www.cs.cmu.edu/~tianchez

Learning End-to-End Dialog Systems with Latent Actions

Tiancheng Zhao (Tony)
Advisor: Dr. Maxine Eskenazi
Language Technologies Institute, Carnegie Mellon University

Language Technologies Institute

Carnegie Mellon University

NSF

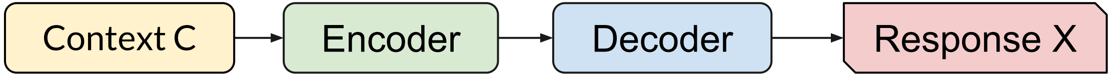
Dialrc

Outline

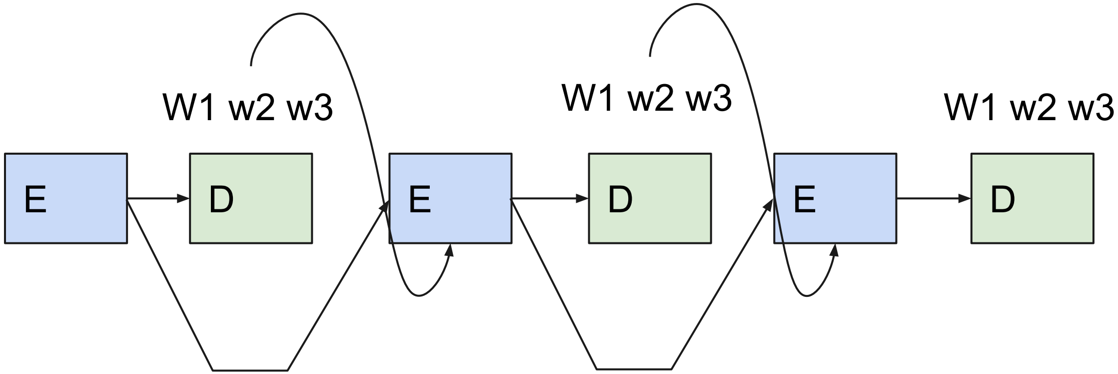
- Introduction to Generative Encoder Decoder Dialog Models
- What are Latent Actions?
- Stochastic Latent Actions (ACL 2017)
- Interpretable Latent Actions (ACL 2018)
- Cross-domain Latent Actions (SIGDIAL 2018 best paper award)

Introduction

- Encoder Decoders are auto-regressive models that model the conditional distribution P(X|C), or $P(w_{1t'}, \dots w_{nt'} \mid w_1 \dots w_k)$.



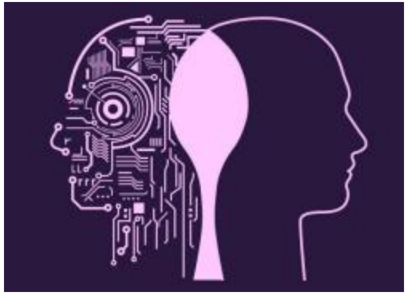
- Many powerful models available if we simply care about learning the distribution.
    - Variations of RNN, Attention Mechanism, Transformers, Wavenet ...



Beyond P(X|C)

- Dialog is a complex process and direct modelling of P(x|c) is data intensive, and suffers from many other issues.
- We need more than distribution modeling in the dialogs, e.g. controllability, interpretability or reliability etc.

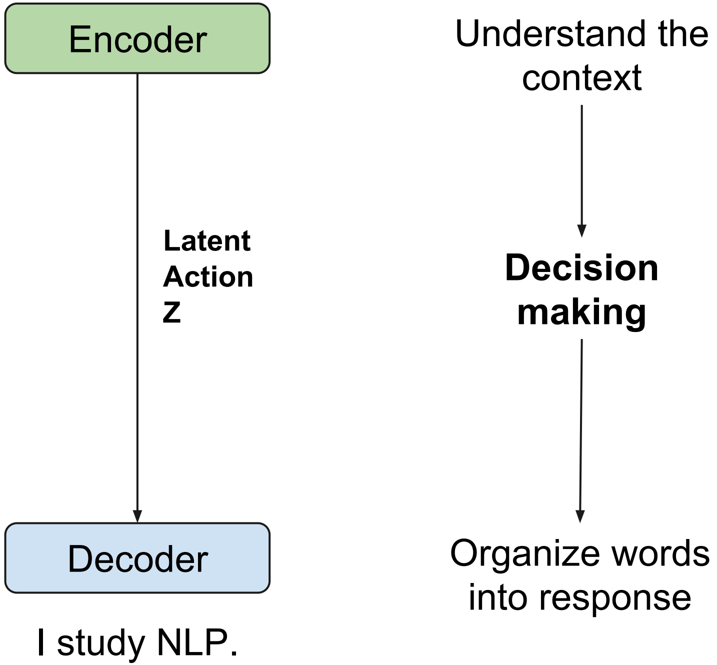
Diversity



Interpretability

Transferability

What are Latent Actions?



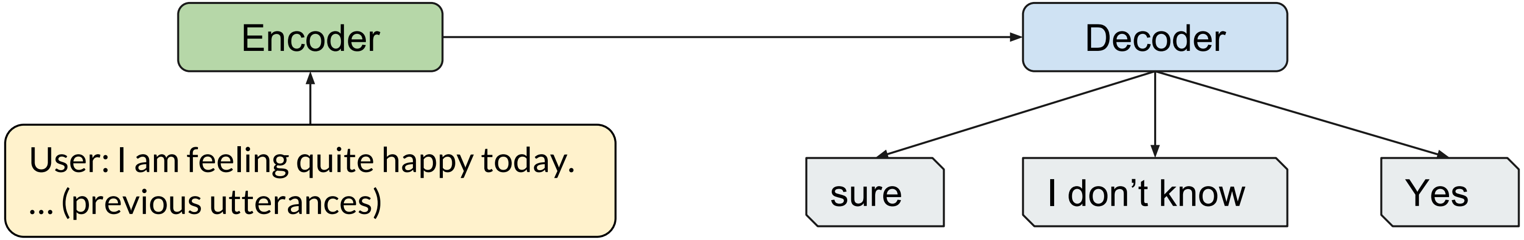
- Model “high level” actions in E2E dialog models.
  - Factorize the generation process from $P(X|C)$ to $P(X|\mathbf{Z}, C)P(\mathbf{Z}|C)$
  - Model intentions & temporal abstraction
  - Enable explainable Inference
  - Enable Knowledge Transfer

Diversity (ACL 2017)

The Dull Response Problem

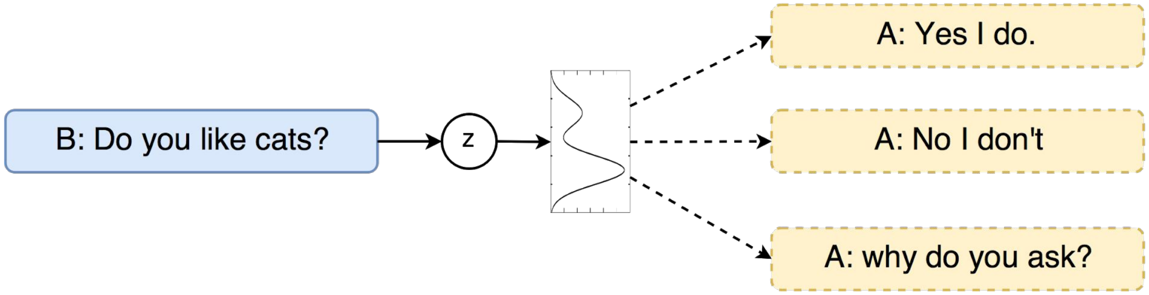
Dull response problem [Li et al 2015, Serban et al. 2016]. Prior solutions include:

- Add more info to the dialog context [Xing et al 2016, Li et al 2016]
- Improve decoding algorithm, e.g. beam search [Wiseman and Rush 2016]



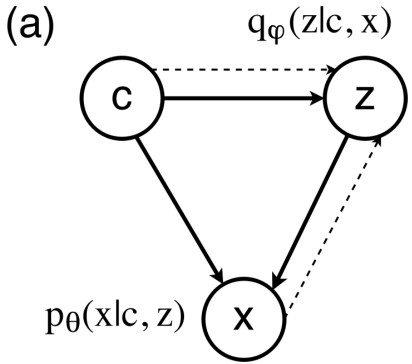
Model the Initial State Distribution [Zhao et al 2017a]

- Response generation in conversation is a **ONE-TO-MANY** mapping problem at the **discourse level**.
- Introduce latent variable **z** (the initial state of the decoder) to represent system's next "high-level action"
- P(Z|C) should be modelled as a **probabilistic distribution** rather than point estimate.



