Structured and Natural Responses Co-generation for Conversational Search

Chenchen Ye
National University of Singapore
chenchenye.ccye@gmail.com

Lizi Liao
Singapore Management University
liaolizi.llz@gmail.com

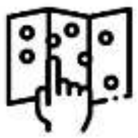
Fuli Feng
University of Science and Technology of China, fulifeng93@gmail.com

Wei Ji
Sea-NExT Joint Lab,
National University of Singapore
jiwei@nus.edu.sg

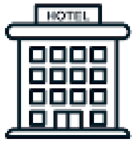
Tat-Seng Chua
Sea-NExT Joint Lab,
National University of Singapore
dcscts@nus.edu.sg

Introduction: Conversational Search Systems

- **Real-life applications**: personal assistants and industrial chatbots
- Such systems aim to facilitate people with services through natural language conversations



attraction recommendation



hotel booking

restaurant reservation

product search in e-commerce

Can you see if there are any **entertainment attractions** in the **centre of town**?

Unfortunately, there are **no entertainment attractions in the centre of town**. Is there a **different area** of town that would work for you?

What about a **museum**?

Yes, there are **11 museums**. I 'd **recommend** Broughton House Gallery. It 's at 98 King Street and has free entrance. Would that work for you ?

Yes, that will work great.

Introduction: Conversational Search Systems

- **Ultimate goals** for a conversational search system:
  - to generate fluent and informative natural responses to the end user
  - to maintain structured internal states such as dialogue acts for search optimization

Can you see if there are any entertainment attractions in the centre of town?

internal search states

Unfortunately, there are no entertainment attractions in the centre of town. Is there a different area of town that would work for you?

Attraction – No offer – Area: *centre of town*
Attraction – No offer – Type: *entertainment*
Attraction – Request – Area: *?*

What about a museum?

Yes, there are 11 museums. I 'd recommend Broughton House Gallery. It 's at 98 King Street and has free entrance. Would that work for you ?

Attraction – Inform – Choice: *11*
Attraction – Inform – Type: *museums*
Attraction – Inform – Name: *Broughton House Gallery*
Attraction – Inform – Address: *98 King street*
Attraction – Inform – Price: *free*

Yes, that will work great.

Natural Responses **Structured Responses**

Structured and Natural Responses Co-Generation (Co-Gen)

Introduction: Conversational Search Systems

- **Ultimate goals** for a conversational search system:
  - to generate fluent and informative natural responses to the end user
  - to maintain structured internal states such as dialogue acts for search optimization

Can you see if there are any entertainment attractions in the centre of town?

internal search states

Why Co-Generation?

What about a museum?

Yes, there are 11 museums. I 'd recommend Broughton House Gallery. It 's at 98 King Street and has free entrance. Would that work for you ?

Attraction – Inform – Choice: *11*
Attraction – Inform – Type: *museums*
Attraction – Inform – Name: *Broughton House Gallery*
Attraction – Inform – Address: *98 King street*
Attraction – Inform – Price: *free*

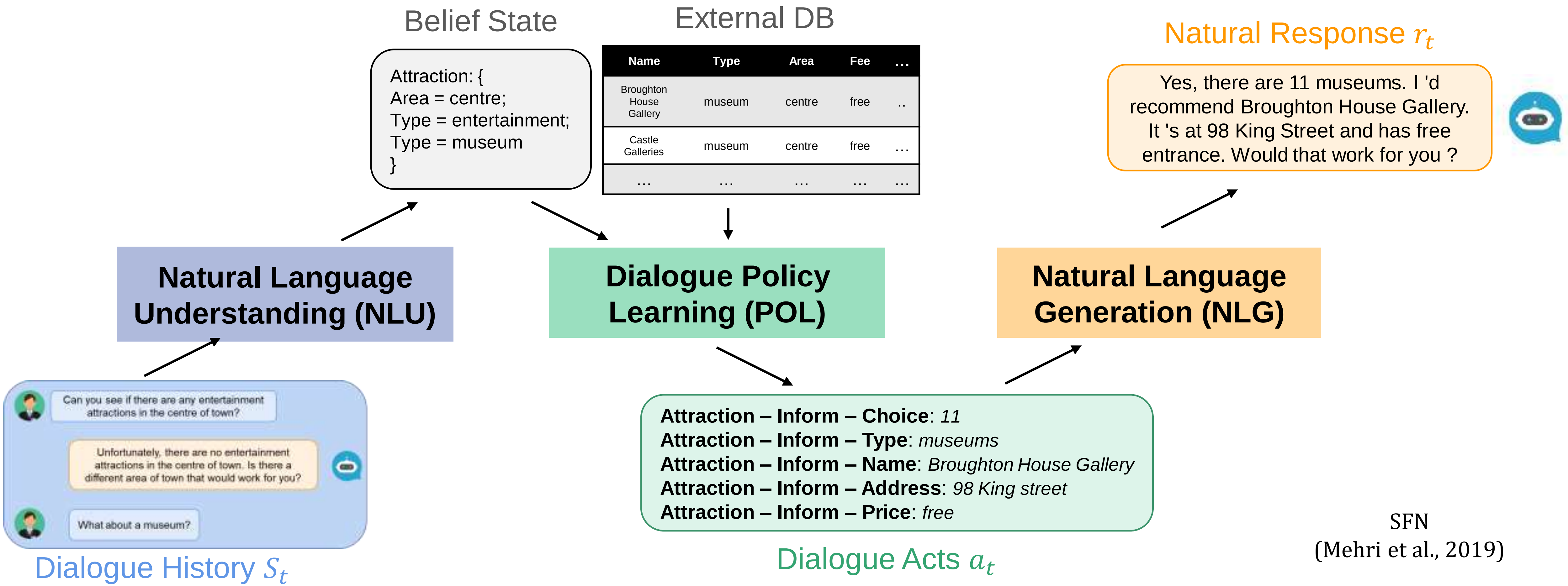
Yes, that will work great.

Natural Responses **Structured Responses**

Structured and Natural Responses Co-Generation (Co-Gen)

Existing methods: Pipelined methods (modular)

