Overview of Deep Reinforcement Learning in a Natural Language Action Space

By
Saptarashmi Bandyopadhyay
First Year PhD Student
Department of Computer Science and Engineering
Pennsylvania State University, University Park

Outline

- Overview of the paper by Ji He et al. [1]
- Focus on Game features
- Comparison and contrast of the games in Ji He et al. paper to our possible system
- Other learning approaches to games

Introduction

- A conference paper 'Deep Reinforcement Learning with a Natural Language Action Space' [1]
- Authors:
  - Dr. Ji He (currently at Citadel) and Prof. Mari Ostendorf, University of Washington, Seattle
  - Dr. Jianshu Chen (currently at Tencent), Dr. Xiaodong He (currently Deputy M.D. at JD AI Research), Dr. Jianfeng Gao, Dr. Lihong Li (currently at Google) and Dr. Li Deng (currently chief A.I. officer at Citadel), Microsoft Research, Redmond, U.S.A

Citations and publication

- 35 citations till date in Google Scholar
- Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1621–1630,Berlin, Germany, August 7-12, 2016

Overview

- New architecture for reinforcement learning with Deep Neural Networks (DNNs)
- Deep Reinforcement Relevance Network (DRRNs)
- Separate DNNs to map state and action strings to embedding vectors
- Combination by an interaction function (e.g. inner dot product) to approximate the Q function in Reinforcement Learning (RL)
- Conclusion: Model is extracting meaning, not simply remembering strings of text.

System

- Sequential decision-making system
- Action taken at a particular state to maximize the long term reward.
- In text-based game, player or system given a text string (current state) and several text strings (possible action).
- Selection of one action updates the environmental state as a new text string.

Learning objective

- Reward given at each transition or at the end
- Understanding the state text and all possible action text strings to pick up most relevant action string
- Exploring the sequence of strings to obtain highest long-term reward
- Action text string considered more relevant (to a state) among all actions if it gives a higher long-term reward
- Output of interaction function defining Q-function value for the current state-action pair,characterizing optimal long-term reward by pairing these 2 texts

Novel aspects of the architecture

- Learning of 2 different meaning representation types
  - Tendency of state texts to describe scenes
  - Tendency of action texts to describe possible actions from the user.
- Learning of continuous space representation of actions to generalize paraphrased descriptions of unknown actions

Text games and Q learning

$$Q^{\pi}(s,a) = \mathbb{E}\left\{\sum_{k=0}^{+\infty} \gamma^{k} r_{t+k} \middle| s_t = s, a_t = a \right\}$$

- s_t is current state context
- a_t is the potential action
- $\pi(a_t|s_t)$ is the probability of taking action a_t given state s_t called policy of the agent
- γ is discount factor
- $Q^{\pi}(s,a)$ is expected return starting from s, taking action a with policy $\pi(a|s)$

Soft-max selection strategy

- Exploration policy during learning
- Selection of action at at state st by the probability of

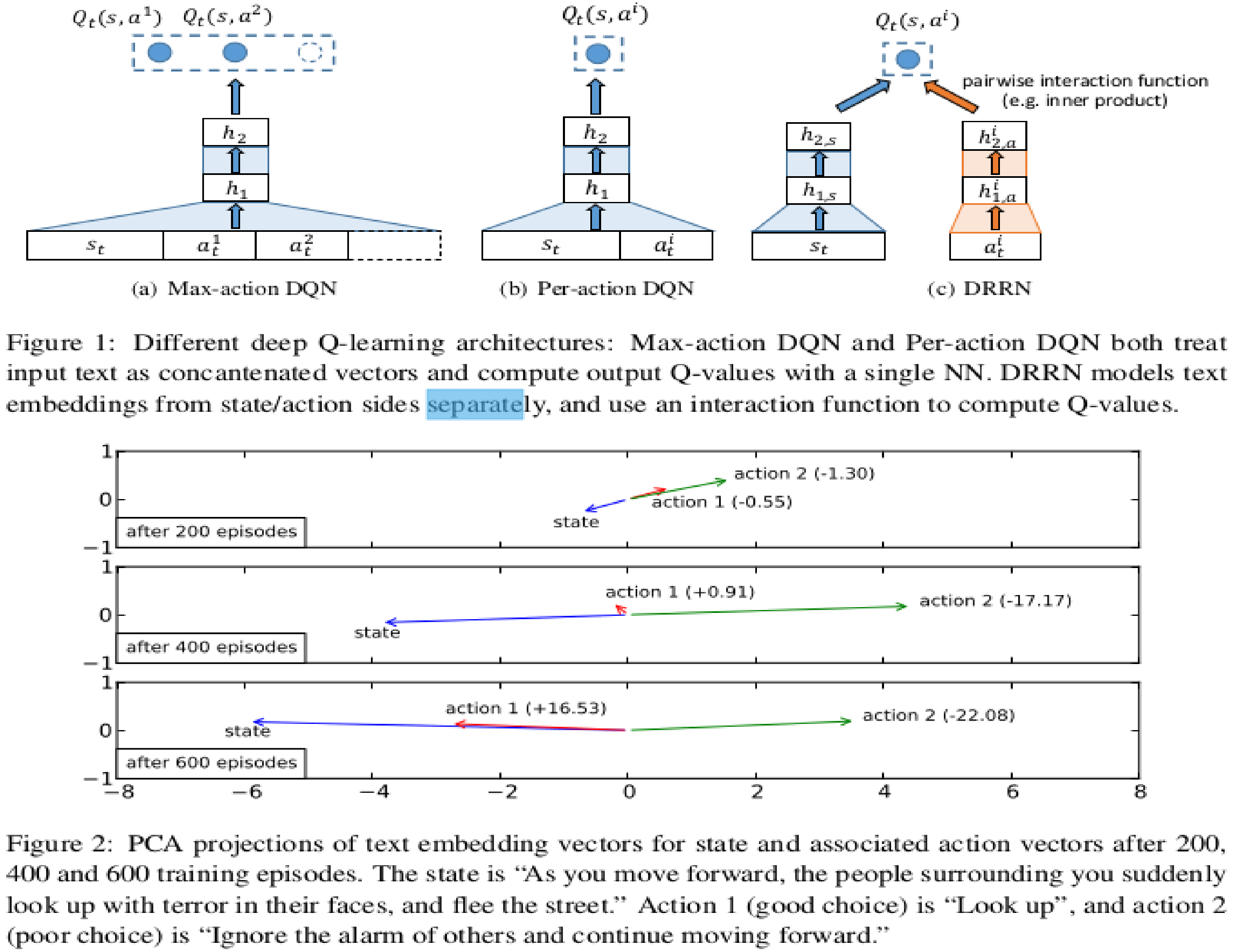
$$\pi(a_t = a_t^i | s_t) = \frac{\exp(\alpha \cdot Q(s_t, a_t^i))}{\sum_{j=1}^{|A_t|} \exp(\alpha \cdot Q(s_t, a_t^j))}$$

- A_t is the set of feasible actions at state s_t
- a_t^i is the i-th feasible action
- | · | means cardinality of the set
- α is scaling factor

Overview of selection strategy

- All methods initialized with small random weights making Q learning explorative
- Better approximation will lead to an α value put high probability on optimal action with low exploration probability

Overview of DRRN architecture


$Q_t(s,a^1)$ $Q_t(s,a^2)$
h_2
h_1
s_t a_t^1 a_t^2
$Q_t(s,a^i)$
h_2
h_1
s_t a_t^i
$Q_t(s,a^i)$
pairwise interaction function
(e.g. inner product)
$h_{2,s}$
$h_{1,s}$
s_t
$h_{2,a}^i$
$h_{1,a}^i$
a_t^i

(a) Max-action DQN (b) Per-action DQN (c) DRRN

Figure 1: Different deep Q-learning architectures: Max-action DQN and Per-action DQN both treat input text as concantenated vectors and compute output Q-values with a single NN. DRRN models text embeddings from state/action sides separately, and use an interaction function to compute Q-values.

action 2 (-1.30)
action 1 (-0.55)
state
after 200 episodes
action 1 (+0.91)
action 2 (-17.17)
state
after 400 episodes
action 1 (+16.53)
action 2 (-22.08)
state
after 600 episodes
1
0
−1
−8
−6
−4
−2
0
2
4
6
8

Figure 2: PCA projections of text embedding vectors for state and associated action vectors after 200, 400 and 600 training episodes. The state is "As you move forward, the people surrounding you suddenly look up with terror in their faces, and flee the street." Action 1 (good choice) is "Look up", and action 2 (poor choice) is "Ignore the alarm of others and continue moving forward."

