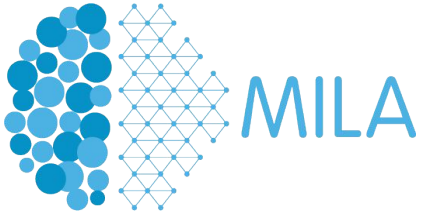
A Deep Reinforcement Learning Chatbot

Iulian Vlad Serban

Department of Computer Science and Operations Research

University of Montreal

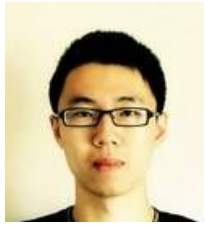
Montreal, Quebec, Canada



McGill

MILA Team

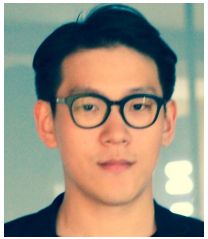
Chinnadhurai Sankar



Saizheng Zhang

Zhouhan Lin

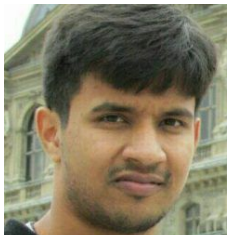
Sandeep Subramanian



Taesup Kim

Sarath Chandar

Rosemary Nan Ke



Sai Mudumba

Alexandre de Brebisson

Jose Sotelo

Dendi Suhubdy

Vincent Michalski

Alexandre Nguyen

Yoshua Bengio

Staff & Special Thanks

Mathieu Germain

Michael Pieper

Joelle Pineau

Aaron Courville

Ansona On Yi Ching

Michael Noseworthy

Prasanna Parthasarathi

Nicolas Angelard-Gontier

Peter Henderson

Ryan Lowe

Amazon Alexa Prize Competition

University Challenge

- Develop conversational agent (socialbot), which can discuss a wide range of topics

- Millions will soon interact with bots through Alexa platform

University of Montreal is one of 12 sponsored teams from +100 team applications.

Amazon Selects Teams to Compete for Inaugural $2.5 Million Alexa Prize

Amazon will sponsor 12 university teams to compete in the 2016-2017 Alexa Prize. This year's inaugural competition focuses on the grand challenge of building a socialbot that can converse coherently and engagingly with humans on popular topics for 20 minutes. The sponsored teams will receive a $100,000 stipend, Alexa-enabled devices, free Amazon Web Services (AWS) services to support their development efforts, and support from the Alexa Skills Kit (ASK) team.

Amazon received over one hundred applications from leading universities across 22 countries.

All applications were reviewed and evaluated based on the following criteria: the potential scientific contribution to the field, the technical merit of the approach, the novelty of the idea, and the team's ability to execute against their plan.

