Code & Data: github.com/snakeztc/NeuralDialog-LAED

Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation

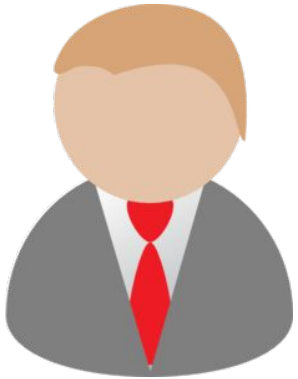
Tiancheng Zhao, Kyusong Lee and Maxine Eskenazi

Language Technologies Institute, Carnegie Mellon University

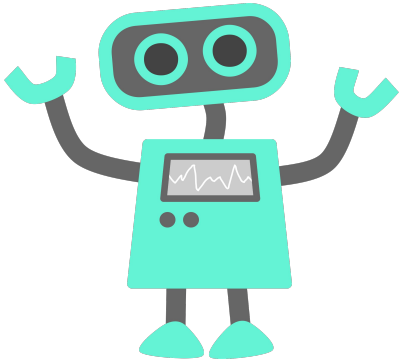
Language Technologies Institute

Carnegie Mellon University

Sentence Representation in Conversations



- Traditional System: **hand-crafted** semantic frame
  - [**Inform** location=Pittsburgh, time=now]
  - Not scalable to complex domains



- Neural dialog models: **continuous** hidden vectors
  - Directly output system responses in words
  - Hard to interpret & control

[Ritter et al 2011, Vinyals et al 2015, Serban et al 2016, Wen et al 2016, Zhao et al 2017]

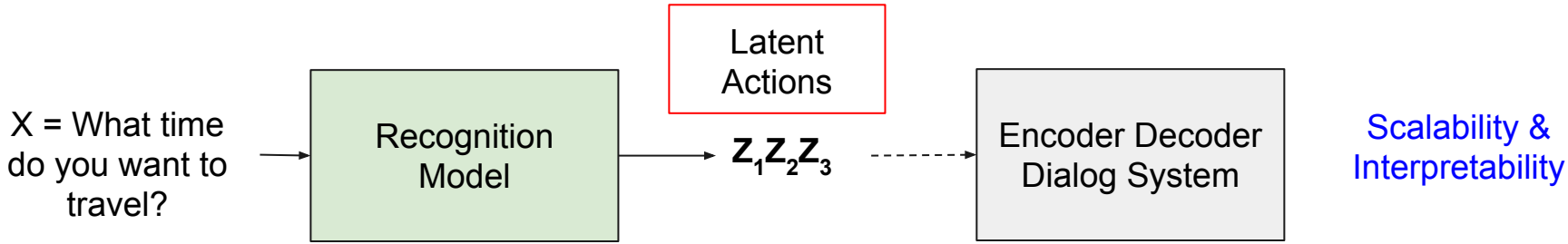
Why discrete sentence representation?

1. Inrepteablity & controbility & multimodal distribution
2. Semi-supervised Learning [Kingma et al 2014 NIPS, Zhou et al 2017 ACL]
3. Reinforcement Learning [Wen et al 2017]

Why discrete sentence representation?

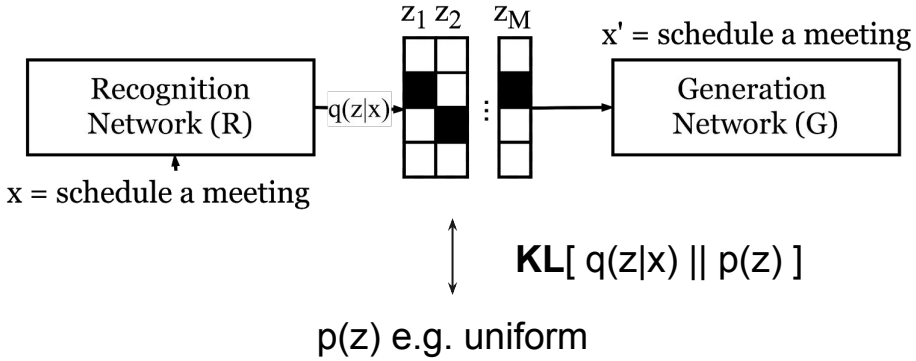
1. Inrepteablity & controbility & multimodal distribution
2. Semi-supervised Learning [Kingma et al 2014 NIPS, Zhou et al 2017 ACL]
3. Reinforcement Learning [Wen et al 2017]