Text-based game

- Can be of three types
  - Parser-based
  - Choice-based
  - Hyper-text
- Study on two types of games
  - Deterministic text games (Saving John)
  - Larger scale stochastic text game (Machine of Death)

Feasible actions in the games

- $|A_t|$ can be a maximum of 4 in Saving John game
- 9 maximum feasible actions in the Machine of Death game
- An episode in Machine of Death is restricted to no longer than 500 steps [1]

Reward of the games

- Manual annotation of final rewards for distinct endings
- Magnitude of the scores provide sentiment polarity of good/bad endings
- Small negative reward in each non-terminating step to encourage the user to complete the game quickly

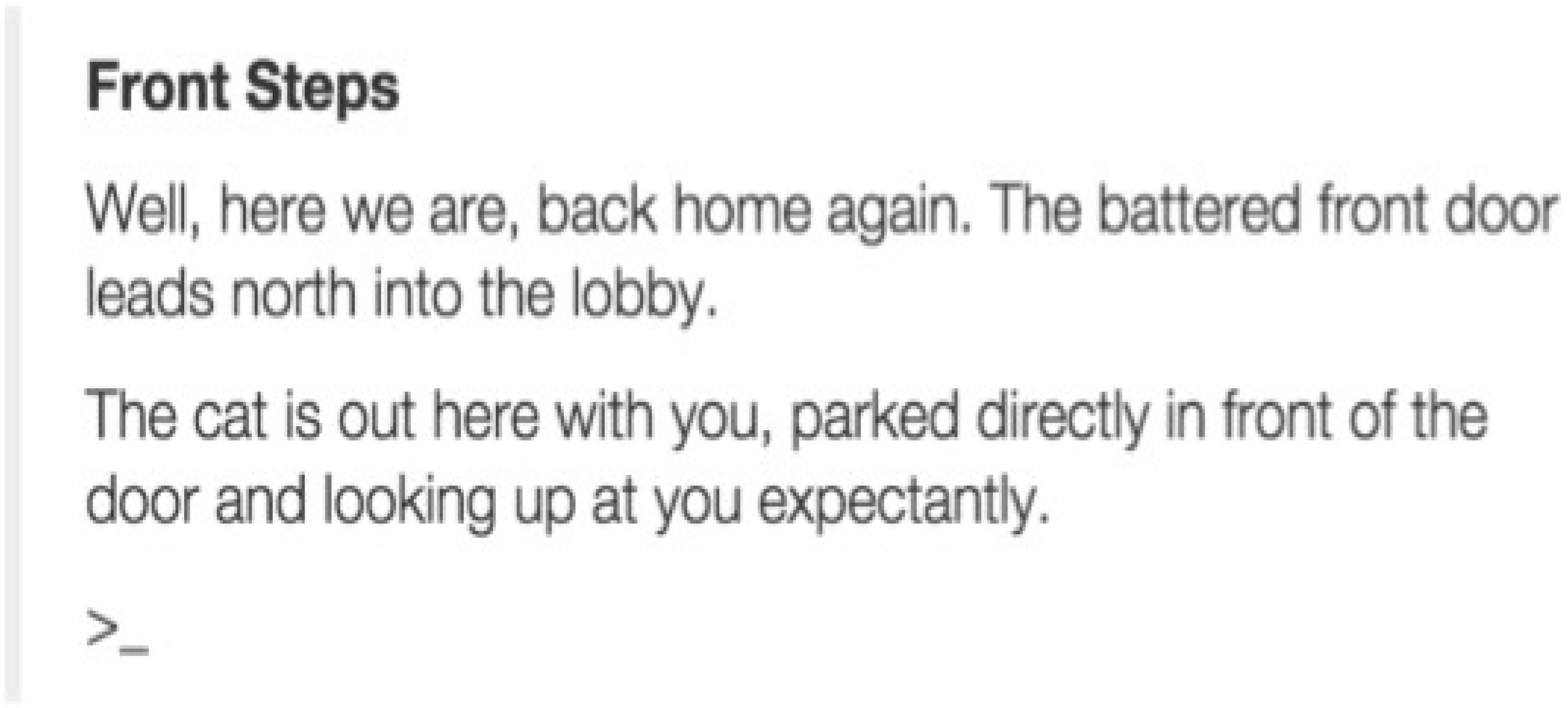
Features of the game

- Raw bag of words
- Different vocabularies for the state side and the action side

Parser-based games

- Inputs are typed-in commands from players
  - e.g. verb phrases like 'eat apple', 'go east'
- Least complex action language

Illustration of parser-based games



Front Steps

Well, here we are, back home again. The battered front door leads north into the lobby.

The cat is out here with you, parked directly in front of the door and looking up at you expectantly.

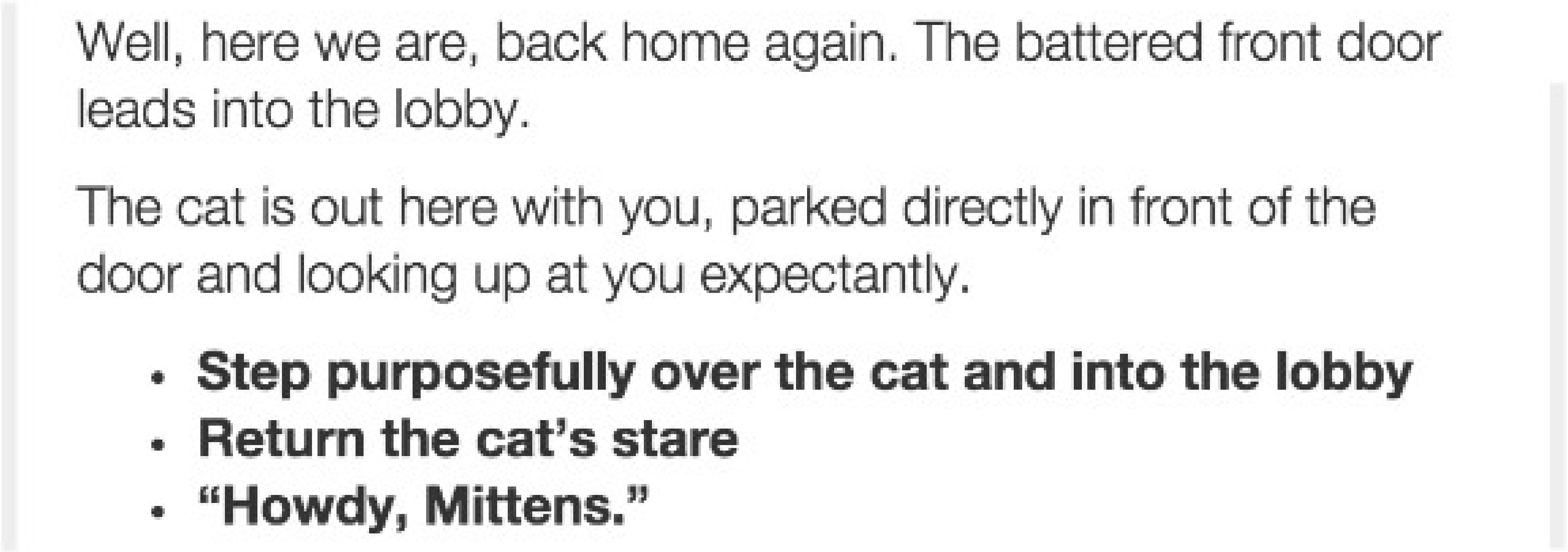
```
>_
```

(a) Parser-based

Choice-based and hypertext games

- Actions presented after or embedded within the state text
- Player selects an action
- Game continues based on selected action at the particular state
- More popular with development of web browsing and HTML display [1]
  - 8% in 2010
  - 62% in 2014

Illustration of choice-based games



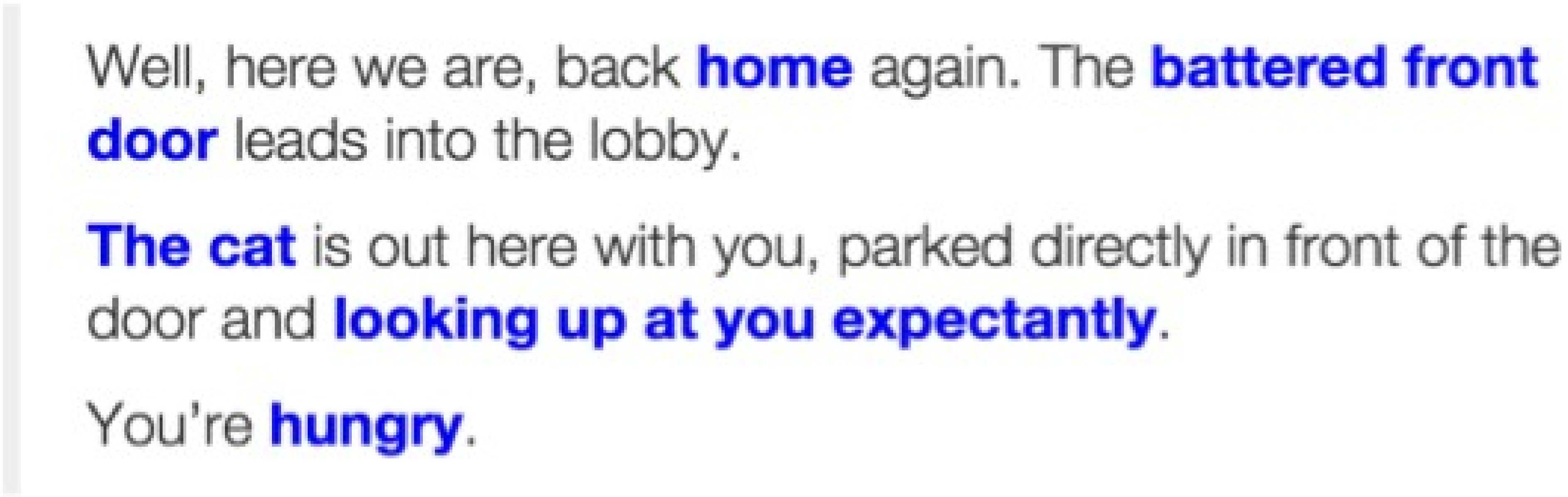
Well, here we are, back home again. The battered front door leads into the lobby.