Conditional Variational Auto Encoder (CVAE)

- **C** is dialog context
  - B: Do you like cats? A: Yes I do
- **Z** is the latent variable (gaussian)
- **X** is the next response
  - B: So do I.
- Trained by Stochastic Gradient Variational Bayes (SGVB) [Kingma and Welling 2013]

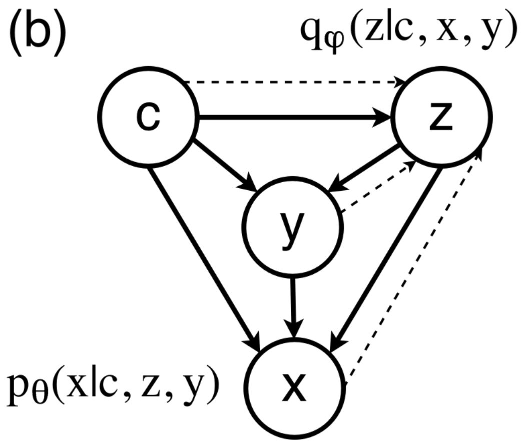


$$\begin{aligned}\mathcal{L}(\theta, \phi; x, c) &= -KL(q_\phi(z|x,c) \| p_\theta(z|c)) \\ &\quad + \mathbf{E}_{q_\phi(z|c,x)}[\log p_\theta(x|z,c)] \qquad (1) \\ &\leq \log p(x|c)\end{aligned}$$

Knowledge-Guided CVAE (kgCVAE)

- **Y** is linguistic features extracted from responses
  - Dialog act: statement -> “So do I”.
- Use **Y** to guide the learning of latent **Z**

$$\mathcal{L}(\theta, \phi; x, c, y) = -KL(q_\phi(z|x,c,y) \| P_\theta(z|c)) + \mathbf{E}_{q_\phi(z|c,x,y)}[\log p(x|z,c,y)] + \mathbf{E}_{q_\phi(z|c,x,y)}[\log p(y|z,c)] \quad (4)$$



Posterior Collapse of Z

Training CVAE with RNN decoder is hard due to the ***posterior collapse problem*** [Bowman et al., 2015]

- RNN decoder can cheat by using LM information and ignore **Z**!

Bowman et al. [2015] described two methods to alleviate the problem :

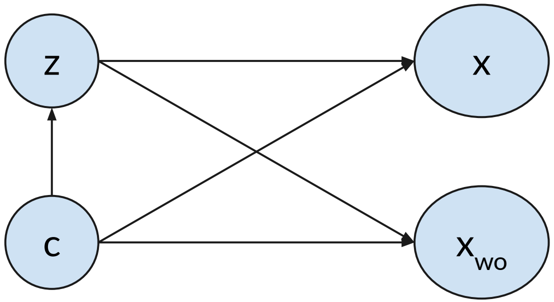
1. KL annealing (KLA): gradually increase the weight of KL term from 0 to 1 (need early stop).
2. Word drop decoding: setting a proportion of target words to 0 (need careful parameter picking).

BOW Loss

- Predict the bag-of-words in the responses **X** at once (word counts in the response)
- Break the dependency between words and eliminate the chance of cheating based on LM.

$$\log p(x_{bow}|z,c) = \log \prod_{t=1}^{|x|} \frac{e^{f_{x_t}}}{\sum_j^V e^{f_j}}$$

$$\mathcal{L}'(\theta,\phi;x,c) = \mathcal{L}(\theta,\phi;x,c) + \mathbf{E}_{q_\phi(z|c,x,y)}[\log p(x_{bow}|z,c)]$$



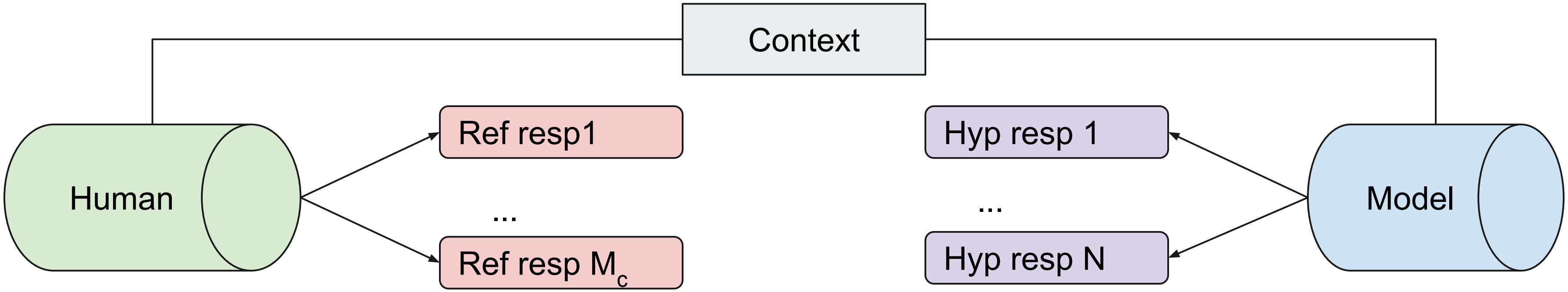
RNN Loss

FF Bag-of-word Loss

Dataset

Data Name	Switchboard Release 2
Number of dialogs	2,400 (2316/60/62 - train/valid/test)
Number of context-response pairs	207,833/5,225/5,481
Vocabulary Size	Top 10K
Dialog Act Labels	42 types, tagged by SVM and human
Number of Topics	70 tagged by humans

Quantitative Metrics



$$\text{precision(c)} = \frac{\sum_{i=1}^{N} max_{j\in[1,M_c]} d(r_j, h_i)}{N}$$

Appropriateness

$$\text{recall(c)} = \frac{\sum_{j=1}^{M_c} max_{i\in[1,N]} d(r_j, h_i))}{M_c}$$

Diversity

d(r, h) is a distance function [0, 1] to measure the similarity between a reference and a hypothesis.

