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

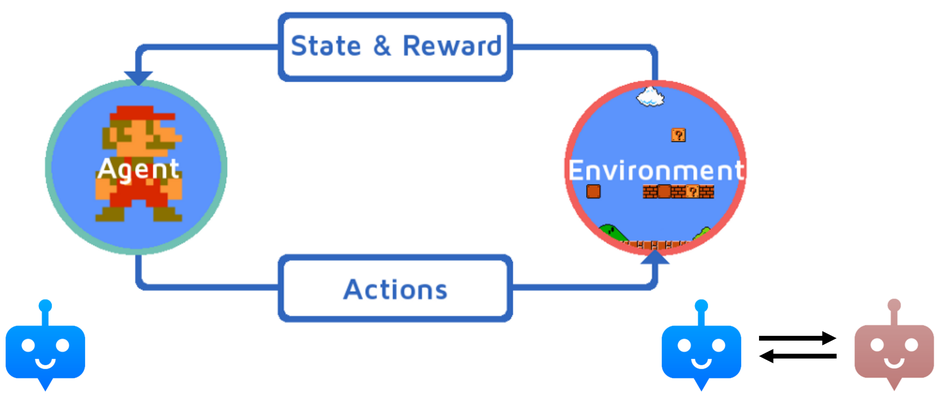
Maximum likelihood training has **limitations**:

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

Repetitive	Incoherent	Inappropriate
[Usr]: hello! [Bot]: I don't know. [Usr]: how are you? [Bot]: I don't know what that means.	[Usr]: what's your favorite food? [Bot]: I love food [Usr]: you like food? [Bot]: no, I don't	[Usr]: I am studying computer science. [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)



Dialog System

Self-play

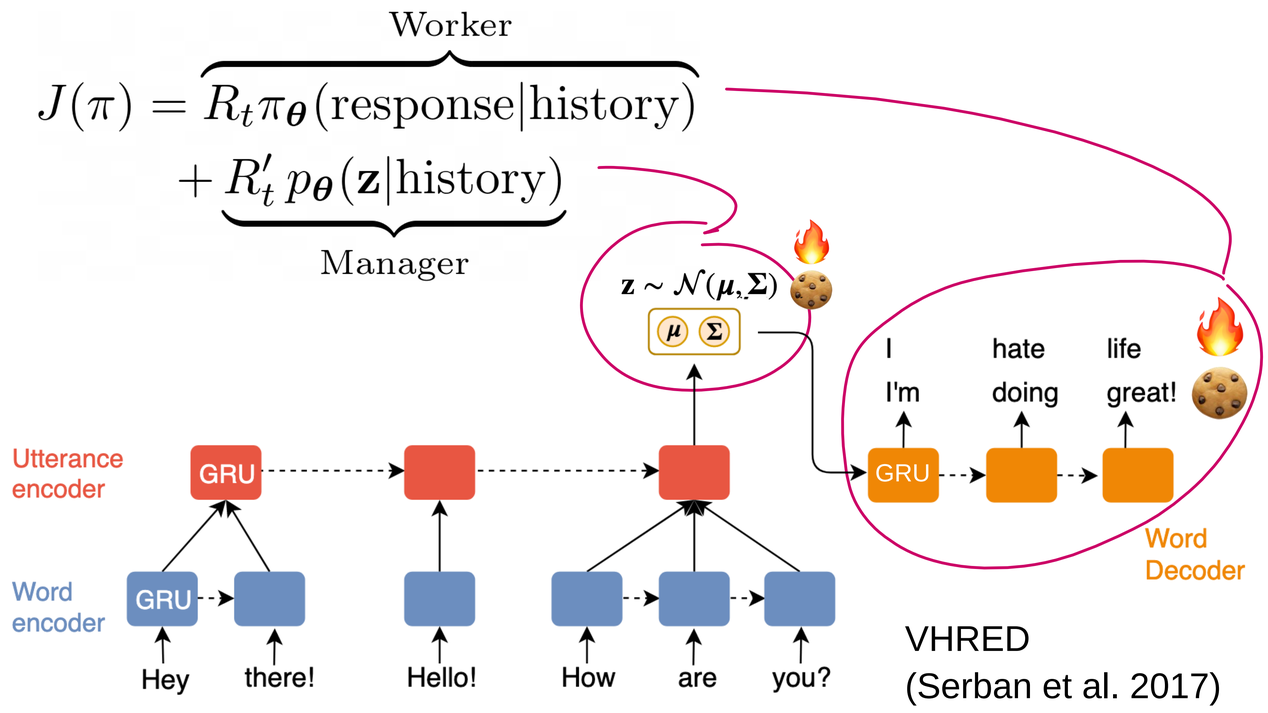
All previous approaches only tune the **word level. However:**

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.

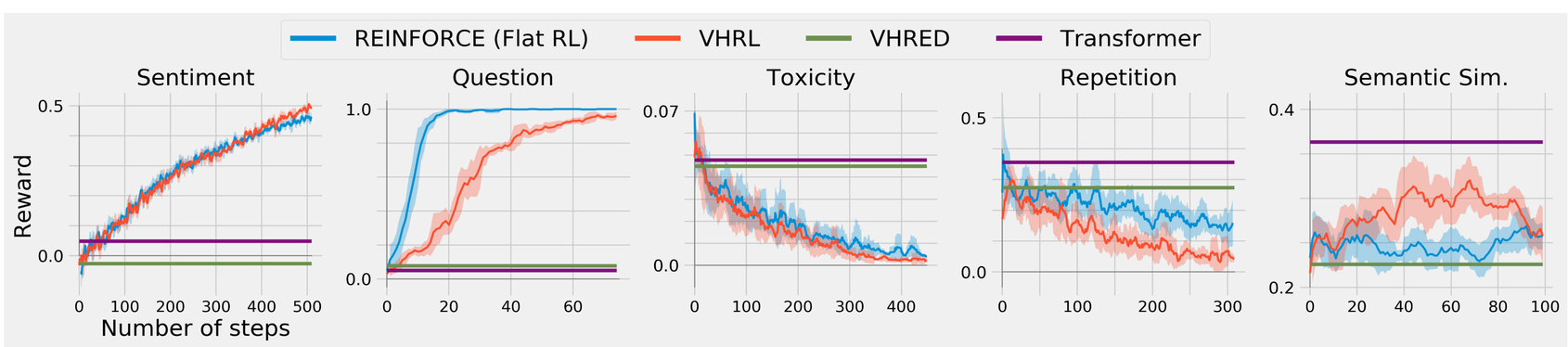
$$J(\pi) = \overbrace{R_t \pi_{\boldsymbol{\theta}}(\text{response}|\text{history})}^{\text{Worker}} + \underbrace{R'_t\, p_{\boldsymbol{\theta}}(\mathbf{z}|\text{history})}_{\text{Manager}}$$



VHRED
(Serban et al. 2017)

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**.



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

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**