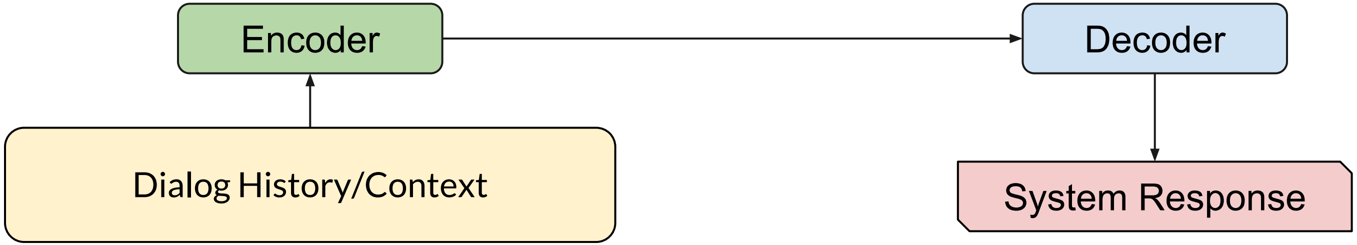
Learning Discourse-level Diversity for Neural Dialog Models Using Conditional Variational Autoencoders

Tiancheng Zhao, Ran Zhao and Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University

Code&Data: https://github.com/snakeztc/NeuralDialog-CVAE

Introduction

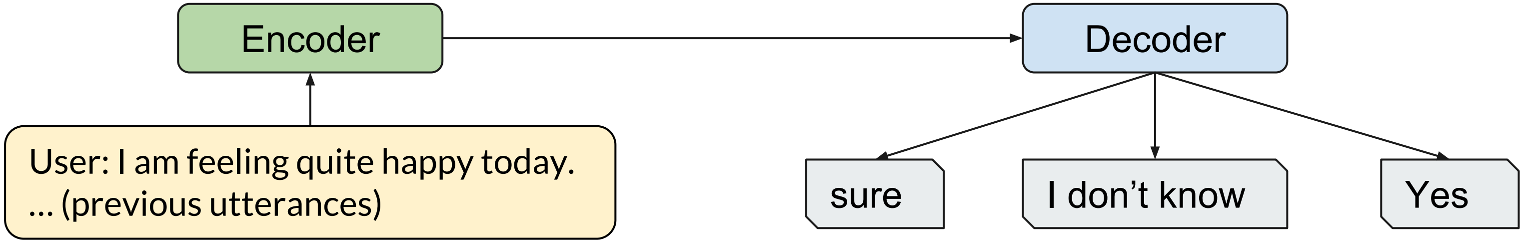
- End-to-end dialog models based on encoder-decoder models have shown great promises for modeling open-domain conversations, due to its flexibility and scalability.



Introduction

However, **dull response problem!** [Li et al 2015, Serban et al. 2016]. Current 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]

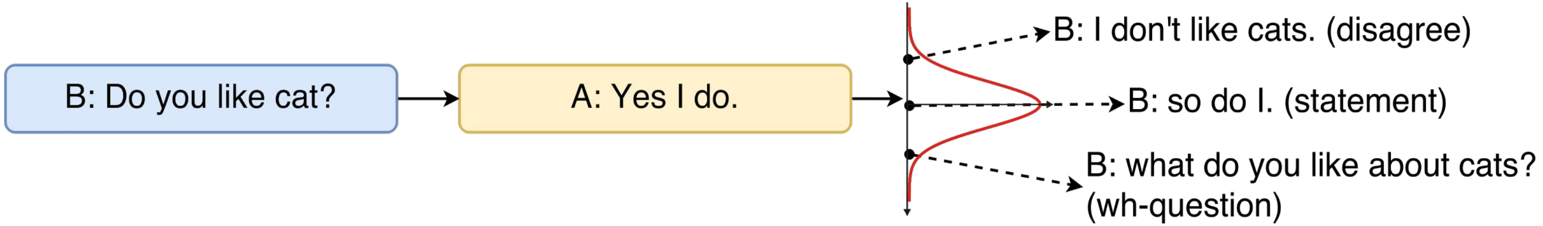


Our Key Insights

- Response generation in conversation is a **ONE-TO-MANY** mapping problem at the **discourse level**.
- A similar dialog context can have many different yet valid responses.
- Learn a **probabilistic distribution** over the valid responses instead of only keep the most likely one.