Our goal:



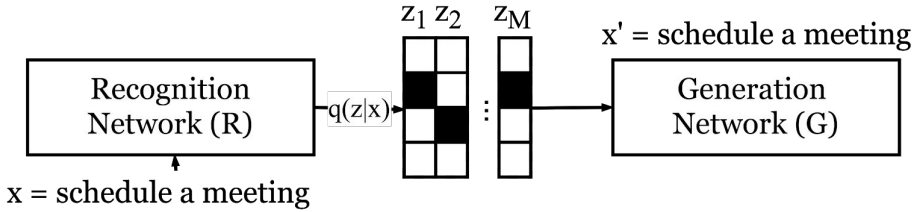
Baseline: Discrete Variational Autoencoder (VAE)

- **M** discrete **K**-way latent variables **z** with RNN recognition & generation network.
- Reparametrization using **Gumbel-Softmax** [Jang et al., 2016; Maddison et al., 2016]



Baseline: Discrete Variational Autoencoder (VAE)

- **M** discrete **K**-way latent variables **z** with GRU encoder & decoder.
- Reparametrization using **Gumbel-Softmax** [Jang et al., 2016; Maddison et al., 2016]



- **FAIL** to learn meaningful **z** because of **posterior collapse** (**z** is constant regardless of x)
- MANY prior solution on continuous VAE, e.g. (not exhaustive), yet still open-ended question
  - KL-annealing, decoder word dropout [Bowman et a2015] Bag-of-word loss [Zhao et al 2017] Dilated CNN decoder [Yang, et al 2017] Wake-sleep [Shen et al 2017]

Anti-Info Nature in Evidence Lower Bound (ELBO)

- Write ELBO as an expectation over the whole dataset

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})}[\log p_{\mathcal{G}}(\mathbf{x}|\mathbf{z})] - \text{KL}(q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] \quad (1)$$

Anti-Info Nature in Evidence Lower Bound (ELBO)

- Write ELBO as an expectation over the whole dataset

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})}[\log p_{\mathcal{G}}(\mathbf{x}|\mathbf{z})] - \text{KL}(q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] \quad (1)$$

- Expand the KL term, and plug back in:

$$\mathbb{E}_{\mathbf{x}}[\text{KL}(q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] = I(Z, X) + \text{KL}(q(\mathbf{z})||p(\mathbf{z})) \quad (2)$$

Maximize ELBO
→ Minimize **I(Z, X)** to 0
→ Posterior collapse with powerful decoder.

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})p(\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - I(Z, X) - \text{KL}(q(\mathbf{z})||p(\mathbf{z})) \quad (3)$$

Discrete Information VAE (DI-VAE)

- A natural solution is to maximize both data log likelihood & mutual information.

$$\mathcal{L}_{\text{VAE}} + I(Z, X) = \mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})p(\mathbf{x})}[\log p_{\mathcal{G}}(\mathbf{x}|\mathbf{z})] - \text{KL}(q(\mathbf{z})\|p(\mathbf{z})) \quad (4)$$

- Match prior result for continuous VAE. [Mazhazni et al 2015, Kim et al 2017]

Discrete Information VAE (DI-VAE)

- A natural solution is to maximize both data log likelihood & mutual information.

$$\mathcal{L}_{\text{VAE}} + I(Z, X) = \mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})p(\mathbf{x})}[\log p_{\mathcal{G}}(\mathbf{x}|\mathbf{z})] - \text{KL}(q(\mathbf{z})\|p(\mathbf{z})) \tag{4}$$

- Match prior result for continuous VAE. [Mazhazni et al 2015, Kim et al 2017]
- Propose **Batch Prior Regularization (BPR)** to minimize KL [q(z)||p(z)] for discrete latent variables:

N: mini-batch size.

$$q(\mathbf{z}) \approx \frac{1}{N}\sum_{n=1}^{N} q(\mathbf{z}|\mathbf{x}_n) = q'(\mathbf{z}) \tag{5}$$

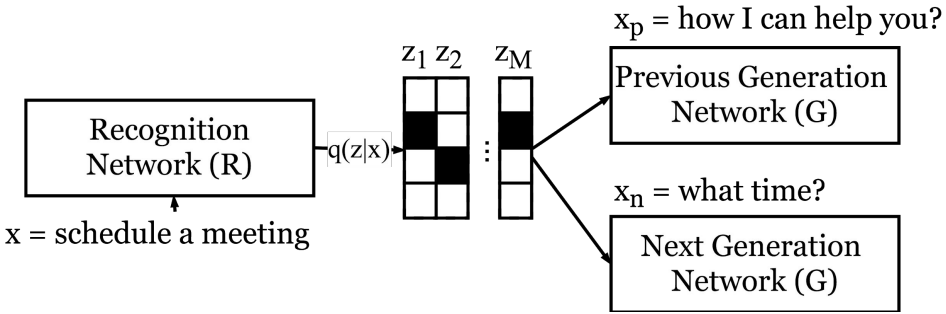
$$\text{KL}(q'(\mathbf{z})\|p(\mathbf{z})) = \sum_{k=1}^{K} q'(\mathbf{z}=k)\log\frac{q'(\mathbf{z}=k)}{p(\mathbf{z}=k)} \tag{6}$$

Fundamentally different from KL-annealing, since BPR is non-linear.

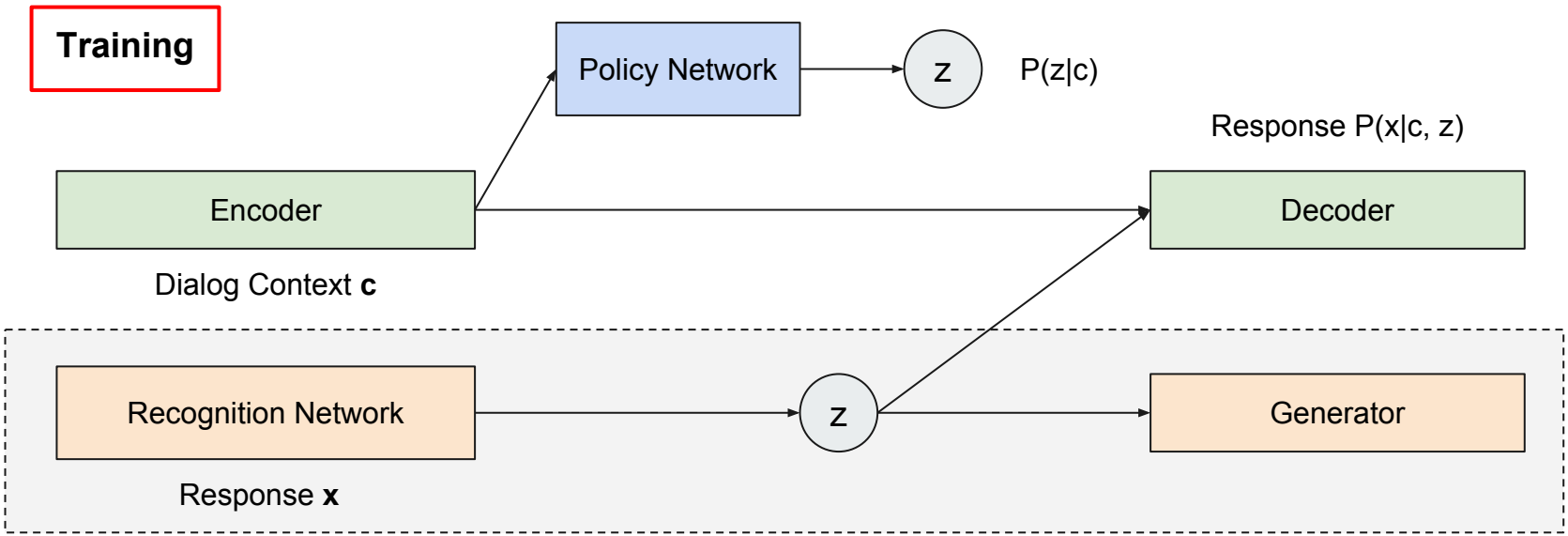
Learning from Context Predicting (DI-VST)