- **Traditional methods: a pipeline with several separated modules.**

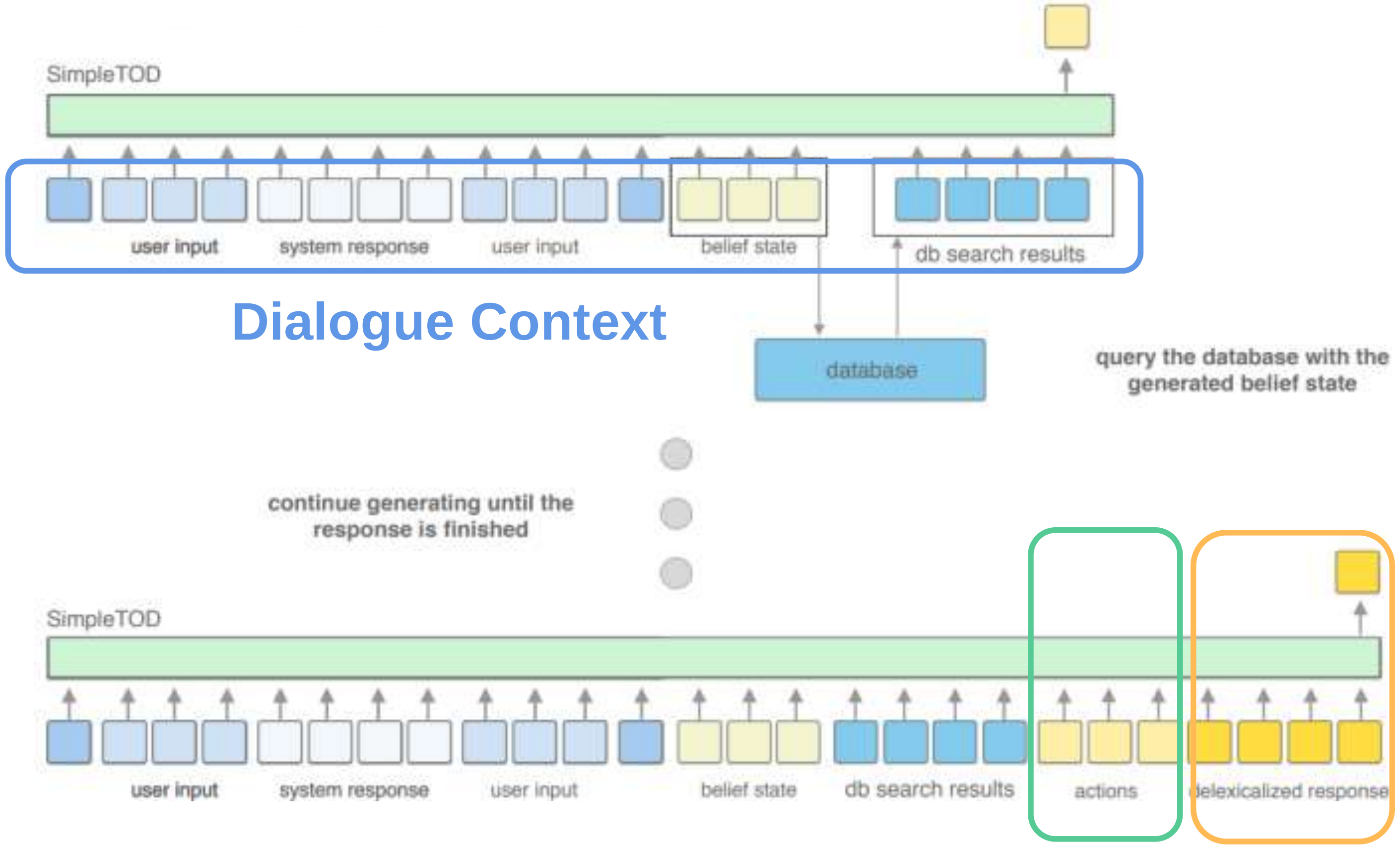


SFN
(Mehri et al., 2019)

- **Information loss**: e.g., utterances in history are no longer available when generating natural responses
- **Error propagation**: e.g., error occurs in dialogue acts prediction may mislead following response generation

Existing methods: Pipelined methods (language modeling)

- **Language modeling methods:** benefit from **pretrained transformer-based models.**
  - e.g., BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019)
  - connect dialogue history, states, dialogue acts, and responses to a long sequence

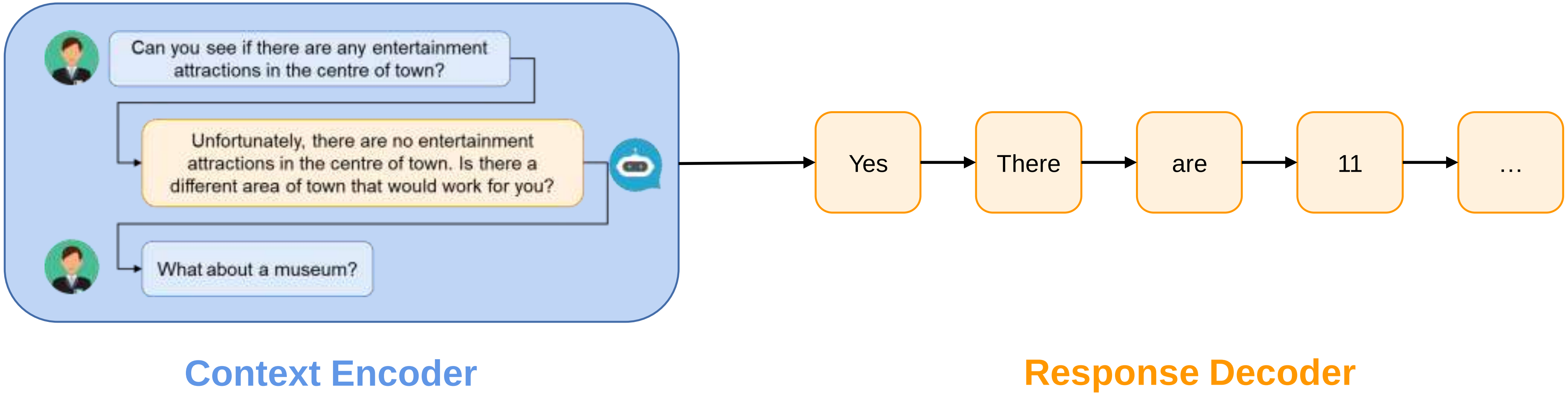


SimpleTOD
(Hosseini-Asl et al., 2020)

- **Error propagation** still exists from dialogue acts to response
- **Uni-direction**: the model is unable to foresee the future of the conversation.
  e.g., The reflections from the latter-generated responses are unable to benefit the former-generated acts.

Existing methods: End-to-end methods

- **Sequence-to-sequence models**
  - learn **the mapping from dialogue context to natural response** in an end-to-end manner
  - from basic neural recurrent models e.g., RNN, Bi-LSTM to advanced ones e.g., DialoGPT (Zhang et al,. 2020)



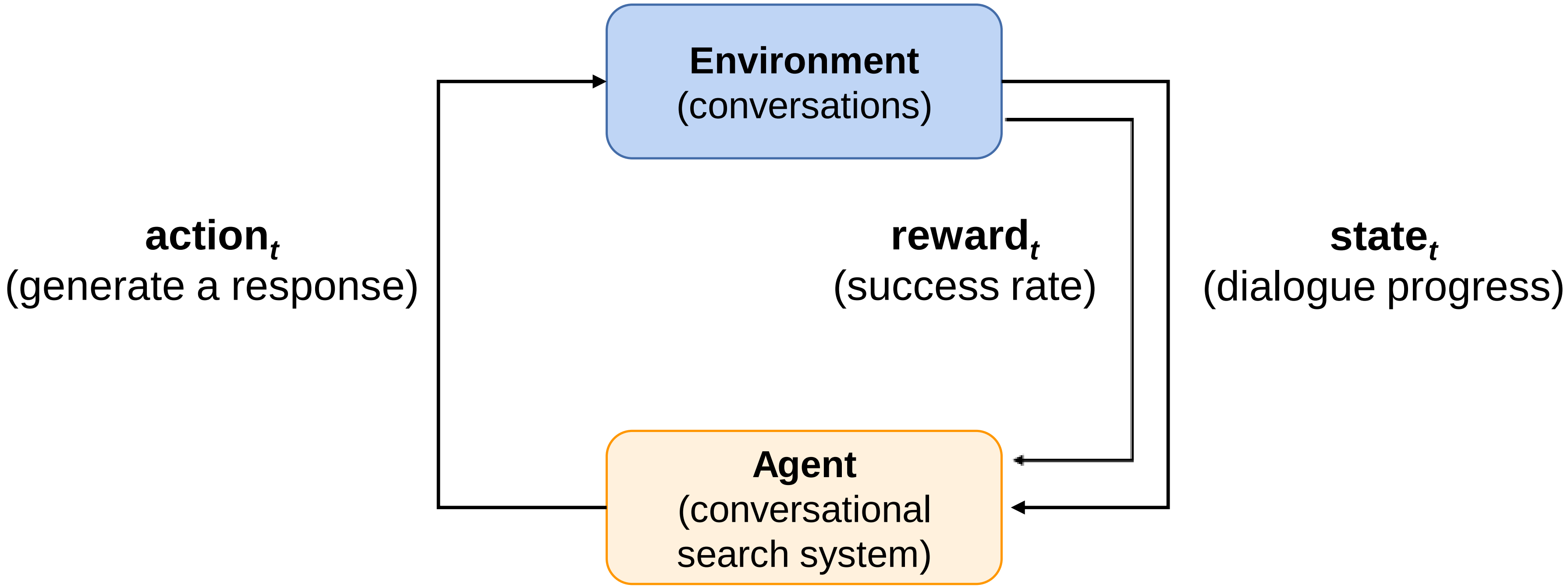
- **Omit dialogue acts**: fail to maintain search states, and this usually results in **low <u>inform rate</u>** & **<u>success rate</u>**.

