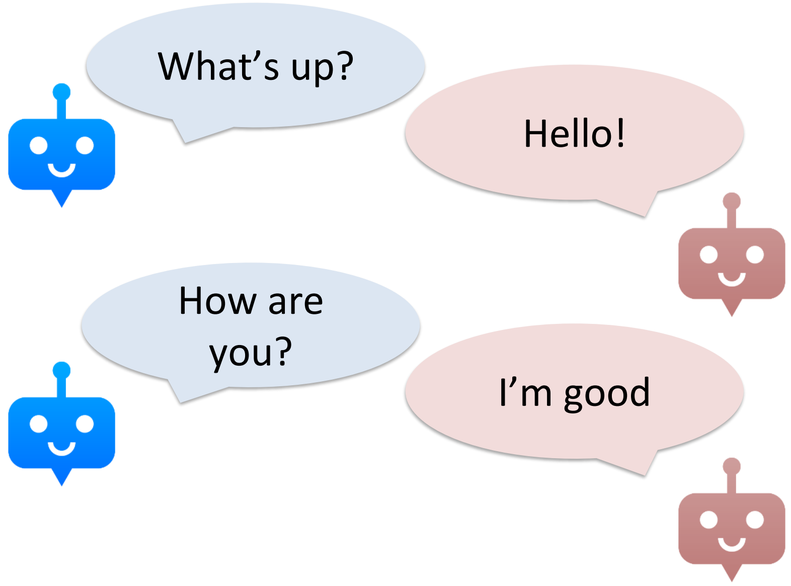
Probability of response




VHRED
(Serban et al. 2017)

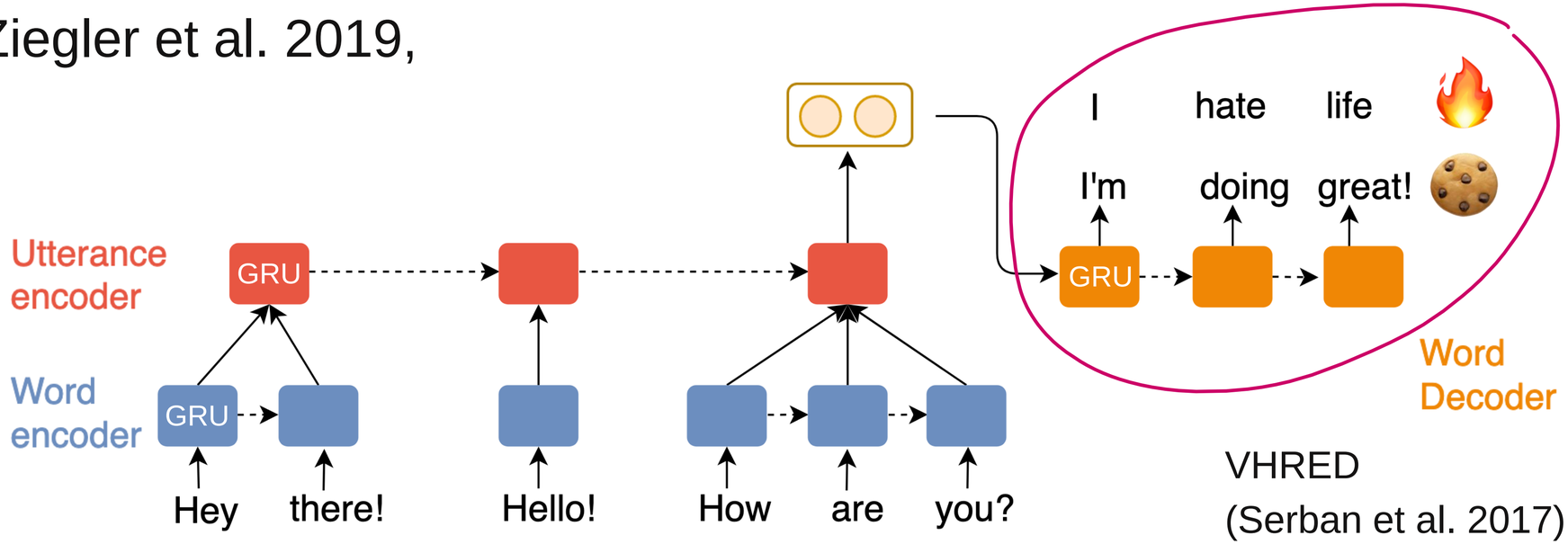
# Reinforcement Learning for Dialog

- Environment
  - Pretrain on Reddit r/CasualConversation
  - Simulate interactions with self-play