The cat is out here with you, parked directly in front of the door and looking up at you expectantly.

- **Step purposefully over the cat and into the lobby**
- **Return the cat's stare**
- **"Howdy, Mittens."**

(b) Choiced-based

Illustration of hypertext games



Well, here we are, back **home** again. The **battered front door** leads into the lobby.

The cat is out here with you, parked directly in front of the door and **looking up at you expectantly**.

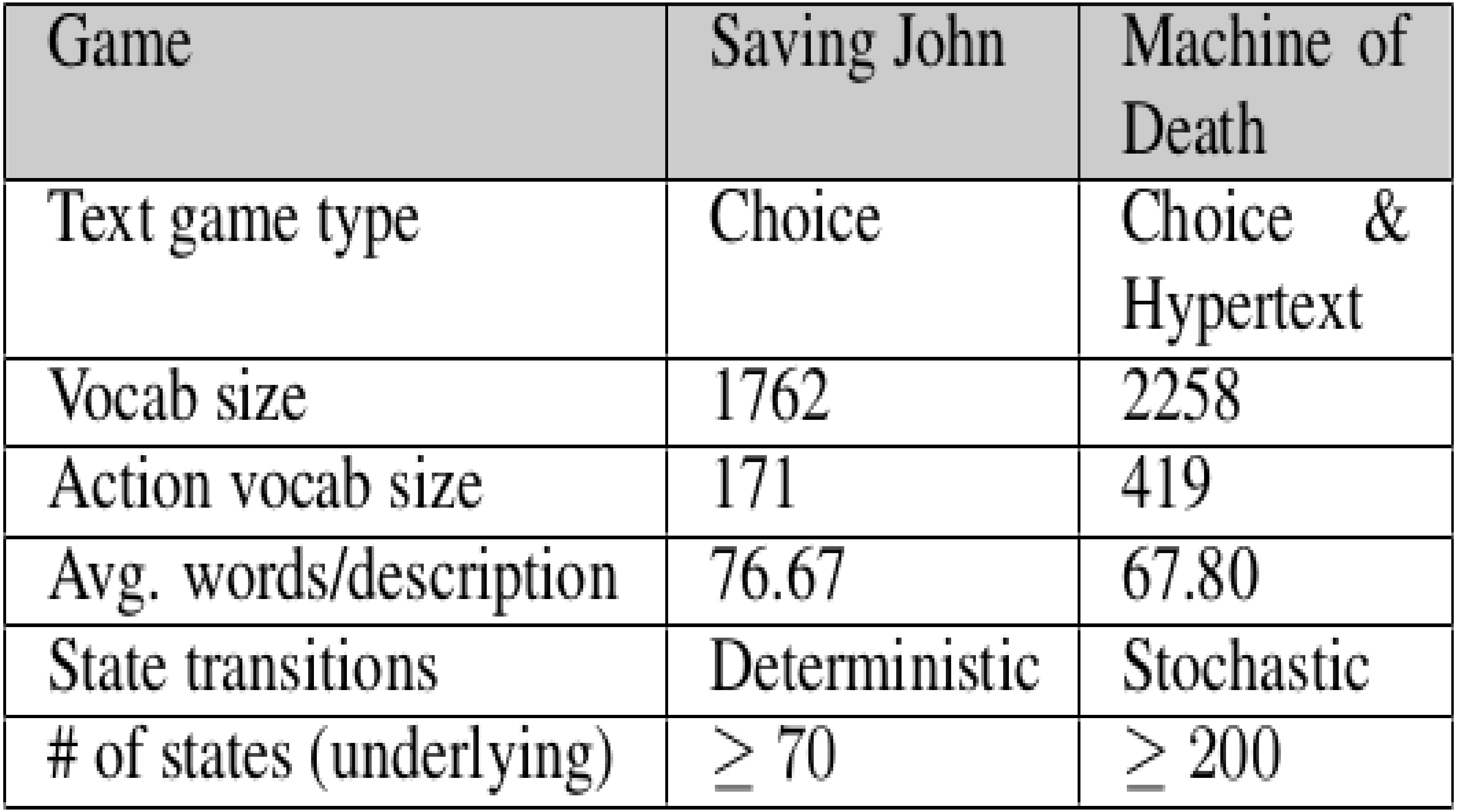
You're **hungry**.

(c) Hypertext-based

Action spaces of the games

- Parser-based games can be reduced to a fixed action-set DQN
- In choice-based and hypertext games, size of action space could be exponential with the length of action sentences
  - Continuous representation of the action space

Game statistics

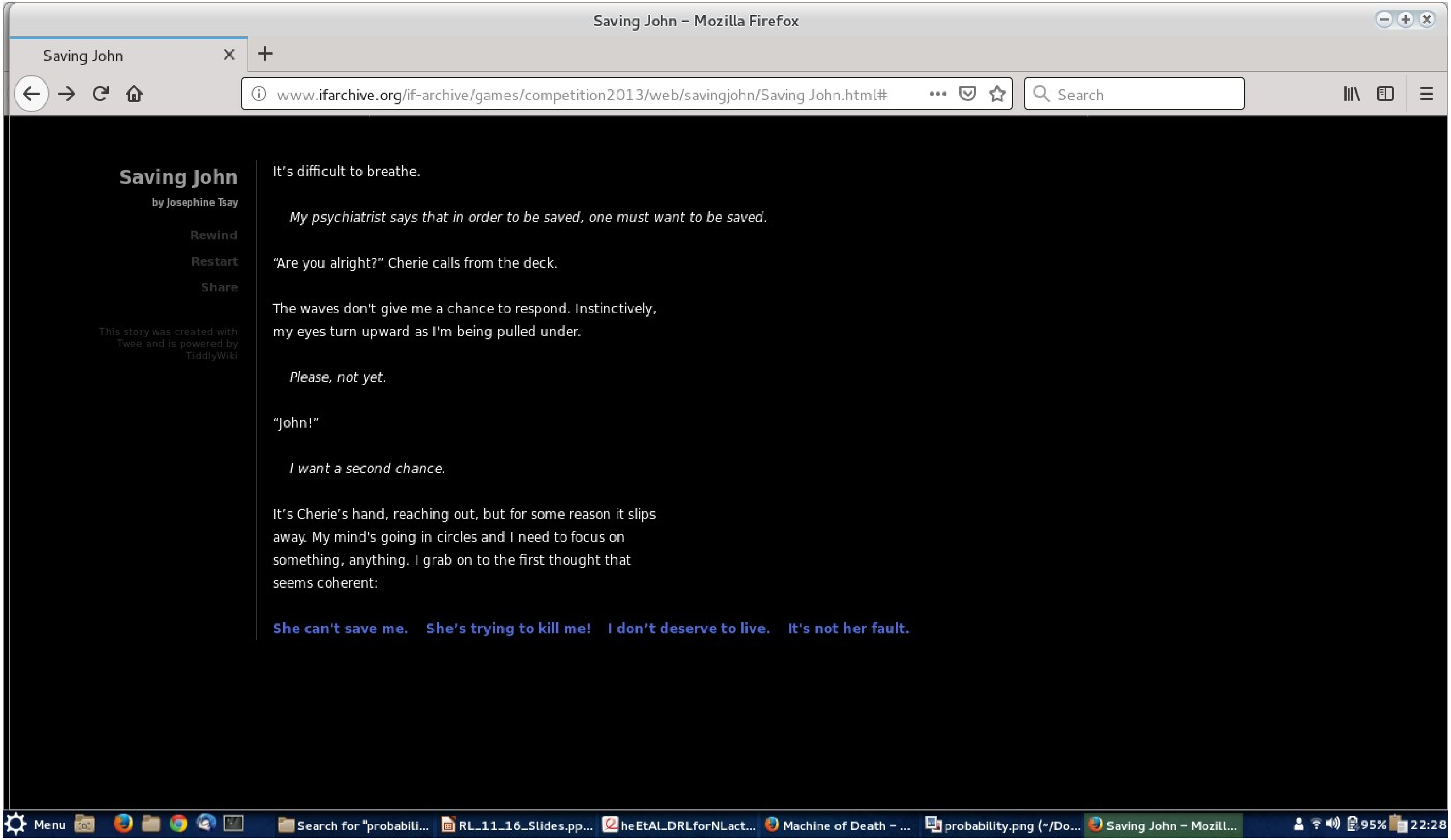


Game	Saving John	Machine of Death
Text game type	Choice	Choice & Hypertext
Vocab size	1762	2258
Action vocab size	171	419
Avg. words/description	76.67	67.80
State transitions	Deterministic	Stochastic
# of states (underlying)	≥ 70	≥ 200

Saving John game [2,4]

- Choice-based games
- Mapping from text strings to a_t is clear

Saving John game



Saving John

by Josephine Tsay

Rewind

Restart

Share

This story was created with Twee and is powered by TiddlyWiki

It's difficult to breathe.