<u>Inform Rate</u>: measures whether <u>correct entities</u> has been provided (e.g., correct museum)

<u>Success Rate</u>: measures whether the <u>requested information</u> has been fulfilled (e.g., the address of the museum)

Existing methods: End-to-end methods (RL–applied)

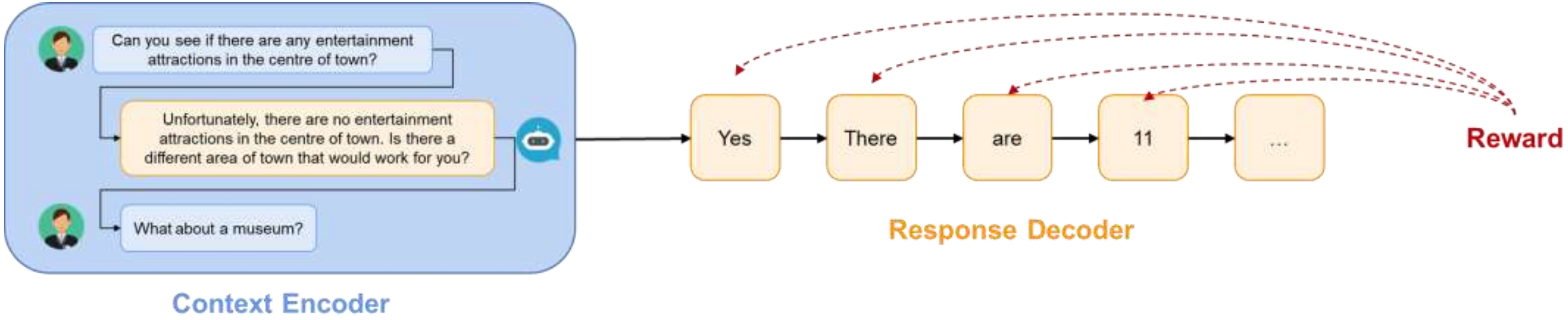
- **Reinforcement learning (RL): finetune Supervised learning (SL)-pretrained model**
  - **RL** (Williams, 1992) is a general-purpose framework for decision-making
  - **Goal**: let an agent to select actions that maximize the future rewards received from the environment



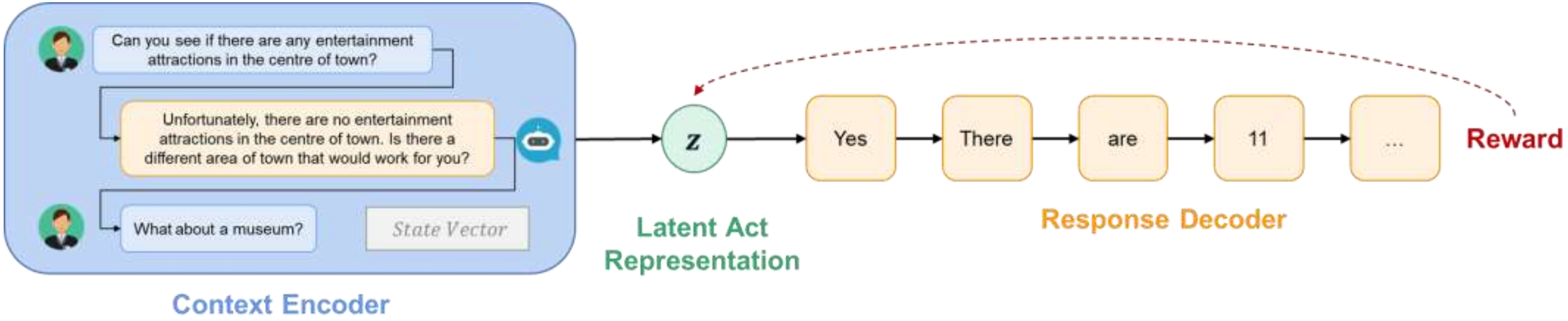
  - Use **success rate** as the **reward** to optimize the model **towards task completion**

Existing methods: End-to-end methods (RL–applied)

- **Word-level RL** (Li et al., 2016)
- **Large action space** (whole vocabulary) **and long trajectory**



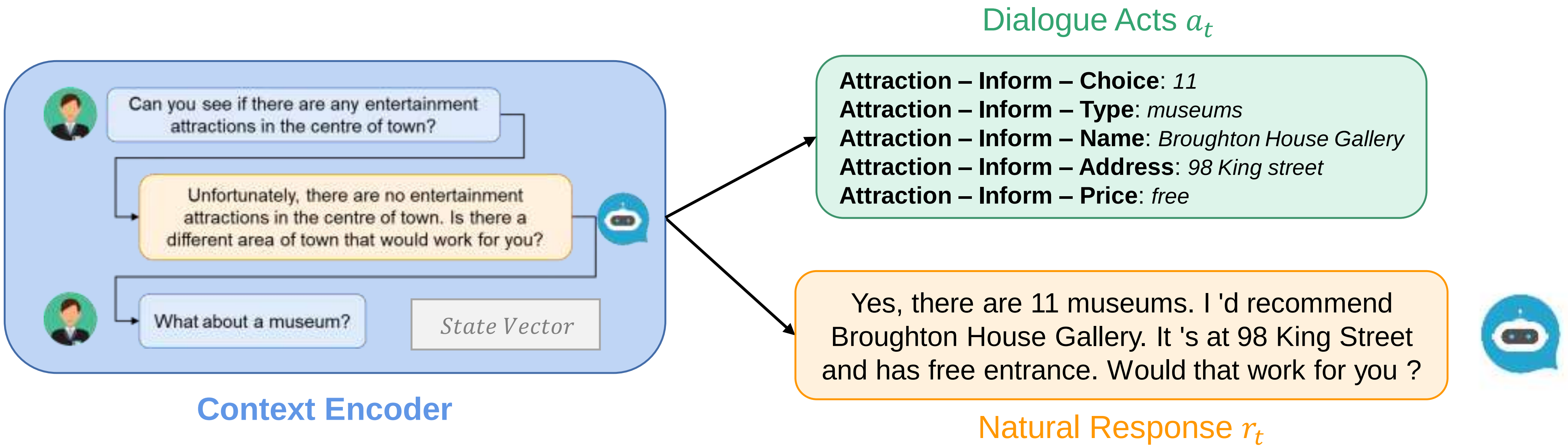
- **Latent Act RL: variational autoencoder (VAE)** -LaRL, LAVA, HDNO (Zhao et al., 2019; Lubis et al., 2020; Wang et al., 2020)