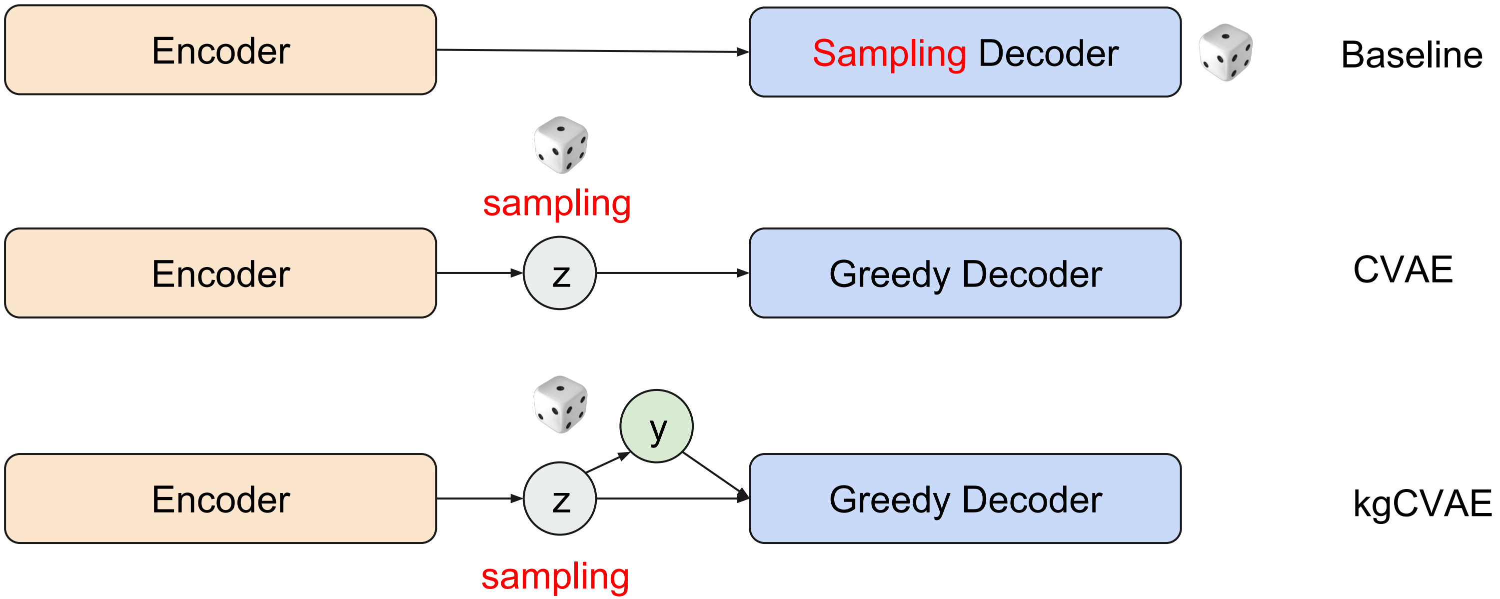
Distance Functions used for Evaluation

1. Smoothed Sentence-level BLEU (1/2/3/4): lexical similarity
2. Cosine distance of Bag-of-word Embeddings: distributed semantic similarity.

   (pre-trained Glove embedding on twitter)

   a. Average of embeddings (A-bow)
   b. Extrema of embeddings (E-bow)
3. Dialog Act Match: illocutionary force-level similarity
   a. (Use pre-trained dialog act tagger for tagging)

Models (trained with BOW loss)



Quantitative Analysis Results

Metrics	Perplexity (KL)	BLEU-1 (p/r)	BLEU-2 (p/r)	BLEU-3 (p/r)	BLEU-4 (p/r)	A-bow (p/r)	E-bow (p/r)	DA (p/r)
Baseline (sample)	35.4 (n/a)	0.405/ 0.336	0.3/ 0.281	0.272/ 0.254	0.226/ 0.215	0.387/ 0.337	0.701/ 0.684	**0.736**/ 0.514
CVAE (greedy)	20.2 (11.36)	0.372/ 0.381	0.295/ 0.322	0.265/ 0.292	0.223/ 0.248	**0.389/ 0.361**	0.705/ 0.709	0.704/ **0.604**
kgCVAE (greedy)	16.02 (13.08)	**0.412/ 0.411**	**0.350/ 0.356**	**0.310/ 0.318**	**0.262/ 0.272**	0.373/ 0.336	**0.711/ 0.712**	0.721/ 0.598

Note: BLEU are normalized into [0, 1] to be valid precision and recall distance function

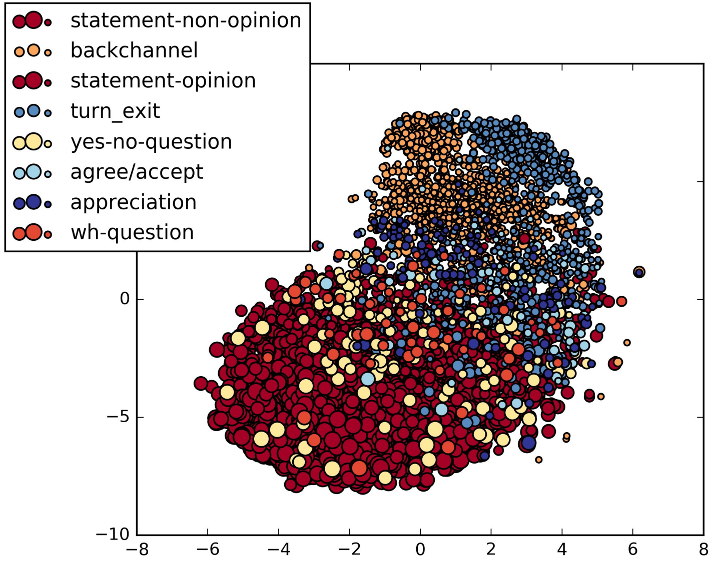
Qualitative Analysis

Topic: Recycling **Context**: **A**: are they doing a lot of recycling out in Georgia?
Target (statement): well at my workplace we have places for aluminium cans

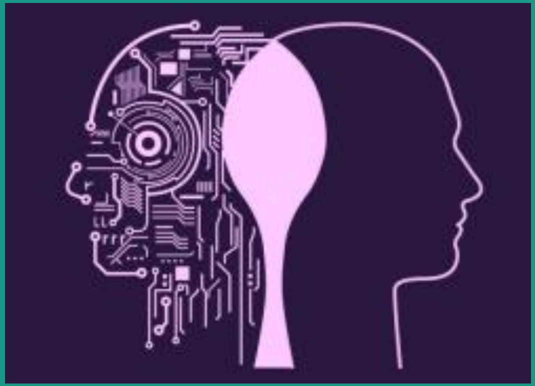
Baseline + Sampling	**kgCVAE + Greedy**
1. well I’m a graduate student and have two kids.	1. (non-understand) pardon.
2. well I was in last year and so we’ve had lots of recycling.	2. (statement) oh you’re not going to have a curbside pick up here.
3. I’m not sure.	3. (statement) okay I am sure about a recycling center.
4. well I don’t know I just moved here in new york.	4. (yes-answer) yeah so.

Latent Space Visualization

- Visualization of the posterior **Z** on the test dataset in 2D space using t-SNE.
- Assign different colors to the top 8 frequent dialog acts.
- The size of circle represents the response length.
- Exhibit clear clusterings of responses w.r.t the dialog act



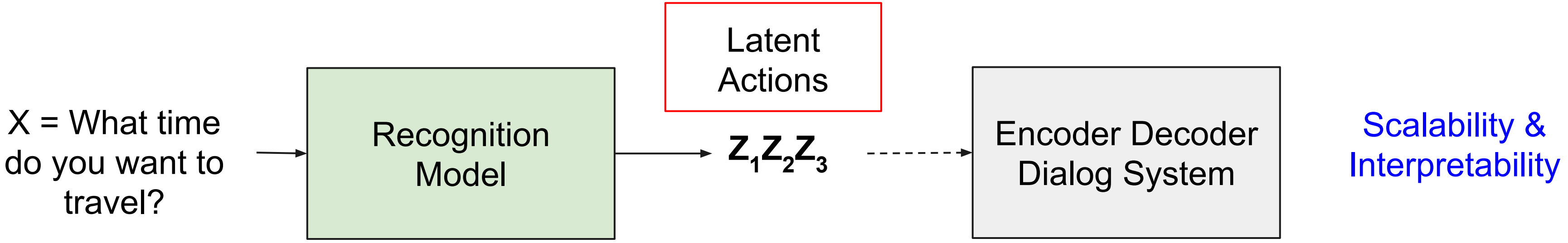
Interpretability (ACL 2018)



Why discrete sentence representation?