Industry

Google

BOTLER

Microsoft

slack

SAMSUNG

AUTOMAT

Apple

facebook

amazon

IBM

ETS

interactions

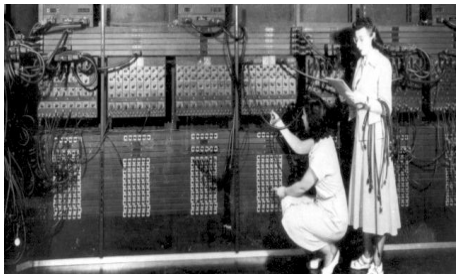
Baidu 百度

snapchat

Human-Computer Interaction: Chronology

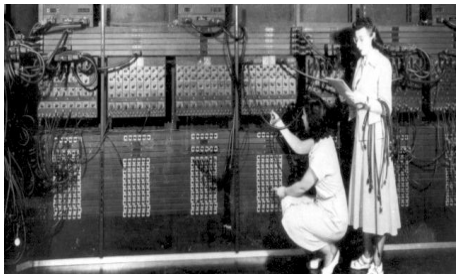
Human-Computer Interaction: Chronology

Patch Cords & Punch Cards



Human-Computer Interaction: Chronology

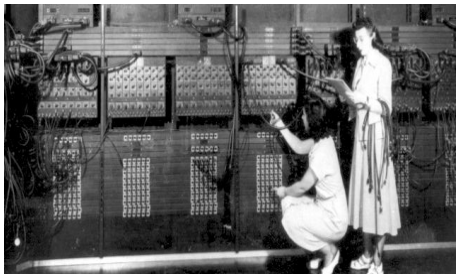
Patch Cords & Punch Cards



Terminals

Human-Computer Interaction: Chronology

Patch Cords & Punch Cards

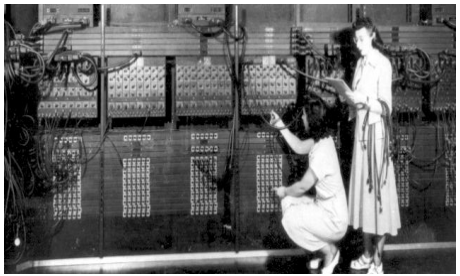


Terminals

Graphical User Interfaces

Human-Computer Interaction: Chronology

Patch Cords & Punch Cards



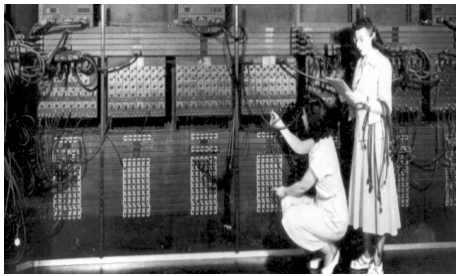
Terminals

Graphical User Interfaces

Touch User interfaces

Human-Computer Interaction: Chronology

Patch Cords & Punch Cards

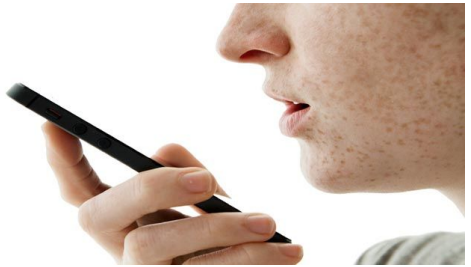


Terminals

Graphical User Interfaces

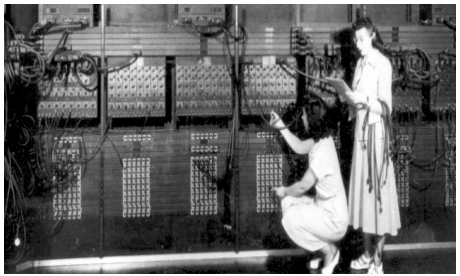
Touch User interfaces

Spoken & Chat Interfaces



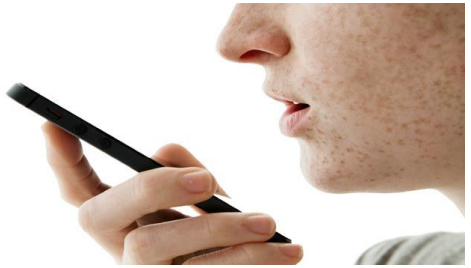
Human-Computer Interaction: Chronology

Patch Cords & Punch Cards



Touch User interfaces

Spoken & Chat Interfaces



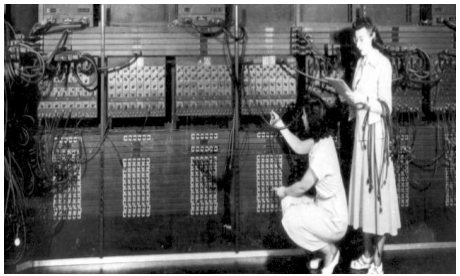
Terminals

Graphical User Interfaces

Embodied Interfaces

Human-Computer Interaction: Chronology

Patch Cords & Punch Cards



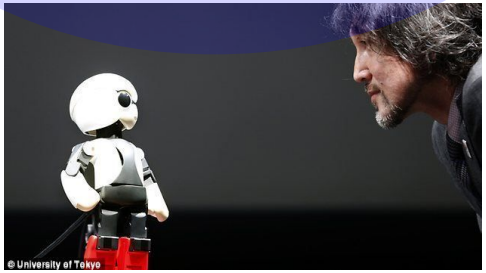
Terminals

Graphical User Interfaces

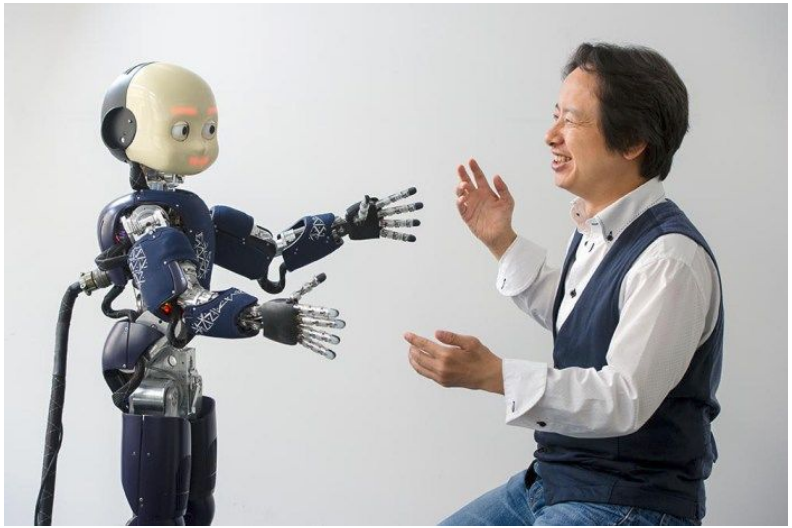
Touch User interfaces

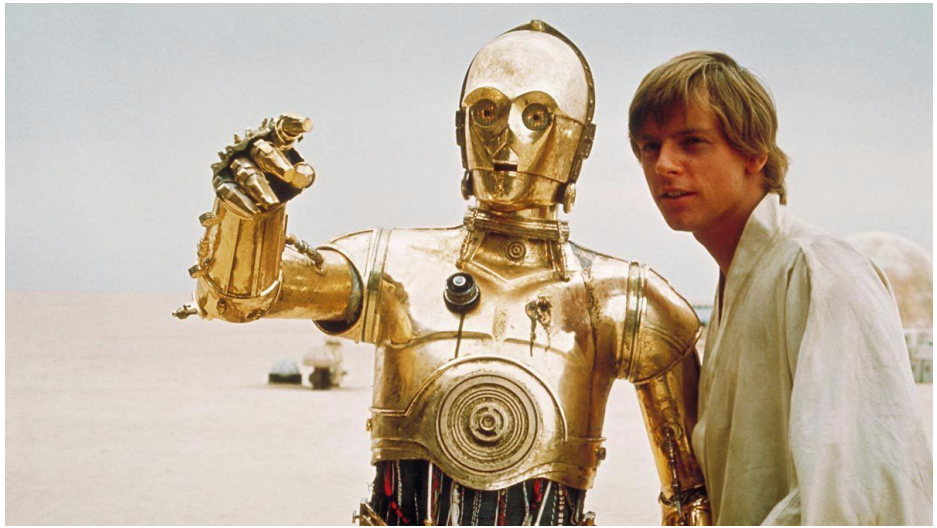
Spoken & Chat Interfaces

Embodied Interfaces

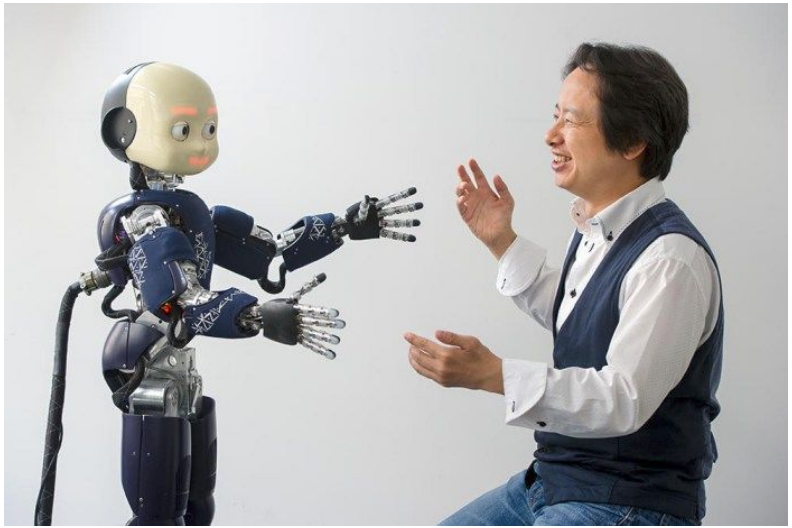


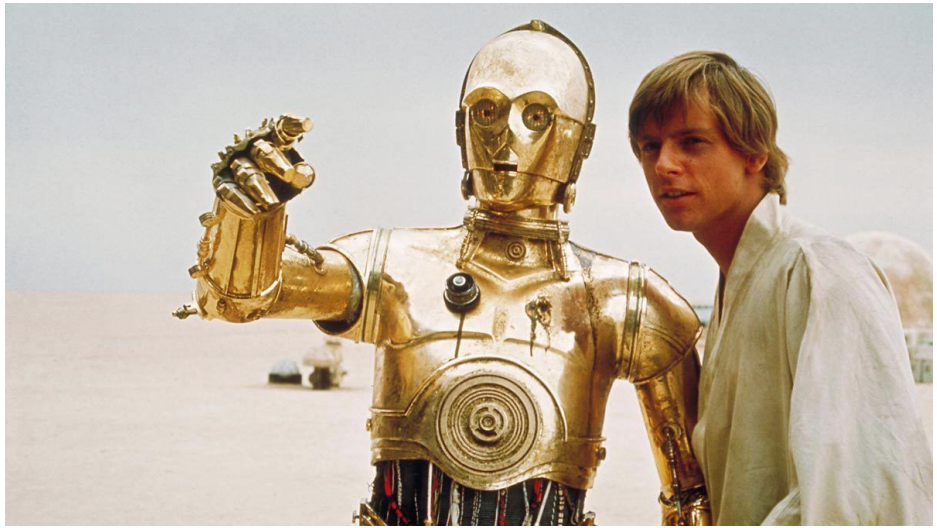
Research Vision: Why We Care





Research Vision: Why We Care

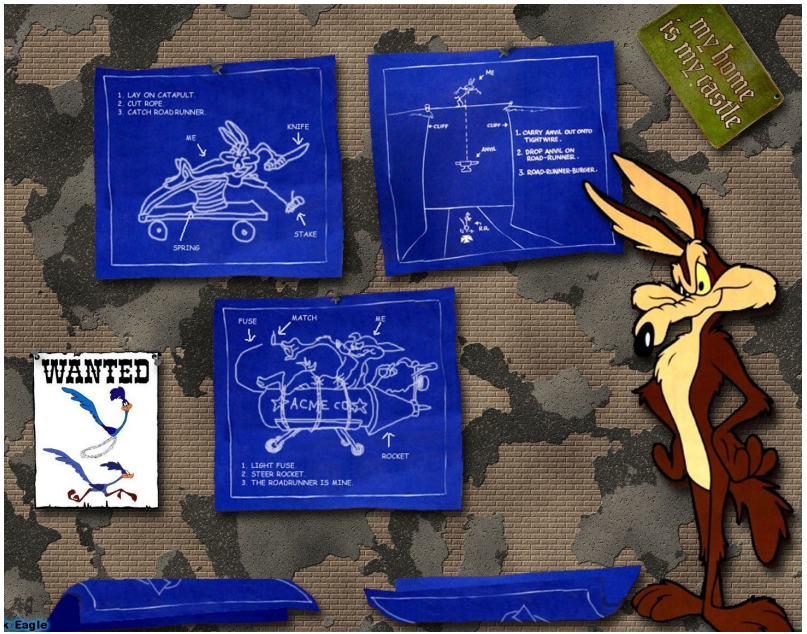




Research Vision: Why We Care

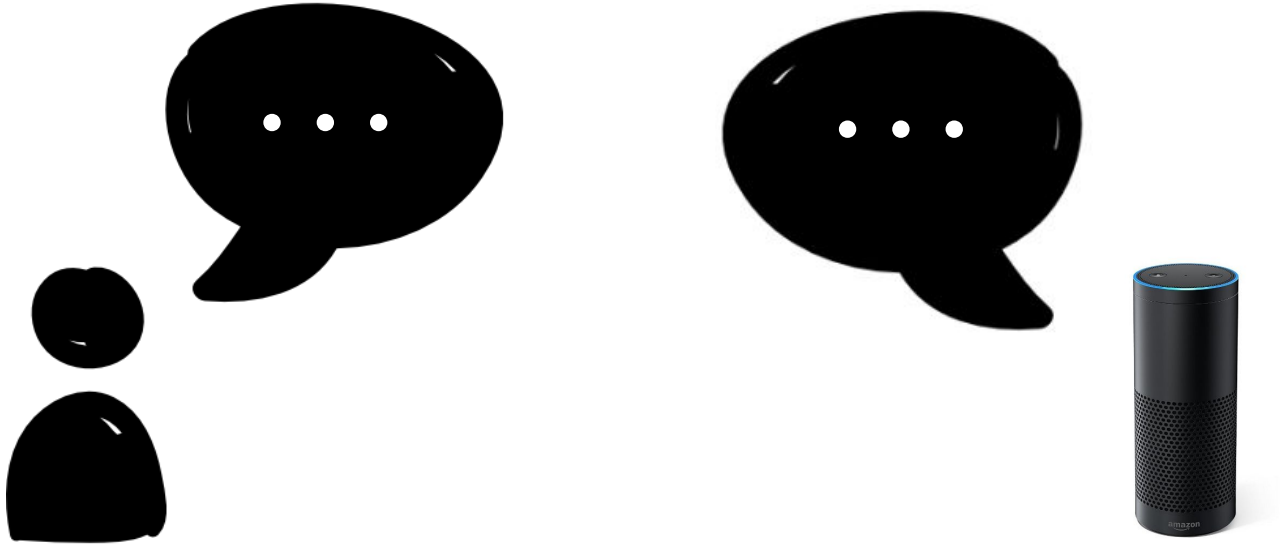
Natural Language Processing

Sequential Decision Making

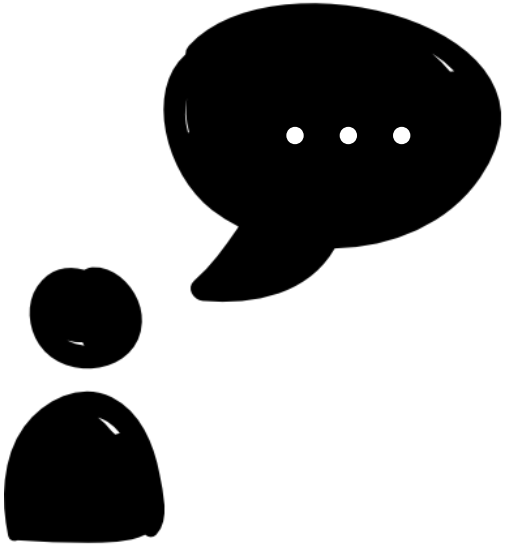


System Overview

System Overview



System Overview

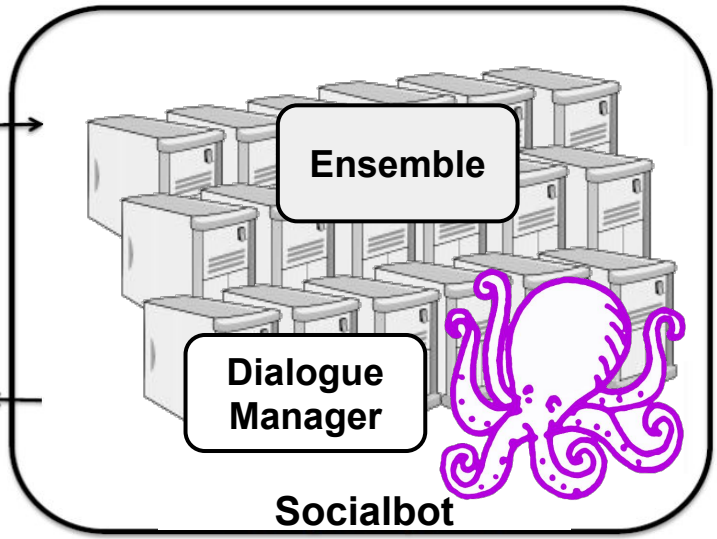


Automatic Speech Recognizer

Text-To-Speech Synthesizer

Web Services / APIs:

- Search Engines
- Question Answering
- Knowledge Bases



Demo

Response Models

Response Models

22 Response Models:

- 4 Template and Rule-based Models
- 15 Neural Network Models
- 2 Knowledge Base Question Answering Models
- 1 Logistic Regression Model

Example Response Models

Response Model: Alicebot

- All-purpose model
- Uses pattern matching rules to generate response

Example

```
<category>
<pattern>WHAT IS YOUR FAVORITE TV *</pattern>
<template>My favorite show is <bot name="favoriteshow"/>.</template>
</category>
```

What is your favorite TV episode? → My favorite show is The Simpsons.

Response Model: Evibot

- Amazon's proprietary question-answering web-service: www.evi.com
- Evi handles mainly factual questions

Examples

How many people live in Greenland?	→	The population of Greenland is about 56,300.
What do bears eat?	→	Bears eat ice cream, honey, meat (excluding seafood), fish, salmon, trout...
What's the best cheese in the world?	→	Sorry, I don't yet have an answer to that question.

Response Model: Initiatorbot

- Conversation starter phrases
- Model takes priority when user gives greeting

"I find it a very good conversation starter."

Examples

Hello!	→	What did you do today?
Hi. How are you?	→	Tell me about your hobbies?

Response Model: BoWFactGenerator

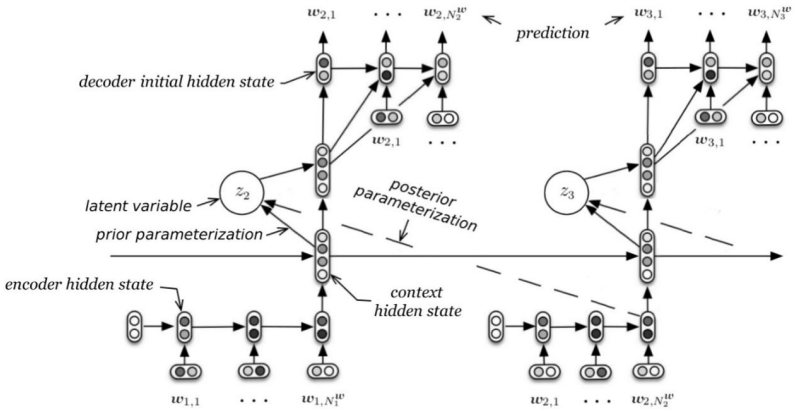
- Outputs interesting and funny facts
- Bag-of-words Word2Vec embedding similarity applied to find most relevant fact

Examples

I love dogs!	→	Here's a funny fact! Dogs have lived with humans for over 14,000 years.
Do you have a brain?	→	I don't know, but did you know that the human brain is about 75% water.
Have you ever been to the Himalayas?	→	I'm not quite sure, but did you know that the Himalayas cover one-tenth of the Earth's surface.

Response Model: VHREDWashingtonPost

- Outputs news comments:
- **Step 1**: Retrieves K=20 comments from Washington Post using word embeddings
- **Step 2:** Returns response with highest log-likelihood under VHRED (Serban et al., 2017)



Examples

Let's talk about Game of Thrones?	→	You forgot Tyrion's scene in which he releases the dragons.
I don't like cats. I like dogs.	→	No, you are a cat lady. nothing crazy about it.

Model Selection Policy

Crowdsourcing + Reinforcement Learning

Model Selection Policy

- Model selection policy selects response from candidate responses
- Learning policy is a reinforcement learning problem
- System is an agent; given state h_t, system takes action:

$$a_t \in \{a_t^1, \ldots, a_t^K\}$$

- System must maximize the *expected cumulative return*:

$$R = \sum_{t=1}^{T} \gamma^t r_t$$

Model Selection Policy

Action-value Parametrization:

- Estimate expected return for any {state, action} pair:

$$Q_\theta(h_t, a_t^k) \quad \text{for } k = 1, \ldots, K$$

- Select action maximizing (estimated) expected return:

$$\pi_\theta(h_t) = \arg\max_k Q_\theta(h_t, a_t^k)$$

Model Selection Policy

Stochastic Policy Parametrization:

- Discrete distribution over actions:

$$\pi_\theta(a_t^k|h_t) = \frac{e^{\lambda^{-1} f_\theta(h_t, a_t^k)}}{\sum_{a_t'} e^{\lambda^{-1} f_\theta(h_t, a_t')}} \quad \text{for } k = 1, \ldots, K$$

where $f_\theta(h_t, a_t^k)$ outputs score for each {state, action} pair, and λ is temperature

Model Selection Policy

Stochastic Policy Parametrization:

- Select action according to probability $\pi_\theta(a_t^k|h_t)$

- Also possible to use greedy variant:

$$\pi_\theta^{\text{greedy}}(h_t) = \arg\max_k \pi_\theta(a_t^k|h_t)$$

Model Selection Policy

- For both parametrizations, must learn a scoring function:

Action-value parametrization: $Q_\theta(h_t, a_t^K)$

Stochastic policy parametrization: $f_\theta(h_t, a_t^k)$

Model Selection Policy

Scoring Model Architecture:

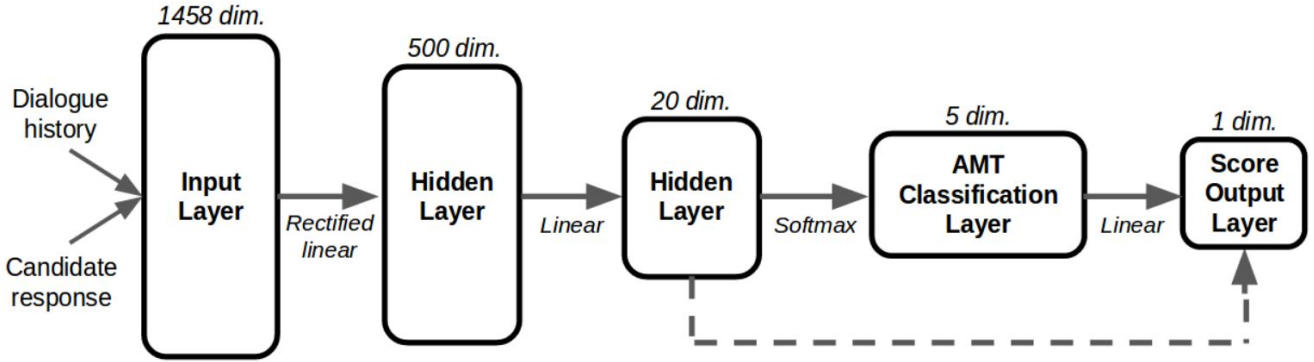


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Model Selection Policy

Three approaches:

- Supervised AMT
- Off-policy REINFORCE
- Q-learning AMT (Abstract Discourse MDP)

Supervised AMT

Crowdsourcing:

- ~200,000 labels on Amazon Mechanical Turk (AMT)
- Annotators rate dialogue responses on scale 1-5:

1: “Very Poor” response
2: “Poor” response
3: “Acceptable” response
4: “Good” response
5: “Excellent” response

Conversation	Response 1	Response 2	Response 3	Response 4
A: you need to work on your English B: Why do you say that about me? A: Well your English is very poor	But English is my native language.	What other reasons come to mind?	Here's a funny fact! Go. is the shortest complete sentence in the English language.	bye doggie
Score	4	3	3	2

Instructions:

Rate the appropriateness of the response between **1** (inappropriate, does not make any sense) and **5** (highly appropriate and interesting). The score **3** indicates neutral (acceptable, but not interesting). Remember to take into account the previous conversation.

Next

3/28

Press "Next" after filling in all the text boxes.

Supervised AMT

Training:

- Train scoring model on AMT labels using log-likelihood
- Fix last layer as: [1.0, …, 5.0]

Conversation	Response 1	Response 2	Response 3	Response 4
A: you need to work on your English B: Why do you say that about me? A: Well your English is very poor	But English is my native language.	What other reasons come to mind?	Here's a funny fact! Go. is the shortest complete sentence in the English language.	bye doggie
Score	4 ▾	3 ▾	3 ▾	2 ▾

Instructions:

Rate the appropriateness of the response between **1** (inappropriate, does not make any sense) and **5** (highly appropriate and interesting). The score **3** indicates neutral (acceptable, but not interesting). Remember to take into account the previous conversation.

Next

3/28

Press "Next" after filling in all the text boxes.

Supervised AMT

Training:

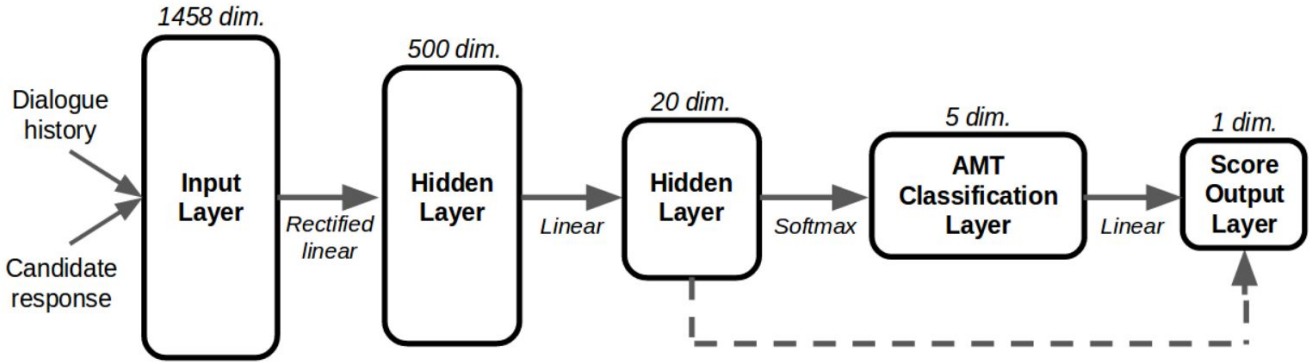


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Supervised AMT

Training:

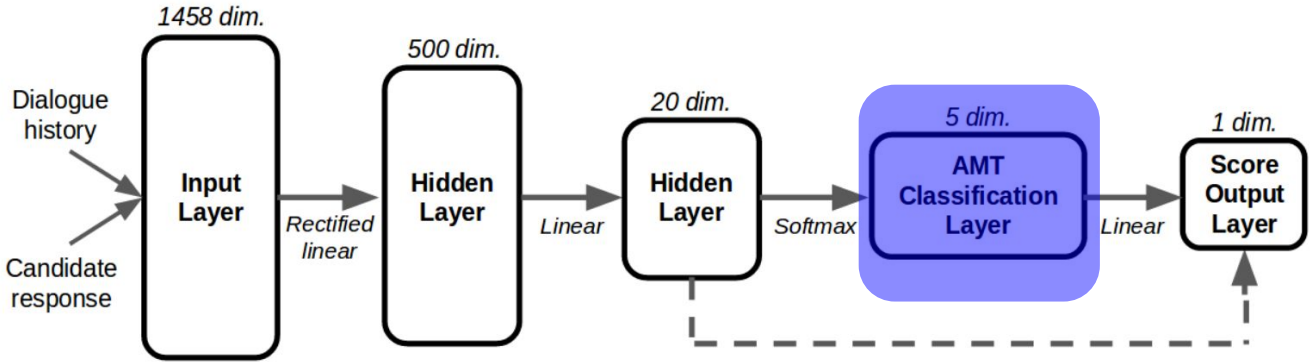


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Supervised AMT

Training:

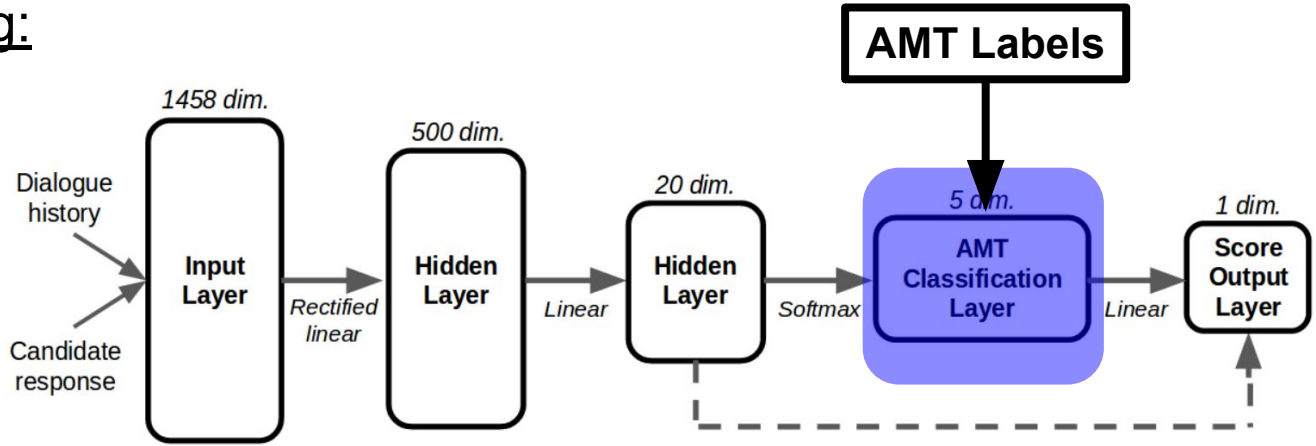


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Supervised AMT

Training:

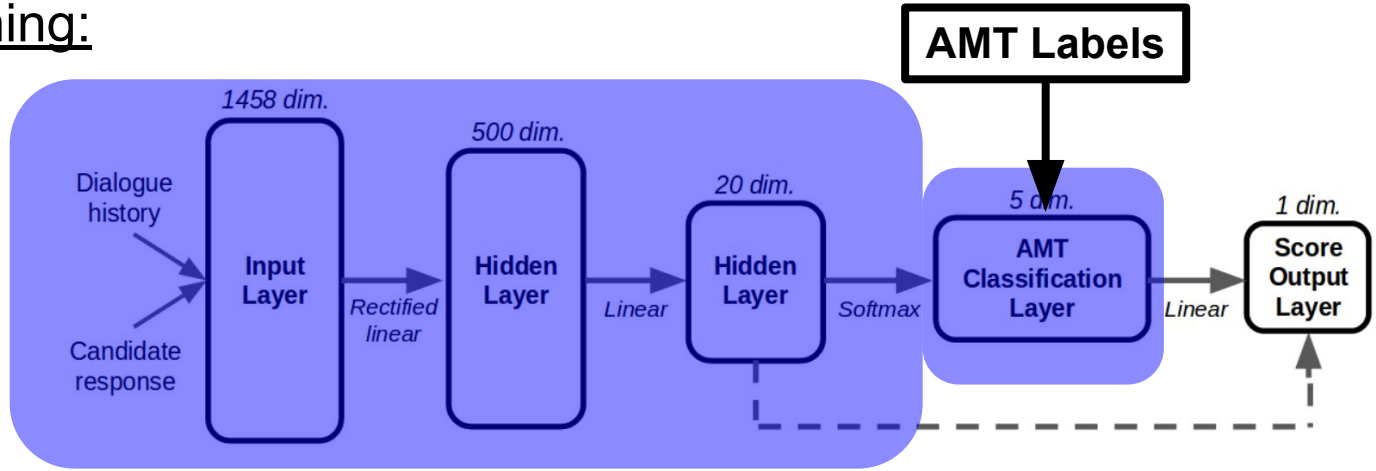


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Supervised AMT

Training:

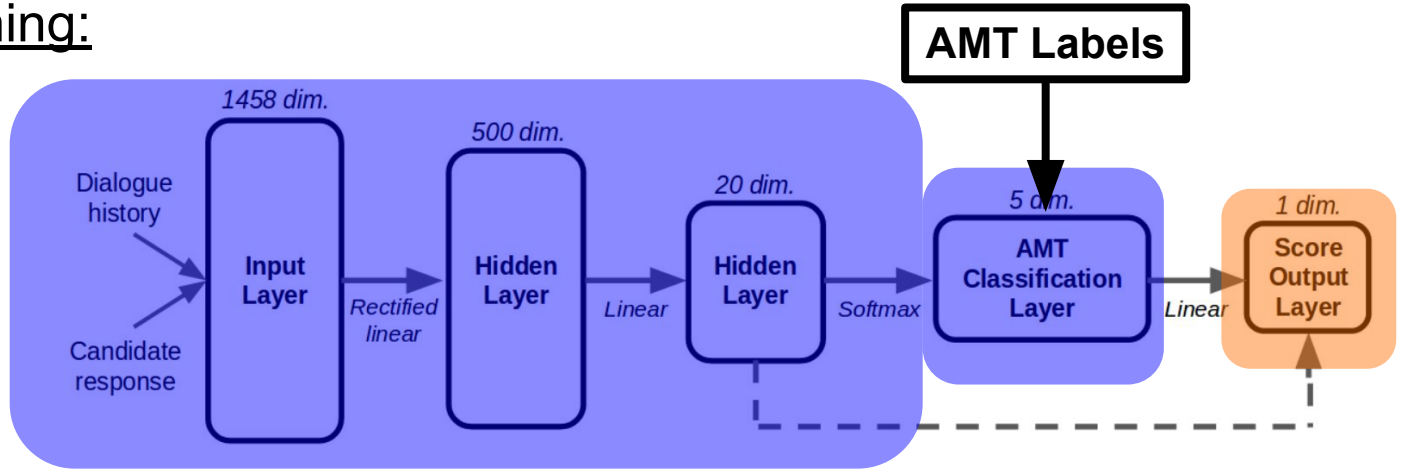


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Supervised AMT

Training:

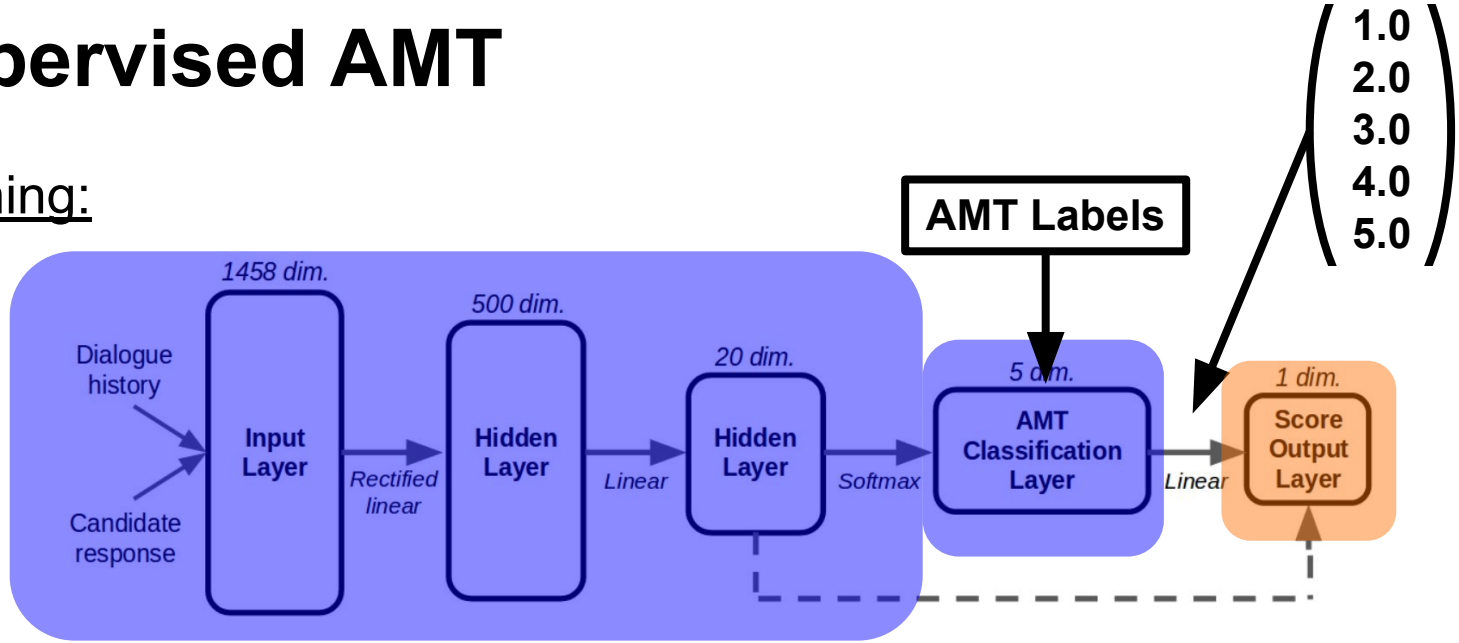


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Supervised AMT

Training:

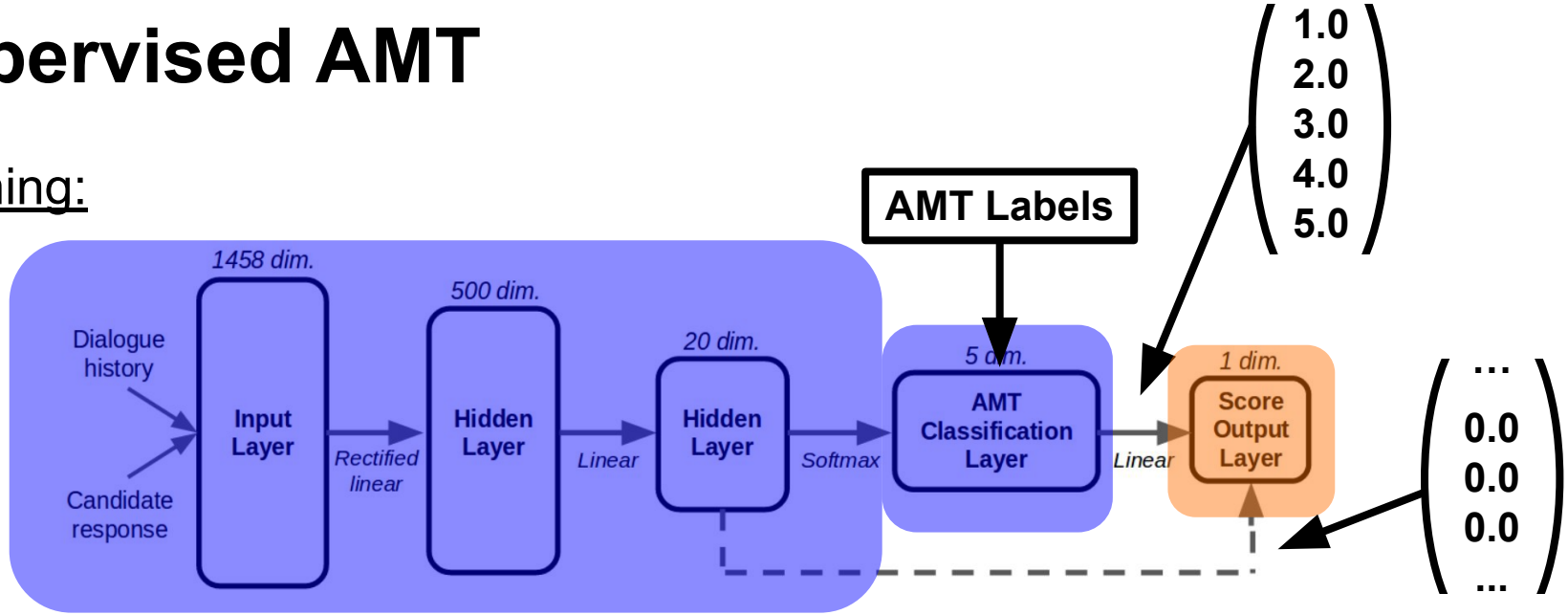


Figure 2: Computational graph for scoring model, which defines the response model selection policies based on both an action-value function parametrization and stochastic policy parametrization.

Off-policy REINFORCE

Learning from real-world users:

- Real-world Alexa users interact with our socialbot and give scores 1-5 at the end of their dialogues
- We want to use these scores to improve the system

Off-policy REINFORCE

REINFORCE algorithm:

- Updates parameters according to trial-and-error:
  - When system got a high score, learn to repeat those actions
  - When system got a low score, learn to avoid those actions

Off-policy REINFORCE

REINFORCE algorithm:

- Updates parameters according to trial-and-error:
  - When system got a high score, learn to repeat those actions
  - When system got a low score, learn to avoid those actions
- REINFORCE parameter updates (Williams, 1992):

$$\Delta\theta \propto \nabla_\theta \log \pi_\theta(a_t^d | h_t^d) R^d$$

h_t^d: dialogue history for dialogue *d* at turn *t*

a_t^d: dialogue action taken at dialogue *d* at turn *t*

R^d: Alexa user score for dialogue *d*

π_θ: stochastic policy, with parameters θ

Off-policy REINFORCE

REINFORCE algorithm:

- We use **importance weighted** variant, which is able to utilize data collected under different policies (Precup, 2000; Precup et al., 2001):

$$\Delta\theta \propto \frac{\pi_\theta(a_t^d|h_t^d)}{\pi_{\theta_d}(a_t^d|h_t^d)} \nabla_\theta \log \pi_\theta(a_t^d|h_t^d) R^d$$

h_t^d: dialogue history for dialogue *d* at turn *t*

a_t^d: dialogue action taken at dialogue *d* at turn *t*

R^d: Alexa user score for dialogue *d*

π_θ: stochastic policy, with parameters θ

π_{θ_d}: stochastic policy used during dialogue *d*, with parameters θ_d

Off-policy REINFORCE

Training:

- Pretrain the policy as Supervised AMT policy
- Train Off-policy REINFORCE on ~5000 recorded dialogues
- Reward shaping based on sentiment (Ng et al., 1999)

Abstract Discourse MDP

Fitting Markov decision process (MDP):

- Alternative to REINFORCE, which also:
  1) leverages transitions between turns
  2) leverages AMT labels during training
- Define the MDP *state* as discrete latent variable z_t, representing three abstract discourse properties:
  - *User dialogue act (e.g. greeting, question, statement)*
  - *User sentiment (positive, neutral, negative)*
  - *User generic utterance (binary variable)*

Abstract Discourse MDP

Fitting Markov decision process (MDP):

- Given state z_t, MDP samples dialogue history h_t recorded from a previous dialogue, subject to {dialogue act, sentiment, generic}
- Agent takes an action a_t, and receives a reward r_t and AMT label y_t
- MDP then transitions to new state z_{t+1}, conditioned on $\{z_t, h_t, a_t, y_t\}$