- **Explicitly ignore the structured dialogue acts**
- Hinder search optimization, e.g., query execution or error debugging

Co-Generation

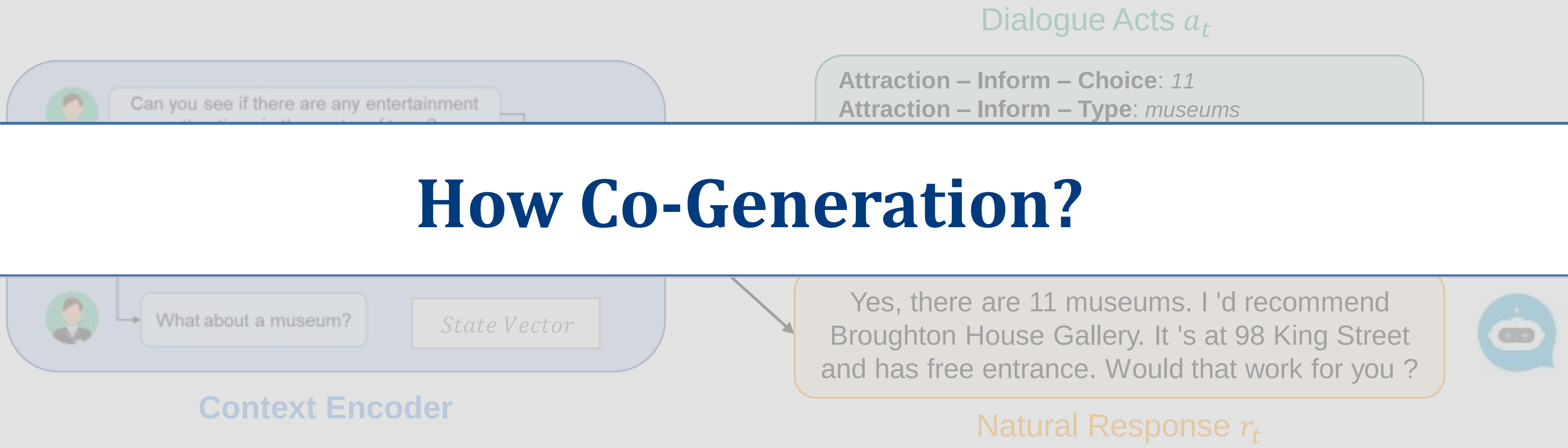
- Treat structured dialogue act as **another sequence generation task**
- Combine it with natural response generation in **multitask learning** setting



- **Obtain structured dialogue acts** as internal search states
- **Reduce error propagation** on natural response generation from dialogue acts
- **Decease information loss** from context to both responses

Co-Generation

- Treat structured dialogue act as **another sequence generation task**
- Combine it with natural response generation in **multitask learning** setting

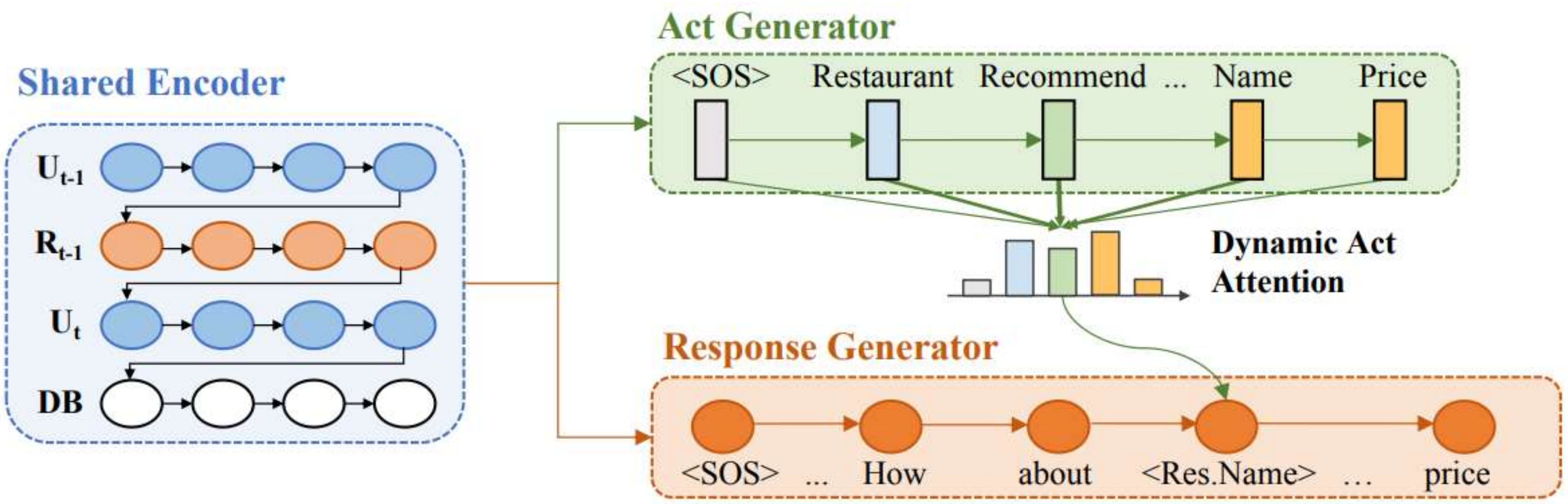


How Co-Generation?

- **Obtain structured dialogue acts** as internal search states
- **Reduce error propagation** on natural response generation from dialogue acts
- **Decease information loss** from context to both responses

Co-Generation: MarCo

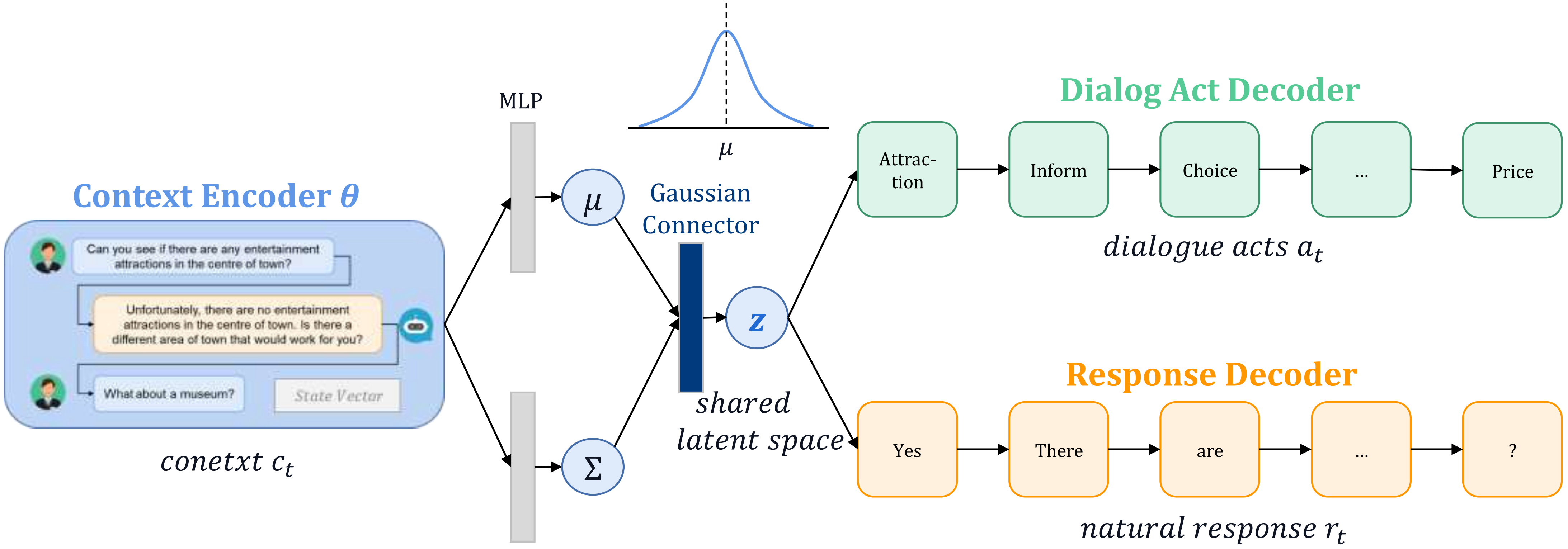
- MarCo (Wang et al., 2020): uses dynamic attentions from dialogue acts to natural response