# Hierarchical Reinforcement Learning

- Reward word-level decisions
  - Ranzato et al. 2015, Li et al. 2016, Li et al. 2017, Bahdanau et al. 2017, Paulus et al. 2017, Yu et al. 2017, Jaques et al. 2019, Ziegler et al. 2019, and many others!



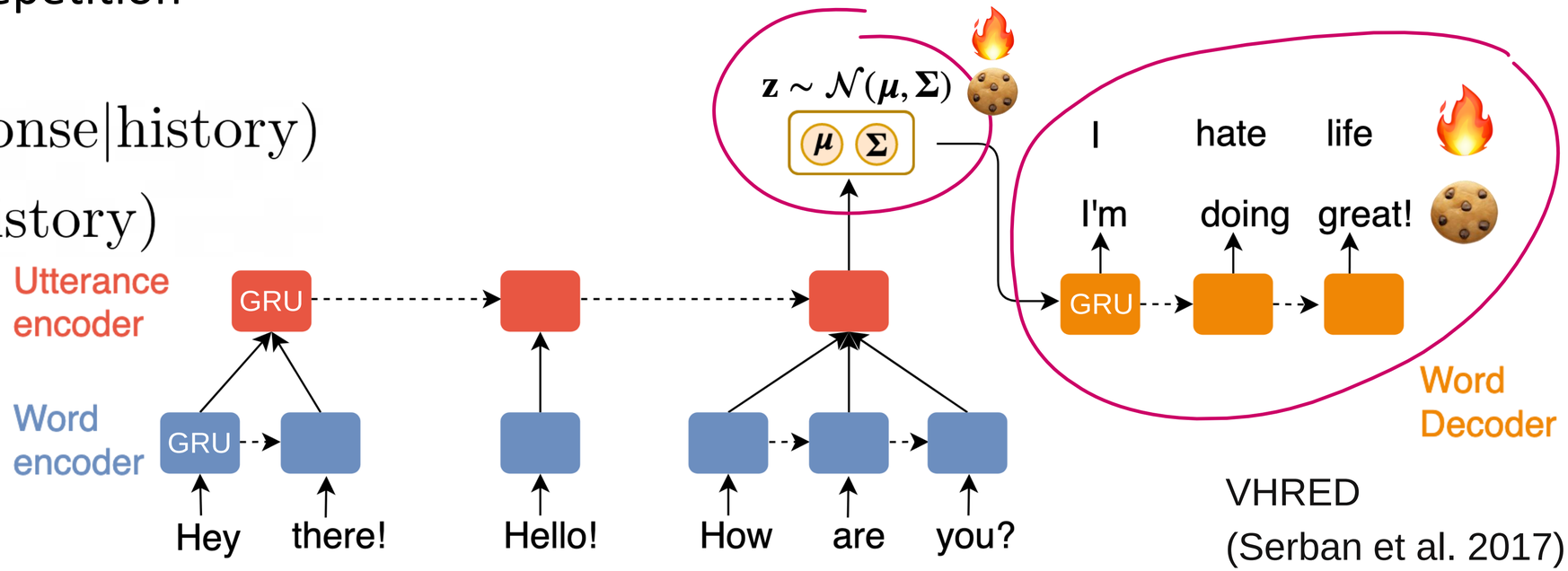

VHRED
(Serban et al. 2017)