- Skip-Thought (ST) is well-known distributional sentence representation [Hill et al 2016]
- The meaning of sentences in dialogs is highly contextual, e.g. dialog acts.
- We extend DI-VAE to Discrete Information Variational Skip Thought (DI-VST).

$$\mathcal{L}_{\text{DI-VST}} = \mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})p(\mathbf{x}))}[\log(p^n_{\mathcal{G}}(\mathbf{x}_n|\mathbf{z})p^p_{\mathcal{G}}(\mathbf{x}_p|\mathbf{z}))] - \text{KL}(q(\mathbf{z})\|p(\mathbf{z})) \quad (7)$$



Integration with Encoder-Decoders

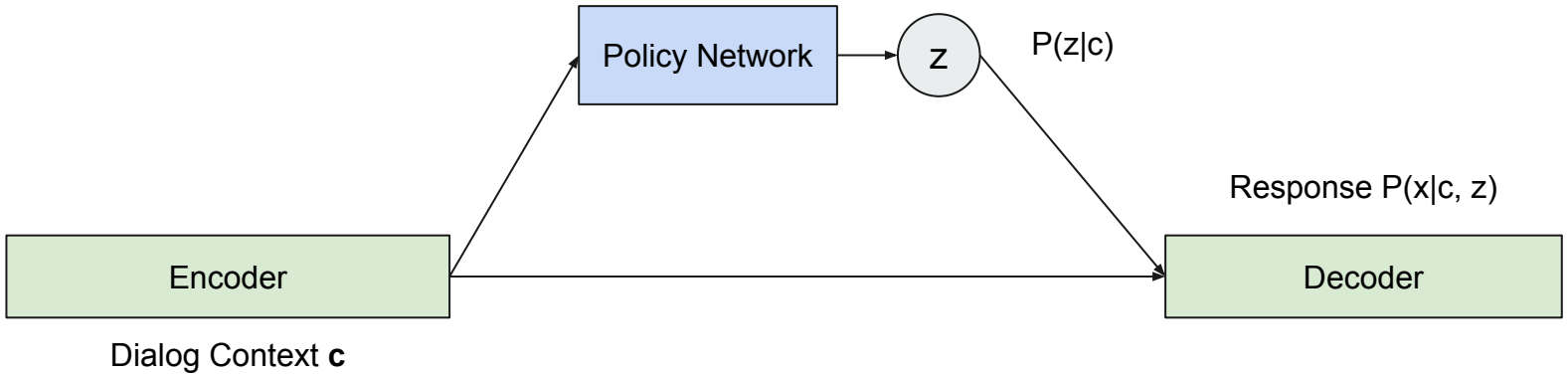


Optional: penalize decoder if generated **x** not exhibiting **z**
[Hu et al 2017]

$$\mathcal{L}_{\text{Attr}}(\theta_{\mathcal{F}}) = \mathbb{E}_{q_{\mathcal{R}}(\mathbf{z}|\mathbf{x})p(\mathbf{c},\mathbf{x})}[\log q_{\mathcal{R}}(\mathbf{z}|\mathcal{F}(\mathbf{c},\mathbf{z}))] \quad (9)$$

Integration with Encoder-Decoders

Testing



Evaluation Datasets

1. Penn Tree Bank (PTB) [Marcus et al 1993]:
    a. Past evaluation dataset for text VAE [Bowman et al 2015]
2. Stanford Multi-domain Dialog Dataset (SMD) [Eric and Manning 2017]
    a. 3,031 Human-Woz dialog dataset from 3 domains: weather, navigation & scheduling.
3. Switchboard (SW) [Jurafsky et al 1997]
    a. 2,400 human-human telephone non-task-oriented dialogues about a given topic.
4. Daily Dialogs (DD) [Li et al 2017]
    a. 13,188 human-human non-task-oriented dialogs from chat room.

