What have we done so far?

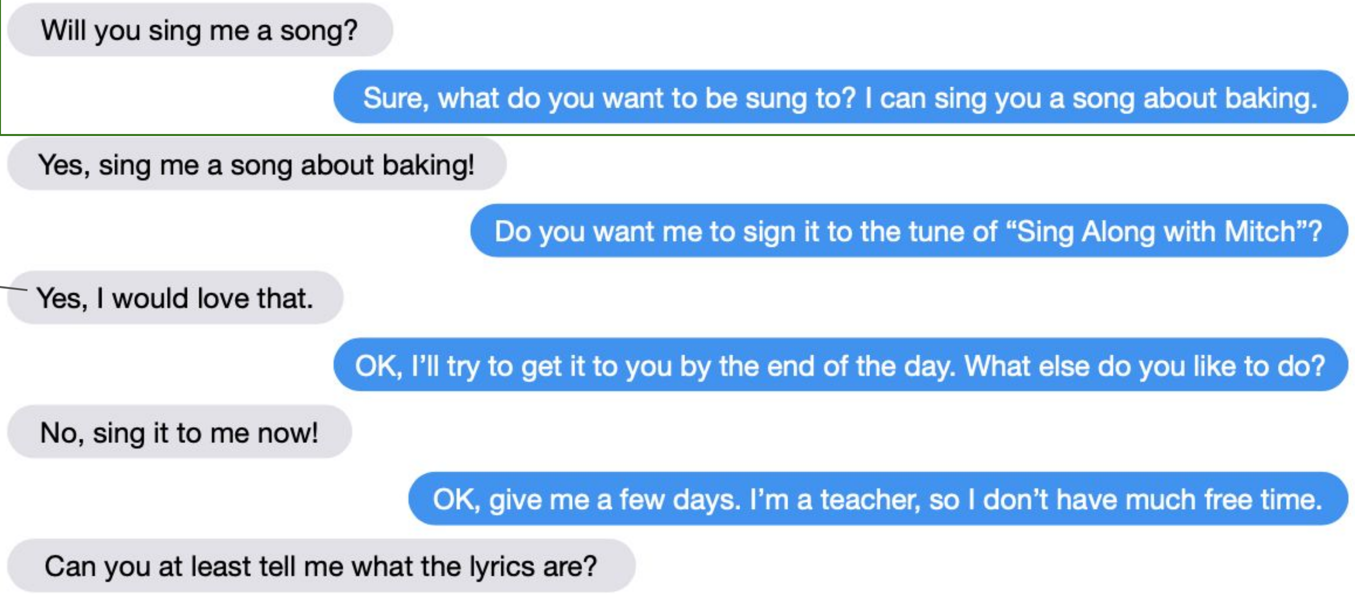
Summaries

- Jurafsky’s chapter: Chatbots & Dialog Systems
- Blog: Learning through human feedback
- Ziegler 2020
- Jaques 2019
- Hancock 2019
- Gained additional Information from Tutorials

Jurafsky - What’s a Dialog?

User (has the *Initiative*)

Conversational agent: Chatbot



Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sign it to the tune of “Sing Along with Mitch”?

Yes, I would love that.

OK, I’ll try to get it to you by the end of the day. What else do you like to do?

No, sing it to me now!

OK, give me a few days. I’m a teacher, so I don’t have much free time.

Can you at least tell me what the lyrics are?