Abstract Discourse MDP

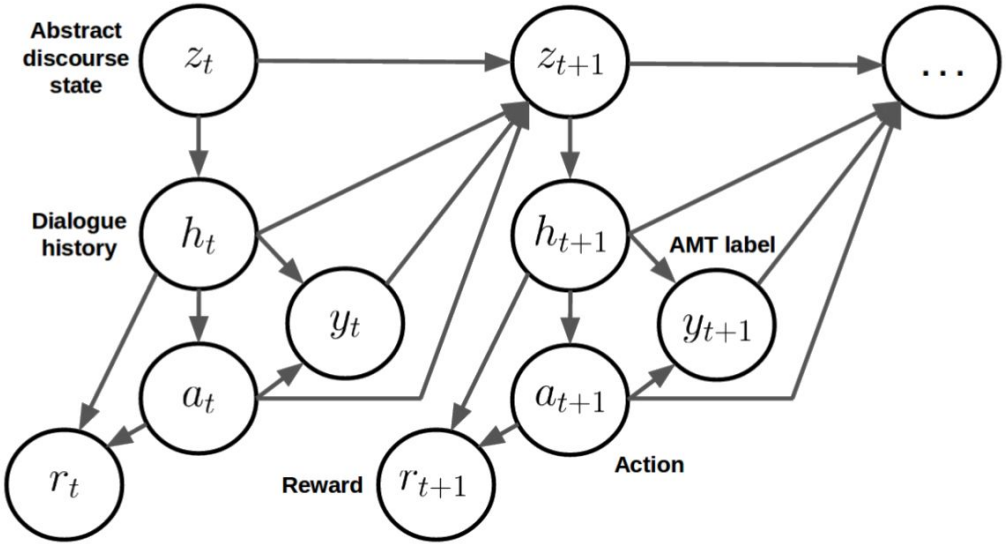


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

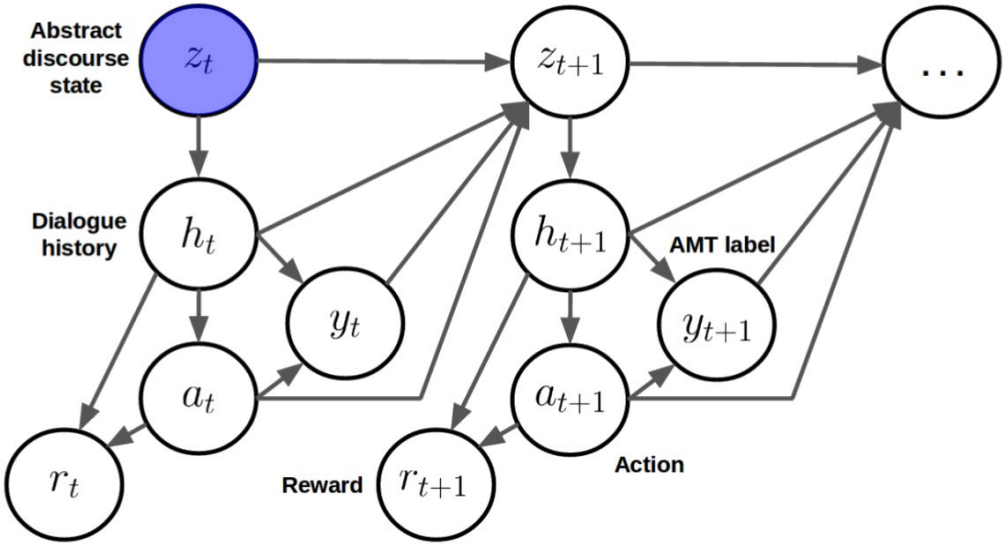


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

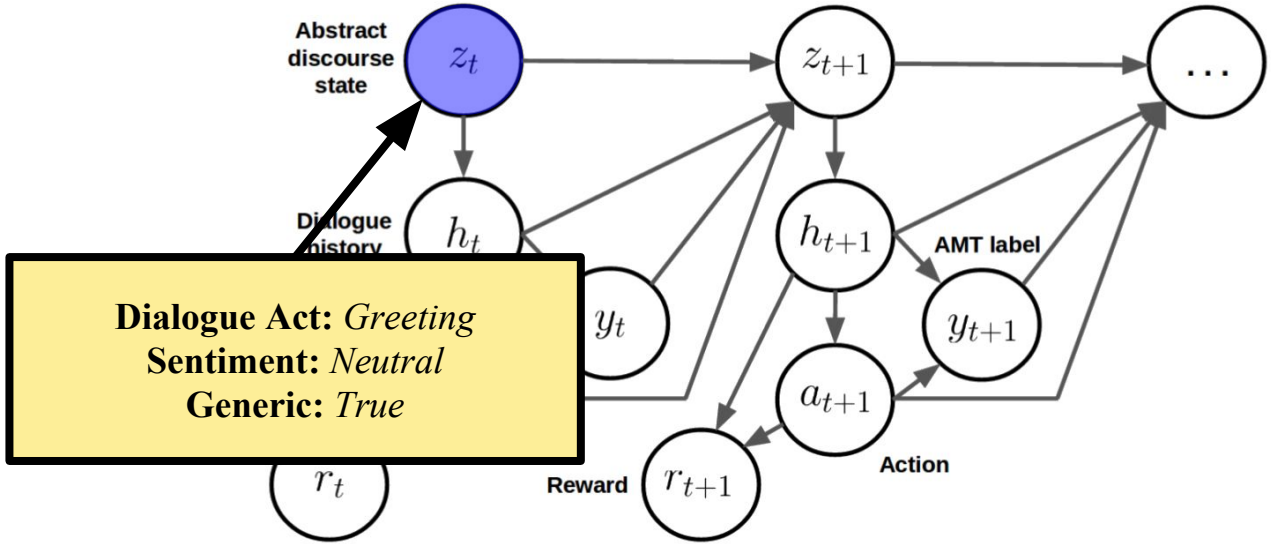


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

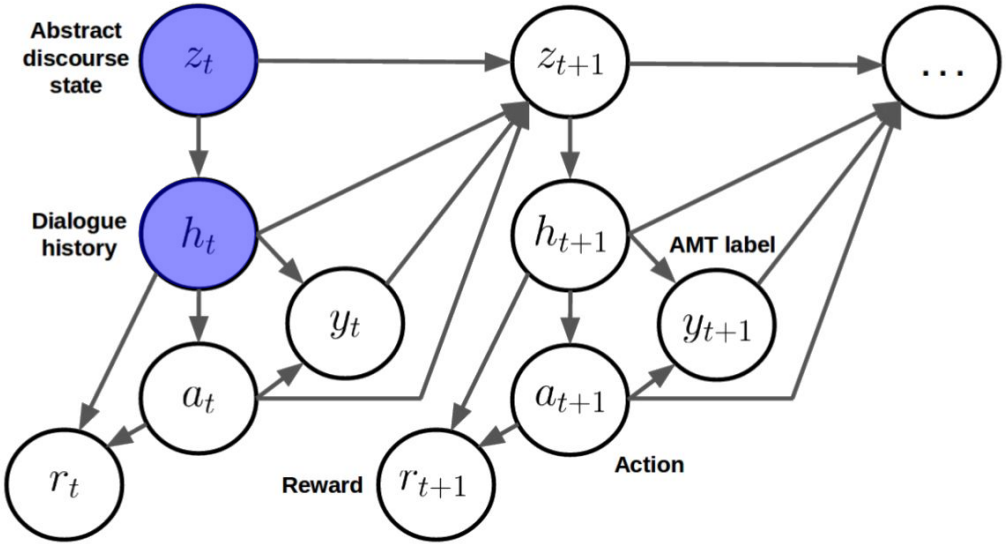


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

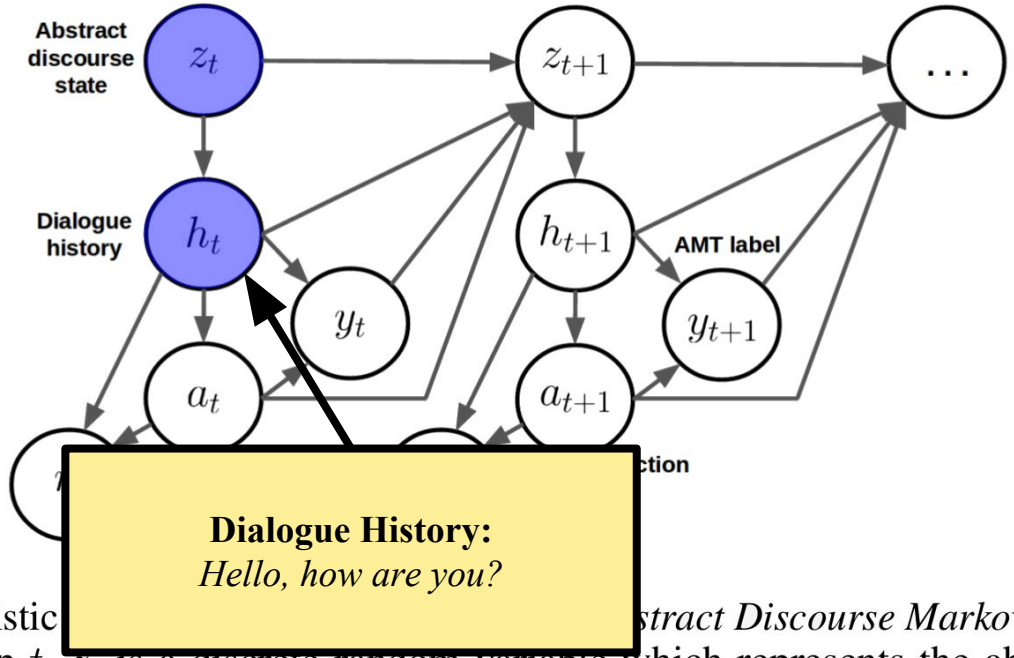


Figure 6: Probabilistic ... *stract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

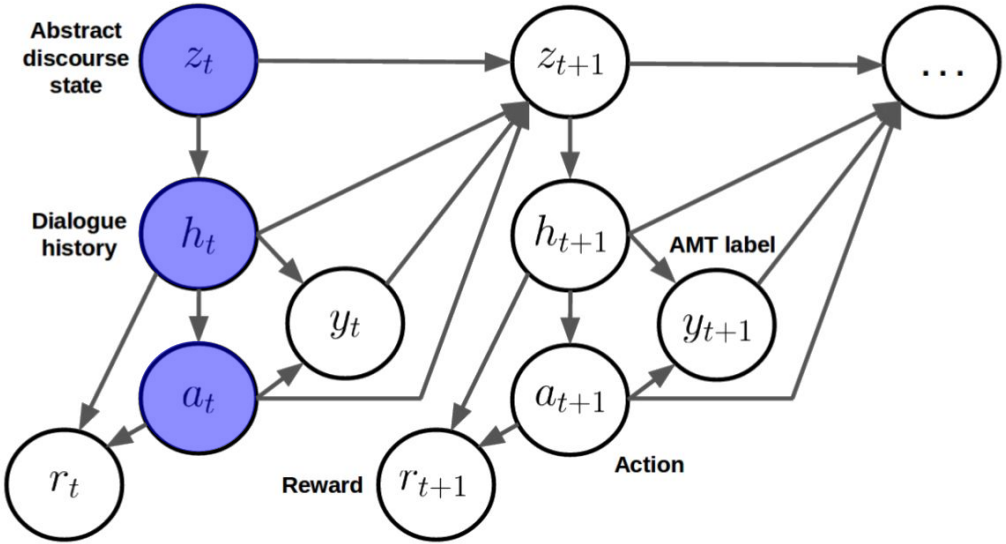


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

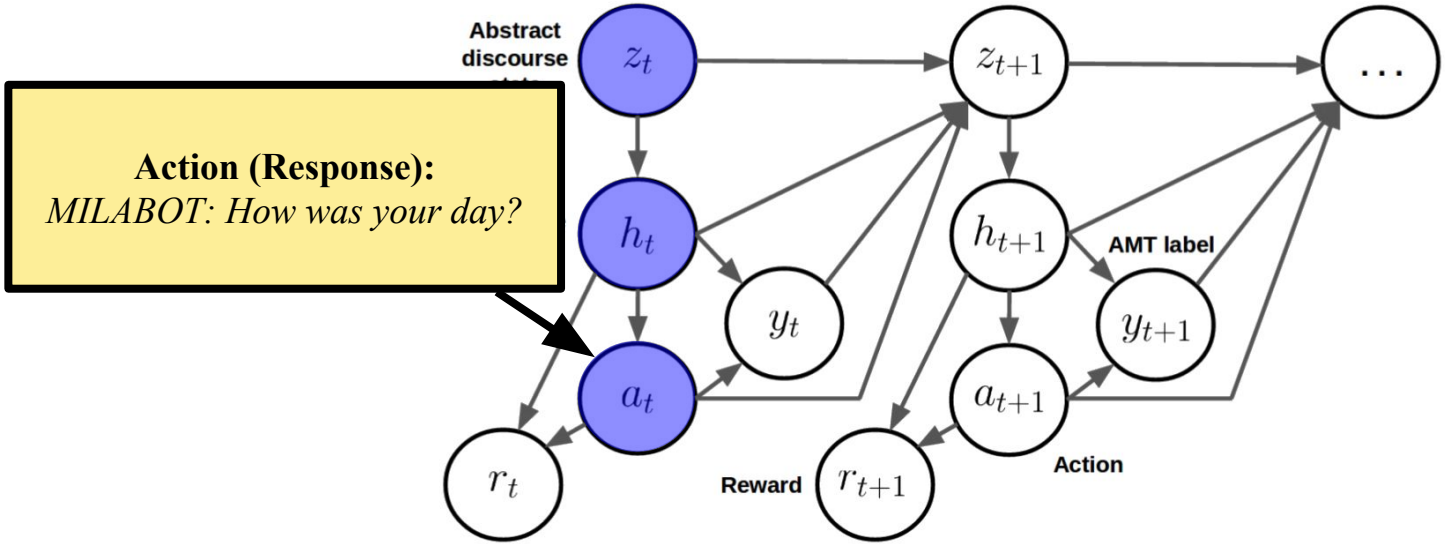


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

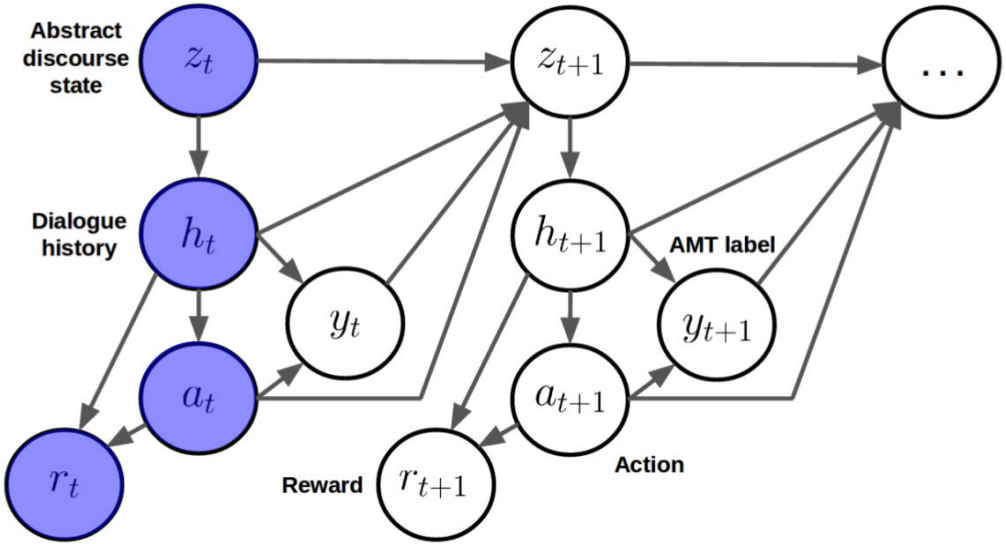


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

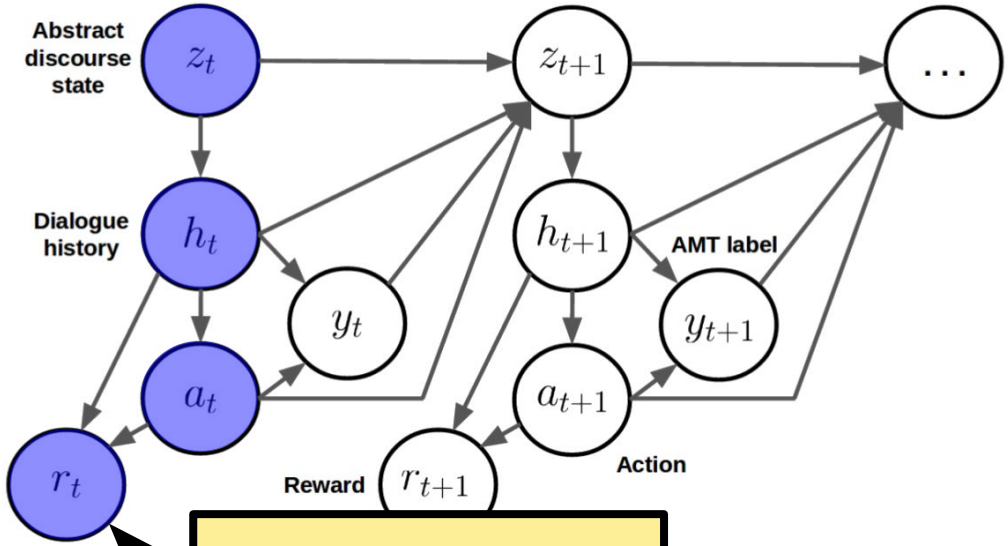


Figure 6: Probabilistic directed ... *ract Discourse Markov Decision Process.* For each time step t, z_t is a ... ich represents the abstract state of the dialogue, h_t represents the di... the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