Our Key Insights

- Response generation in conversation is a **ONE-TO-MANY** mapping problem at the **discourse level**.
    - A similar dialog context can have many different yet valid responses.
- Learn a **probabilistic distribution** over the valid responses instead of only keep the most likely one.

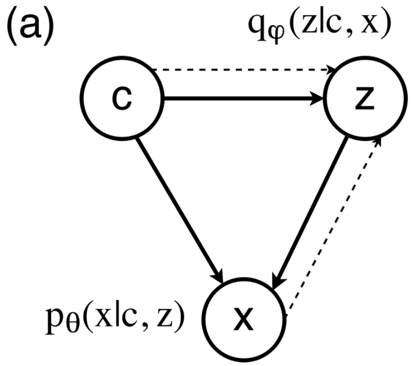


Our Contributions

1. Present an E2E dialog model adapted from Conditional Variational Autoencoder (CVAE).
2. Enable integration of **expert knowledge** via knowledge-guided CVAE**.**
3. Improve the training method of optimizing CVAE/VAE for text generation.

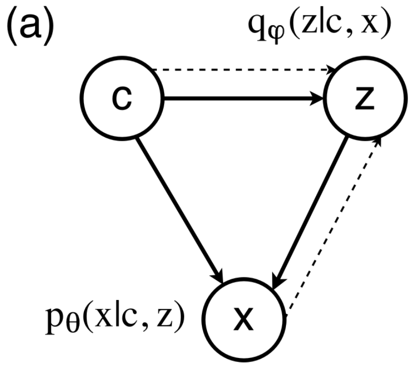
Conditional Variational Auto Encoder (CVAE)

- **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.



Conditional Variational Auto Encoder (CVAE)

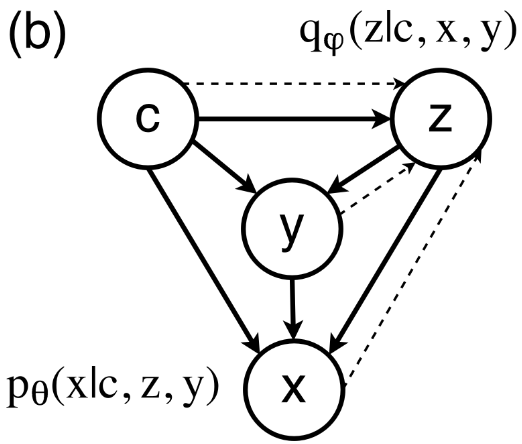
- **C** is dialog context
  - B: Do you like cats? A: Yes I do
- **Z** is the latent variable (gaussian)
- **X** is the next response
  - B: So do I.
- Trained by Stochastic Gradient Variational Bayes (SGVB) [Kingma and Welling 2013]