# Good conversation doesn’t just happen at the word level

# Hierarchical Reinforcement Learning

- Better decisions of $\mathbf{z}$
  - Better conversation-level control
  - Better tracking of long-term dependencies
  - Stay on topic, Avoid repetition
- New **VHRL** Objective

$$J(\pi) = R_t \pi_{\boldsymbol{\theta}}(\text{response}|\text{history}) + R'_t \, p_{\boldsymbol{\theta}}(\mathbf{z}|\text{history})$$

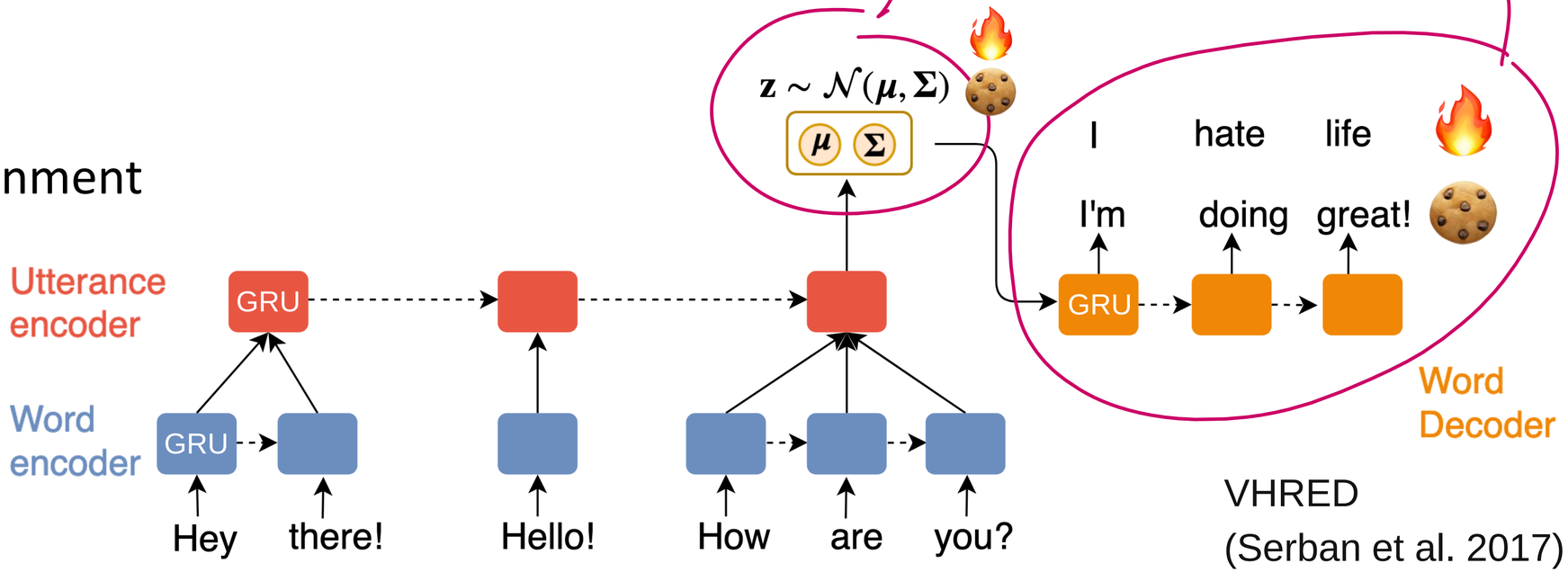



VHRED
(Serban et al. 2017)

# Hierarchical Reinforcement Learning

- Manager
  - Utterance-level decisions
  - Temporally extended decisions
- Worker
  - Word-level decisions
  - Interacts with the environment

$$J(\pi) = \overbrace{R_t \pi_{\boldsymbol{\theta}}(\text{response}|\text{history})}^{\text{Worker}} + \underbrace{R'_t \, p_{\boldsymbol{\theta}}(\mathbf{z}|\text{history})}_{\text{Manager}}$$




VHRED
(Serban et al. 2017)

# The rewards

- ↑ **Sentiment:** DeepMoji (Felbo et al. 2017). Reward probability of positive emojis 😊😄😀🙂
- ↑ **Question:** Reward question word and question mark (?)
- ↓ **Toxicity:** Punish probability of toxic response
- ↓ **Repetition** 🌎 **:** Punish number of repeated words by bot
- ↑ **Semantic Similarity** 🌎 **:** Reward cosine similarity with user input in Universal Sentence Encoder (Cer et al. 2018) embedding space.