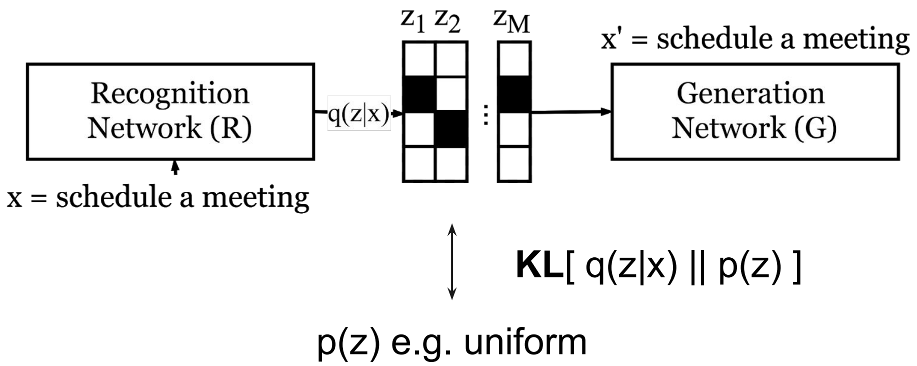
1. Inrepteablity & controbility
2. Multimodal distribution
3. Semi-supervised Learning [Kingma et al 2014 NIPS, Zhou et al 2017 ACL]

Our goal:



Baseline: Discrete Variational Autoencoder (VAE)

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



Anti-Info Nature in Evidence Lower Bound (ELBO)

- Write ELBO as an expectation over the whole dataset

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

- Expand the KL term, and plug back in:

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

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

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

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]
- Propose **Batch Prior Regularization (BPR)** to minimize KL [q(z)||p(z)] for discrete latent variables:

N: mini-batch size.

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

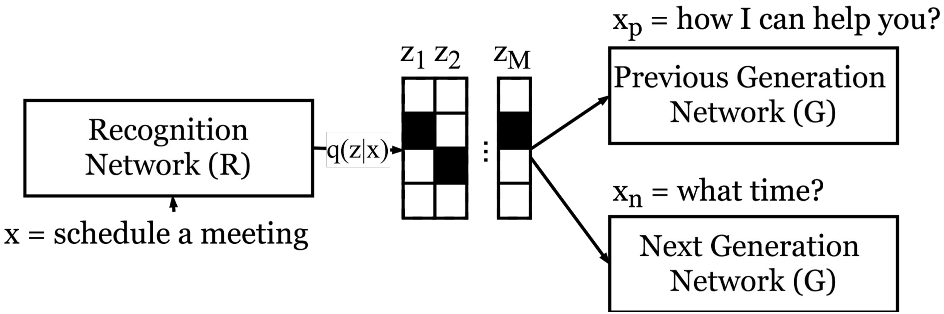
$$\text{KL}(q'(\mathbf{z})\|p(\mathbf{z})) = \sum_{k=1}^{K} q'(\mathbf{z}=k)\log\frac{q'(\mathbf{z}=k)}{p(\mathbf{z}=k)} \quad (6)$$

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

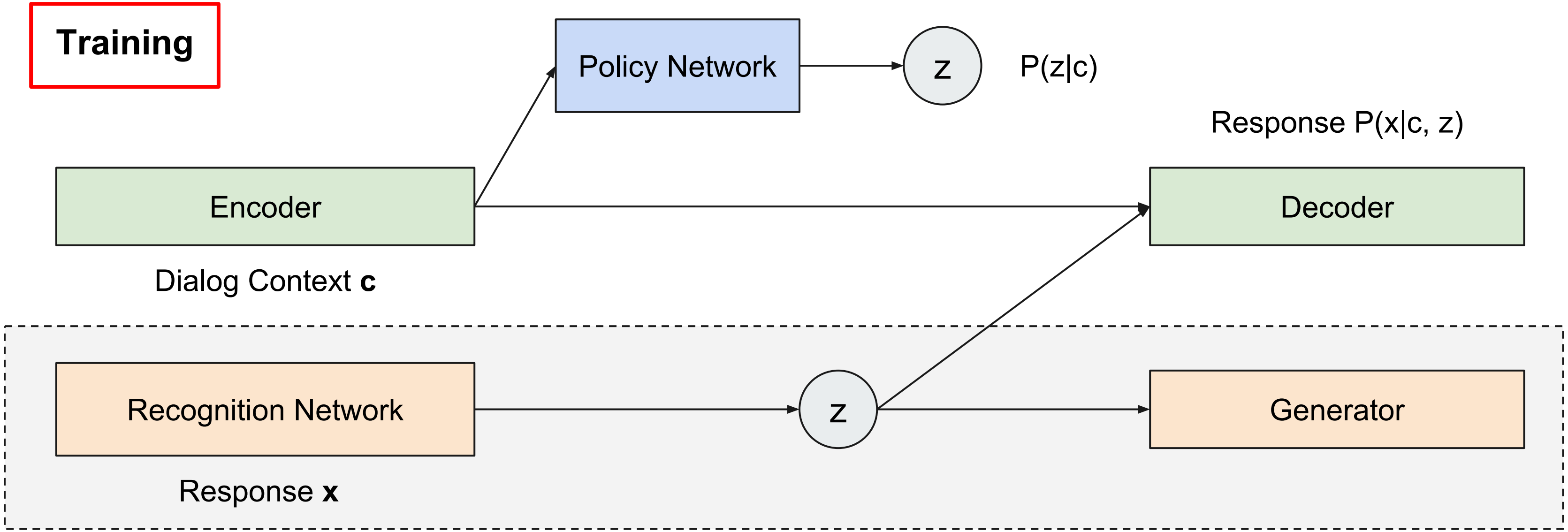
Learning from Context Predicting (DI-VST)

- 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).

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



Integration with Encoder-Decoders

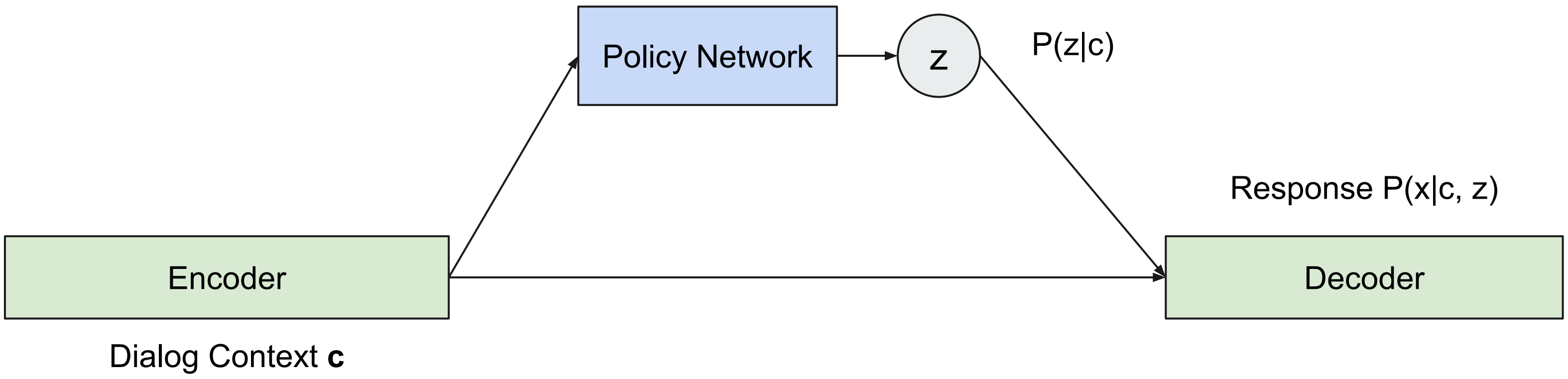


Optional: penalize decoder if generated **x** not exhibiting **z**
[Hu et al 2017]

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

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	$\text{KL}(q\|p)$	$I(\mathbf{x},\mathbf{z})$
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.