My psychiatrist says that in order to be saved, one must want to be saved.

"Are you alright?" Cherie calls from the deck.

The waves don't give me a chance to respond. Instinctively, my eyes turn upward as I'm being pulled under.

Please, not yet.

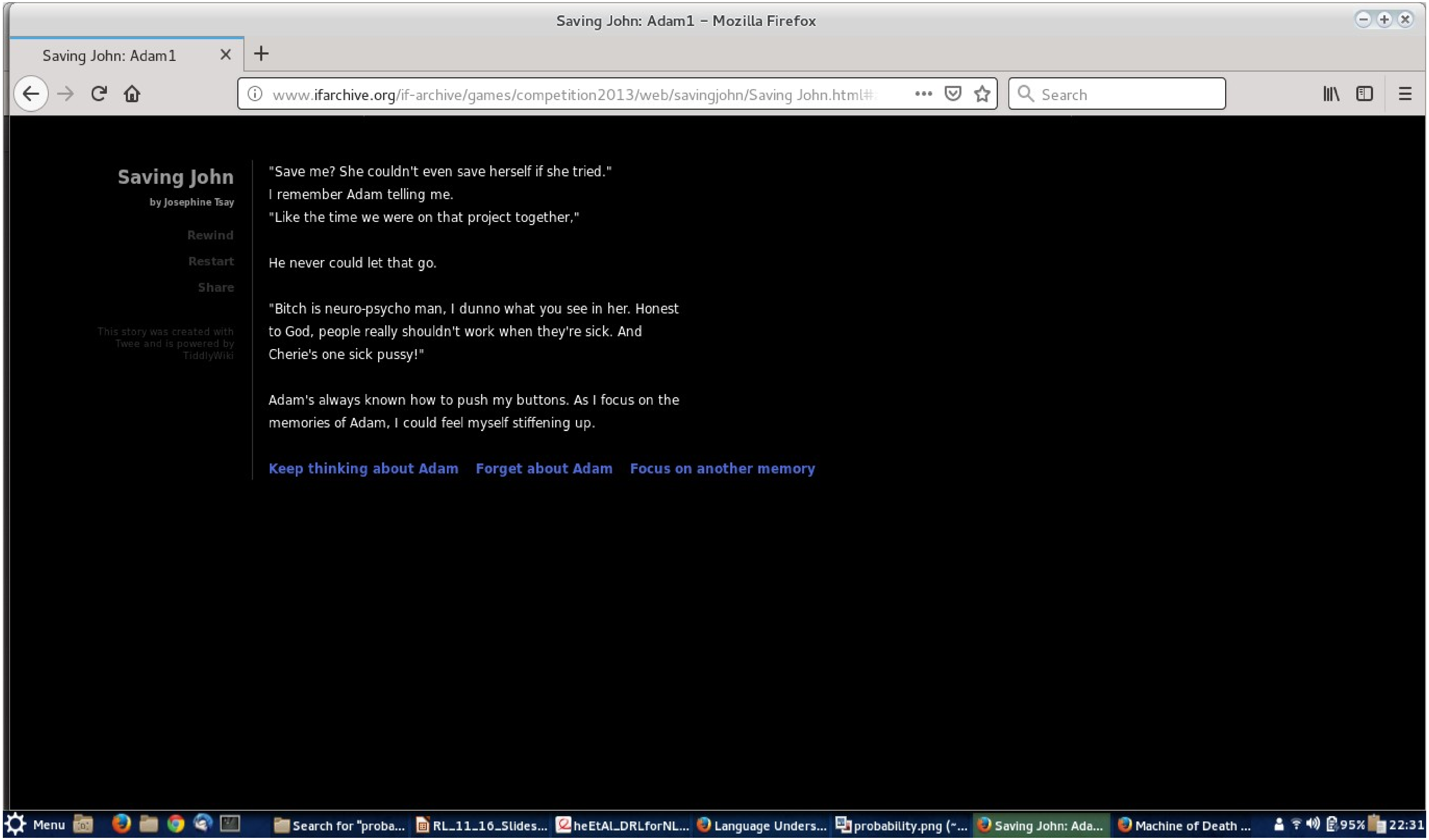
"John!"

I want a second chance.

It's Cherie's hand, reaching out, but for some reason it slips away. My mind's going in circles and I need to focus on something, anything. I grab on to the first thought that seems coherent:

She can't save me. **She's trying to kill me!** **I don't deserve to live.** **It's not her fault.**

State transition in Saving John game



"Save me? She couldn't even save herself if she tried."
I remember Adam telling me.
"Like the time we were on that project together,"

He never could let that go.

"Bitch is neuro-psycho man, I dunno what you see in her. Honest to God, people really shouldn't work when they're sick. And Cherie's one sick pussy!"

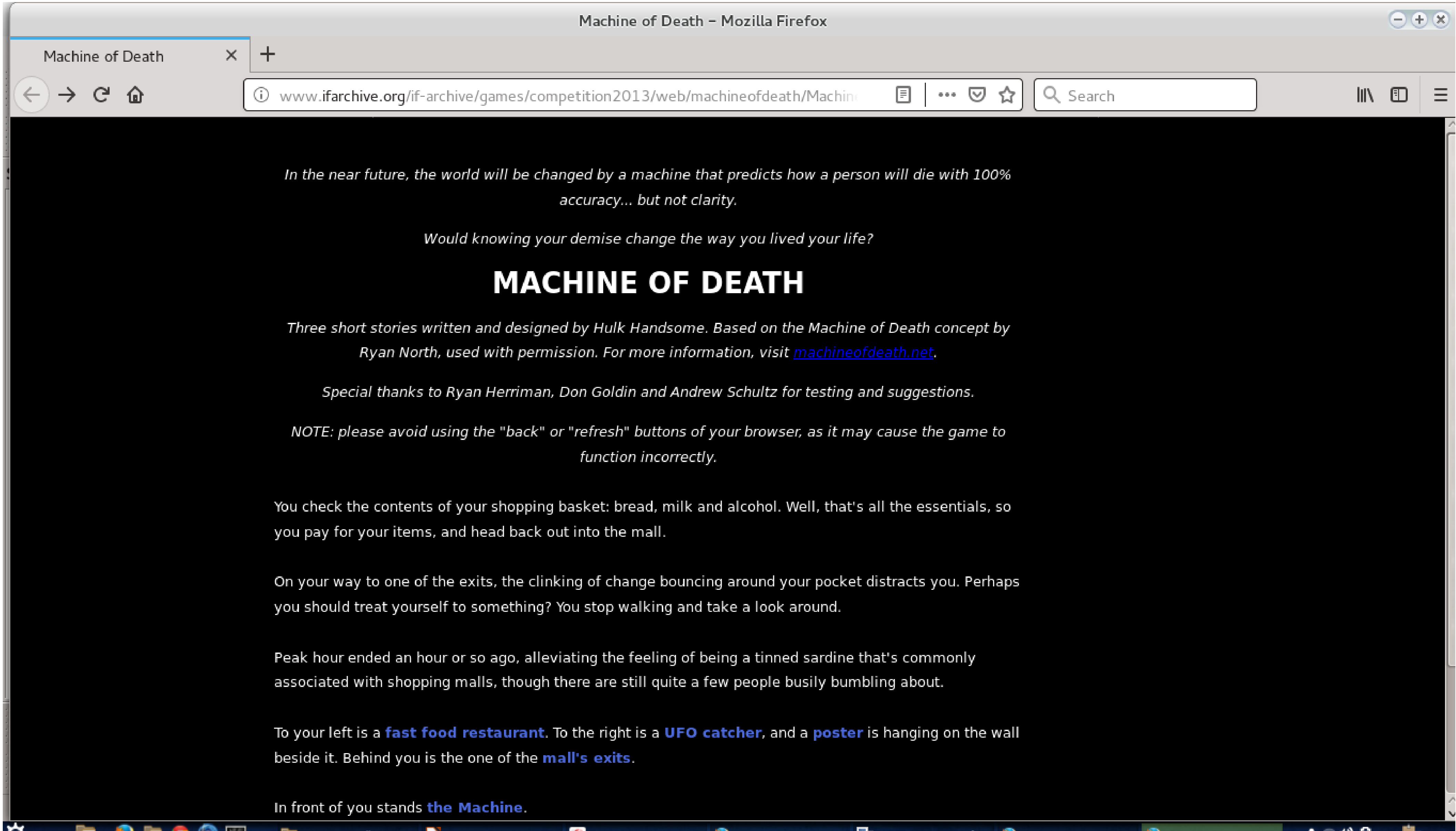
Adam's always known how to push my buttons. As I focus on the memories of Adam, I could feel myself stiffening up.

Keep thinking about Adam **Forget about Adam** **Focus on another memory**

Machine of Death (MoD) game [3,4]

- Hypertext game
- Actions are sub-strings of the state
- s_t is associated with full state description
- a_t is represented by sub-strings without any surrounding context

MoD game



Machine of Death – Mozilla Firefox
Machine of Death
www.ifarchive.org/if-archive/games/competition2013/web/machineofdeath/Machin
Search

In the near future, the world will be changed by a machine that predicts how a person will die with 100% accuracy... but not clarity.