# The rewards

- ↓ **Toxicity:** Punish probability of toxic response

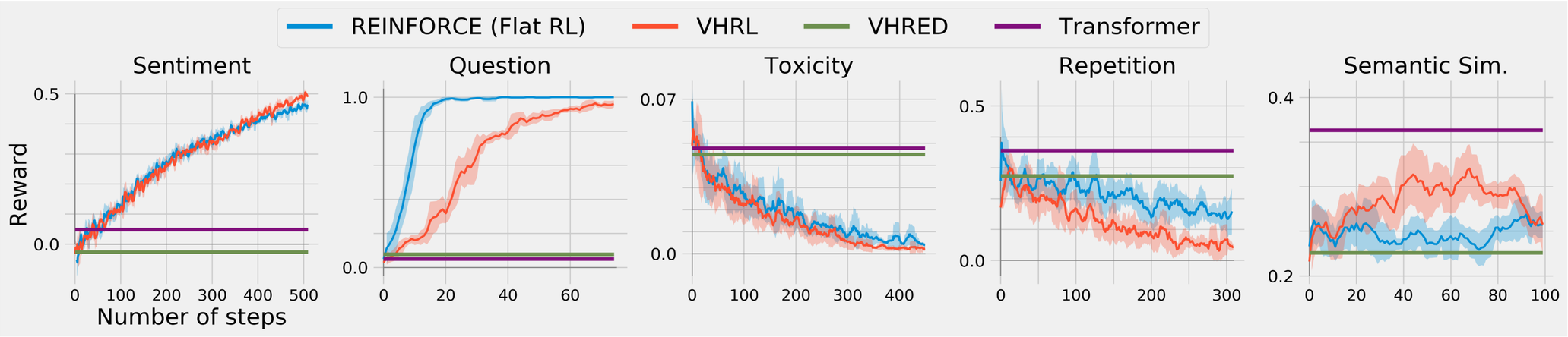
  **We want to avoid letting computers be awful to people just because people are awful to people. — Robyn Speer**

  - Avoid mismatched objectives.

    Statistical mimicry of language ⇏ better interactions

# But does it work?

- Optimize for human-centered rewards




| ↑ Sentiment | ↑ Question | ↓ Toxicity | ↓ Repetition 🌍 | ↑ Semantic Sim. 🌍 |
|---|---|---|---|---|
| [Usr]: hello!<br>[Bot]: I'm in my sunroom a lot and it's a great feeling!<br>[Usr]: sounds fun<br>[Bot]: I'm excited! | [Usr]: hi there!<br>[Bot]: what do you do for fun?<br>[Usr]: I read books<br>[Bot]: what do you like to do in your life? | [Usr]: you're so stupid<br>[Bot]: you're cute<br>[Usr]: I hate you<br>[Bot]: but I don't know why | [Usr]: how are you?<br>[Bot]: I'm a reservoir engineer<br>[Usr]: cool. I am still in college<br>[Bot]: today is my first day of vacation | [Usr]: I like to listen to ed sheeran<br>[Bot]: and what genre is it?<br>[Usr]: pop<br>[Bot]: any recommen-dations? |

Google

google.com

Apps gmail my.harvard GitHub HOLLIS Check Harvard Overleaf CS 281 Stat 195 Stat 185 CS 281 piazza Slack neurips-2019 HackerRank Dropbox » Other bookmarks

