Rethinking Action Spaces for RL

Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, **T. Zhao** et. al., **NAACL-HLT 2019**

09/07/2020

Presented by: Mojtaba Valipour

PhD student of Computer Science at Data Analytics Lab

CS 885 – Reinforcement Learning – Pascal Poupart

UNIVERSITY OF WATERLOO

ANALYTICS LAB DATA

Outline

IMAGE CREDIT: PIXAR Wall-E

IMAGE CREDIT: https://www.theguardian.com

Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

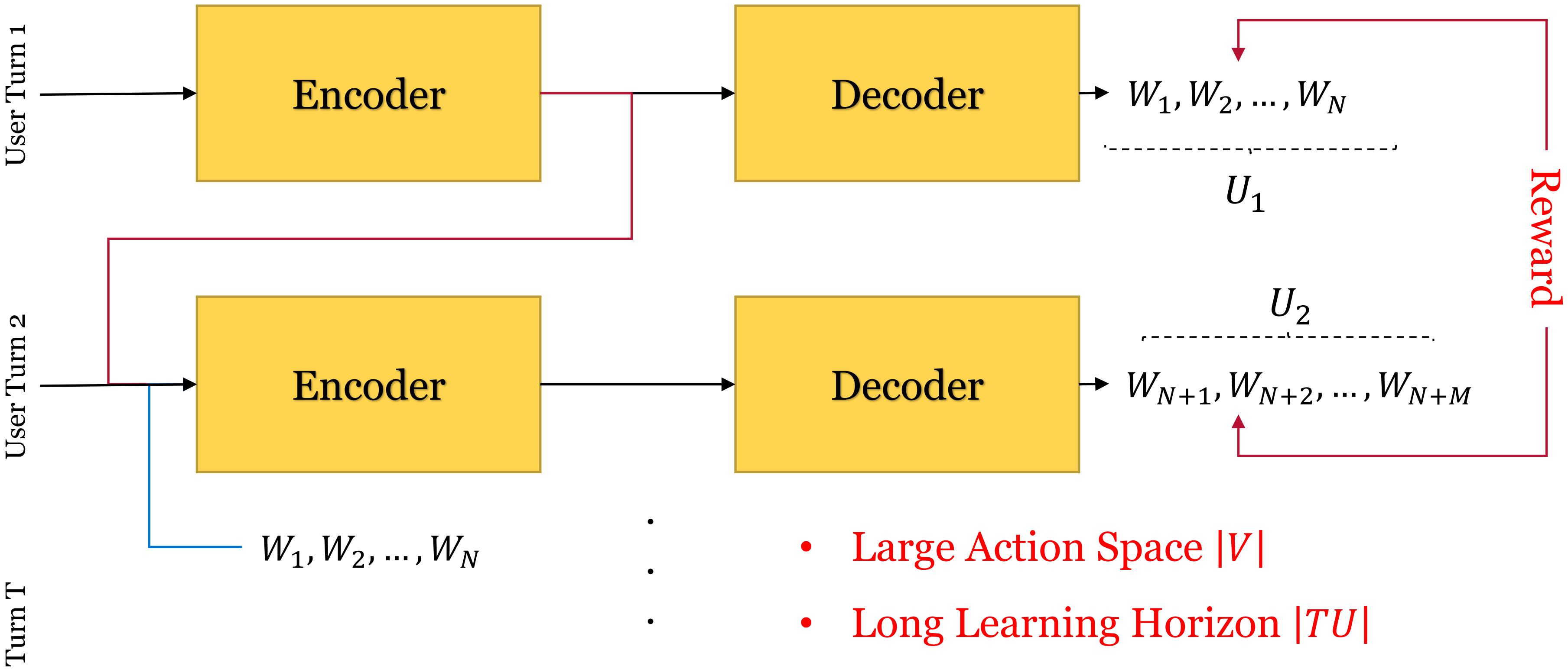
PROBLEM

What and Why?

END2END DIALOG AGENTS

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta\left[\sum_{0}^{T}\sum_{j=0}^{U_t} R_{tj}\nabla_\theta log p_\theta(w_{tj}|w_{<tj}, c_t)\right]$$

$$L_{SL}(\theta) = \mathbb{E}_{x,c}[log p_\theta(x|c)]$$



- Large Action Space $|V|$
- Long Learning Horizon $|TU|$

Scenario 1
1 book, 1 hat and 3 balls available.
Value Function
Agent: book=1, hat=6, ball=1.
User: book=1, hat=6, ball=1.
Dialog
Agent: I'd like the hat and 2 balls.
User: Deal.
Agent: SELECTION
Outcome: Agreement
Agent: 1 hat and 2 balls; User: 1 book and 1 ball
Agent reward = 8; User reward = 2

Scenario 2
4 books, 1 hat and 1 ball available.
Value Function
Agent: book=0, hat=7, ball=3;
User: book=1, hat=6, ball=0.
Dialog
User: I can offer you 1 hat and 2 books.
Agent: I need the hat and ball.
User: That won't work for me. I can offer you one book and the hat.
Agent: I need the hat and ball.
User: I have to have the hat.
Agent: I need the hat and ball.
User: Okay.
Agent: SELECTION
Outcome: Agreement
Agent: 1 hat and 1 ball; User: 4 books
Agent reward = 10; User reward = 4

Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

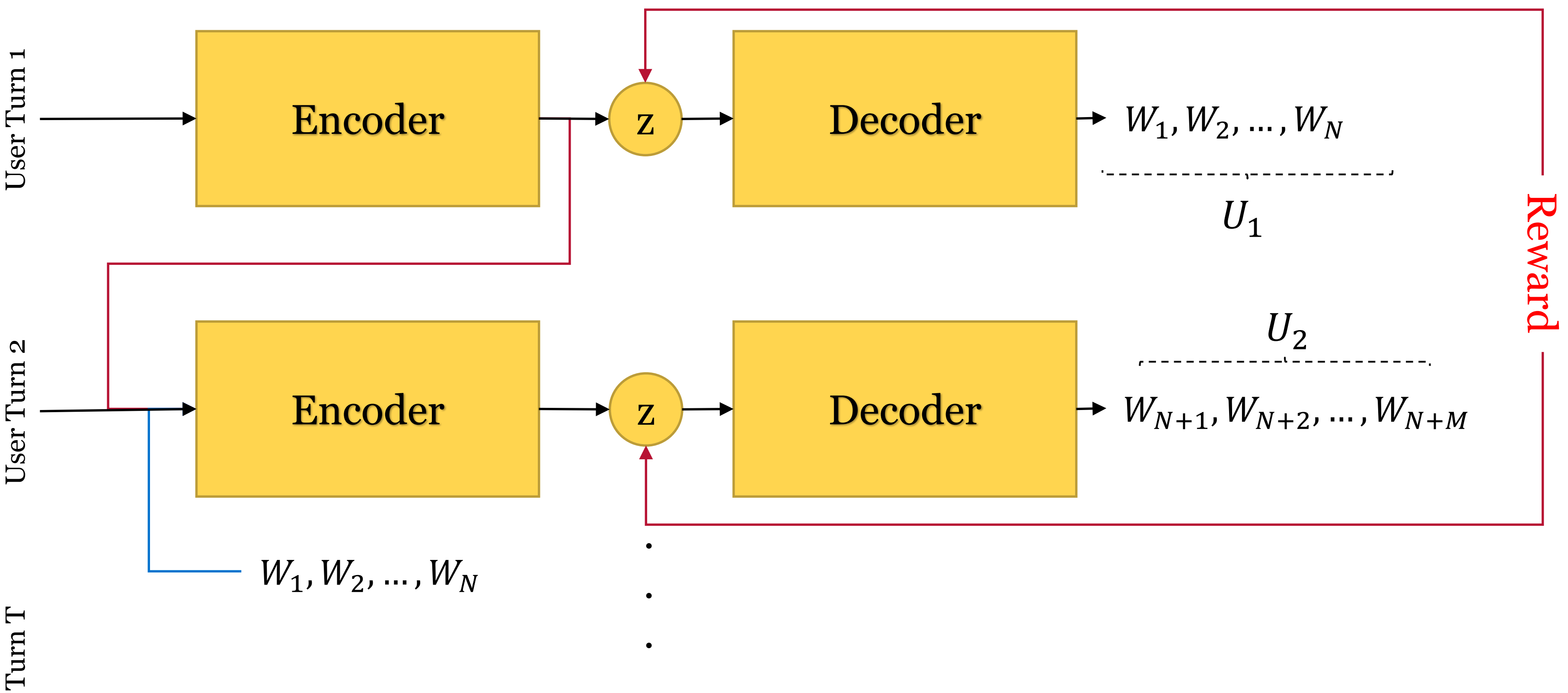
PROPOSED METHOD

LARL? How to discretize the action space?

PROPOSED MODEL

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta\left[\sum_0^T R_t log p_\theta(z|c_t)\right]$$

$$p(x|c) = p(x|z)\, p(z|c)$$



Now the **question** is what kind of **latent actions** is more **suitable** for this task:

- Gaussian
- Categorical

?

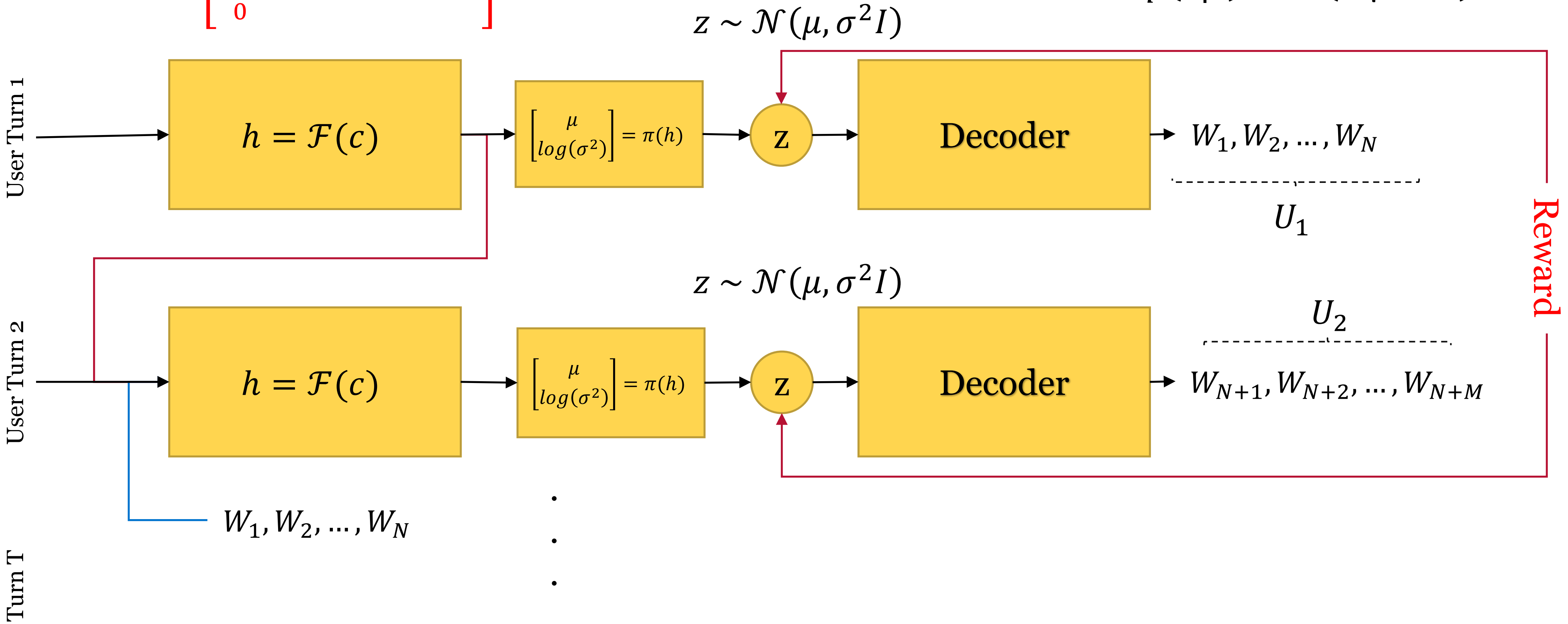
Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