Would knowing your demise change the way you lived your life?

MACHINE OF DEATH

Three short stories written and designed by Hulk Handsome. Based on the Machine of Death concept by Ryan North, used with permission. For more information, visit machineofdeath.net.

Special thanks to Ryan Herriman, Don Goldin and Andrew Schultz for testing and suggestions.

NOTE: please avoid using the "back" or "refresh" buttons of your browser, as it may cause the game to function incorrectly.

You check the contents of your shopping basket: bread, milk and alcohol. Well, that's all the essentials, so you pay for your items, and head back out into the mall.

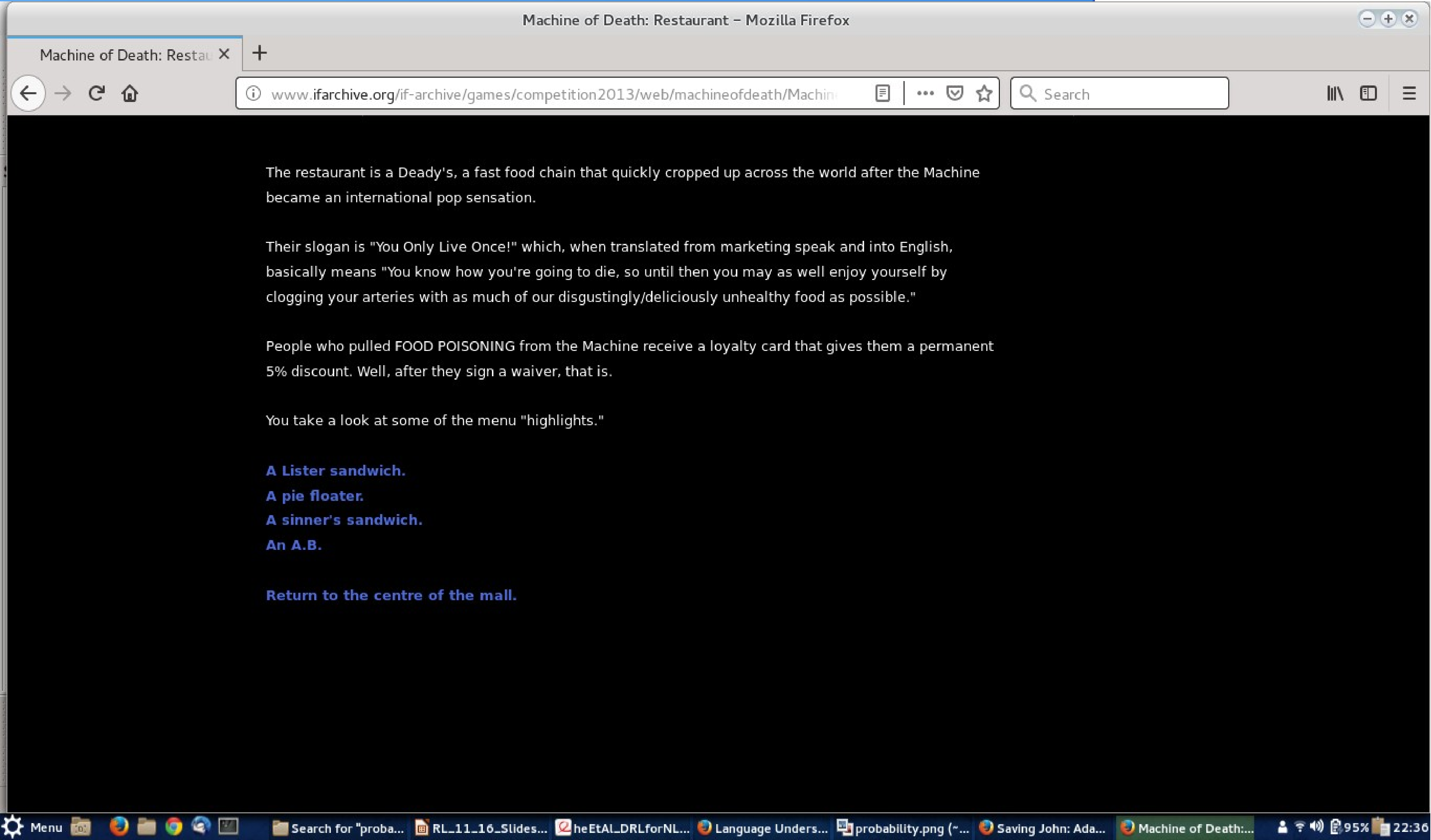
On your way to one of the exits, the clinking of change bouncing around your pocket distracts you. Perhaps you should treat yourself to something? You stop walking and take a look around.

Peak hour ended an hour or so ago, alleviating the feeling of being a tinned sardine that's commonly associated with shopping malls, though there are still quite a few people busily bumbling about.

To your left is a **fast food restaurant**. To the right is a **UFO catcher**, and a **poster** is hanging on the wall beside it. Behind you is the one of the **mall's exits**.

In front of you stands **the Machine**.

Possible action in MoD game



The restaurant is a Deady's, a fast food chain that quickly cropped up across the world after the Machine became an international pop sensation.

Their slogan is "You Only Live Once!" which, when translated from marketing speak and into English, basically means "You know how you're going to die, so until then you may as well enjoy yourself by clogging your arteries with as much of our disgustingly/deliciously unhealthy food as possible."

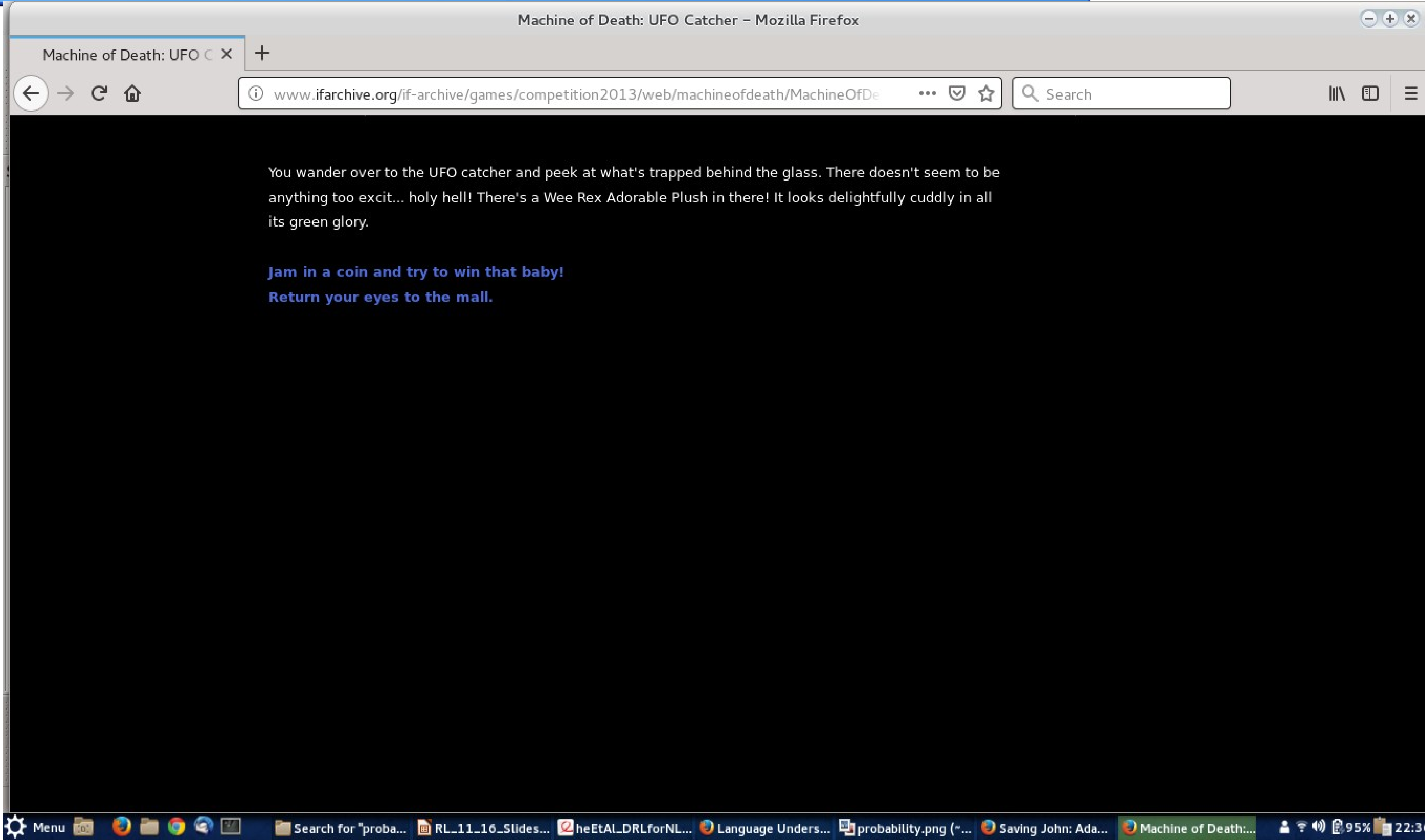
People who pulled FOOD POISONING from the Machine receive a loyalty card that gives them a permanent 5% discount. Well, after they sign a waiver, that is.