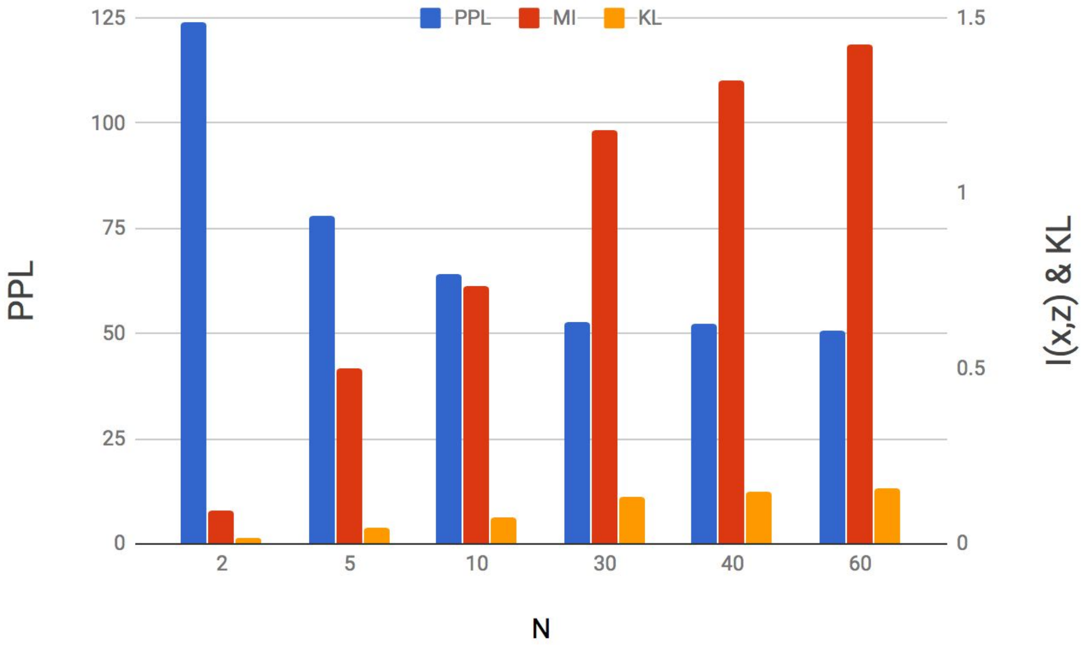
How large should the batch size be?

> 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

Language Technologies Institute

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.

Semantic Consistency of the Generation

Domain	AE-ED	$+\mathcal{L}_{\text{attr}}$	ST-ED	$+\mathcal{L}_{\text{attr}}$
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

Language Technologies Institute

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

Transferability (SIGDIAL 2018)

Problem: Data Scarcity & Poor Generalization

- GEDMs require **LARGE** training data
- **Impractical** since data are often NOT available:
  - Booking, recommendation, entertainment etc
- **Goal**:
  - Exploit GEDMs flexibility and let one model simultaneously learn many domains. **(Multi-task)**
  - Transfer knowledge from related domains with data to new domains without data. **(Zero-shot)**

Example: a customer service agent in shoe department can begin to work in the clothing department after reading training materials, without the need for example dialogs.

Define Zero-shot Dialog Generation (ZSDG)

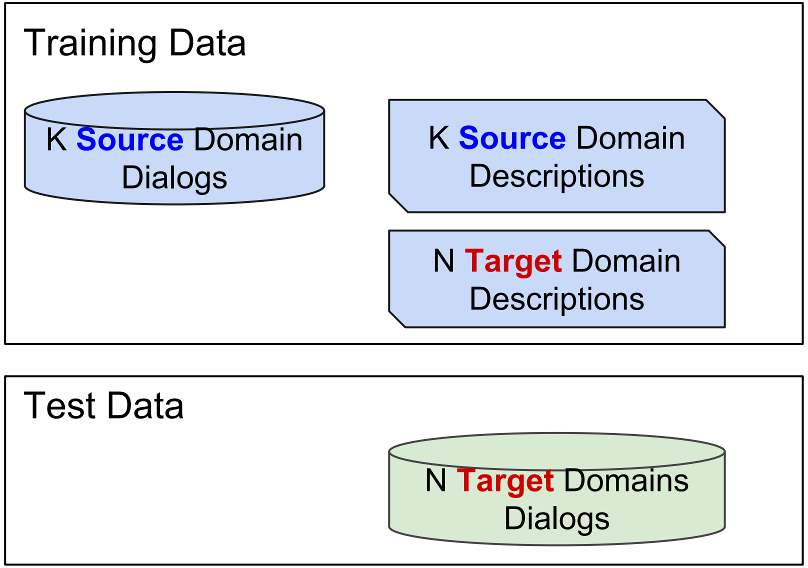
- Source domains: D_{source} is a set of dialog domain with dialog training data.
- Target domains: D_{target} is a set of dialog domains without data.
- Domain description: φ(d) captures domain-specific information about d
- Context is **c** and response is **x**

$$\text{Train Data: } \{\mathbf{c}, \mathbf{x}, d\} \sim p_{\text{source}}(\mathbf{c}, \mathbf{x}, d)$$

$$\{\phi(d)\}, d \in D$$

$$\text{Test Data: } \{\mathbf{c}, \mathbf{x}, d\} \sim p_{\text{target}}(\mathbf{c}, \mathbf{x}, d)$$

$$\text{Goal: } \mathcal{F} : C \times D \rightarrow X$$

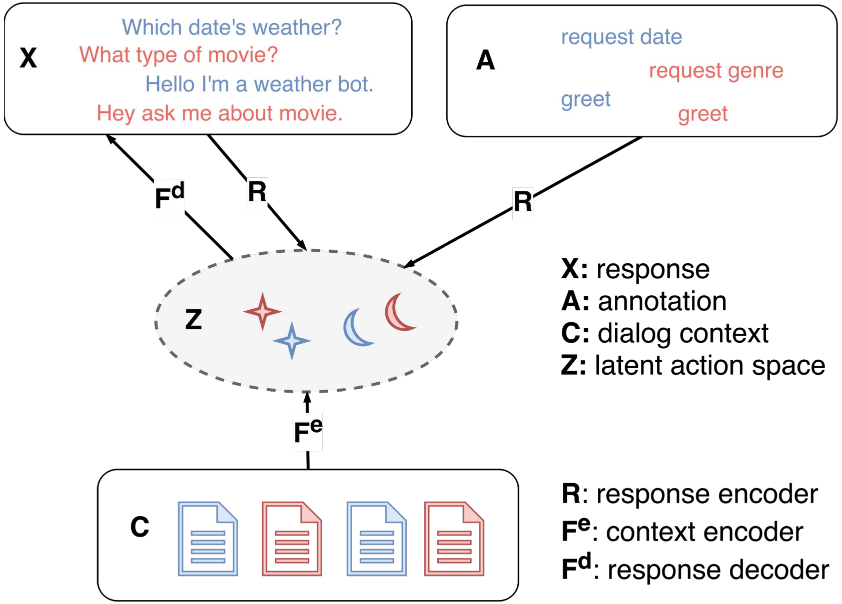


Seed Response (SR) as Domain Description

- Define SR(d) as a set of tuples
    - Each tuple contains utterances with annotations for a domain: $\{\mathbf{x}, \mathbf{a}, \mathbf{d}\}_{\mathbf{seed}}$
    - **x** is an example utterance, **a** is annotation, **d** is domain index.
- **Assumption**: Shared state tracking & policy <-->domain-specific NLU & NLG

x	**a**	**d**
x = the weather in New York is raining	[Inform, location=New York, weather_type=Rain]	weather
x=what’s the location?	[request location]	weather