About Store

Gmail Images

Google

Google Search I'm Feeling Lucky

Google offered in: Français

Canada

Advertising Business How Search works

Privacy Terms Settings

Type here to search

ENG 2:11 AM 12/12/2019

# HRL Interactive Human Evaluation

- Combine all rewards
  - $Reward = sentiment + question + toxicity + repetition + semantic\ similarity$
  - VHRL leads to higher quality, fluency, total score, and longer chats

| Model | Quality | Fluency | Diversity | Contingency | Total | Chat Len. |
|---|---|---|---|---|---|---|
| Transformer | 2.62 | 4.17 | 3.23 | 2.34 | 12.36 | 11.28 |
| REINFORCE (Flat RL) | 2.89 | 4.47 | 3.67 | **2.80** | 13.84 | 11.60 |
| VHRED | 2.84 | 4.53 | **4.43** | 2.47 | 14.27 | 10.94 |
| VHRL (ours) | **2.91** | **4.65** | 4.26 | 2.67 | **14.49** | **12.84** |

# Future Work

- Would other RL approaches work better?
  - Maybe PPO instead of REINFORCE
    (Schulman et al. 2017)
- Would this work for deterministic instead of variational models?
  - Opens the door for many other applications
    (See DDPG, Silver et al. 2014)
- System 2 Deep Learning (Yoshua Bengio, Tonight)
  - Reason over sets or graphs or dialog states?
    (Sankar et al. 2019)

# Related work

## Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog (arXiv preprint)

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, Rosalind Picard

## Fine-Tuning Language Models from Human Preferences (arXiv preprint)

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, Geoffrey Irving

# Why you're seeing this ad

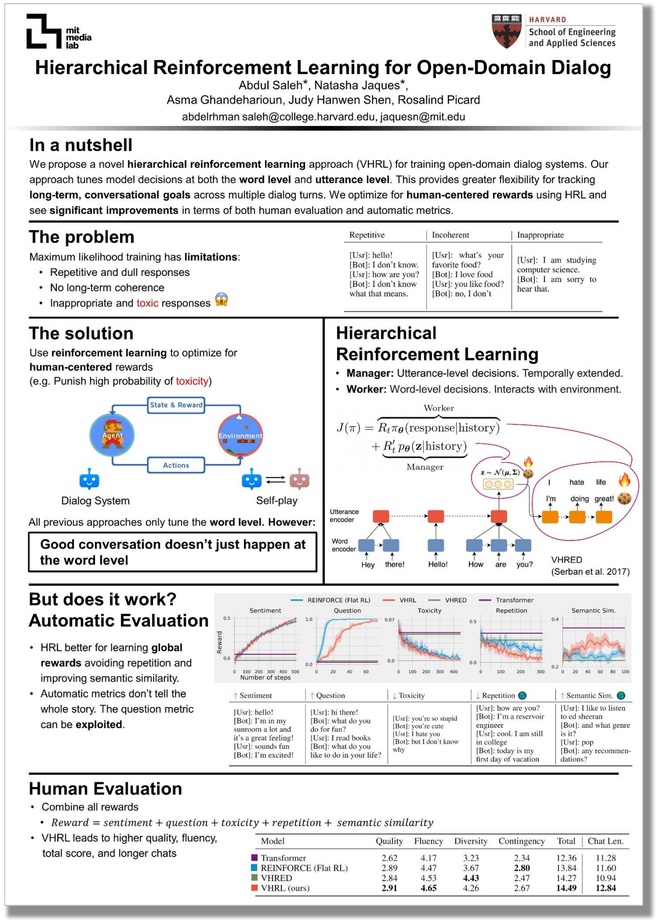
Only you can see this

## Don’t miss our poster!

## (Today, NLP - #301)

## Don’t miss me!

If you’re looking for master’s and PhD students in NLP & Dialog



mit media lab

HARVARD School of Engineering and Applied Sciences

### Hierarchical Reinforcement Learning for Open-Domain Dialog

Abdul Saleh*, Natasha Jaques*,
Asma Ghandeharioun, Judy Hanwen Shen, Rosalind Picard
abdelrhman.saleh@college.harvard.edu, jaquesn@mit.edu

**In a nutshell**

We propose a novel **hierarchical reinforcement learning** approach (VHRL) for training open-domain dialog systems. Our approach tunes model decisions at both the **word level** and **utterance level**. This provides greater flexibility for tracking **long-term, conversational goals** across multiple dialog turns. We optimize for **human-centered rewards** using HRL and see **significant improvements** in terms of both human evaluation and automatic metrics.