GAUSSIAN LATENT ACTION

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta \left[\sum_0^T R_t log p_\theta(z|c_t) \right]$$

$$p(x|z) = p_{\theta_d}(z)$$

$$p(z|c) = \mathcal{N}(z; \mu, \sigma^2 I)$$



Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

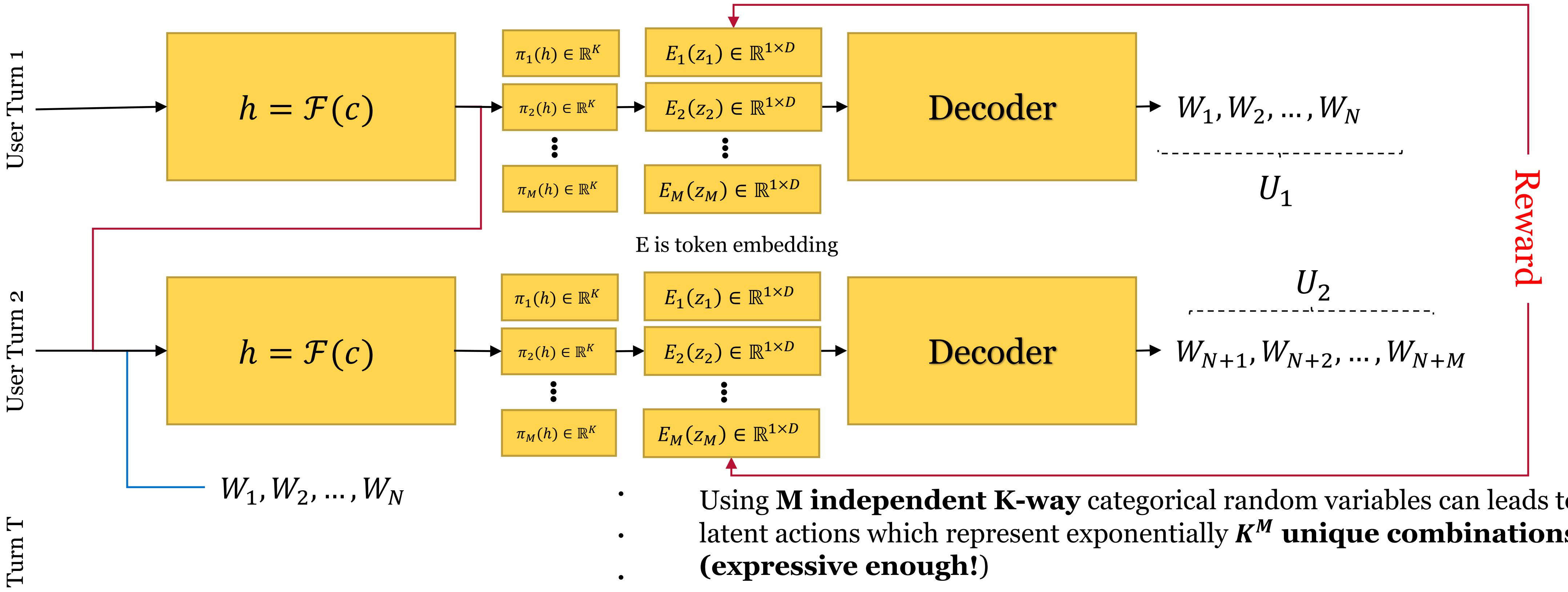
CATEGORICAL LATENT ACTION

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta\left[\sum_0^T R_t log p_\theta(z|c_t)\right]$$

$$z_m \sim p(Z_m|c) = softmax(\pi_m(h))$$

$$p(x|z) = p_{\theta_d}(\mathrm{E}_{1:M}(z_{1:m}) \in \mathbb{R}^{M\times D})$$

$$p_\theta(z|c) = \prod_{m=1}^{M} p(Z_m = z_m|c)$$

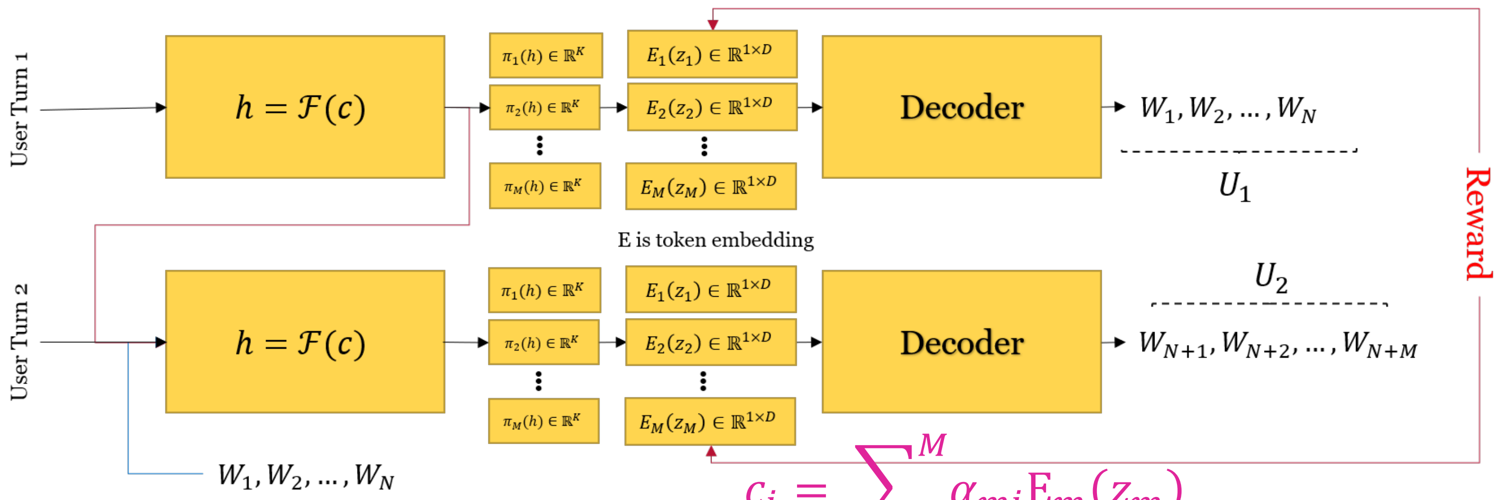


Using **M independent K-way** categorical random variables can leads to latent actions which represent exponentially $\boldsymbol{K^M}$ **unique combinations (expressive enough!)**

Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

ATTENTION FUSION

$\mathrm{E}_{1:M}(z_{1:m}) \in \mathbb{R}^{M \times D}$ Decoder Initial State $\in \mathbb{R}^D$



$$z_m \sim p(Z_m|c) = softmax(\pi_m(h))$$

$$c_i = \sum_1^M \alpha_{mi} \mathrm{E}_m(z_m)$$

$$\alpha_{mi} = softmax(h_{d_i}^T W_a E_m(z_m))$$

$$\widetilde{h_{d_i}} = tanh\left(W_s \begin{bmatrix} h_{d_i} \\ c_i \end{bmatrix}\right)$$

Summation Fusion:

$$\mathrm{x} = p_{\theta_d}\left(\sum_1^M \mathrm{E}_m(z_m)\right) \in \mathbb{R}^D$$

- lose fine-grained order information
- Issues with long responses

Contribution

Attention Fusion:

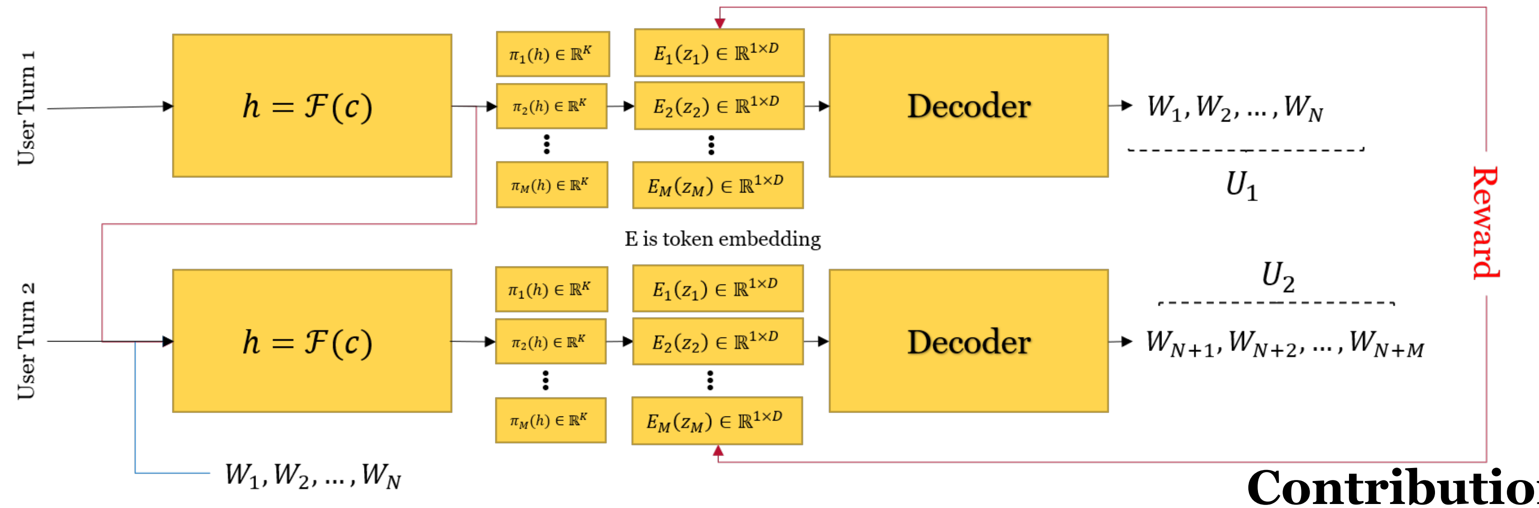
i: step index during decoding

$$p(w_i | h_{d_i}, c_i) = softmax(W_o \widetilde{h_{d_i}})$$

$$h_{d_{i+1}} = RNN(h_{d_i}, w_{i+1}, \widetilde{h_{d_i}})$$

Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

OPTIMIZATION



$$p(z|c) = \pi(\mathcal{F}(\mathrm{c}))$$

$$p(x|z) = p_{\theta_d}(\mathrm{E(z)})$$

Contribution

Full ELBO (Evidence Lower Bound):

$$L_{full}(\theta) = p_{q(z|x,c)}(x|z) - D_{KL}[q(z|x,c)||p(z|c)]$$

Exposure Bias: The decoder only sees z sampled from $q(z|x,c)$, and never experiences z sampled from $p_\theta(z|c)$

Lite ELBO (Evidence Lower Bound):

$$q(z|x,c) = p_{\theta_e}(z|c)$$

$$L_{lite}(\theta) = p_{p(z|c)}(x|z) - \cancel{D_{KL}[p_{\theta_e}(z|c)||p(z|c)]}$$

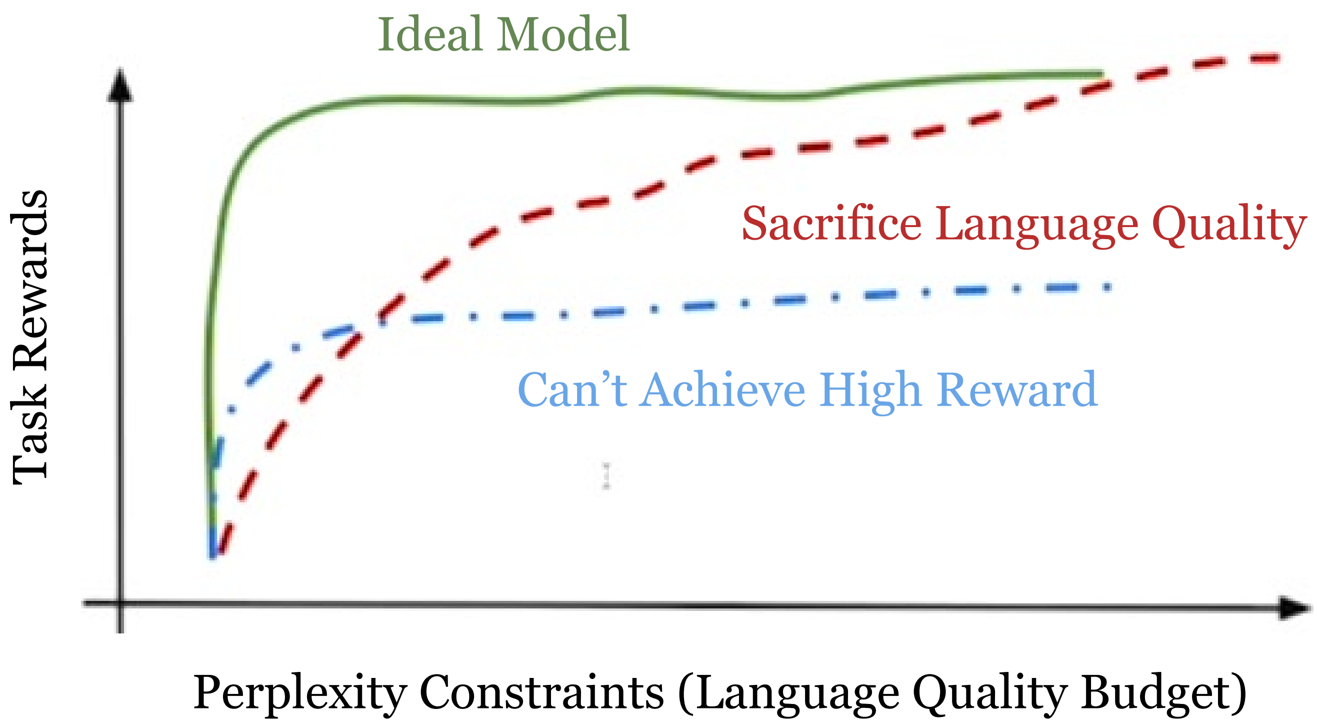
$$L_{lite}(\theta) = p_{p(z|c)}(x|z) - \beta D_{KL}[p(z|c)||p(z)]$$