The Effectiveness of Batch Prior Regularization (BPR)

For auto-encoding

- **DAE**: Autoencoder + Gumbel Softmax
- **DVAE**: Discrete VAE with ELBO loss
- **DI-VAE**: Discrete VAE + BPR

For context-predicting

- **DST**: Skip thought + Gumbel Softmax
- **DVST**: Variational Skip Thought
- **DI-VST**: Variational Skip Thought + BPR

Dom	Model	PPL	$\mathrm{KL}(q\|\|p)$	$I(\mathbf{x}, \mathbf{z})$
PTB	RNNLM	116.22	-	-
	VAE	73.49	15.94*	-
	DAE	66.49	2.20	0.349
	DVAE	70.84	0.315	0.286
	DI-VAE	**52.53**	**0.133**	**1.18**
DD	RNNLM	31.15	-	-
	DST	$\mathbf{x}_p$:28.23	0.588	**1.359**
		$\mathbf{x}_n$:28.16		
	DVST	$\mathbf{x}_p$:30.36	**0.007**	0.081
		$\mathbf{x}_n$:30.71		
	DI-VST	$\mathbf{x}_p$:**28.04**	0.088	1.028
		$\mathbf{x}_n$:**27.94**		

Table 1: Results for various discrete sentence representations.

The Effectiveness of Batch Prior Regularization (BPR)

For auto-encoding

- **DAE**: Autoencoder + Gumbel Softmax
- **DVAE**: Discrete VAE with ELBO loss
- **DI-VAE**: Discrete VAE + BPR

For context-predicting

- **DST**: Skip thought + Gumbel Softmax
- **DVST**: Variational Skip Thought
- **DI-VST**: Variational Skip Thought + BPR

Dom	Model	PPL	KL$(q\|\|p)$	$I(\mathbf{x}, \mathbf{z})$
PTB	RNNLM	116.22	-	-
	VAE	73.49	15.94*	-
	DAE	66.49	2.20	0.349
	DVAE	70.84	0.315	0.286
	DI-VAE	**52.53**	**0.133**	**1.18**
DD	RNNLM	31.15	-	-
	DST	$\mathbf{x}_p$:28.23 $\mathbf{x}_n$:28.16	0.588	**1.359**
	DVST	$\mathbf{x}_p$:30.36 $\mathbf{x}_n$:30.71	**0.007**	0.081
	DI-VST	$\mathbf{x}_p$:**28.04** $\mathbf{x}_n$:**27.94**	0.088	1.028

Table 1: Results for various discrete sentence representations.

The Effectiveness of Batch Prior Regularization (BPR)

For auto-encoding

- **DAE**: Autoencoder + Gumbel Softmax
- **DVAE**: Discrete VAE with ELBO loss
- **DI-VAE**: Discrete VAE + BPR

For context-predicting

- **DST**: Skip thought + Gumbel Softmax
- **DVST**: Variational Skip Thought
- **DI-VST**: Variational Skip Thought + BPR

Dom	Model	PPL	KL$(q\|\|p)$	$I(\mathbf{x}, \mathbf{z})$
PTB	RNNLM	116.22	-	-
	VAE	73.49	15.94*	-
	DAE	66.49	2.20	0.349
	DVAE	70.84	0.315	0.286
	DI-VAE	**52.53**	**0.133**	**1.18**
DD	RNNLM	31.15	-	-
	DST	$\mathbf{x}_p$:28.23 $\mathbf{x}_n$:28.16	0.588	**1.359**
	DVST	$\mathbf{x}_p$:30.36 $\mathbf{x}_n$:30.71	**0.007**	0.081
	DI-VST	$\mathbf{x}_p$:**28.04** $\mathbf{x}_n$:**27.94**	0.088	1.028

Table 1: Results for various discrete sentence representations.