You take a look at some of the menu "highlights."

A Lister sandwich.
A pie floater.
A sinner's sandwich.
An A.B.

Return to the centre of the mall.

Another possible state transition in MoD game

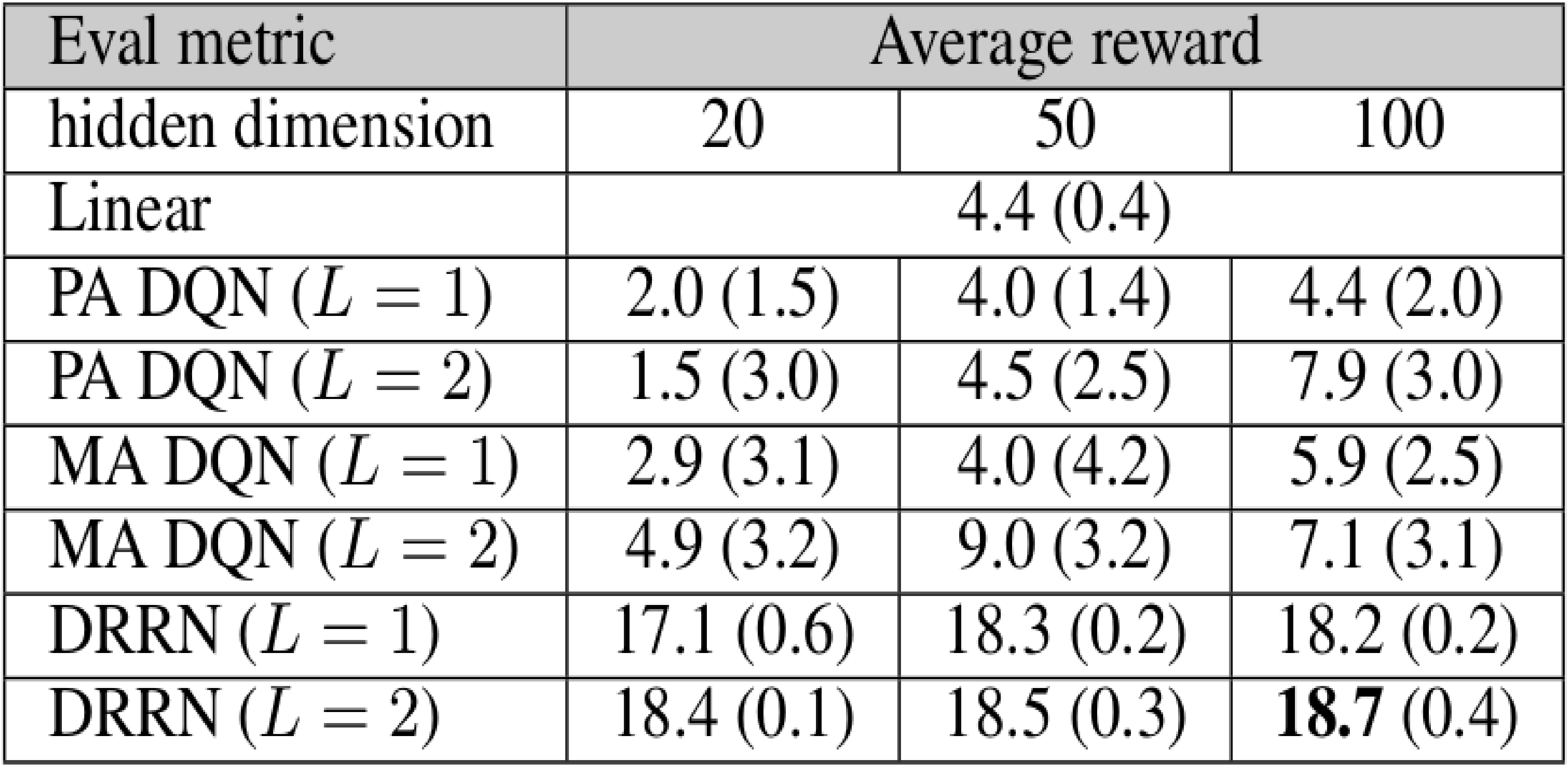


You wander over to the UFO catcher and peek at what's trapped behind the glass. There doesn't seem to be anything too excit... holy hell! There's a Wee Rex Adorable Plush in there! It looks delightfully cuddly in all its green glory.

Jam in a coin and try to win that baby!

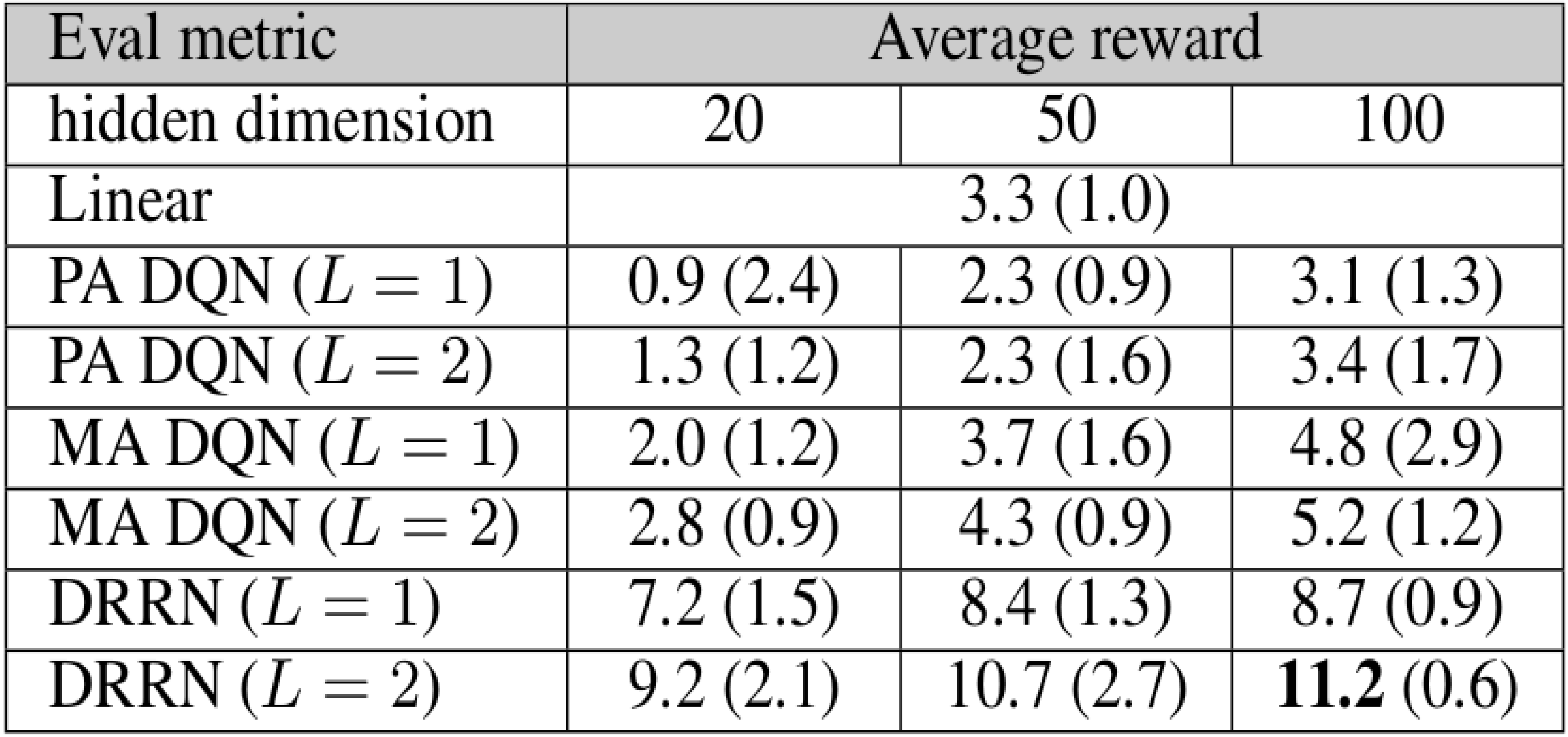
Return your eyes to the mall.