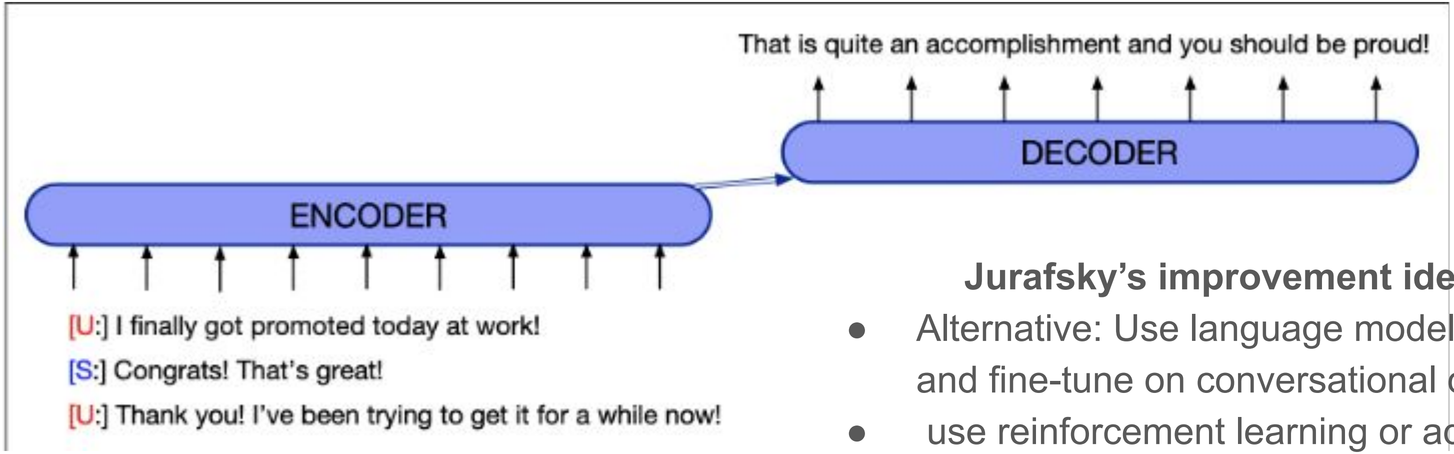
Adjacency pair:
Proposal &
Commitment

Turn

[1] Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2020.

Corpus based chatbots with response by generation

- Transformer Network => generate each token at a time based on query q and the response so far r_1 - r_{t-1} $\hat{r}_t = \text{argmax}_{w \in V} P(w|q, r_1 ... r_{t-1})$



[U:] I finally got promoted today at work!

[S:] Congrats! That's great!

[U:] Thank you! I've been trying to get it for a while now!

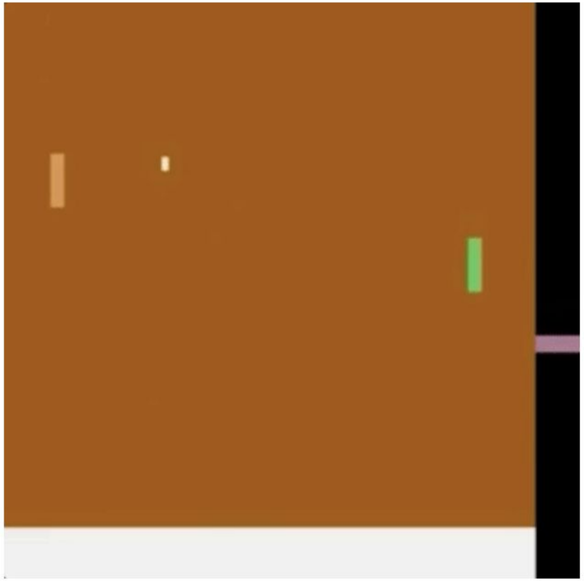
Jurafsky's improvement ideas

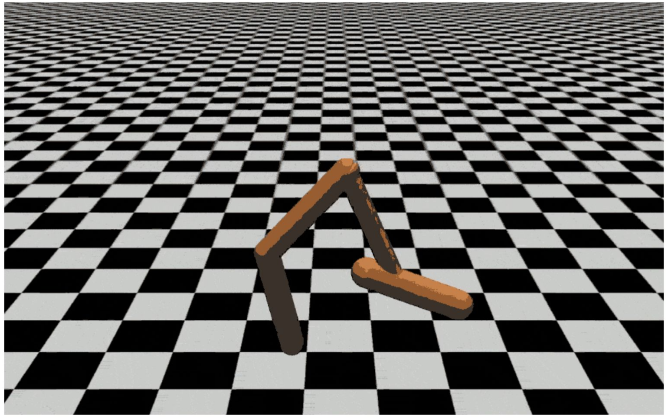
- Alternative: Use language model as generator and fine-tune on conversational dataset
- use reinforcement learning or adversarial networks => learn making the conversation more natural in the long run

[2] Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2020.