Action Matching Algorithm

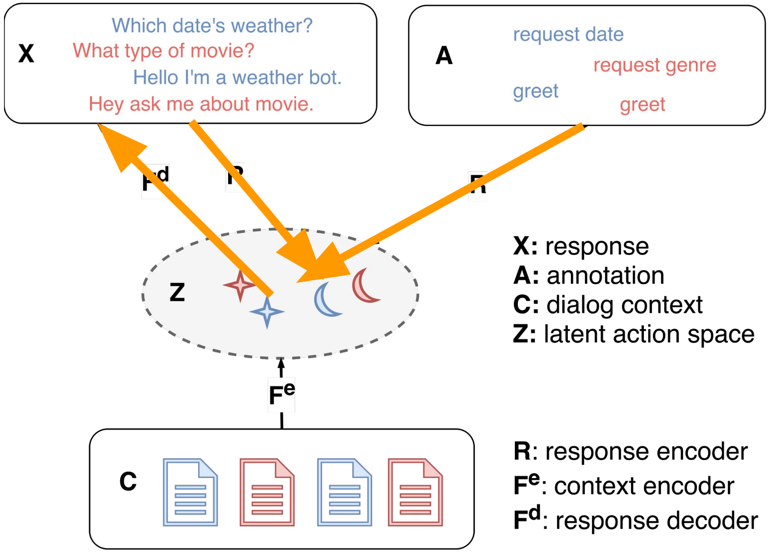


- **R**: encode utterances/annotations into latent actions
  - $z^d_x = R(x, d)$
  - $z^d_a = R(a, d)$
- **F**^{e}: predict latent action given the context
  - $z^d_c = F^e(c, d)$
- **F**^{d}: generates the response from latent action
  - $x = F^d(z)$

For Seed Response Data

Objective 1: $\mathcal{L}_{dd} = -\log(P_{\mathcal{F}^d}(x|\mathcal{R}(a,d))) + \lambda\mathbf{D}[\mathcal{R}(x,d)||\mathcal{R}(a,d)]$

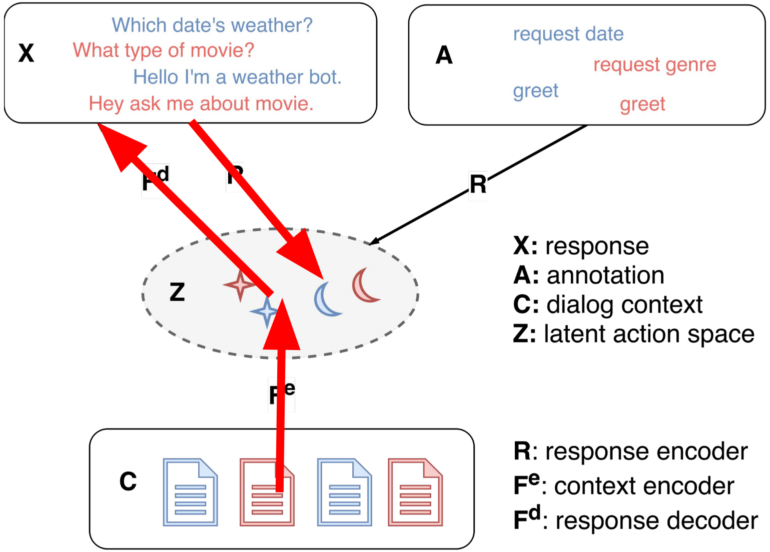
Warm-up decoder for both source & target utterances



Cluster cross-domain latent actions based on annotations.

For Source Dialog Data

Objective 2: $\mathcal{L}_{dialog} = -\log(P_{\mathcal{F}^d}(x|\mathcal{F}^e(c,d))) + \lambda \mathbf{D}[\mathcal{R}(x,d)||\mathcal{F}^e(c,d)]$



Train decoder in source domains with a lot of data

Train encoder to predict **Z** well

Optimization by Alternating these 2 losses

- $\mathcal{L}_{dd} = -\log(P_{\mathcal{F}^d}(x|\mathcal{R}(a,d))) + \lambda \mathbf{D}[\mathcal{R}(x,d)||\mathcal{R}(a,d)]$
- $\mathcal{L}_{dialog} = -\log(P_{\mathcal{F}^d}(x|\mathcal{F}^e(c,d))) + \lambda \mathbf{D}[\mathcal{R}(x,d)||\mathcal{F}^e(c,d)]$

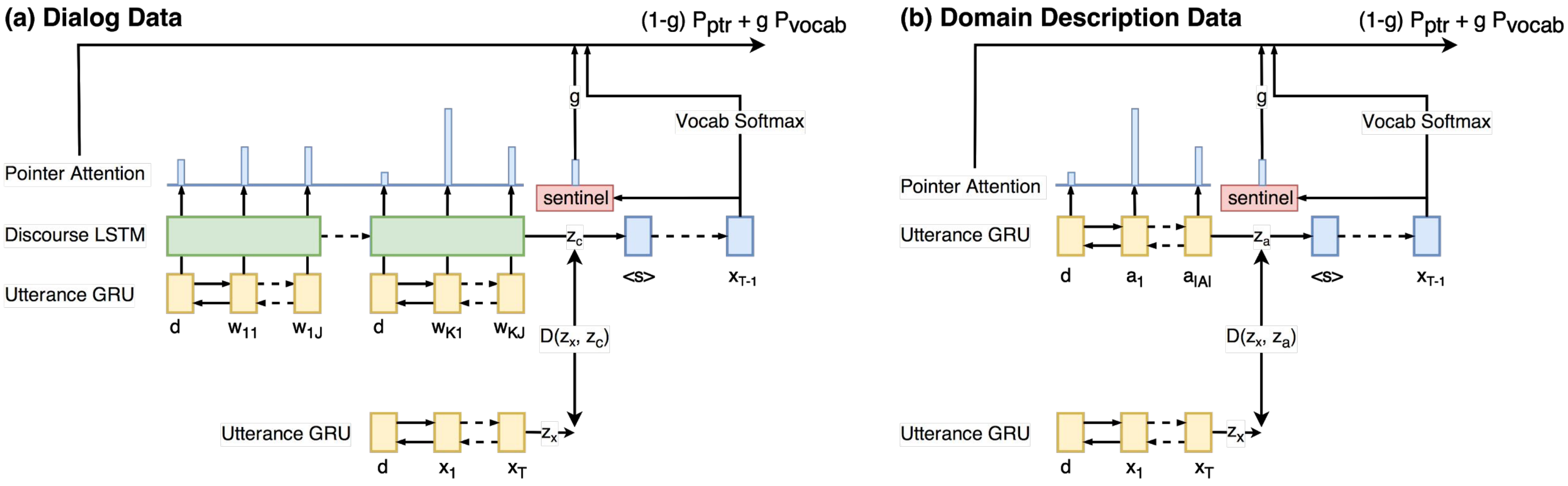
Algorithm 1: Action Matching Training

Initialize weights of $\mathcal{F}^e$, $\mathcal{F}^d$, $\mathcal{R}$;
Data = $\{\mathbf{c}, \mathbf{x}, d\} \bigcup \{\mathbf{x}, \mathbf{a}, d\}_{seed}$
while *batch* $\sim$ *Data* **do**
    if *batch in the form* $\{\mathbf{c}, \mathbf{x}, d\}$ **then**
        Backpropagate loss $\mathcal{L}_{\text{dialog}}$
    else
        Backpropagate loss $\mathcal{L}_{\text{dd}}$
    end
end

Implementation

- Recognition Network **R:** Bidirectional GRU
- Encoder $\mathbf{F}^{e}$: Hierarchical Recurrent LSTM Encoder (HRE) [Li et al 2015]
- Decoder $\mathbf{F}^{d}$**:**
  - LSTM Attention decoder
    - Attention over every words in the context
    - Standard baseline.
  - LSTM Pointer-sentinel Mixture (PSM) Decoder (Copy mechanism) [Merity et al 2016]
    - Can copy any words from the context
    - Proven to show good performance in generating OOV tokens.