- **Interrelationships** between two responses are **only modeled locally**
- **Hard to realize the required synchronization** between the two decoder branches due to different sequence lengths

Co-Generation: Co-Gen (Our solution)

- **Training Scheme**: 1) SL Pretraining 2)RL Finetuning
- In the **main stream**: Encode context: $s_t = \text{Bi-GRU}(\ S_t\ ),\ c_t = [s_t; d_t]$
- We assume the latent variable $\boldsymbol{z}$ is sampled from a multivariate **Gaussian distribution**: $p_\Theta(z|c) = \mathbb{N}(z|\mu, \Sigma)$
- **Two decoder branches from $\boldsymbol{z}$**, for which we also use GRU cells: $\boldsymbol{c} \rightarrow \boldsymbol{z} \rightarrow \boldsymbol{a}$ $\boldsymbol{c} \rightarrow \boldsymbol{z} \rightarrow \boldsymbol{r}$



Co-Generation: Co-Gen (Our solution)

- **Training Scheme**: 1) SL Pretraining 2)RL Finetuning
- Two **auxiliary streams**: two **VAE tasks** for a and r
- For example, on **the Act VAE stream**: $a \rightarrow z^a \rightarrow a$ **to regularize** $c \rightarrow z \rightarrow a$

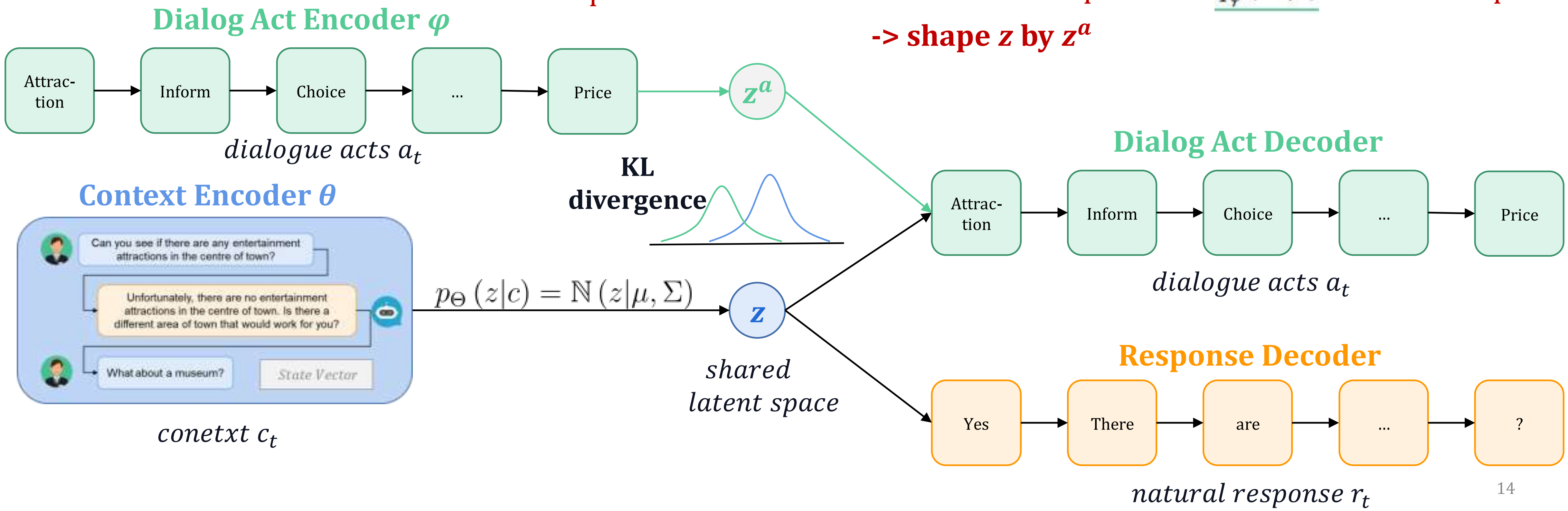
$$L_{VAE}(\psi) = \mathbb{E}_{q_\psi(z^a|a)}[\log p_\psi(a|z^a)] - D_{KL}[\underline{q_\psi(z^a|a)}||p(z^a)],$$

$$L_{lite\_A}(\theta) = p_{p_\theta(z|c)}(a|z) - \beta \cdot D_{KL}[p_\theta(z|c)||\underline{q_\psi(z^a|a)}].$$