Blog: Learning through human feedback

Tasks like Atari gameplay or a simulated robot learning a bagflip

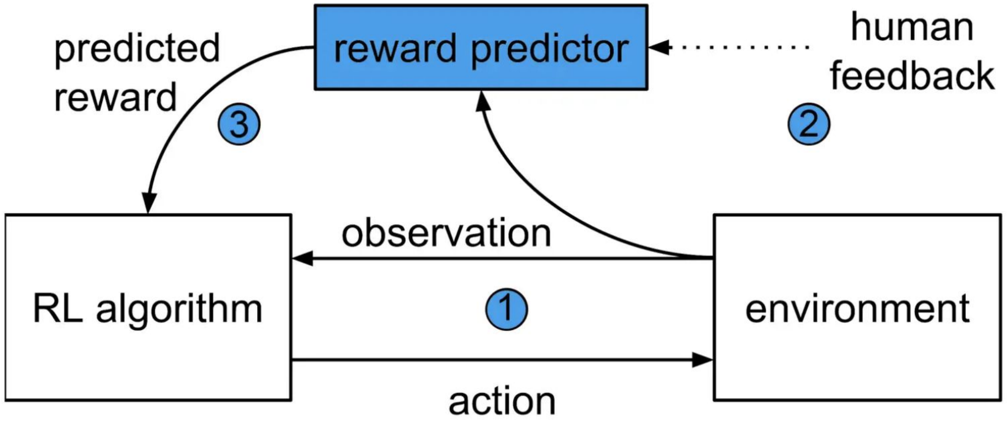




[2] Jan Leike, Mijan Martic, Shane Legg. Blog - Learning through human feedback
https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

Learning through human feedback

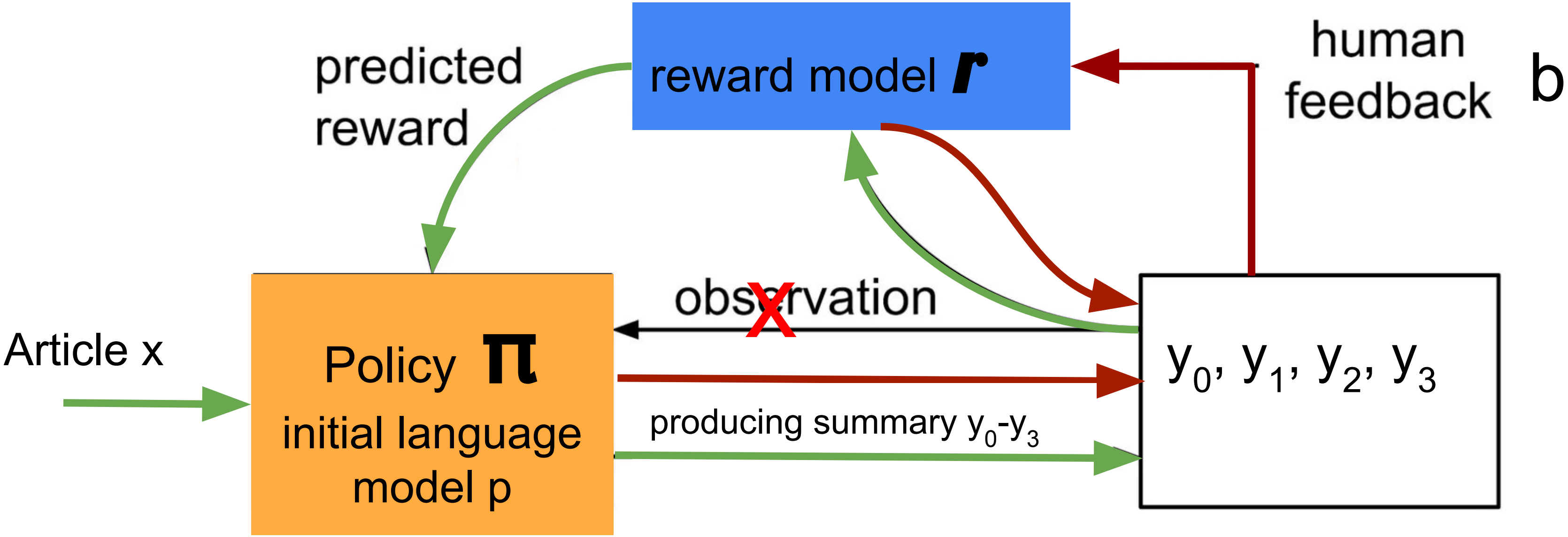
1. RL agent explores environm.
2. Periodically a human selects one of two behaviours
3. The human choice is trained to use the **reward predictor** which trains the agent