Implementation with PSM decoder



Data

1. CMU SimDial: simulated dataset

2. Stanford Multi-domain Dialog (SMD) Dataset:

   Human-Woz dataset

CMU SimDial

- A open-source multi-domain dialog generator with complexity control.
- Source Domains (900 training, 100 validation dialogs for each domain):
  - Restaurant, Bus, Weather
- Target Domains (500 testing dialogs for each domain)
  - Restaurant **(in-domain)**
  - Restaurant-slot **(unseen slot)**: introduce new slot values
  - Restaurant-style **(unseen NLG)**: same slot values but different NLG templates
  - Movie **(new-domain)**: completely new domains
- Seed Response (SR):
  - 100 unique random utterances from each domain, annotations are semantic frames used by the simulator.
  - I believe you said Boston. Where are you going?” → [implicit_confirm location=Boston; request location]

Stanford Multi-domain Dialog (SMD)

- 3031 human-Woz data about 3 domains [Eric and Manning 2017]
  - Schedule, Navigation, Weather
- Leave-one-out to rotate among each domain as the target domain.
- Random sample 150 unique utterances from each domain as SR
- An expert annotated the 150 utterances in SR (available online)
  - All right, I've set your next dentist appointment for 10am. Anything else? → [ack; inform goal event=dentist appointment time=10am ; request needs].
- All the target data that we need is the 150 utterances with annotations - No large dialog corpus is needed!

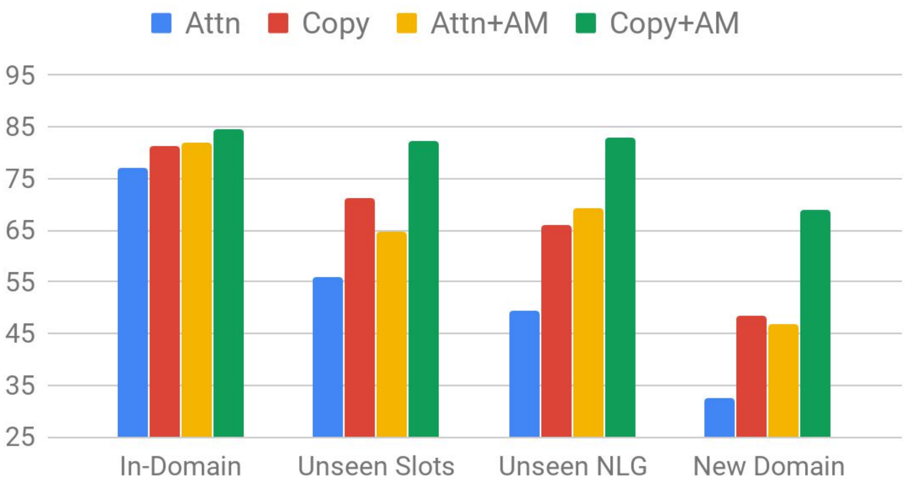
Metrics and Compared Models

1. **BLEU-4**: corpus-level BLUE-4 between the generated responses and references.
2. **Entity F1**: checks if the generated responses contains the correct entities (slot values)
3. **Act F1**: checks if the generated responses exhibits the correct dialog acts (using a classifier)
4. **KB F1**: check if the generated API call has all correct command tokens.
5. **BEAK**: geometric mean of the above 4 scores. BEAK = (bleu × ent × act × kb)^(¼)
   a. **BE (for SMD)**: BE = (bleu x ent)^(½)

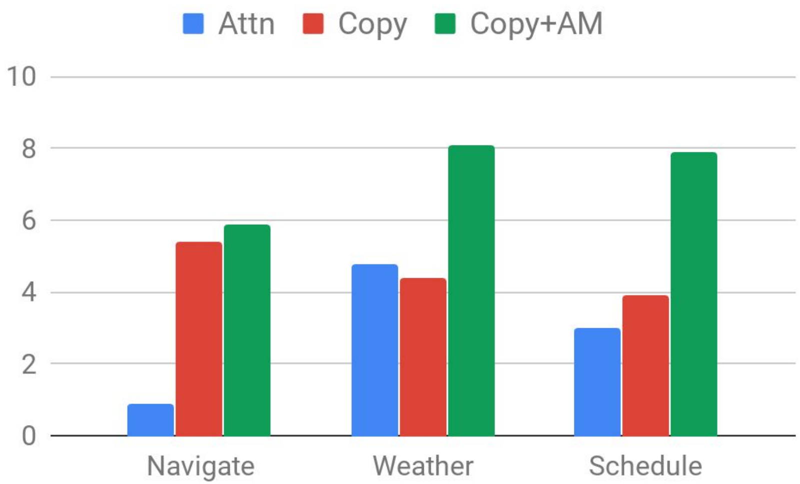
Four models are compared:

1. HRE + Attention Decoder (+Attn)
2. HRE + PSM Decoder (+Copy)
3. HRE + Attention Decoder + AM training (+Attn+AM)
4. HRE + PSM Decoder + AM training (+Copy+AM)

Overall Performance



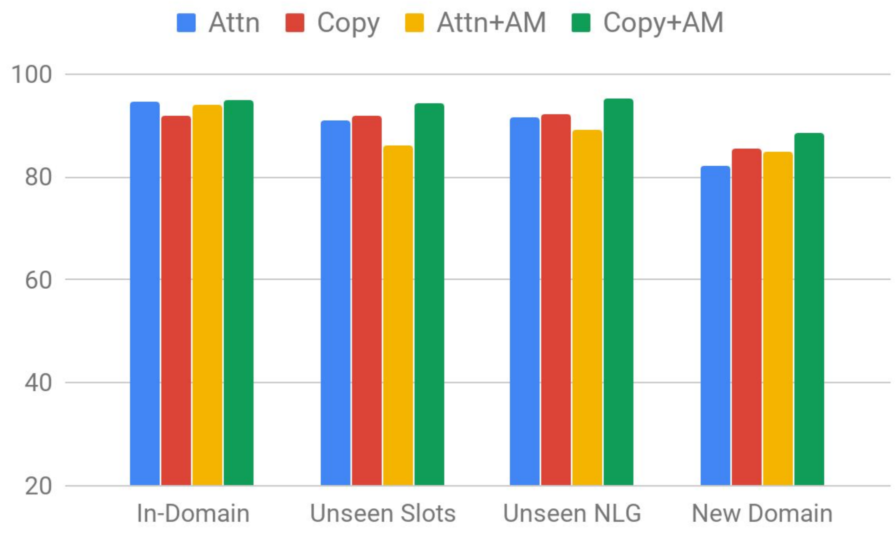
BEAK on SimDial



BE on SMD

1. What fails when testing on new domain?
2. What problem does Copy solve?
3. What problem does AM solve?
4. How does the size of SR affect AM's performance?

What Fails on New Domains?



Dialog Act F1 on SimDial

Answer: fail to generate the correct **entity** as well as the correct **overall sentence**. Dialog acts are okay.