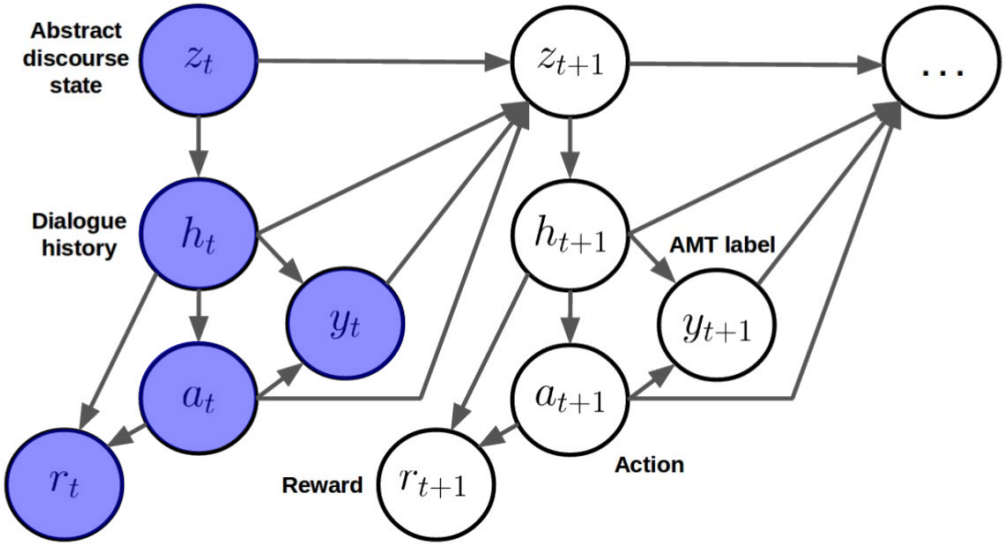


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

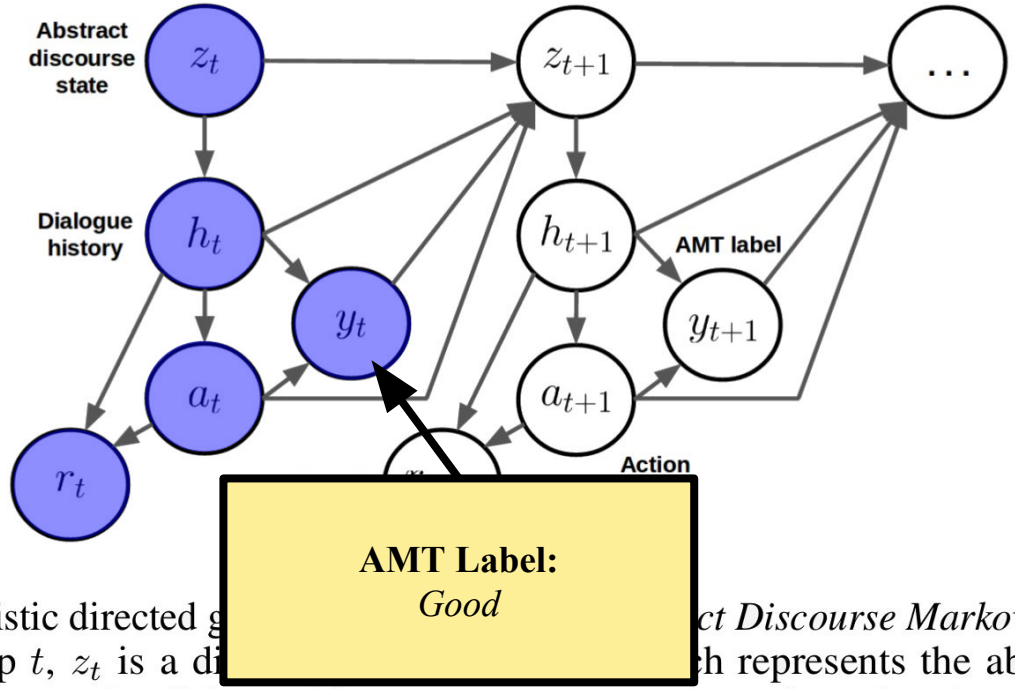


Figure 6: Probabilistic directed g ct Discourse Markov Decision Process. For each time step t, z_t is a di h represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