[2] Jan Leike, Mijan Martic, Shane Legg. Blog - Learning through human feedback
https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

… transferred to Zieglers model

1. Gather samples from initial π
2. Train *r* on human samples with loss (1)
3. Train π with reward R (2)



[3] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, Geoffrey Irving: Fine-Tuning Language Models from Human Preferences. CoRR abs/1909.08593 (2019)

Jaques

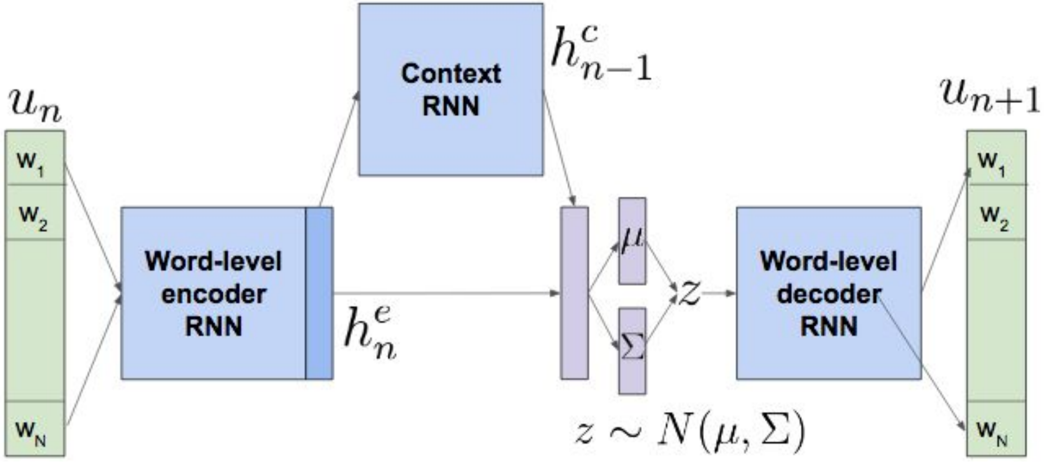
Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog

New RL algorithms able to effectively learn offline, without exploring, from a fixed batch of human interaction data

How it works:

- models pre trained on data as a strong prior → use KL control to penalize divergence
- use humans’ implicit reactions (sentiments, length etc.)
- collecting of data with an interactive online platform

RL for open domain dialog generation



- construction of a response utterance by iteratively choosing an action as the next token
- human response used to compute a reward signal to train the model

  → at the last token of the bot’s utterance, the estimated future reward must include the human’s response

  → goal of using reward functions: Production of positive reactions

Psychology of human conversation

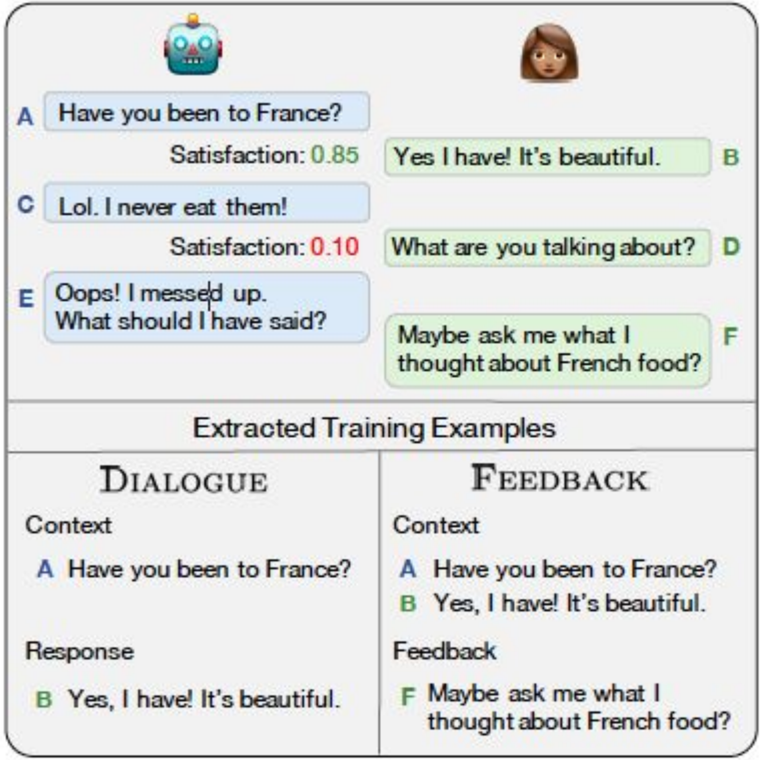
- Eliciting positive sentiment → sense of understanding
- Eliciting longer conversations → signal of engagement
- Eliciting laughter → building solidarity
- High semantic similarity → paraphrasing and style matching
- Asking questions → active listening skill