posterior

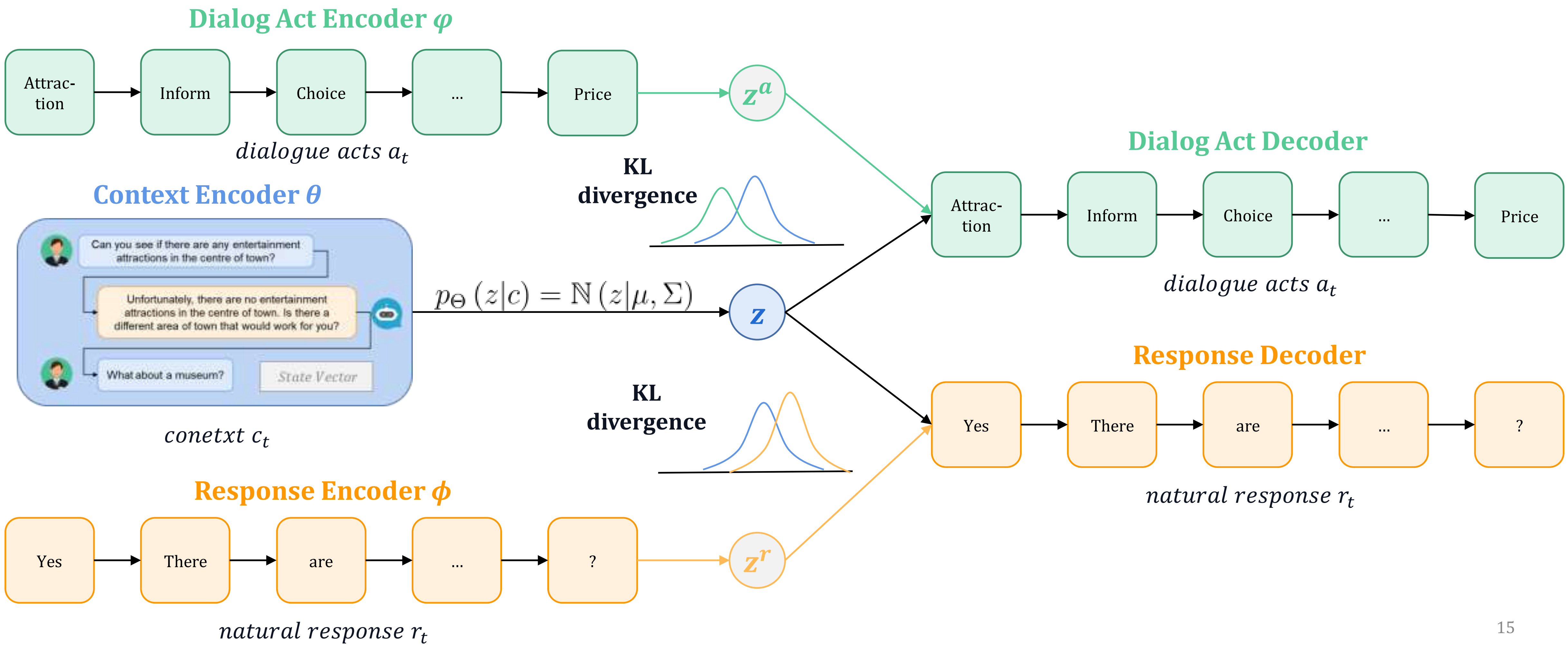
Use the learned posterior $q_\psi(z^a|a)$ as informed prior

-> shape z by z^a



Co-Generation: Co-Gen (Our solution)

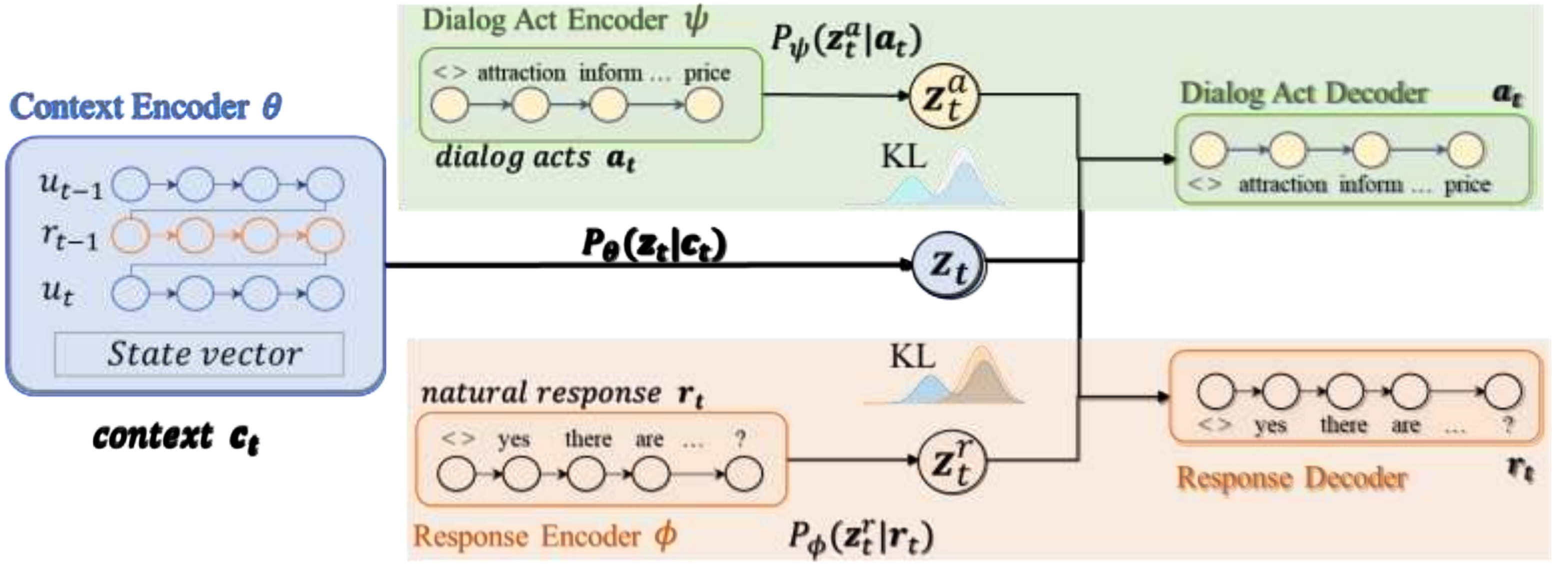
- **Training Scheme:** 1) SL Pretraining 2)RL Finetuning
- Two **auxiliary streams**: two **VAE tasks** for a and r $a \rightarrow z^a \rightarrow a$ $r \rightarrow z^r \rightarrow r$



Co-Generation: Co-Gen (Our solution)

- **Training Scheme**: 1) SL Pretraining 2)RL Finetuning
- **Main stream** from context and two **auxiliary streams** for a and r

$a \to z^a \to a$



$c \to z \to a$

$c \to z \to r$

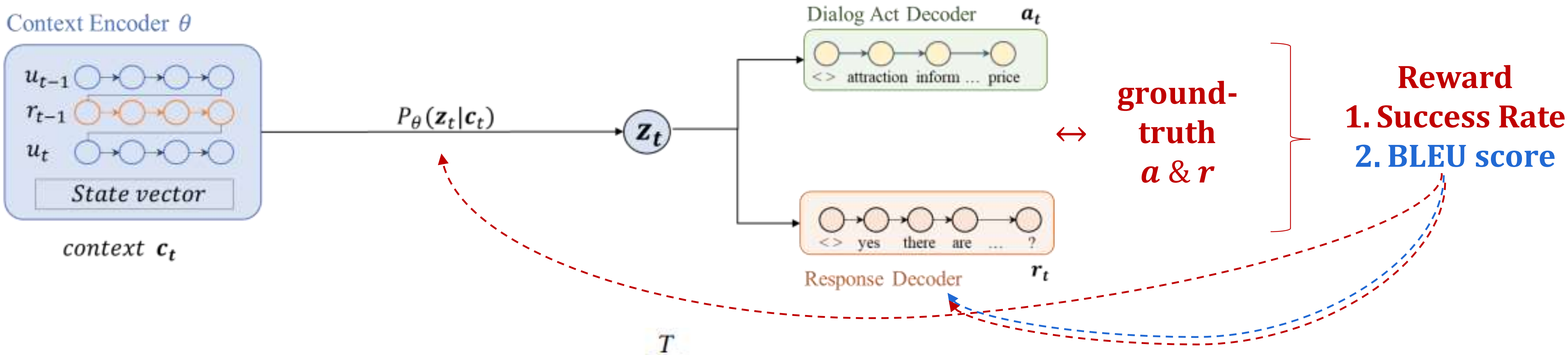
$r \to z^r \to r$

- We want the VAEs to capture **global generative factors** from:
  - **Dialogue acts**: e.g., intent and domain information
  - **Natural responses**: e.g., useful utterance patterns
- **Global semantic associations**

Co-Generation: Co-Gen (Our solution)