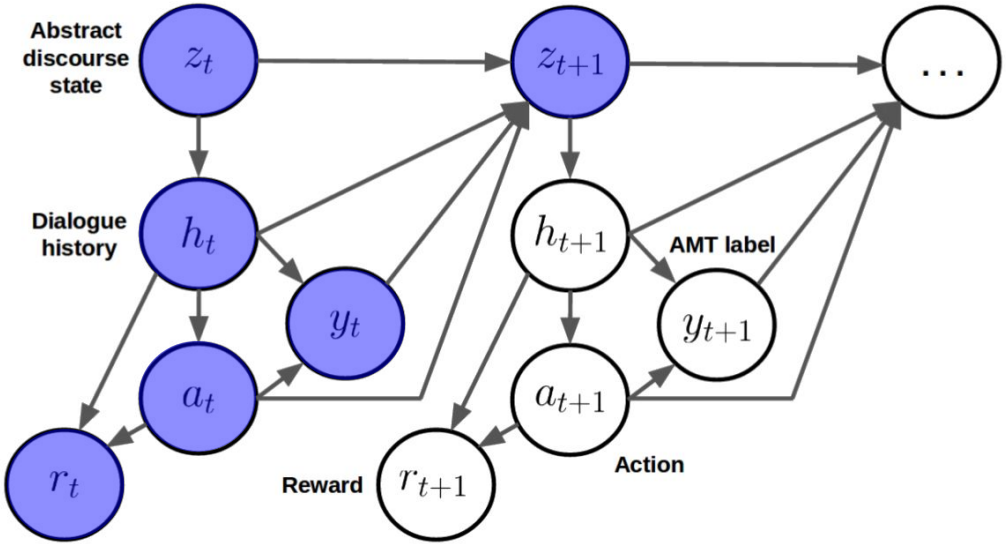


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

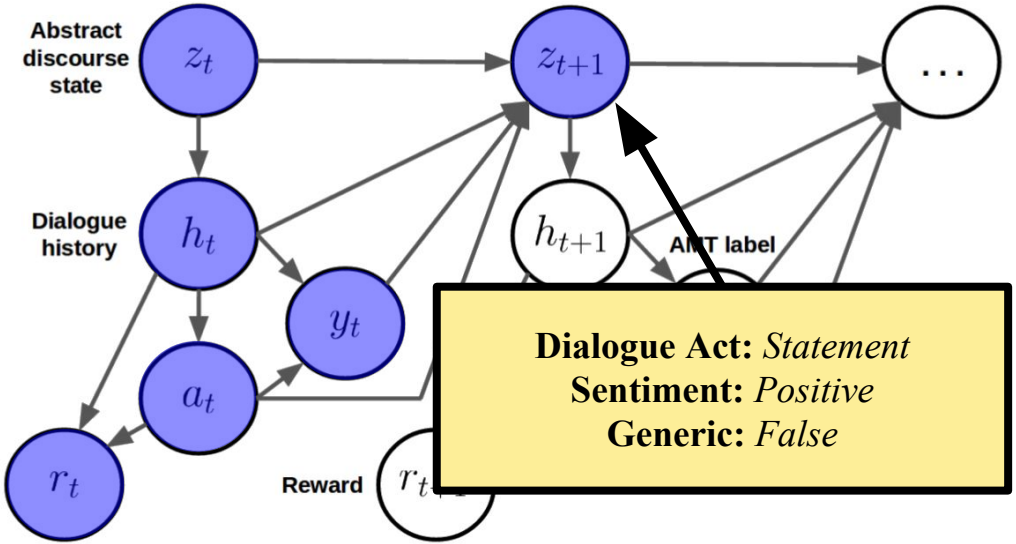


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

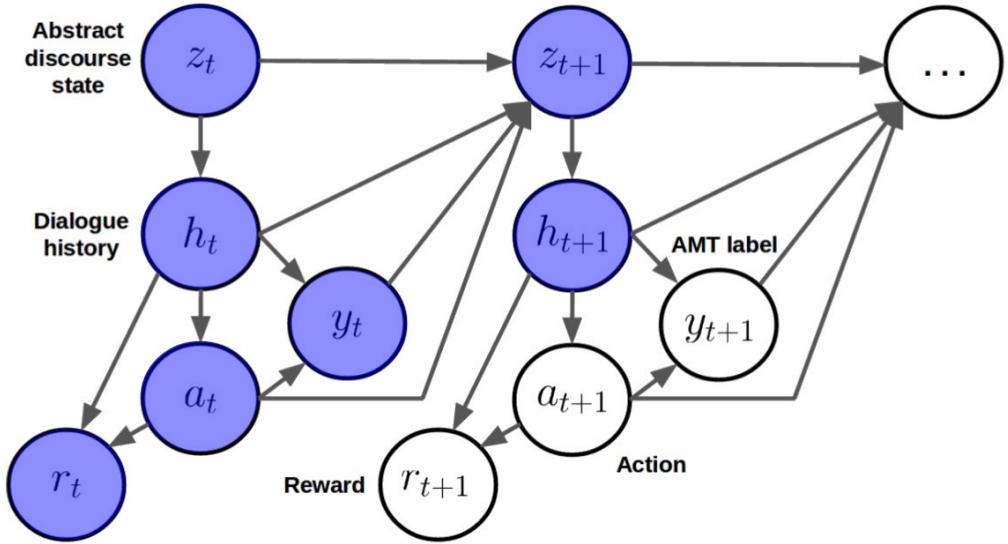


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

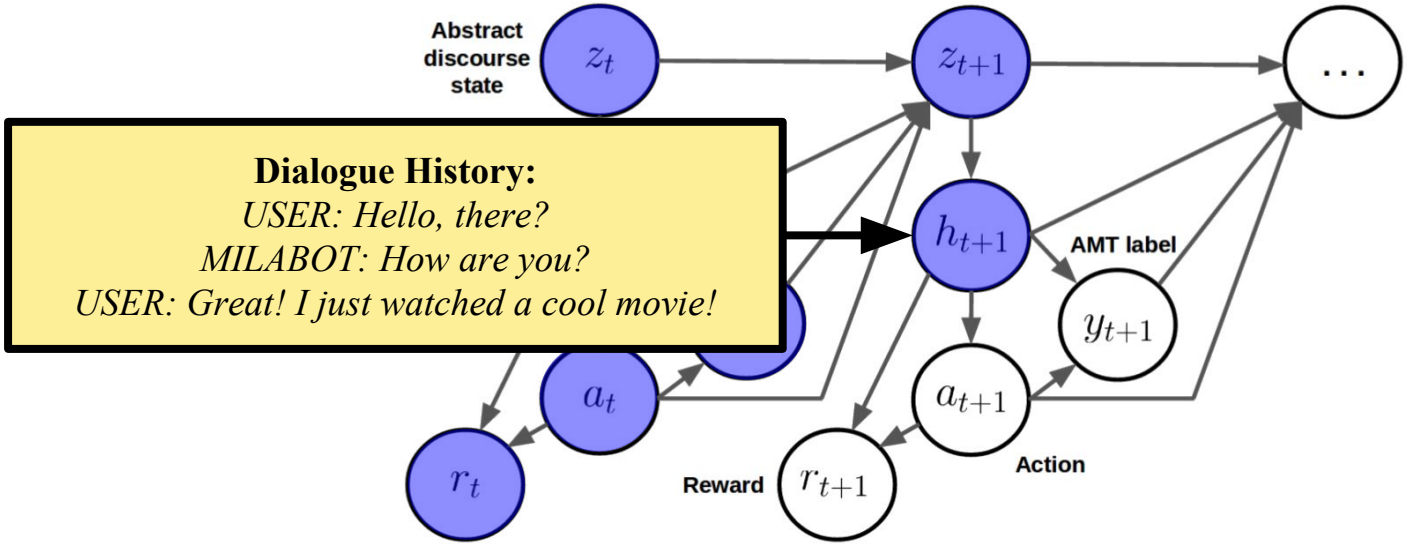


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

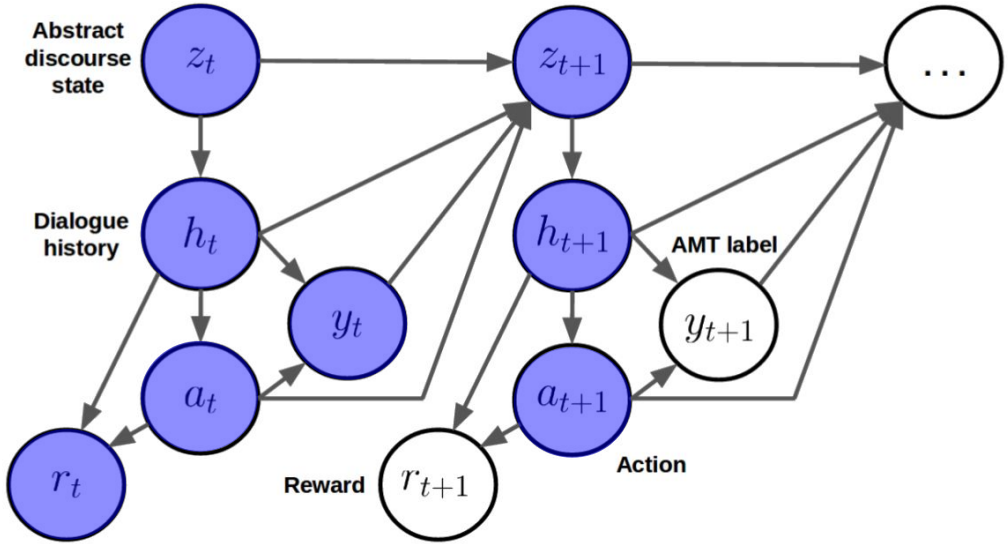


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

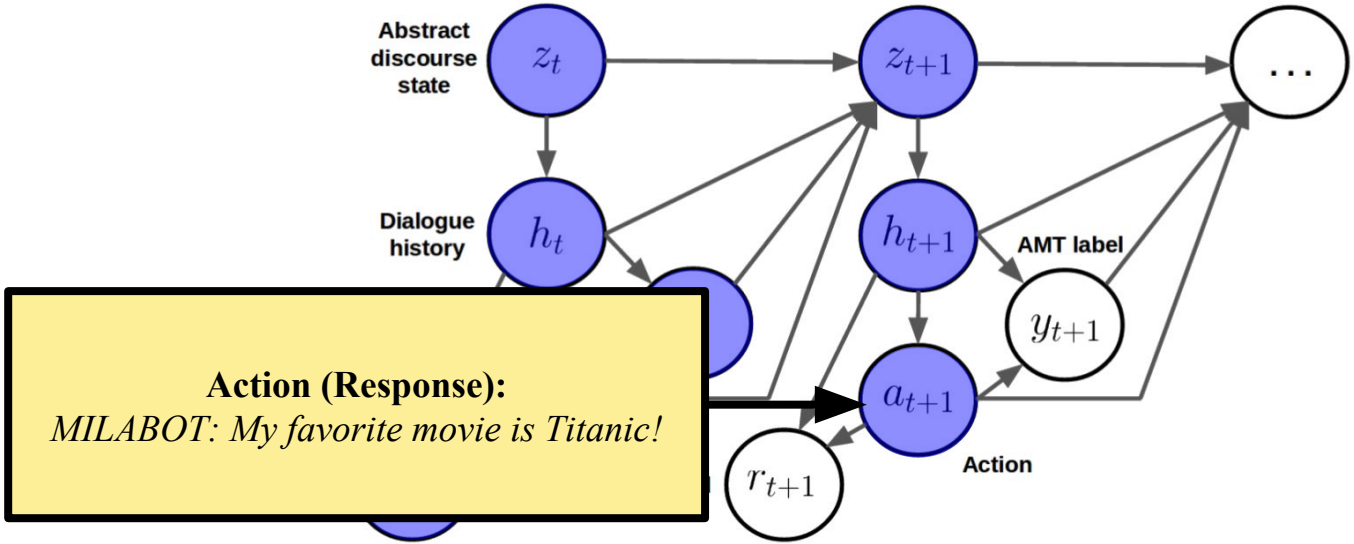


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

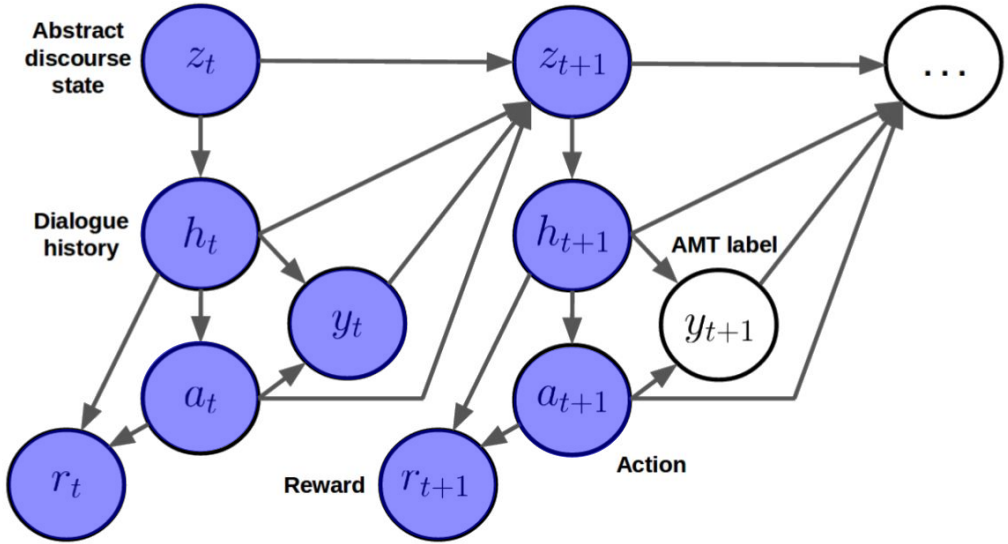
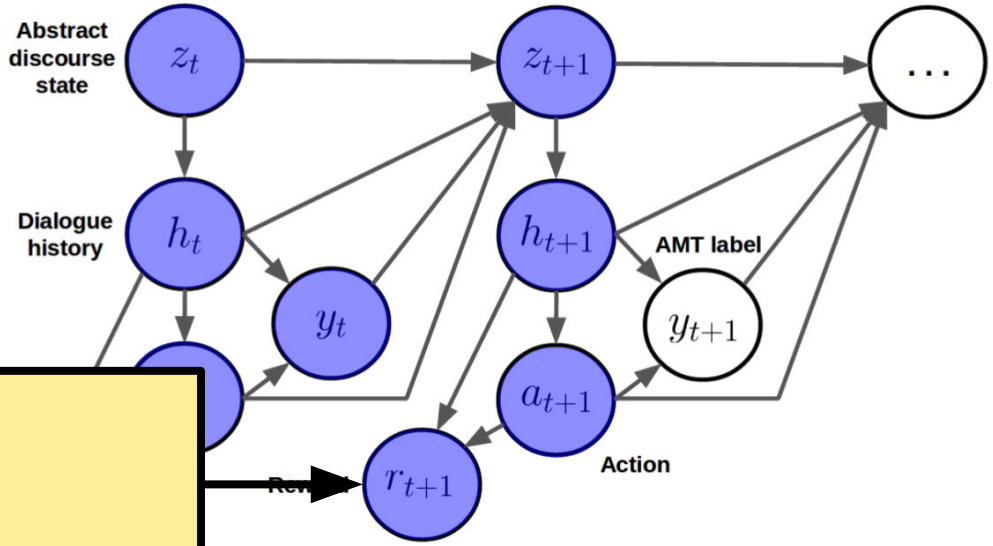


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP



Reward:
3.2

Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

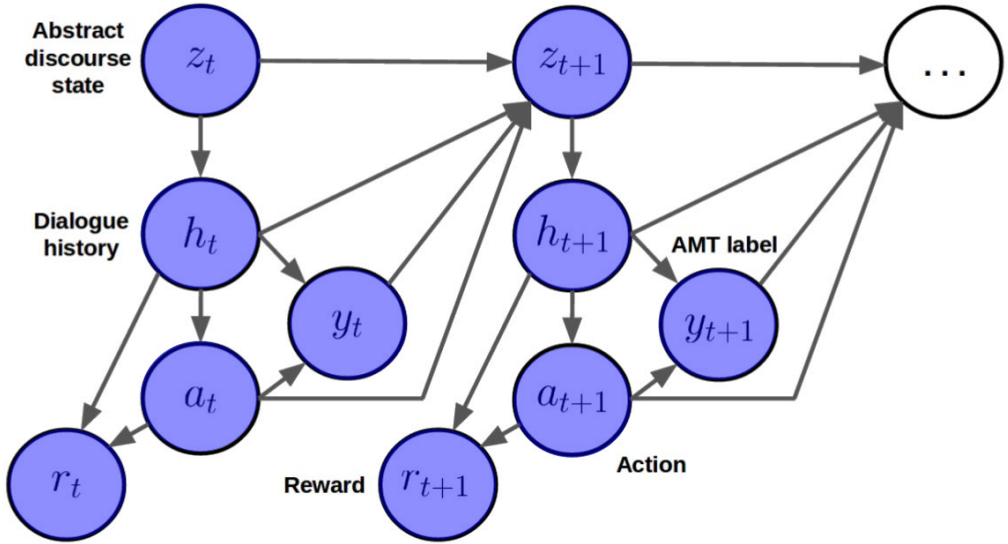


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

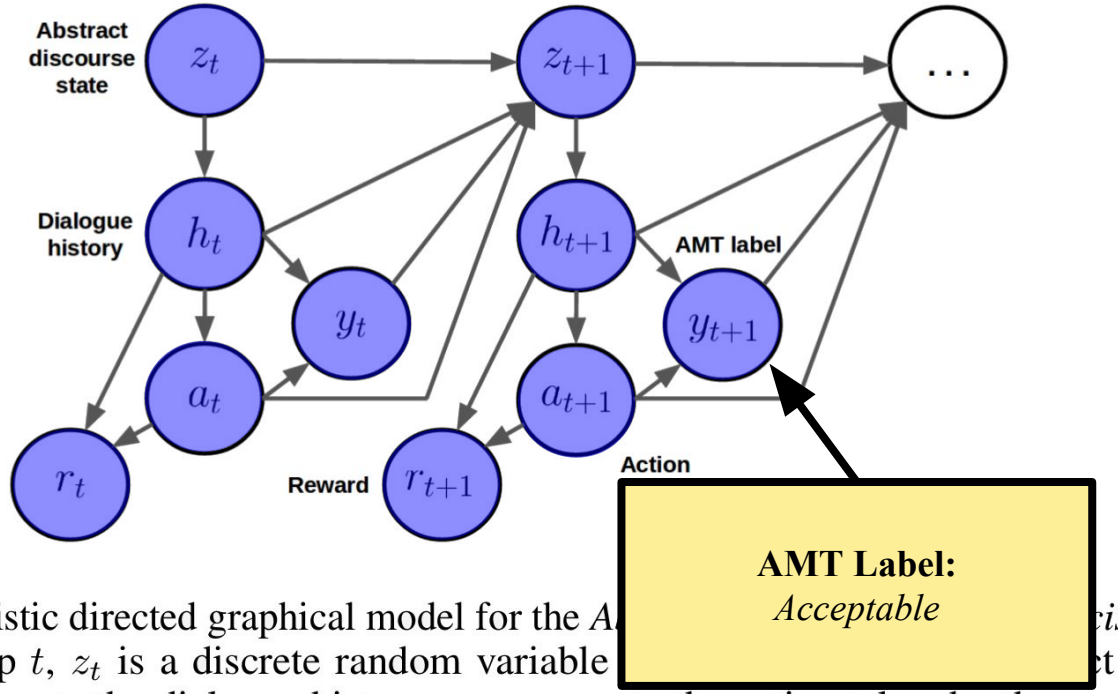


Figure 6: Probabilistic directed graphical model for the *A* *cision Process.* For each time step t, z_t is a discrete random variable t state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

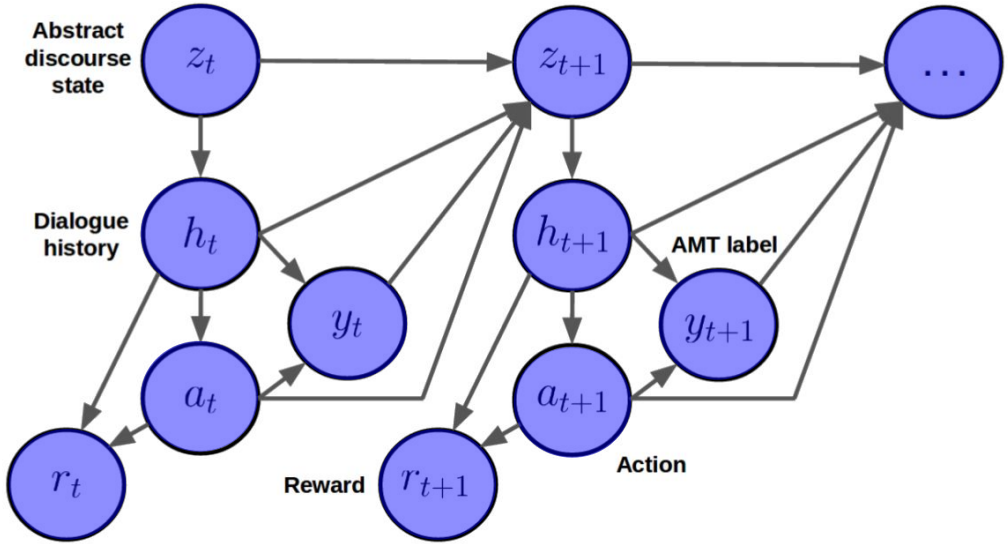


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

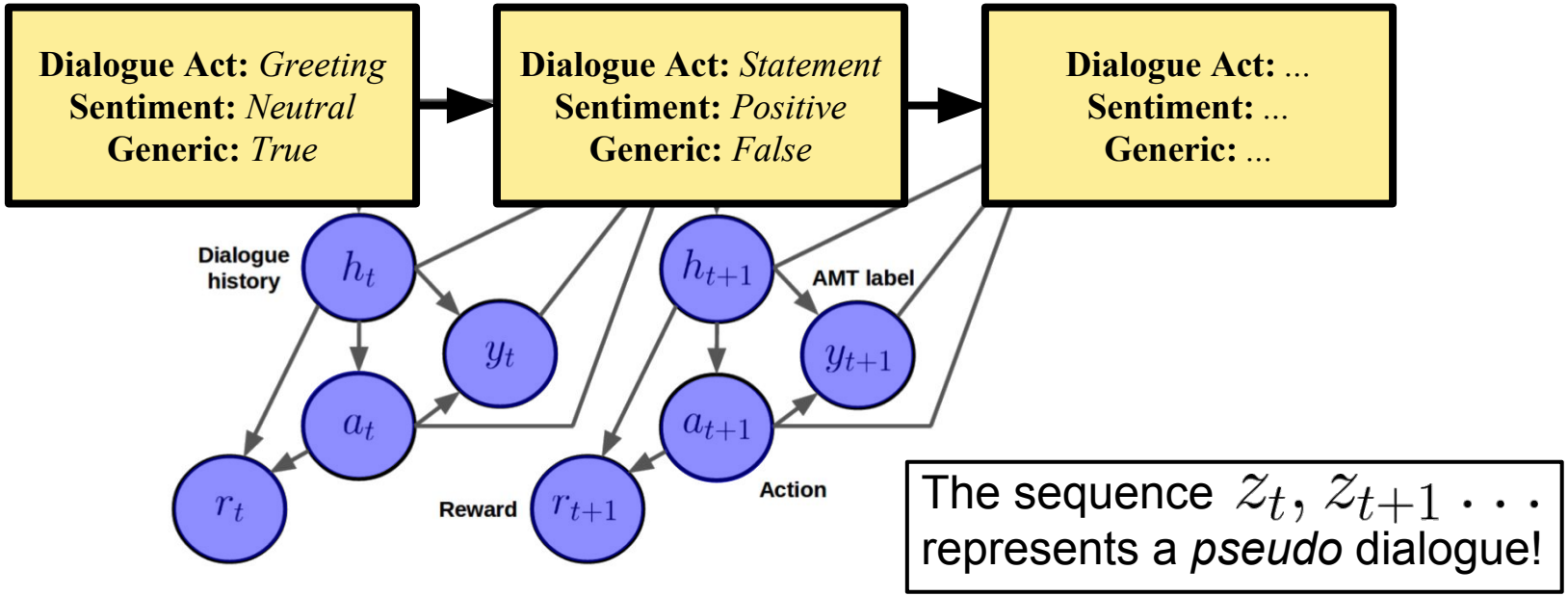


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process*. For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