$p(z) = 1/K$ OR $p(z) = \mathcal{N}(0, I)$

Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

LANGUAGE CONSTRAINED REWARD CURVE (LCR)

Contribution



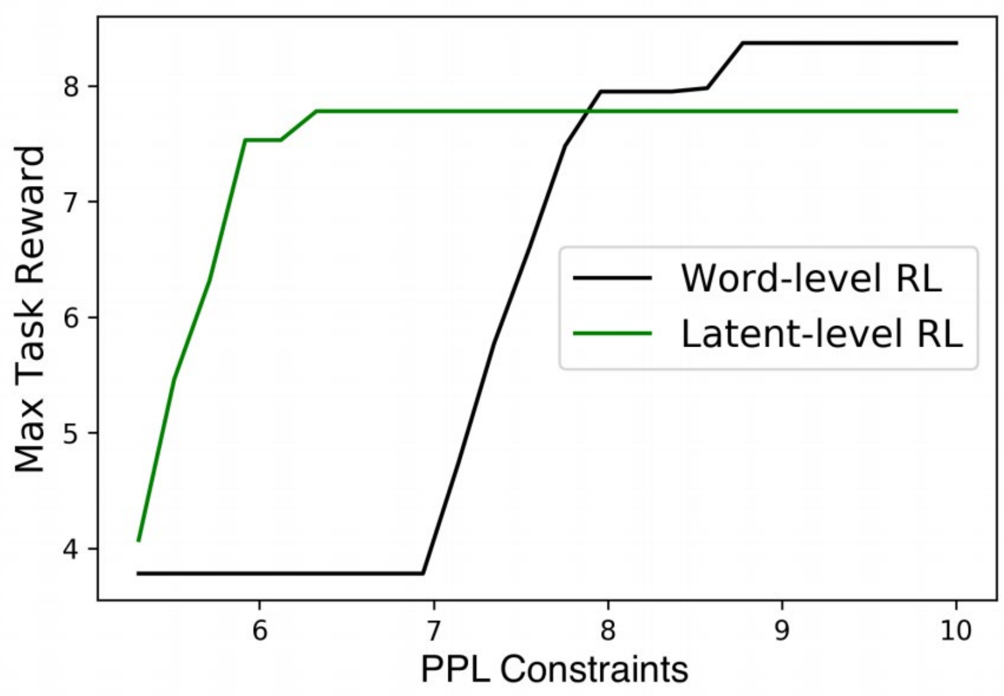
Past **metrics** can’t quantify the **balance** between **task reward** and **language** generation **quality** well

Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.
2- https://vimeo.com/360620730

RESULTS: DEAL OR NO DEAL

DealOrNoDeal is a **negotiation** dataset that contains **5805 dialogs** based **on 2236 unique scenarios**

252 scenarios for testing environment and randomly **sample 400 scenarios** from the training set for validation



	PPL	Reward	Agree%	Diversity
Baseline	5.23	3.75	59	109
LiteCat	5.35	2.65	41	58
Baseline +RL	8.23	**7.61**	86	5
LiteCat +RL	**6.14**	7.27	**87**	**202**

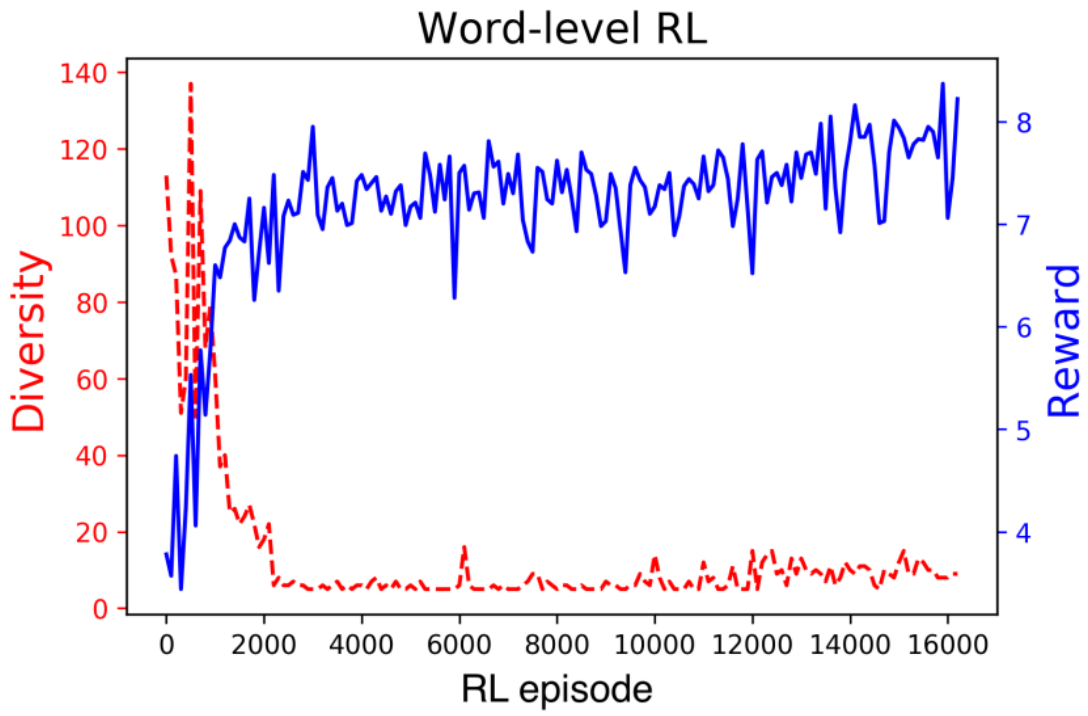
Table 2: Results on DealOrNoDeal. Diversity is measured by the number of unique responses the model used in all scenarios from the test data.