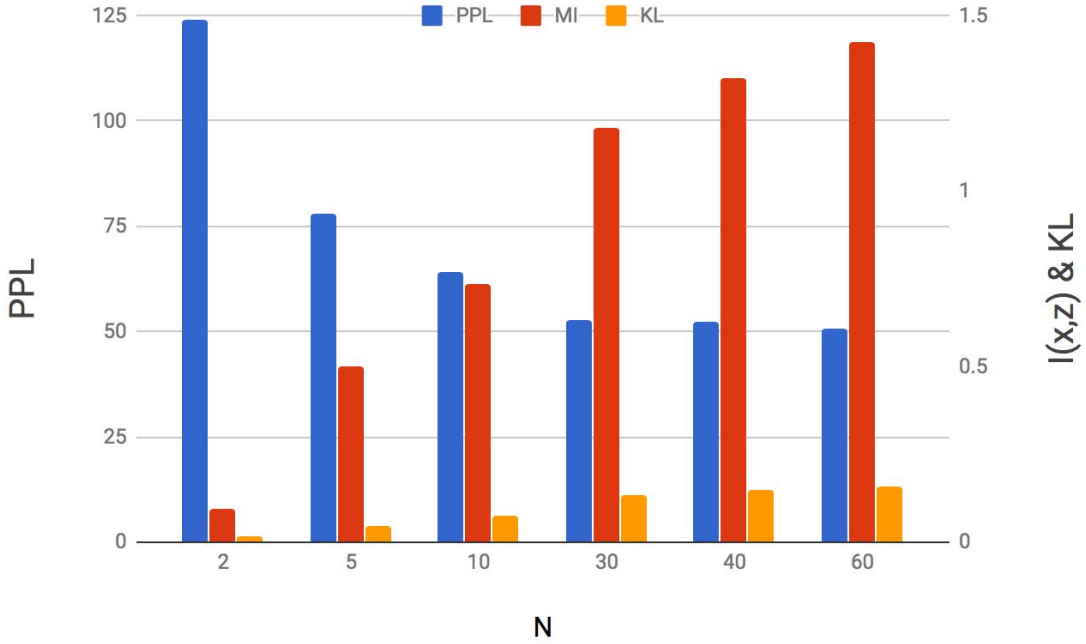
How large should the batch size be?

> When batch size N = 0

- = normal ELBO

> A large batch size leads to more meaningful latent action z

- Slowly increasing KL
- Improve PPL
- I(x,z) is not the final goal



Intropolation in the Latent Space

So you can keep record of all the checks you write.
So you can get all kinds of information and credit cards.
So you can keep track of all the credit cards.
So you kind of look at the credit union.
So you know of all the credit cards.
Yeah because you know of all the credit cards.
Right you know at least a lot of times.
Right you know a lot of times.

Table 9: Interpolating from the source sentence (top) to a target sentence (bottom) by sequentially setting the source latent code to the target code.

Differences between DI-VAE & DI-VST

- DI-VAE cluster utterances based on the words:
  - More fine-grained actions
  - More error-prone since harder to predict
- DI-VST cluster utterances based on the context:
  - Utterance used in the similar context
  - Easier to get agreement.

Model	Action	Sample utterance
DI-VAE	scheduling	- sys: okay, scheduling a yoga activity with Tom for the 8th at 2pm. - sys: okay, scheduling a meeting for 6 pm on Tuesday with your boss to go over the quarterly report.
	requests	- usr: find out if it 's supposed to rain - usr: find nearest coffee shop
DI-VST	ask schedule info	- usr: when is my football activity and who is going with me? - usr: tell me when my dentist appointment is?
	requests	- usr: how about other coffee? - usr: 11 am please

Interpreting Latent Actions

M=3, K=5. The trained R will map any utterance into a_1-a_2-a_3. E.g. How are you? → 1-4-2

- Automatic Evaluation on SW & DD
- Compare latent actions with human-annotations.
- Homogeneity [Rosenberg and Hirschberg, 2007].
  - The higher the more correlated

	SW		**DD**	
	Act	Topic	Act	Emotion
DI-VAE	0.48	0.08	0.18	0.09
DI-VST	0.33	0.13	0.34	0.12

Table 3: Homogeneity results (bounded [0, 1]).