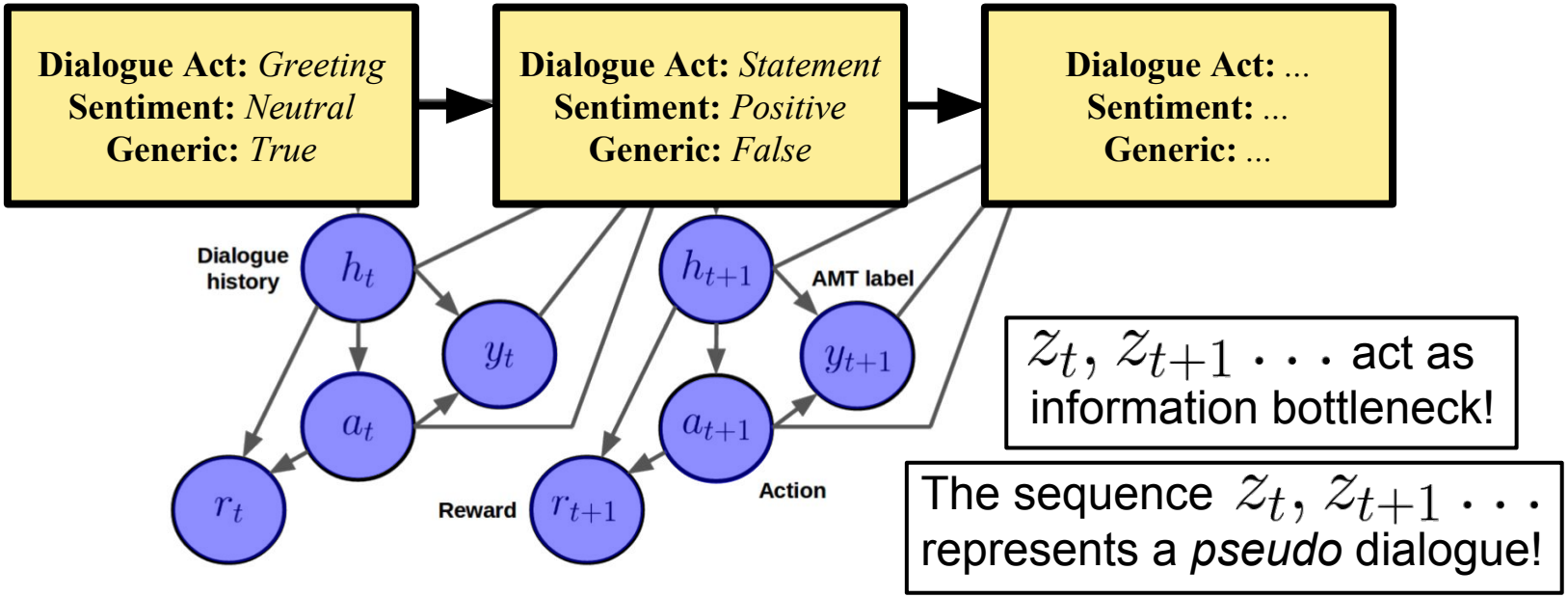


Figure 6: Probabilistic directed graphical model for the *Abstract Discourse Markov Decision Process.* For each time step t, z_t is a discrete random variable which represents the abstract state of the dialogue, h_t represents the dialogue history, a_t represents the action taken by the system (i.e. the selected response), y_t represents the sampled AMT label and r_t represents the sampled reward.

Abstract Discourse MDP

Training:

- Supervised AMT as reward function and for sampling y_t
- MDP transition model fitted using ~0.5M recorded transitions
- Pretrain policy as Supervised AMT policy
- Learn policy using Q-learning with experience replay (Mnih et al., 2013; Lin, 1993)

A/B Experiments with Alexa Users

Machine learning experiments in the wild!

A/B Experiments with Alexa Users

Setup:

- During A/B experiments, Alexa users are directed at random to one policy, which will be used through the entire dialogue
- Policies evaluated w.r.t. average user scores and other statistics

A/B Experiments with Alexa Users

A/B testing results ($\pm$ 95% confidence intervals). Star * indicates statistical significance at 95% level.

Experiment	Policy	User score	Dialogue length	Pos. utterances	Neg. utterances
Exp. #1	*Evibot + Alicebot*	2.86 ± 0.22	31.84 ± 6.02	$2.80\% \pm 0.79$	$5.63\% \pm 1.27$
	Supervised AMT	2.80 ± 0.21	34.94 ± 8.07	$\mathbf{4.00\% \pm 1.05}$	$8.06\% \pm 1.38$
	Off-policy REINFORCE	2.86 ± 0.21	$\mathbf{37.51 \pm 7.21}$	$3.98\% \pm 0.80$	6.25 ± 1.28
	*Q-learning AMT**	$\mathbf{3.15 \pm 0.20}$	30.26 ± 4.64	$3.75\% \pm 0.93$	$\mathbf{5.41\% \pm 1.16}$

A/B Experiments with Alexa Users

A/B testing results ($\pm$ 95% confidence intervals). Star * indicates statistical significance at 95% level.

Experiment	Policy	User score	Dialogue length	Pos. utterances	Neg. utterances
Exp. #1	*Evibot + Alicebot*	2.86 ± 0.22	31.84 ± 6.02	$2.80\% \pm 0.79$	$5.63\% \pm 1.27$
	Supervised AMT	2.80 ± 0.21	34.94 ± 8.07	$\mathbf{4.00\% \pm 1.05}$	$8.06\% \pm 1.38$
	Off-policy REINFORCE	2.86 ± 0.21	$\mathbf{37.51 \pm 7.21}$	$3.98\% \pm 0.80$	6.25 ± 1.28
	*Q-learning AMT**	$\mathbf{3.15 \pm 0.20}$	30.26 ± 4.64	$3.75\% \pm 0.93$	$\mathbf{5.41\% \pm 1.16}$

Reinforcement learning approaches consistently improve the system!

A/B Experiments with Alexa Users

A/B testing results ($\pm$ 95% confidence intervals). Star * indicates statistical significance at 95% level.

Experiment	Policy	User score	Dialogue length	Pos. utterances	Neg. utterances
Exp. #1	*Evibot + Alicebot*	2.86 ± 0.22	31.84 ± 6.02	$2.80\% \pm 0.79$	$5.63\% \pm 1.27$
	Supervised AMT	2.80 ± 0.21	34.94 ± 8.07	$\mathbf{4.00\% \pm 1.05}$	$8.06\% \pm 1.38$
	Off-policy REINFORCE	2.86 ± 0.21	$\mathbf{37.51 \pm 7.21}$	$3.98\% \pm 0.80$	6.25 ± 1.28
	*Q-learning AMT**	$\mathbf{3.15 \pm 0.20}$	30.26 ± 4.64	$3.75\% \pm 0.93$	$\mathbf{5.41\% \pm 1.16}$
Exp. #2	*Off-policy REINFORCE*	$\mathbf{3.06 \pm 0.12}$	$\mathbf{34.45 \pm 3.76}$	$3.23\% \pm 0.45$	$7.97\% \pm 0.85$
	Q-learning AMT	2.92 ± 0.12	31.84 ± 3.69	$\mathbf{3.38\% \pm 0.50}$	$\mathbf{7.61\% \pm 0.84}$
Exp. #3	*Off-policy REINFORCE*	3.03 ± 0.18	30.93 ± 4.96	2.72 ± 0.59	7.36 ± 1.22
	Q-learning AMT	$\mathbf{3.06 \pm 0.17}$	$\mathbf{33.69 \pm 5.84}$	$\mathbf{3.63 \pm 0.68}$	$\mathbf{6.67 \pm 0.98}$

A/B Experiments with Alexa Users

A/B testing results ($\pm$ 95% confidence intervals). Star * indicates statistical significance at 95% level.

Experiment	Policy	User score	Dialogue length	Pos. utterances	Neg. utterances
Exp. #1	*Evibot + Alicebot*	2.86 ± 0.22	31.84 ± 6.02	$2.80\% \pm 0.79$	$5.63\% \pm 1.27$
	Supervised AMT	2.80 ± 0.21	34.94 ± 8.07	$\mathbf{4.00\% \pm 1.05}$	$8.06\% \pm 1.38$
	Off-policy REINFORCE	2.86 ± 0.21	$\mathbf{37.51 \pm 7.21}$	$3.98\% \pm 0.80$	6.25 ± 1.28
	*Q-learning AMT**	$\mathbf{3.15 \pm 0.20}$	30.26 ± 4.64	$3.75\% \pm 0.93$	$\mathbf{5.41\% \pm 1.16}$
Exp. #2	*Off-policy REINFORCE*	$\mathbf{3.06 \pm 0.12}$	$\mathbf{34.45 \pm 3.76}$	$3.23\% \pm 0.45$	$7.97\% \pm 0.85$
	Q-learning AMT	2.92 ± 0.12	31.84 ± 3.69	$\mathbf{3.38\% \pm 0.50}$	$\mathbf{7.61\% \pm 0.84}$
Exp. #3	*Off-policy REINFORCE*	3.03 ± 0.18	30.93 ± 4.96	2.72 ± 0.59	7.36 ± 1.22
	Q-learning AMT	$\mathbf{3.06 \pm 0.17}$	$\mathbf{33.69 \pm 5.84}$	$\mathbf{3.63 \pm 0.68}$	$\mathbf{6.67 \pm 0.98}$

Reinforcement learning is still superior, but it's unclear which approach is better!

A/B Experiments with Alexa Users

A/B testing results (± 95% confidence intervals). Star * indicates statistical significance at 95% level.

Experiment	Policy	User score	Dialogue length	Pos. utterances	Neg. utterances
Exp. #1	*Evibot + Alicebot*	2.86 ± 0.22	31.84 ± 6.02	$2.80\% \pm 0.79$	$5.63\% \pm 1.27$
	Supervised AMT	2.80 ± 0.21	34.94 ± 8.07	$\mathbf{4.00\% \pm 1.05}$	$8.06\% \pm 1.38$
	Off-policy REINFORCE	2.86 ± 0.21	$\mathbf{37.51 \pm 7.21}$	$3.98\% \pm 0.80$	6.25 ± 1.28
	*Q-learning AMT**	$\mathbf{3.15 \pm 0.20}$	30.26 ± 4.64	$3.75\% \pm 0.93$	$\mathbf{5.41\% \pm 1.16}$
Exp. #2	*Off-policy REINFORCE*	$\mathbf{3.06 \pm 0.12}$	$\mathbf{34.45 \pm 3.76}$	$3.23\% \pm 0.45$	$7.97\% \pm 0.85$
	Q-learning AMT	2.92 ± 0.12	31.84 ± 3.69	$\mathbf{3.38\% \pm 0.50}$	$\mathbf{7.61\% \pm 0.84}$
Exp. #3	*Off-policy REINFORCE*	3.03 ± 0.18	30.93 ± 4.96	2.72 ± 0.59	7.36 ± 1.22
	Q-learning AMT	$\mathbf{3.06 \pm 0.17}$	$\mathbf{33.69 \pm 5.84}$	$\mathbf{3.63 \pm 0.68}$	$\mathbf{6.67 \pm 0.98}$

Reinforcement learning is still superior, but it's unclear which approach is better!

***Off-policy REINFORCE* has longer dialogues,**
while *Q-learning AMT* has more positive dialogues...

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

Topical specificity

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

Topical coherence

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

Q-learning AMT* has the most *content* and *topic-dependent* dialogues*!

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

Q-learning AMT* has the most *content* and *topic-dependent* dialogues*!
Perhaps this is the most interesting policy in the long-run?

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

Supervised AMT* has high *user engagement*, as well as c*ontent* and *topic-dependent* dialogues*!

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

***Off-policy REINFORCE* has less content and topic-dependent dialogues.**

A/B Experiments with Alexa Users

Table 8: First A/B testing experiment topical specificity and coherence of the six different policies. The columns are average number of noun phrases per user utterance (User NPs), average number of noun phrases per system utterance (System NPs), average number of overlapping words between the user's utterance and the system's response (Word overlap $t \rightarrow t+1$), and average number of overlapping words between the user's utterance and the system's response in the next turn (Word overlap $t \rightarrow t+3$). 95% confidence intervals are also shown. Stop words are excluded.

Policy	**User NPs**	**System NPs**	**Word overlap $t \rightarrow t+1$**	**Word overlap $t \rightarrow t+3$**
Evibot + Alicebot	0.55 ± 0.03	1.05 ± 0.05	7.33 ± 0.21	7.31 ± 0.22
Supervised AMT	$\mathbf{0.62 \pm 0.03}$	1.75 ± 0.07	10.48 ± 0.28	10.65 ± 0.29
Off-policy REINFORCE	0.59 ± 0.02	1.45 ± 0.05	9.05 ± 0.21	9.14 ± 0.22
Q-learning AMT	0.58 ± 0.03	$\mathbf{1.98 \pm 0.08}$	$\mathbf{11.28 \pm 0.30}$	$\mathbf{11.52 \pm 0.32}$

Evibot + Alicebot* has the most *generic and topic-irrelevant dialogues!

Analyzing the Learned Policies

What are the models thinking?

Analyzing the Learned Policies

Response Model Frequency

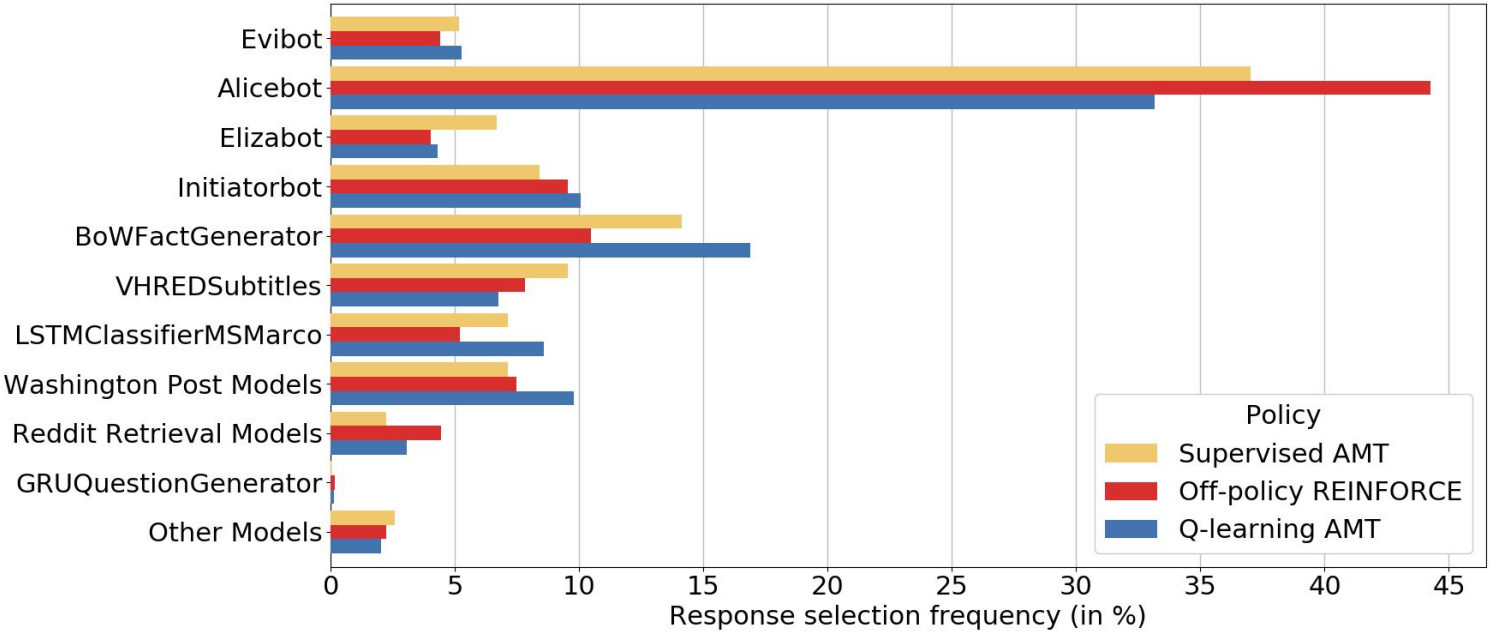


Figure 7: Response model selection probabilities across response models for *Supervised AMT*, *Off-policy REINFORCE* and *Q-learning AMT* on the AMT label test dataset.