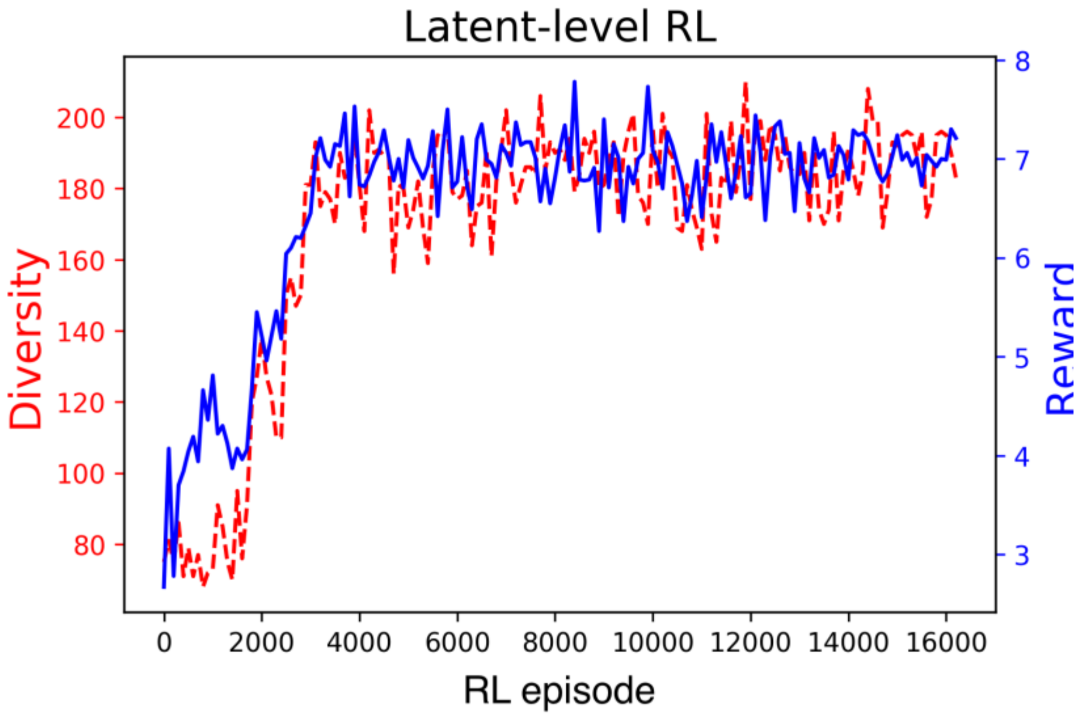
Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

RESULTS: DEAL OR NO DEAL

DealOrNoDeal is a **negotiation** dataset that contains **5805 dialogs** based **on 2236 unique scenarios**

252 scenarios for testing environment and randomly **sample 400 scenarios** from the training set for validation





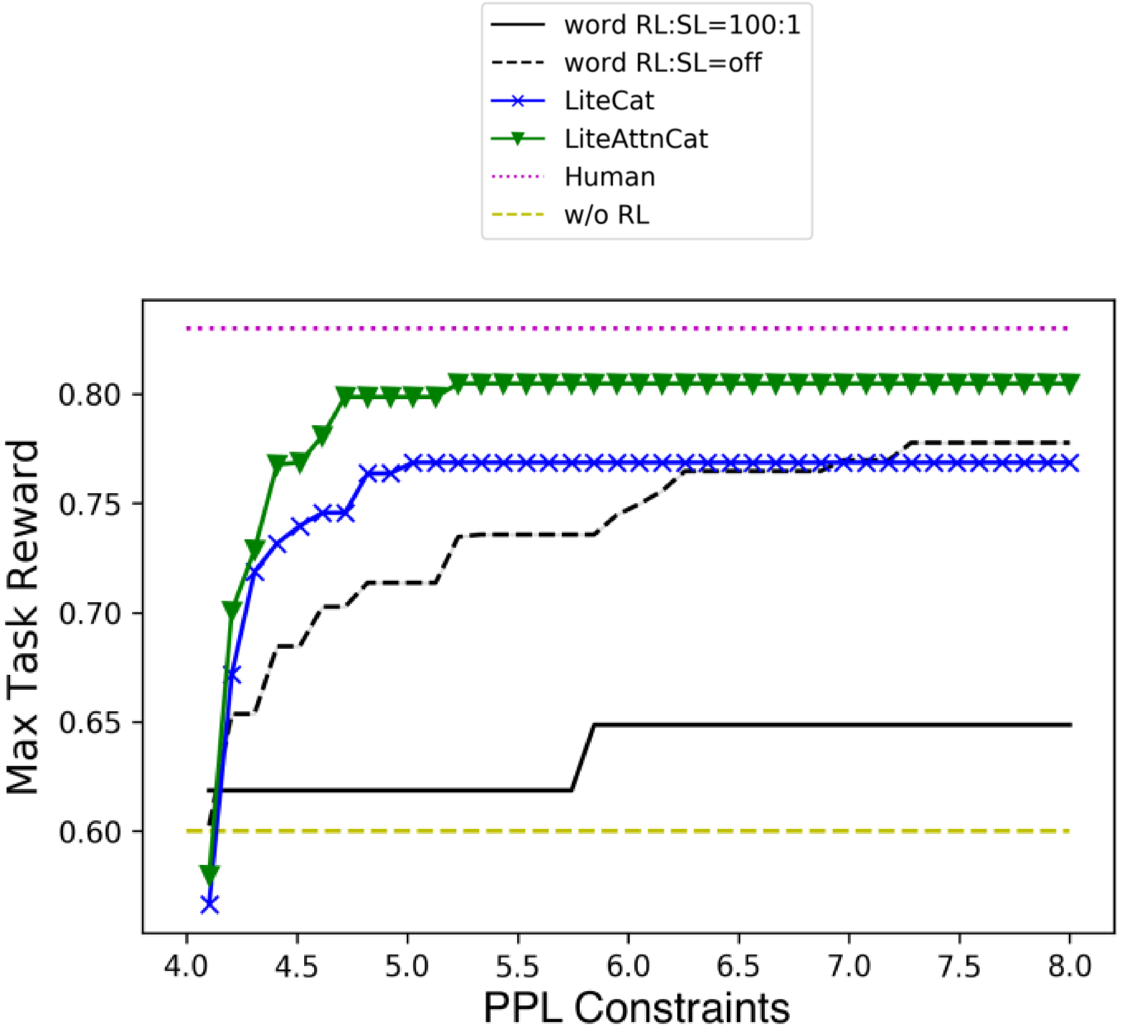
Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