Rewards on Saving John game



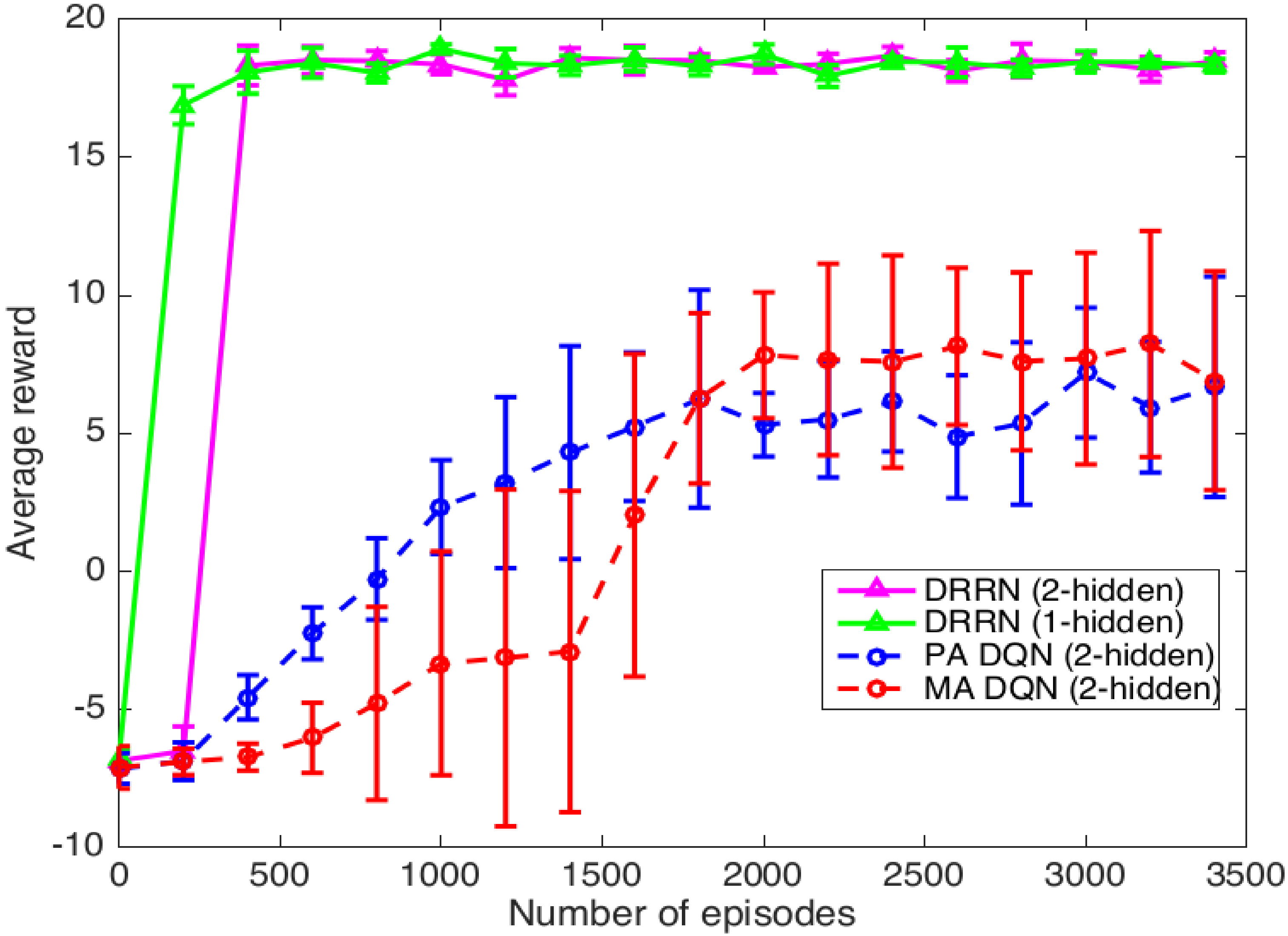
Eval metric	Average reward		
hidden dimension	20	50	100
Linear	4.4 (0.4)		
PA DQN ($L = 1$)	2.0 (1.5)	4.0 (1.4)	4.4 (2.0)
PA DQN ($L = 2$)	1.5 (3.0)	4.5 (2.5)	7.9 (3.0)
MA DQN ($L = 1$)	2.9 (3.1)	4.0 (4.2)	5.9 (2.5)
MA DQN ($L = 2$)	4.9 (3.2)	9.0 (3.2)	7.1 (3.1)
DRRN ($L = 1$)	17.1 (0.6)	18.3 (0.2)	18.2 (0.2)
DRRN ($L = 2$)	18.4 (0.1)	18.5 (0.3)	**18.7** (0.4)

Rewards on MoD game

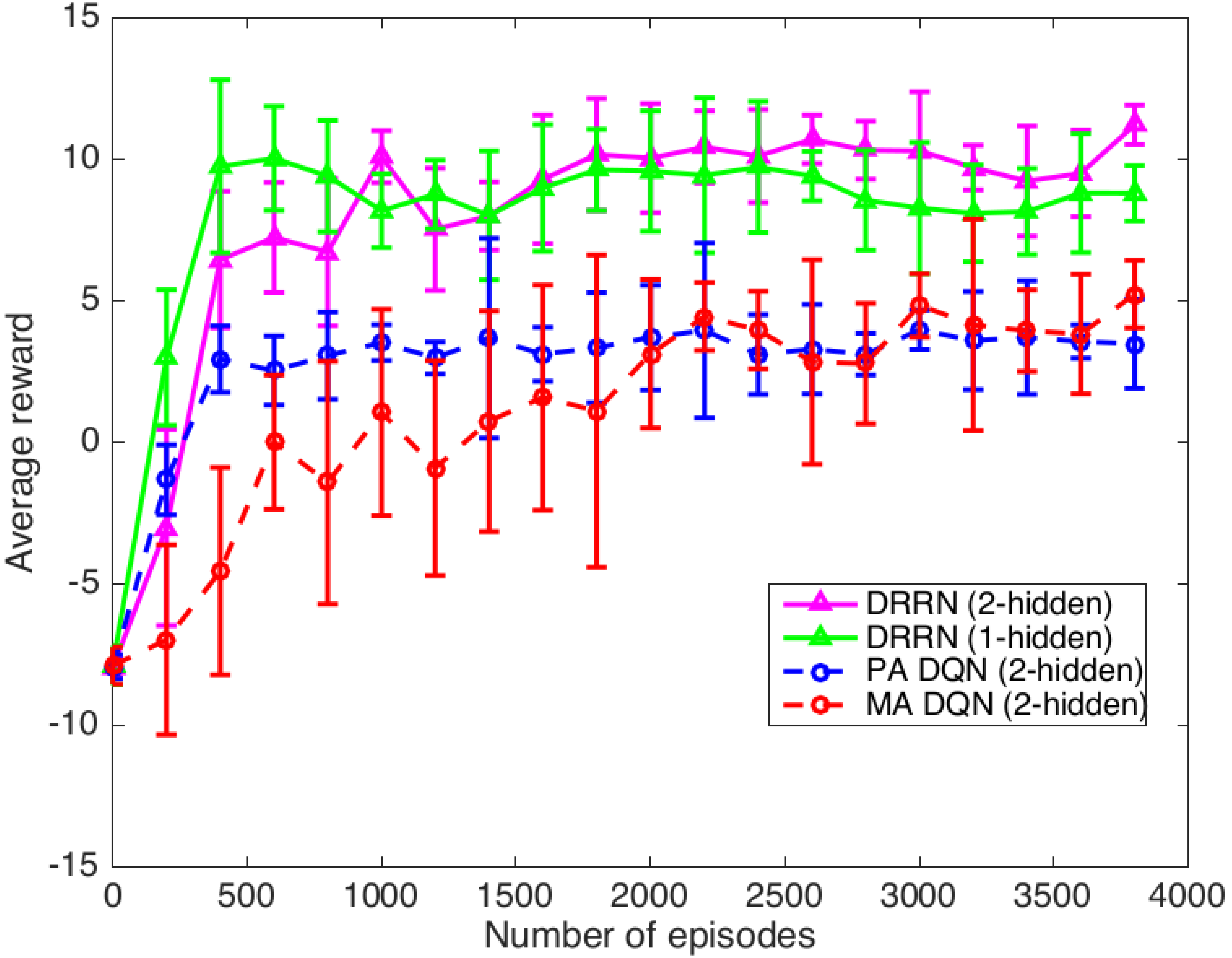


Eval metric	Average reward		
hidden dimension	20	50	100
Linear	3.3 (1.0)		
PA DQN ($L = 1$)	0.9 (2.4)	2.3 (0.9)	3.1 (1.3)
PA DQN ($L = 2$)	1.3 (1.2)	2.3 (1.6)	3.4 (1.7)
MA DQN ($L = 1$)	2.0 (1.2)	3.7 (1.6)	4.8 (2.9)
MA DQN ($L = 2$)	2.8 (0.9)	4.3 (0.9)	5.2 (1.2)
DRRN ($L = 1$)	7.2 (1.5)	8.4 (1.3)	8.7 (0.9)
DRRN ($L = 2$)	9.2 (2.1)	10.7 (2.7)	**11.2** (0.6)