- **Training Scheme:** 1) SL Pretraining 2)RL Finetuning

- **Asynchronous RL:** updates happen for **the encoder parameters** and **the decoder parameters** in turn



- **On the shared latent space z:** $\nabla J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{T} O_t \nabla_\pi \log p_\pi(z_t|c_t)].$

calculated by success rate: whether the requested information has been fulfilled

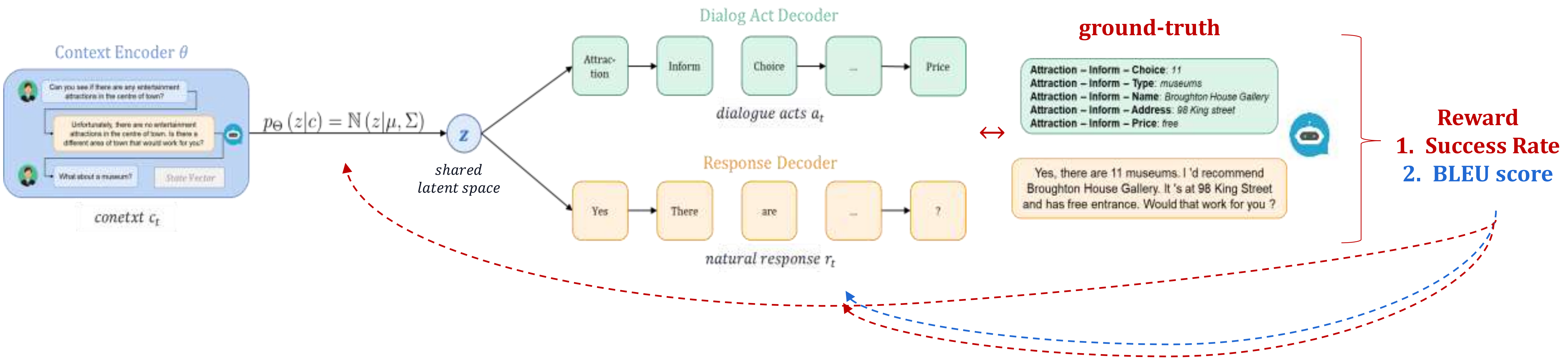
- **On the decoder part:** $\nabla J(\mathrm{G}) = \mathbb{E}_\mathrm{G}[\sum_{t=0}^{T} \sum_{j=0}^{|r_t|} \mathcal{R}_{tj} \nabla \log \mathrm{G}(\mathbf{w}_{tj}|\mathbf{w}_{<tj}, z_t, c_t)],$

further incorporate a weighted BLEU score: n-gram similarity, the fluency of generated responses

(Papineni et al., 2002)

Co-Generation: Co-Gen (Our solution)

- **Training Scheme:** 1) SL Pretraining 2)RL Finetuning



- **Foresee the future** of the conversation
- **Optimize towards both task completion** and **language naturalness**
- BLEU score as additional reward
  - it is less likely that the decoder will be overwhelmed by task completion metrics

Experiment Results: Co-Gen (Our solution)

- Results on **natural language responses**
- Datasets: MultiWoz 2.0 (Budzianowski et al., 2018) and MultiWoz 2.1 (Eric et al., 2019): > 10,000 dialogues & 7 domains
- **Combine Score** = ***BLEU + 0.5 * (Inform + Success)*** (Budzianowski et al., 2018)

Group	Method	MultiWoz 2.0				MultiWoz 2.1			
		Inform	Success	BLEU	Score	Inform	Success	BLEU	Score
Pipe-lined	SFN_SL	90.00	74.20	18.35	100.45	63.10	53.10	17.56	75.66
	SFN	94.40	83.10	16.34	105.09	87.80	76.20	10.57	92.57
	UBAR	94.00	83.60	17.22	106.02	89.6	78.6	17.34	101.44
	HDSA	82.90	68.90	23.60	99.50	86.30	70.60	22.36	100.81
End-to-End	DialoGPT	73.40	48.00	12.16	72.86	72.10	50.10	12.62	73.72
	LaRL	93.49	84.98	12.01	101.25	92.39	85.29	13.72	102.56
	LAVA	97.50	94.80	12.02	108.17	96.39	83.57	14.02	104.00
	HDNO_SL	78.60	70.40	19.26	93.76	78.80	66.70	18.46	91.21
	HDNO	95.80	84.50	18.61	108.76	93.20	81.90	18.35	105.90
Co-Generate	MarCo	92.30	78.60	20.02	105.47	92.50	77.80	19.54	104.69
	Co-Gen_SL	92.10	77.40	20.91	105.66	88.90	80.00	20.67	105.12
	Co-Gen (ours)	94.70	86.70	20.42	**111.12**	91.20	85.20	19.09	**107.29**

- Co-Gen achieves the **best performance regarding the overall performance**
  - **leads the board on the combine score** in the official records[1]

1. https://github.com/budzianowski/multiwoz

Experiment Results: Co-Gen (Our solution)

- Results on **natural language responses**
- Datasets: MultiWoz 2.0 (Budzianowski et al., 2018) and MultiWoz 2.1 (Eric et al., 2019): > 10,000 dialogues & 7 domains
- Combine Score = ***BLEU + 0.5 * (Inform + Success)*** (Budzianowski et al., 2018)

Group	Method	MultiWoz 2.0				MultiWoz 2.1			
		Inform	Success	BLEU	Score	Inform	Success	BLEU	Score
Pipe-lined	SFN_{SL}	90.00	74.20	18.35	100.45	63.10	53.10	17.56	75.66
	SFN	94.40	83.10	16.34	105.09	87.80	76.20	10.57	92.57
	UBAR	94.00	83.60	17.22	106.02	89.6	78.6	17.34	101.44
	HDSA	82.90	68.90	23.60	99.50	86.30	70.60	22.36	100.81
End-to-End	DialoGPT	73.40	48.00	12.16	72.86	72.10	50.10	12.62	73.72
	LaRL	93.49	84.98	12.01	101.25	92.39	85.29	13.72	102.56
	LAVA	97.50	94.80	12.02	108.17	96.39	83.57	14.02	104.00
	$HDNO_{SL}$	78.60	70.40	19.26	93.76	78.80	66.70	18.46	91.21
	HDNO	95.80	84.50	18.61	108.76	93.20	81.90	18.35	105.90
Co-Generate	MarCo	92.30	78.60	20.02	105.47	92.50	77.80	19.54	104.69
	$Co\text{-}Gen_{SL}$	92.10	77.40	20.91	105.66	88.90	80.00	20.67	105.12
	Co-Gen (ours)	94.70	86.70	20.42	**111.12**	91.20	85.20	19.09	**107.29**

- Methods **without RL** process:
  - They obtain relatively lower results especially regarding task completion metrics.