Hancock - Learning from Dialogue after Deployment

- **self training chatbot**

  → asking for feedback & learning to predict it

  → classifying user satisfaction



A Have you been to France?

Satisfaction: 0.85

Yes I have! It's beautiful. B

C Lol. I never eat them!

Satisfaction: 0.10

What are you talking about? D

E Oops! I messed up.
What should I have said?

Maybe ask me what I
thought about French food? F

Extracted Training Examples

DIALOGUE

Context

A Have you been to France?

Response

B Yes, I have! It's beautiful.

FEEDBACK

Context

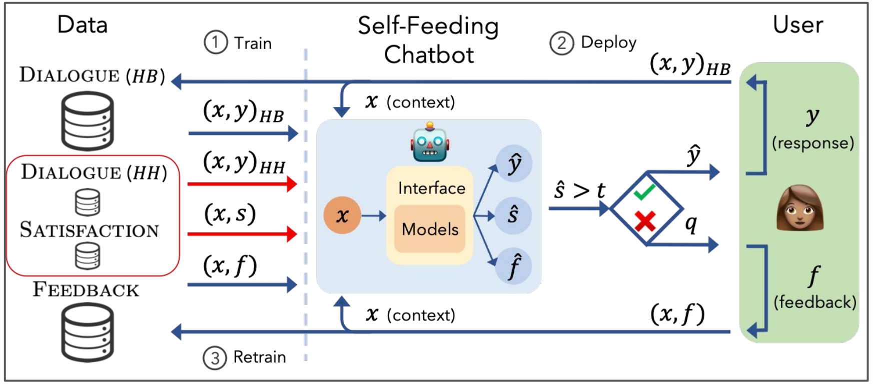
A Have you been to France?

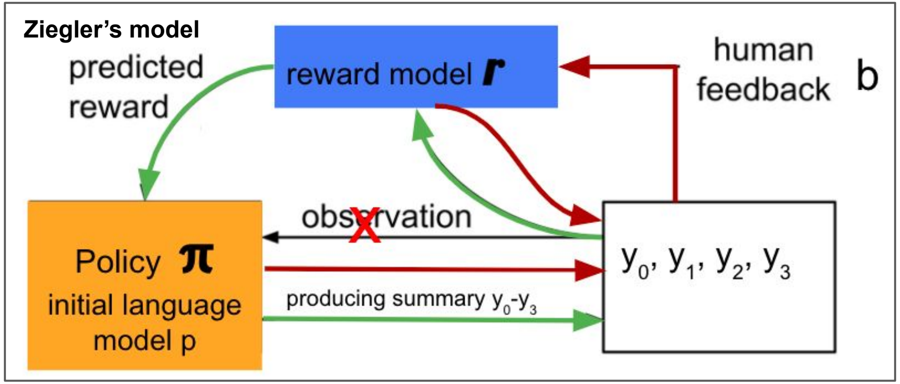
B Yes, I have! It's beautiful.

Feedback

F Maybe ask me what I
thought about French food?

Hancock - Learning from Dialogue after Deployment





Usage of free text input for human feedback maybe a possible more complex extension of Zieglers model