Analyzing the Learned Policies

Off-policy REINFORCE prefers generic, safe Alicebot responses

Response Model Frequency

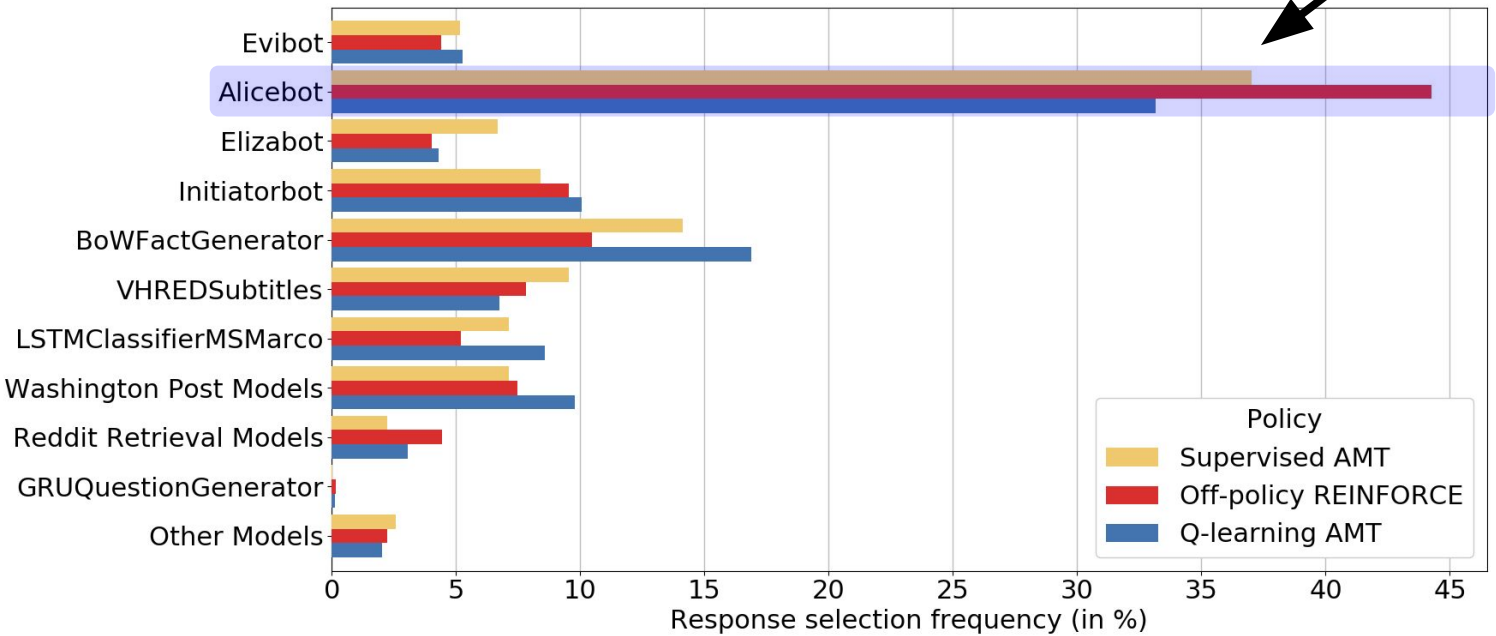


Figure 7: Response model selection probabilities across response models for *Supervised AMT*, *Off-policy REINFORCE* and *Q-learning AMT* on the AMT label test dataset.

Analyzing the Learned Policies

Response Model Frequency

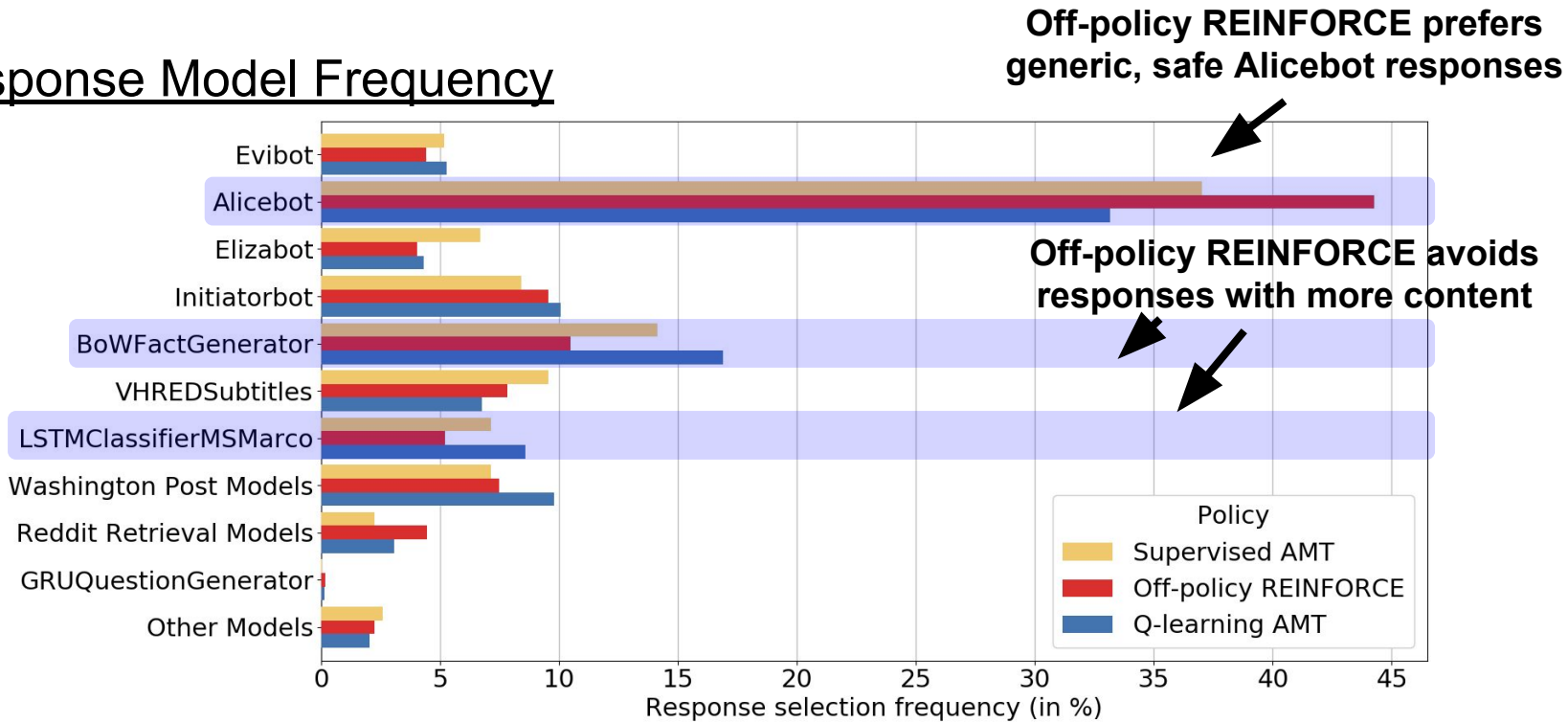


Figure 7: Response model selection probabilities across response models for *Supervised AMT*, *Off-policy REINFORCE* and *Q-learning AMT* on the AMT label test dataset.

Analyzing the Learned Policies

Response Model Frequency

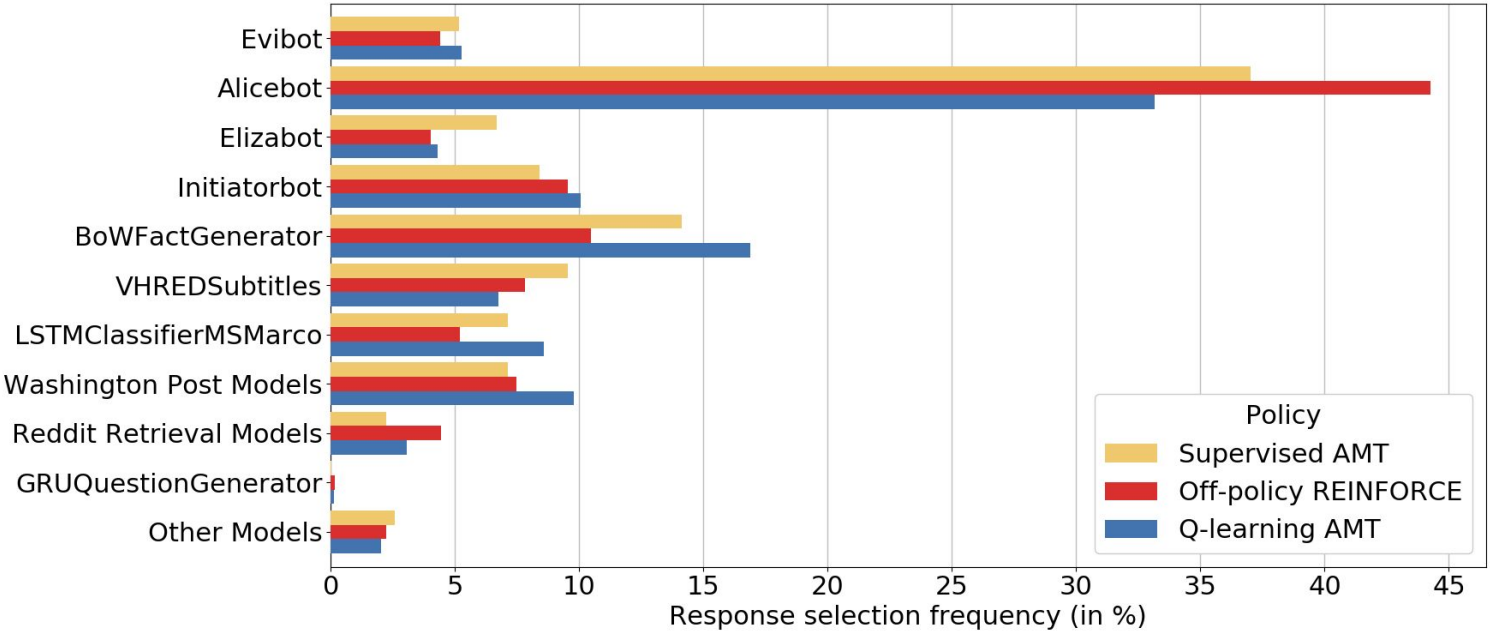


Figure 7: Response model selection probabilities across response models for *Supervised AMT*, *Off-policy REINFORCE* and *Q-learning AMT* on the AMT label test dataset.

Analyzing the Learned Policies

Q-learning prefers to use models with more content (e.g. facts and search engine results)

Response Model Frequency

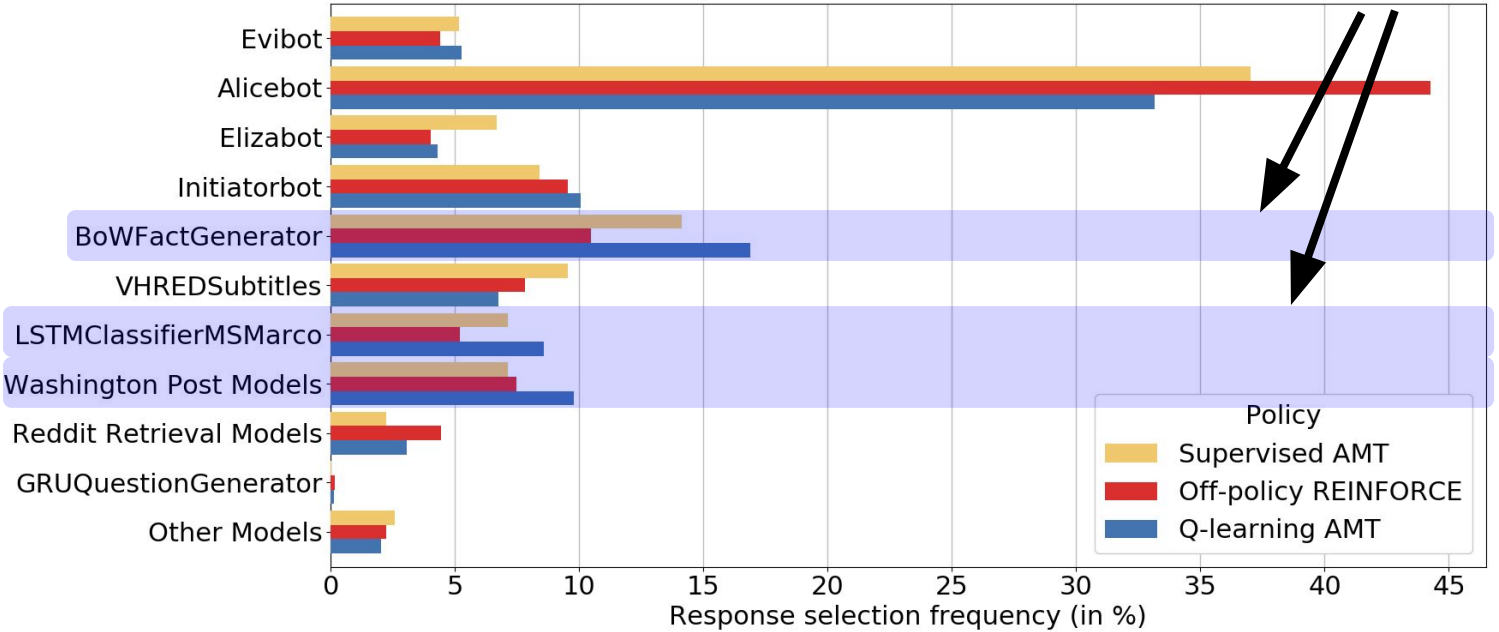


Figure 7: Response model selection probabilities across response models for *Supervised AMT*, *Off-policy REINFORCE* and *Q-learning AMT* on the AMT label test dataset.

Analyzing the Learned Policies

Analysis Conclusion:

- Off-policy REINFORCE has learned a risk averse strategy, which prefers generic responses
- Q-learning AMT has learned a risk tolerant strategy, which prefers contentful responses

Conclusion

We have proposed a deep reinforcement learning chatbot:

- Response model ensemble consisting of neural network retrieval models, neural network generative models and template-based models
- *Model selection policy* trained using crowdsourced labels and reinforcement learning to select best response

We have evaluated the chatbot:

- Evaluation with thousands of real-world Alexa users
- Crowdsourcing and reinforcement learning yielded substantial improvements
- Final system comparable to top teams in Alexa semi-finals

References

MILABOT Paper

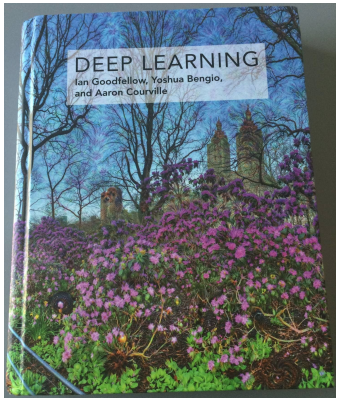
https://arxiv.org/abs/1709.02349

Dataset Survey Paper

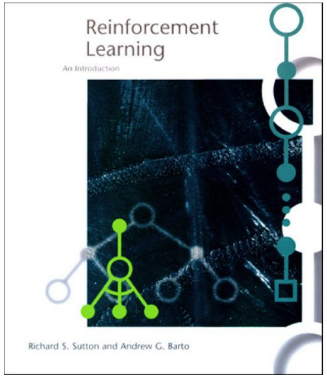
https://arxiv.org/abs/1512.05742

Deep Learning Book

www.deeplearningbook.org



The RL Intro book



Richard Sutton, Andrew Barto
Reinforcement Learning,
An Introduction

http://www.cs.ualberta.ca/~sutton/book/the-book.html

David Silver's online course on RL:

http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html

Questions?