Interpreting Latent Actions

M=3, K=5. The trained R will map any utterance into a_1-a_2-a_3. E.g. How are you? → 1-4-2

- Human Evaluation on SMD
- Expert look at 5 examples and give a name to the latent actions
- 5 workers look at the expert name and another 5 examples.
- Select the ones that match the expert name.

Model	Exp Agree	Worker κ	Match Rate
DI-VAE	85.6%	0.52	71.3%
DI-VST	93.3%	0.48	74.9%

Table 4: Human evaluation results on judging the homogeneity of latent actions in SMD.

Predict Latent Action by the Policy Network

	SMD	
AE-ED	3.045 (51.5% sys 52.4% usr 50.5%)	
ST-ED	1.695 (75.5% sys 82.1% usr 69.2%)	
	DD	SW
AE-ED	4.47 (35.8%)	4.46 (31.68%)
ST-ED	3.89 (47.5%)	3.68 (33.2%)

Table 7: Performance of policy network. $\mathcal{L}_{attr}$ is included in training.

- Provide useful measure about the complexity of the domain.
  - Usr > Sys & Chat > Task
- Predict latent actions from DI-VAE is harder than the ones from DI-VST
- Two types of latent actions has their own pros & cons. Which one is better is application dependent.

Interpretable Response Generation

c	**usr**: Where does my friend live?	
Model	Action	Generated Responses
AE-ED	give loc info $p(\mathbf{z}\|\mathbf{c})$=0.34	- Tom is at 753 University Ave, and a road block. - Comfort Inn is at 7 miles away.
	give user info $p(\mathbf{z}\|\mathbf{c})$=0.22	- Your home address is 5671 barringer street. - Your home is at 10 ames street.
ST-ED	give loc info $p(\mathbf{z}\|\mathbf{c})$=0.93	- Jill's house is 8 miles away at 347 Alta Mesa Ave. - Jill lives at 347 Alta Mesa Ave.

- Examples of interpretable dialog generation on SMD
- First time, a neural dialog system outputs both:
  - target response
  - high-level actions with interpretable meaning

Conclusions & Future Work

- An analysis of ELBO that explains the posterior collapse issue for sentence VAE.
- DI-VAE and DI-VST for learning rich sentence latent representation and integration with encoder-decoders.
- Learn better context-based latent actions
  - Encode human knowledge into the learning process.
  - Learn structured latent action space for complex domains.
  - Evaluate dialog generation performance in human-study.

Thank you!

Code & Data: github.com/snakeztc/NeuralDialog-LAED

Semantic Consistency of the Generation

Domain	AE-ED	+L_{attr}	ST-ED	+L_{attr}
SMD	93.5%	94.8%	91.9%	93.8%
DD	88.4%	93.6%	78.5%	86.1%
SW	84.7%	94.6%	57.3%	61.3%

Table 6: Results for attribute accuracy with and without attribute loss.

- Use the recognition network as a classifier to predict the latent action **z'** based on the generated response **x'**.
- Report accuracy by comparing **z** and **z'**.

What we learned?

- DI-VAE has higher consistency than DI-VST
- L_{attr} helps more in complex domain
- L_{attr} helps DI-VST more than DI-VAE
  - DI-VST is not directly helping generating x
- ST-ED doesn't work well on SW due to complex context pattern
  - Spoken language and turn taking

What defines Interpretable Latent Actions

- **Definition**: Latent action is a set of discrete variable that define the high-level attributes of an utterance (sentence) **X.** Latent action is denoted as **Z**.
- **Two key properties:**
  - **Z** should capture salient sentence-level features about the response **X**.
  - The meaning of latent symbols **Z** should be independent of the context **C**.
- **Why context-independent?**
  - If meaning of **Z** depends on **C**, then often impossible to interpret **Z**
  - Since the possible space of C is huge!
- **Conclusion**: context-independent semantic ensures each assignment of z has the same meaning in all context.