UNIVERSITY OF WATERLOO

RESULTS: MULTI-WOZ

Multi-Woz is a slot-filling dataset that contains **10438 dialogs** on **6 different domains**. **8438 dialogs** are for **training** and **1000 each are for validation and testing**.

	PPL	BLEU	Inform	Success
Human	/	/	90%	82.3%
Baseline	3.98	18.9	71.33%	60.96%
LiteAttnCat	4.05	19.1	67.98%	57.36%
Baseline +RL	17.11	1.4	80.5%	79.07%
LiteAttnCat +RL	**5.22**	**12.8**	**82.78%**	**79.2%**

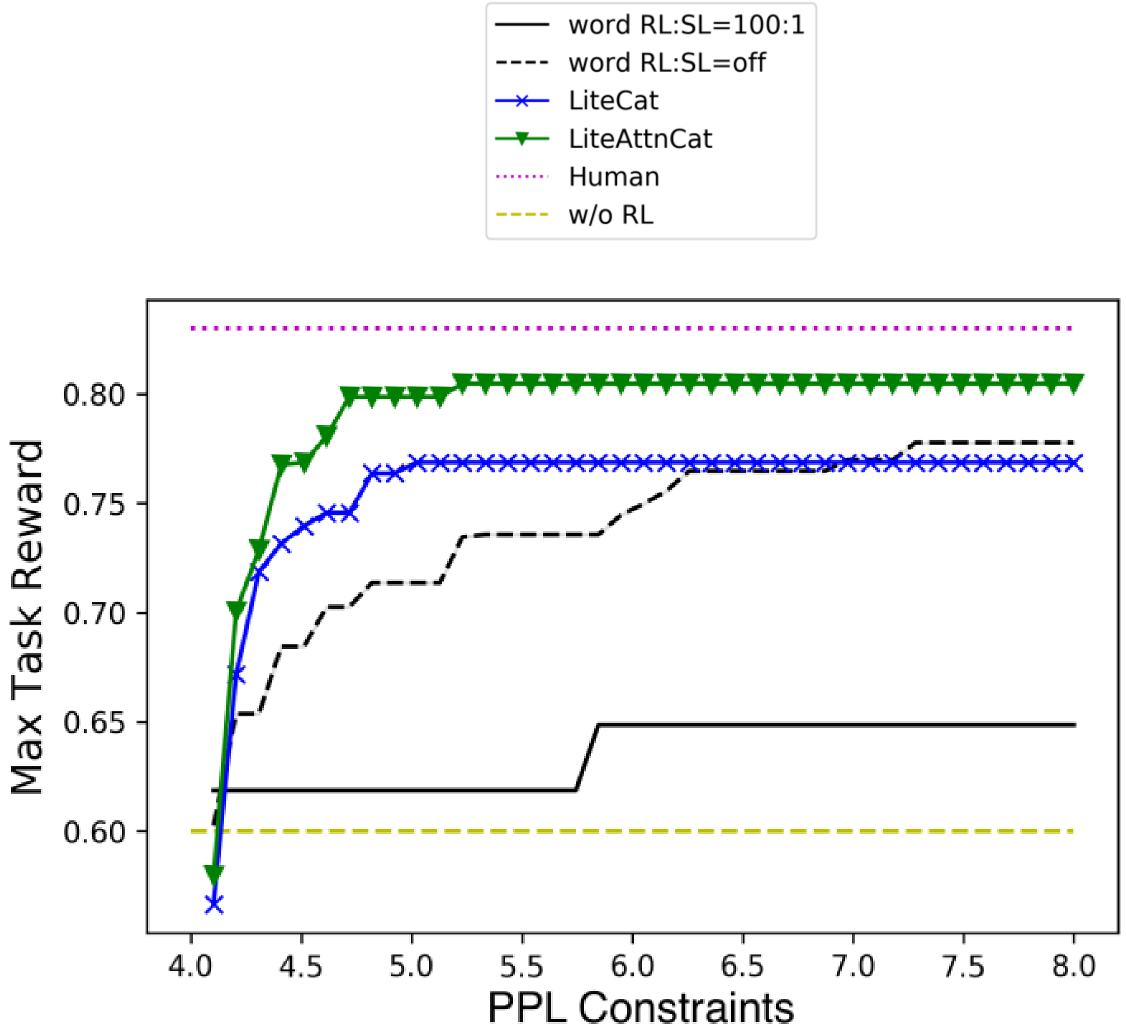


Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

RESULTS: MULTI-WOZ

Multi-Woz is a slot-filling dataset that contains **10438 dialogs** on **6 different domains**. **8438 dialogs** are for **training** and **1000 each are for validation and testing**.

Context	**Sys** I have [value_count] trains matching your request . Is there a specific day and time you would like to travel? **Usr** I would like to leave on [value_day] and arrive by [value_time].
Model	**Generated Response**
word RL:SL=off	[train_id] is leaving [value_place] on [value_day] on [value_day] on [train_id] [train_id] [value_count] [train_id] leaving ...
word RL:SL=100	[train_id] leaves at [value_time] . would you like me to book you a ticket ?
LiteAttnCat	[train_id] leaves [value_place] at [value_time] and arrives in [value_place] at [value_time]. Would you like me to book that for you ?



Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

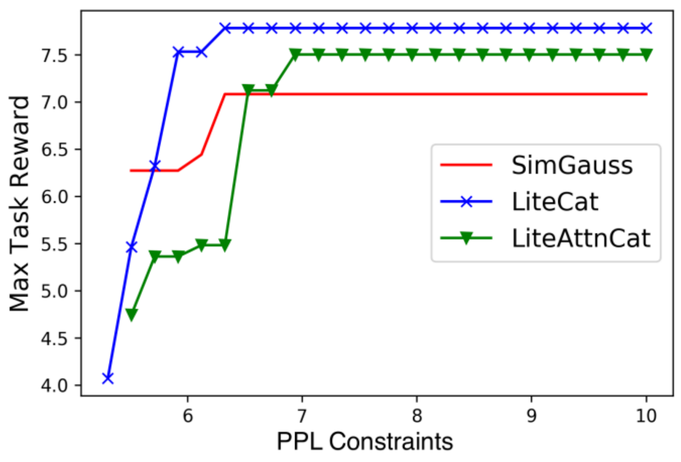
UNIVERSITY OF WATERLOO

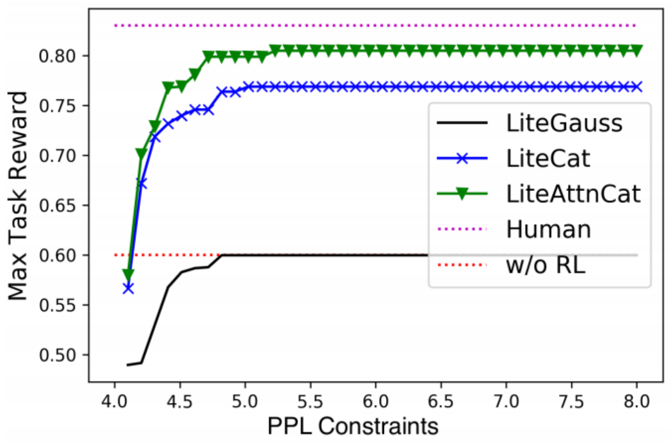
RESULTS

Deal	PPL	Reward	Agree%	Diversity
Baseline	3.23	3.75	59	109
Gauss	110K	2.71	43	176
LiteGauss	5.35	4.48	65	91
Cat	80.41	3.9	62	115
AttnCat	118.3	3.23	51	145
LiteCat	5.35	2.67	41	58
LiteAttnCat	5.25	3.69	52	75
MultiWoz	PPL	BLEU	Inform%	Succ%
Baseline	3.98	18.9	71.33	60.96
Gauss	712.3	7.54	60.5	23.0
LiteGauss	4.06	19.3	56.46	48.06
Cat	7.07	13.7	54.15	42.04
AttnCat	12.01	12.6	63.9	45.8
LiteCat	4.10	19.1	61.56	49.15
LiteAttnCat	4.05	19.1	67.97	57.36

β	0.0	0.01	β	0.0	0.01
LiteCat	4.23	7.27	LiteGauss	4.83	6.67

Table 6: Best rewards in test environments on DealOrNoDeal with various β.





Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

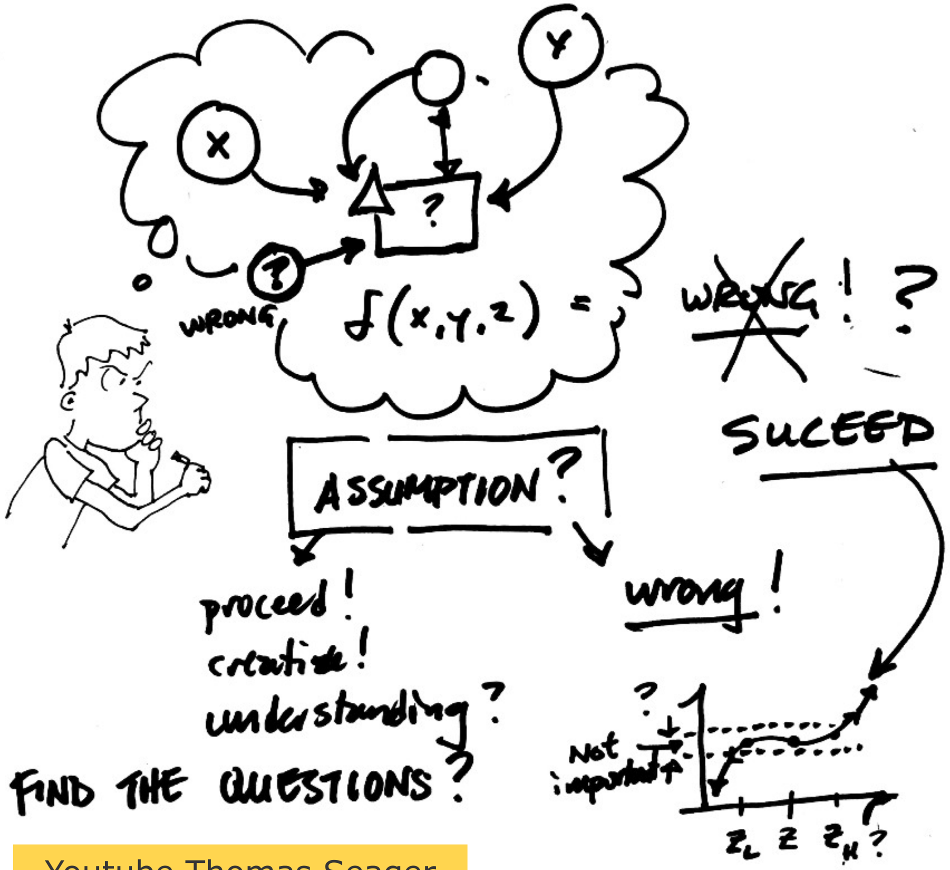
CONCLUSION

- Proposes a **latent action space** for RL in E2E dialog agents
- A regularized ELBO objective (**Exposure Bias**)
- **Attention Fusion** for discrete variables
- Create action abstraction in an **unsupervised** manner
- A new state-of-the-art success rate on **MultiWoz**

Ref:
1-Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, T. Zhao et. al.

Questions

HikingArtist.com



Youtube Thomas Seager

UNIVERSITY OF WATERLOO