Experiment Results: Co-Gen (Our solution)

- Results on **natural language responses**
- Datasets: MultiWoz 2.0 (Budzianowski et al., 2018) and MultiWoz 2.1 (Eric et al., 2019): > 10,000 dialogues & 7 domains
- **Combine Score** = ***BLEU + 0.5 * (Inform + Success)*** (Budzianowski et al., 2018)

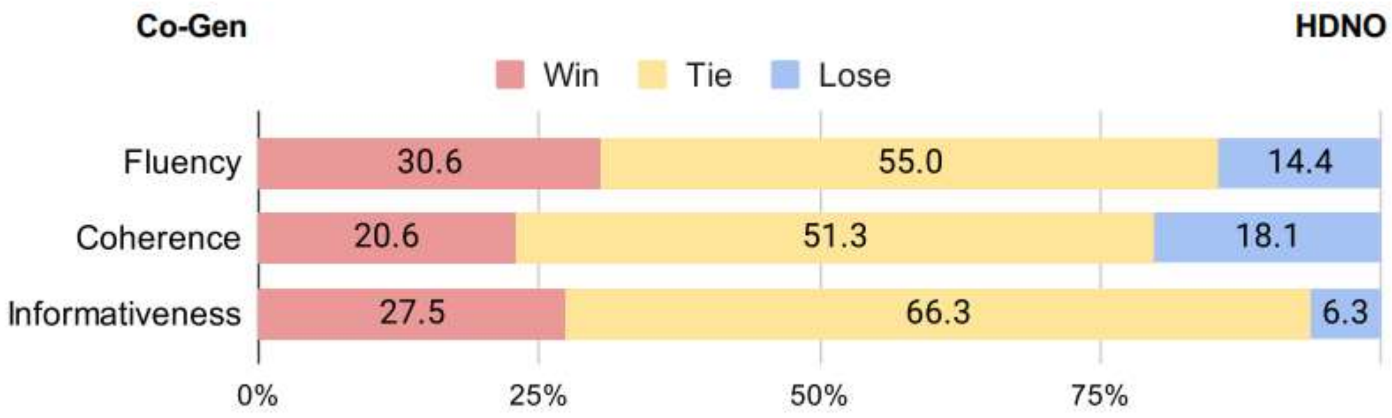
Group	Method	MultiWoz 2.0				MultiWoz 2.1			
		Inform	**Success**	**BLEU**	**Score**	**Inform**	**Success**	**BLEU**	**Score**
Pipe-lined	SFN_$_{SL}$	90.00	74.20	18.35	100.45	63.10	53.10	17.56	75.66
	SFN	94.40	83.10	16.34	105.09	87.80	76.20	10.57	92.57
	UBAR	94.00	83.60	17.22	106.02	89.6	78.6	17.34	101.44
	HDSA	82.90	68.90	23.60	99.50	86.30	70.60	22.36	100.81
End-to-End	DialoGPT	73.40	48.00	12.16	72.86	72.10	50.10	12.62	73.72
	LaRL	93.49	84.98	12.01	101.25	92.39	85.29	13.72	102.56
	LAVA	97.50	94.80	12.02	108.17	96.39	83.57	14.02	104.00
	HDNO_$_{SL}$	78.60	70.40	19.26	93.76	78.80	66.70	18.46	91.21
	HDNO	95.80	84.50	18.61	108.76	93.20	81.90	18.35	105.90
Co-Generate	MarCo	92.30	78.60	20.02	105.47	92.50	77.80	19.54	104.69
	Co-Gen_$_{SL}$	92.10	77.40	20.91	105.66	88.90	80.00	20.67	105.12
	Co-Gen (ours)	94.70	86.70	20.42	**111.12**	91.20	85.20	19.09	**107.29**

- Compared with other **RL-applied** methods:
  - Co-Gen achieves a **more balance results**, with considerable task completion and good BLEU score

Experiment Results: Co-Gen (Our solution)

- Results on **natural language responses**

- **Human evaluation** compared with HDNO (Wang et al., 2020)



Experiment Results: Co-Gen (Our solution)

- Results on dialogue acts as **structured responses**

- **Act F1**: measures **act coverage accuracy** as (domain – action – slot) tuples (Wang et al., 2020, Chen et al., 2019)
- **Entity F1**: measures **entity coverage accuracy** that appears in generated response (Wen et al., 2017)
  - eg., true restaurant name, or its placeholder in delexicalized results

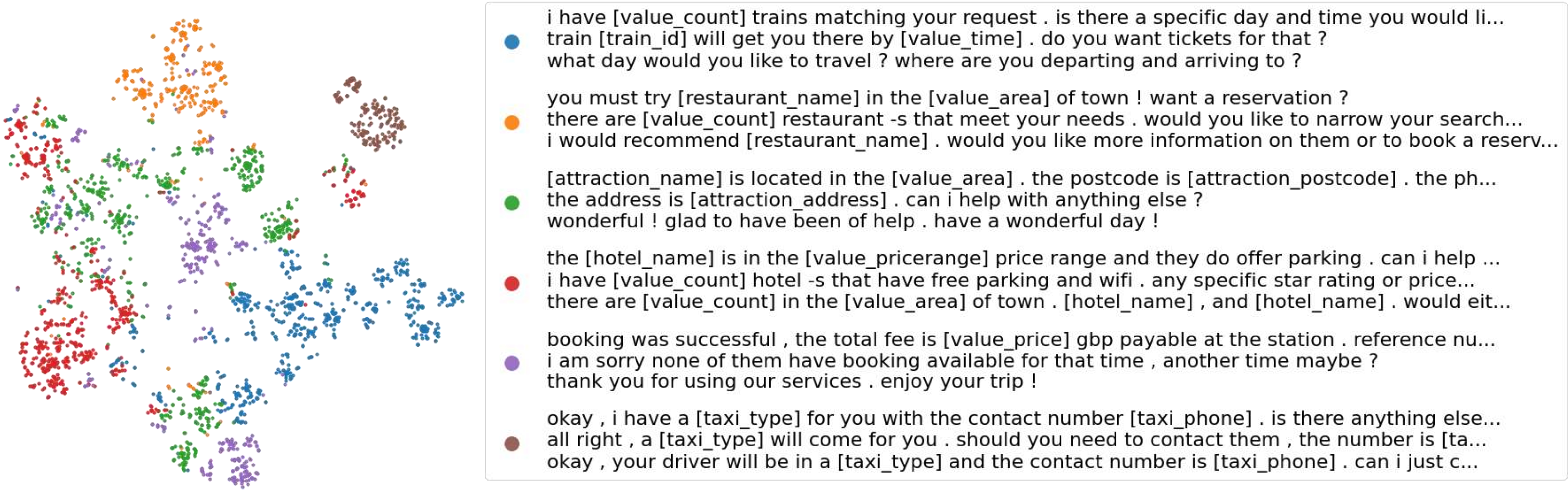
Group	Method	Act F1	Entity F1
Act Prediction Only (Chen et al., 2019)	BiLSTM	71.4	NA
	Word-CNN	71.5	NA
	Transformer	73.1	NA
Pipe-lined	SFN	63.7	77.1
	UBAR	84.6	82.3 [2]
	HDSA	77.3	65.7
Co-Generate	MarCo	73.9	59.9
	Co-Gen	**87.6**	**77.2**

- Co-Gen obtains **more accurate structured responses** as internal search states

2. The evaluation script of UBAR does not differentiate entity domains hence largely inflates the result. This number is thus not comparable to others.

Experiment Results: Co-Gen (Our solution)

- **Semantic meanings** of the **shared latent space *z***
  - cluster the latent vectors in testing dataset
  - project them with t-SNE graph



Summary:

Structured and Natural Responses Co-Generation for Conversational Search

- generating informative and fluent natural responses
- maintaining structured act states for search optimization

- **SL**: auxiliary tasks are used for regularization of the co-generation
- **RL** : towards both search task completion and language fluency

- Thank you for your listening!
- Q & A (Email: chenchenye.ccye@gmail.com)