Learning curve on Saving John game



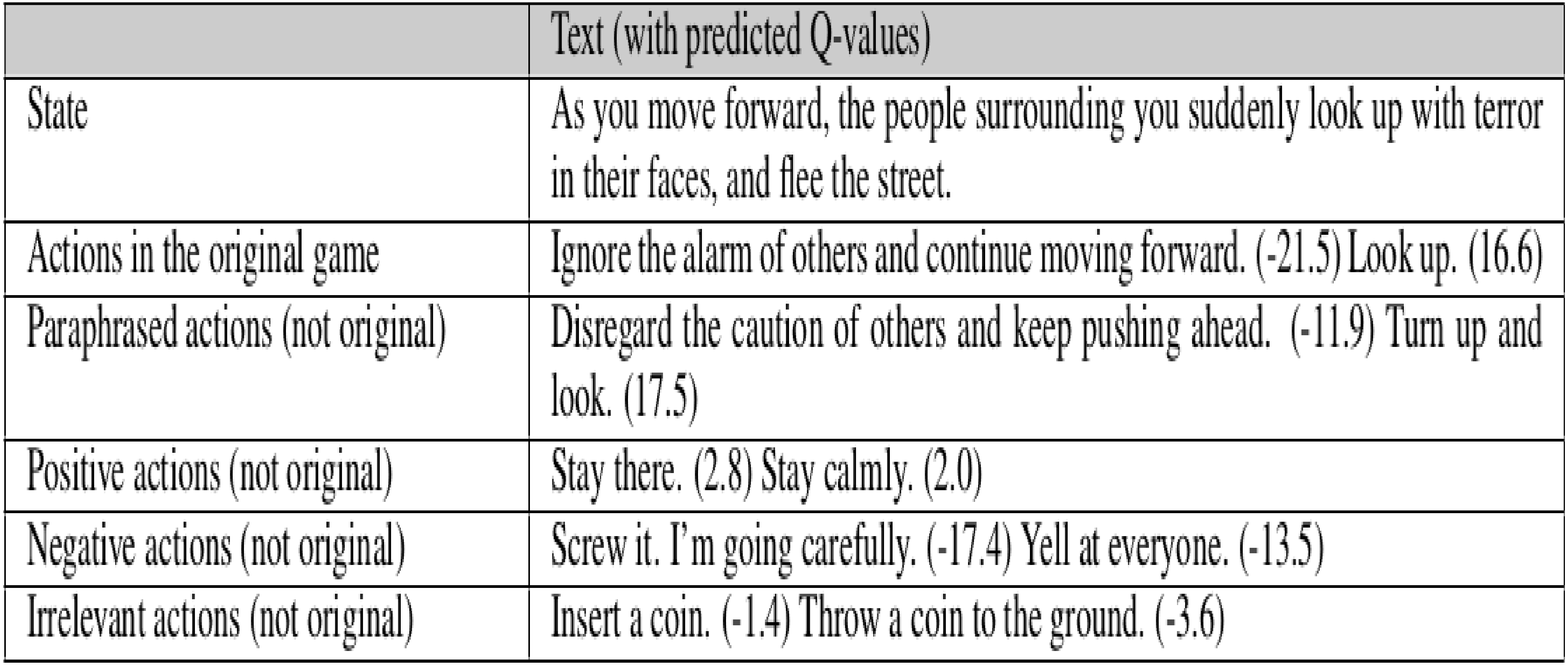
Learning curve on MoD game



Result with paraphrased descriptions

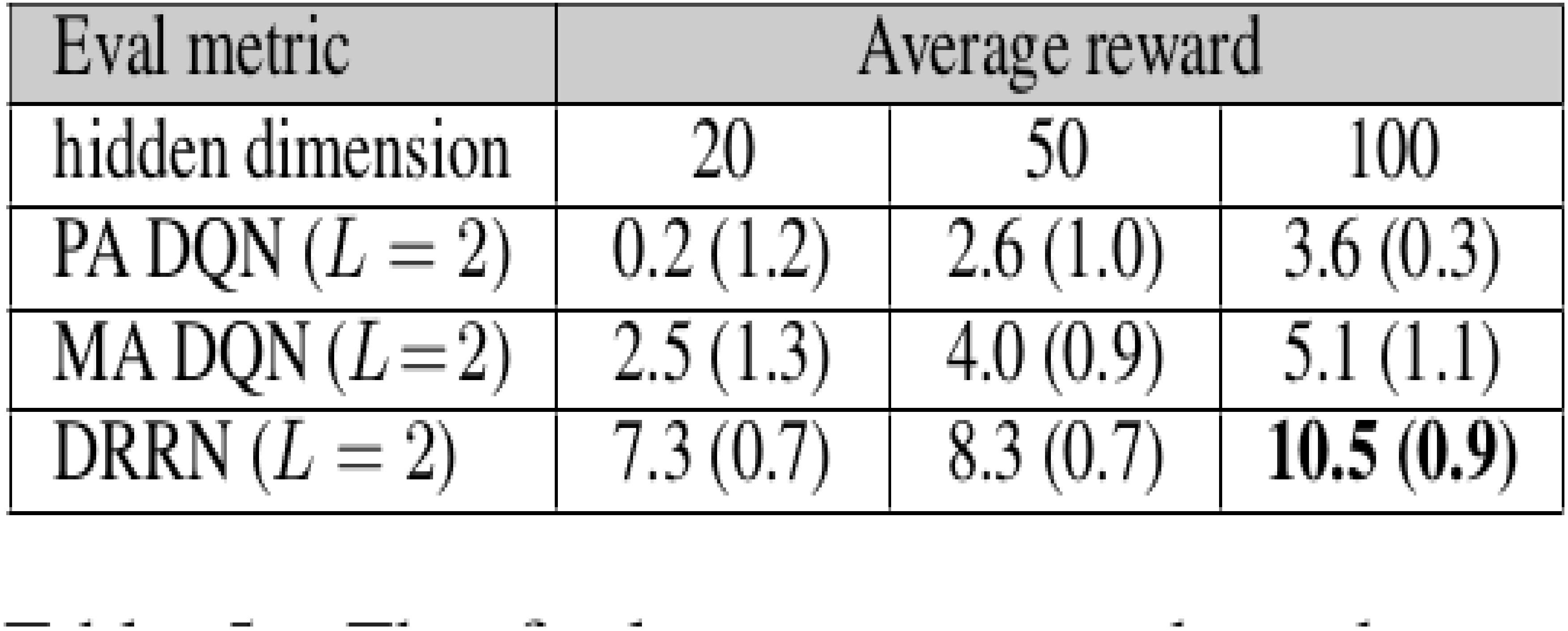
- 2 people paraphrase all actions of MoD game in testing phase
- Dropping some single word actions that are out-of-vocabulary (OOV)
- Word level OOV rate of paraphrased actions is 18.6%
- Q value predicted for each state action pair with fixed model parameters
- Actions more likely to result in good endings presented with high Q values.

Examples of the model on paraphrasing



	Text (with predicted Q-values)
State	As you move forward, the people surrounding you suddenly look up with terror in their faces, and flee the street.
Actions in the original game	Ignore the alarm of others and continue moving forward. (-21.5) Look up. (16.6)
Paraphrased actions (not original)	Disregard the caution of others and keep pushing ahead. (-11.9) Turn up and look. (17.5)
Positive actions (not original)	Stay there. (2.8) Stay calmly. (2.0)
Negative actions (not original)	Screw it. I'm going carefully. (-17.4) Yell at everyone. (-13.5)
Irrelevant actions (not original)	Insert a coin. (-1.4) Throw a coin to the ground. (-3.6)

Reward on paraphrased game of MoD



Eval metric	Average reward		
hidden dimension	20	50	100
PA DQN ($L = 2$)	0.2 (1.2)	2.6 (1.0)	3.6 (0.3)
MA DQN ($L=2$)	2.5 (1.3)	4.0 (0.9)	5.1 (1.1)
DRRN ($L = 2$)	7.3 (0.7)	8.3 (0.7)	**10.5 (0.9)**

Overview of Internatural Harvester (IH) game

- 2 users have to play as a team to earn maximum points
- Bot does not use any A.I. algorithms as the focus is on the dialog generated by collaboration of the 2 users

Contrast between the games in Ji He et al. Paper and IH game

	He paper	IH game
Number of players	1	2
System responses	States and transition of states	Bot provides update messages and the board changes
Environment	Choice-based, hypertext	Game-based, treasures, bots, commands provided
Game objective	Earning points	Earning points
Strategic	Obtaining highest long-term reward.	Till now unknown. Collaboration can be explored.
Observable	Partially	Partially

Interesting paper to follow up

- Language Understanding for Text-based Games using Deep Reinforcement Learning [1,5,6]
- Authors
  - Dr. Karthik Narasimhan (currently at Princeton University), Dr. Tejas Kulkarni (currently at Google DeepMind) and Prof. Regina Berzilay, CSAIL, MIT, U.S.A.
  - Equal contributions by Dr. Narasimhan and Dr. Kulkarni

Overview of the paper

- 133 citations till date in Google Scholar
- Published in the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1–11, Lisbon, Portugal, 17-21 September 2015.
- Motivation behind Ji He paper

Summary of the paper

- Text descriptions mapped to vector representations capturing the semantics of the game state[6]
- Deep reinforcement learning framework to jointly learn state representations and action policies using game rewards as feedbacks
- LSTM-DQN used

Games in the paper

- 2 multi-user dungeon games
- Home World game
- Fantasy World game
- Commands are provided

References

1. Deep Reinforcement Learning with a Natural Language Action Space by Ji He, Jianshu Chen , Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng and Mari Ostendorf, University of Washington, Seattle, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1621–1630, Berlin, Germany, August 7-12, 2016.

2. http://www.ifarchive.org/if-archive/games/competition2013/web/savingjohn/Saving%20John.html#a.5

References

3. http://www.ifarchive.org/if-archive/games/competition2013/web/machineofdeath/MachineOfDeath.html

4. https://github.com/jvking/text-games

5.https://github.com/howardyclo/papernotes/issues/18

References

6. Language Understanding for Text-based Games using Deep Reinforcement Learning by Karthik Narasimhan, Tejas D Kulkarni and Regina Barzilay, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp 1–11, Lisbon, Portugal, 17-21 September 2015