First analyze dialog acts:

Good Examples:

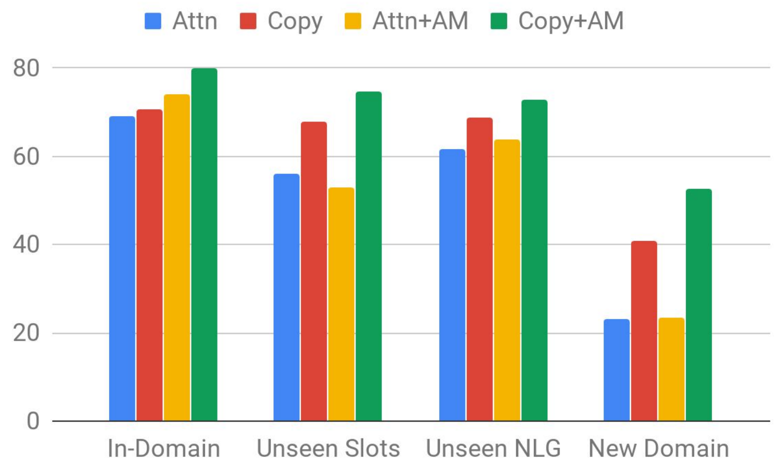
- **Ref**: See you.
- **Generated (Attn)**: See you next time

Bad Examples:

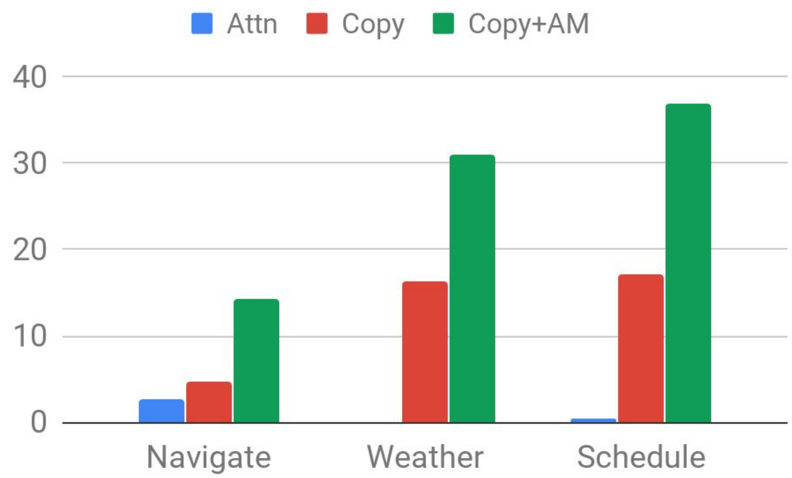
- **Ref**: Hi I am your movie bot. What can I do for you?
- **Generated (Attn)**: Hi this is the restaurant system. How can I help?

- **Ref**: Sci-fi movie. What time's movie?
- **Generated (Attn or Copy)**: Pittsburgh. what kind of restaurant are you looking for?

What Problem Does Copy Solve?



Entity F1 on SimDial



Entity F1 on SMD

Answer: Copy Network improves **entity** score significantly, especially when there are OOV entity

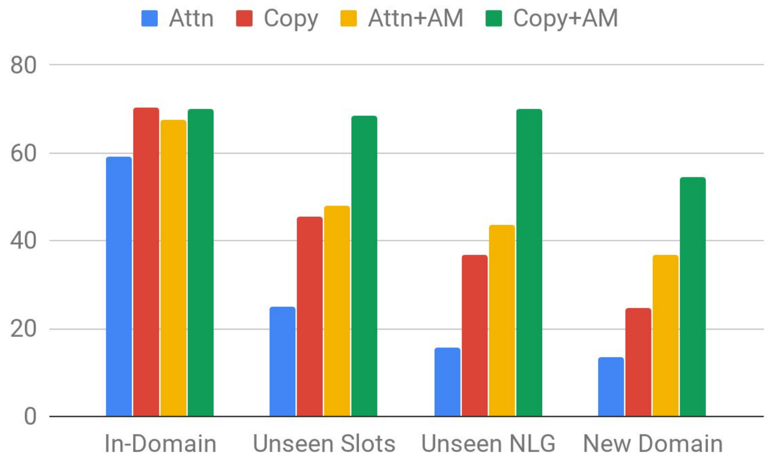
Examples:

- **Ref**: Do you mean sci-fi?
- **Generated (Attn):** Do you mean pizza?
- **Generated (Copy)**: Do you mean sci-fi?

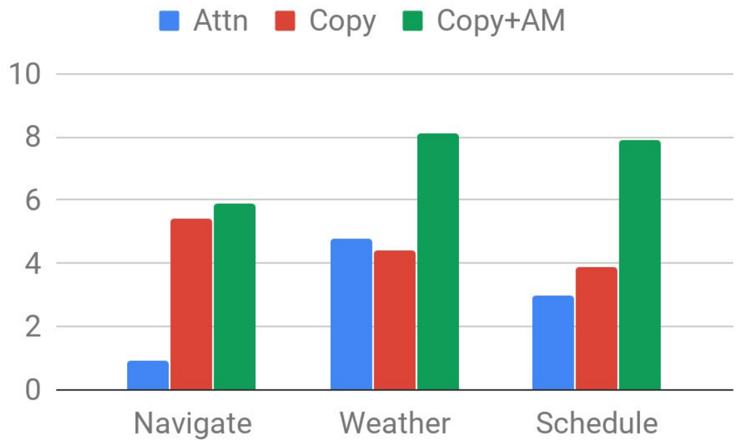
Bad Examples:

- **Ref**: Movie 55 is a good choice.
- **Generated (Copy)**: I would recommend restaurant 55.
- **Ref**: I believe you said comedy movie.
- **Generated (Copy)**: I believe you said comedy food.

What Problem Does AM Solve?



BLEU-4 on SimDial



BLEU-4 on SMD

Answer: AM enables the decoder to generate **overall novel utterances**, not just entities

Examples from SimDial:

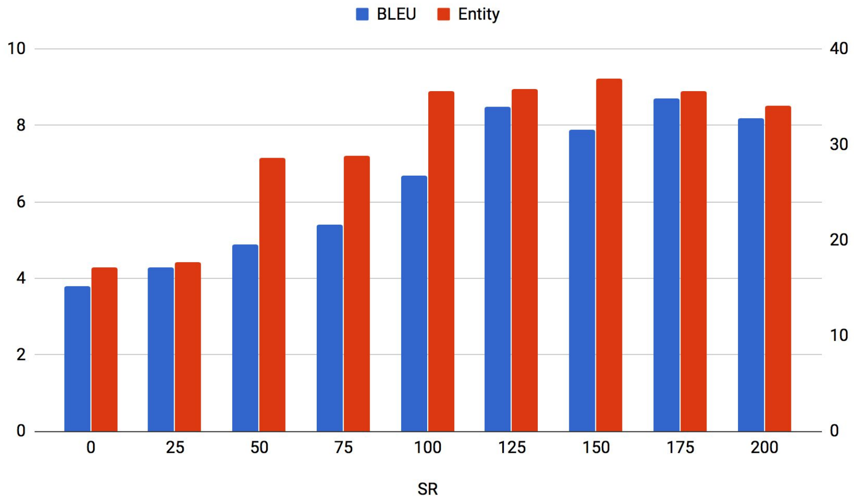
- **Ref**: Movie 55 is a good choice.
- **Generated (Copy+AM)**: Movie 55 is a good choice

Examples from SMD:

- **Ref:** Okay, scheduling Friday dinner with mom at 11 am
- **Generated (Copy+AM)**: scheduling a reminder for dinner on Friday with your 11AM at 10 am

Impact of Seed Response (SR) Size

- Investigate how the size of SR affects the performance of AM algorithm.
- Vary the size of SR from 0 to 200 in the SMD data.
- Use **schedule** as the target domain.



Conclusions

- Better models for latent actions can provide solutions to many problems.
- It’s modeling the turn-level representation for systems (or users)!
- Enable human knowledge to be encoded into turn-level actions.
- Enable knowledge transfers from other domains.
- More to come ...

Thank you!

Code & Data: github.com/snakeztc