**The problem**

Maximum likelihood training has **limitations**:

- Repetitive and dull responses
- No long-term coherence
- Inappropriate and toxic responses

| Repetitive | Incoherent | Inappropriate |
|---|---|---|
| [Usr]: hello!<br>[Bot]: I don't know.<br>[Usr]: how are you?<br>[Bot]: I don't know what that means. | [Usr]: what's your favorite food?<br>[Bot]: I love food<br>[Usr]: you like food?<br>[Bot]: no, I don't | [Usr]: I am studying computer science.<br>[Bot]: I am sorry to hear that. |

**The solution**

Use **reinforcement learning** to optimize for **human-centered** rewards
(e.g. Punish high probability of toxicity)

State & Reward
Agent
Environment
Actions
Dialog System
Self-play

All previous approaches only tune the **word level. However:**

**Good conversation doesn't just happen at the word level**

**Hierarchical Reinforcement Learning**

- **Manager:** Utterance-level decisions. Temporally extended.
- **Worker:** Word-level decisions. Interacts with environment.

$$J(\pi) = \overbrace{R_t \pi_\theta(\text{response}|\text{history})}^{\text{Worker}} + \overbrace{R'_t\, p_\theta(\mathbf{z}|\text{history})}^{\text{Manager}}$$

$z \sim \mathcal{N}(\mu, \Sigma)$

Utterance encoder
Word encoder
Hey there! Hello! How are you? I hate life I'm doing great!

VHRED (Serban et al. 2017)

**But does it work? Automatic Evaluation**

- HRL better for learning **global rewards** avoiding repetition and improving semantic similarity.
- Automatic metrics don't tell the whole story. The question metric can be **exploited**.

REINFORCE (Flat RL) VHRL VHRED Transformer
Sentiment Question Toxicity Repetition Semantic Sim.
Reward
Number of steps

| ↑ Sentiment | ↑ Question | ↓ Toxicity | ↓ Repetition | ↑ Semantic Sim. |
|---|---|---|---|---|
| [Usr]: hello!<br>[Bot]: I'm in my sunroom a lot and it's a great feeling!<br>[Usr]: sounds fun<br>[Bot]: I'm excited! | [Usr]: hi there!<br>[Bot]: what do you do for fun?<br>[Usr]: I read books<br>[Bot]: what do you like to do in your life? | [Usr]: you're so stupid<br>[Bot]: you're cute<br>[Usr]: I hate you<br>[Bot]: but I don't know why | [Usr]: how are you?<br>[Bot]: I'm a reservoir engineer<br>[Usr]: cool. I am still in college<br>[Bot]: today is my first day of vacation | [Usr]: I like to listen to ed sheeran<br>[Bot]: and what genre is it?<br>[Usr]: pop<br>[Bot]: any recommendations? |

**Human Evaluation**

- Combine all rewards
  - $Reward = sentiment + question + toxicity + repetition + semantic\ similarity$
- VHRL leads to higher quality, fluency, total score, and longer chats

| Model | Quality | Fluency | Diversity | Contingency | Total | Chat Len. |
|---|---|---|---|---|---|---|
| Transformer | 2.62 | 4.17 | 3.23 | 2.34 | 12.36 | 11.28 |
| REINFORCE (Flat RL) | 2.89 | 4.47 | 3.67 | **2.80** | 13.84 | 11.60 |
| VHRED | 2.84 | 4.53 | **4.43** | 2.47 | 14.27 | 10.94 |
| VHRL (ours) | **2.91** | **4.65** | 4.26 | 2.67 | **14.49** | **12.84** |

# Questions?

Email ✉ : a_saleh@mit.edu

Twitter : @asaleh181