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

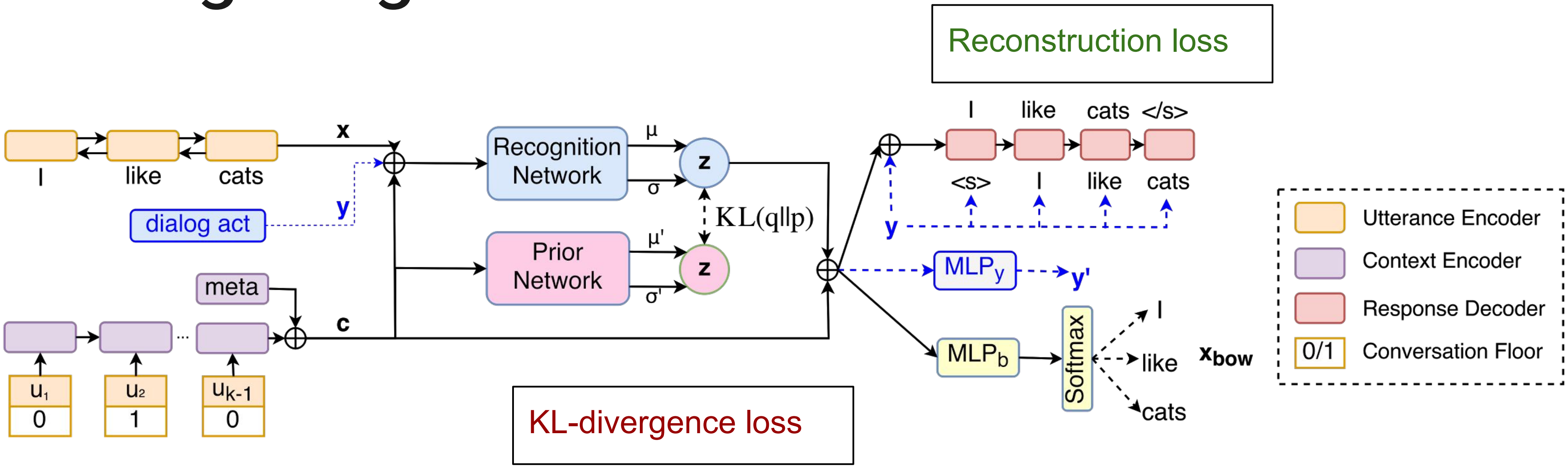
Knowledge-Guided CVAE (kgCVAE)

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

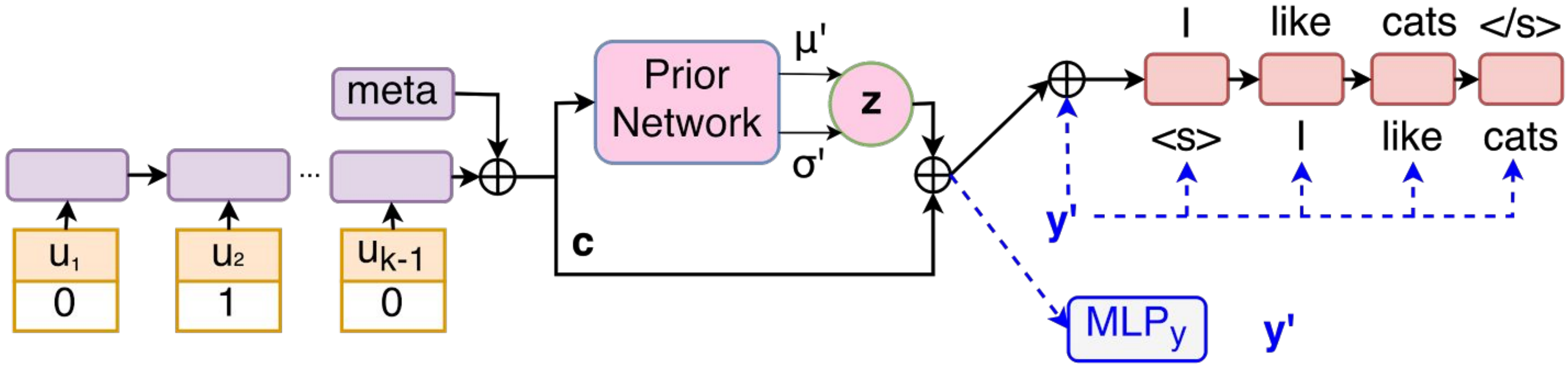


$$\mathcal{L}(\theta, \phi; x, c, y) = -KL(q_\phi(z|x, c, y) \| P_\theta(z|c)) + \mathbf{E}_{q_\phi(z|c,x,y)}[\log p(x|z, c, y)] + \mathbf{E}_{q_\phi(z|c,x,y)}[\log p(y|z, c)] \quad (4)$$

Training of (kg)CVAE



Testing of (kg)CVAE



Optimization Challenge

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

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

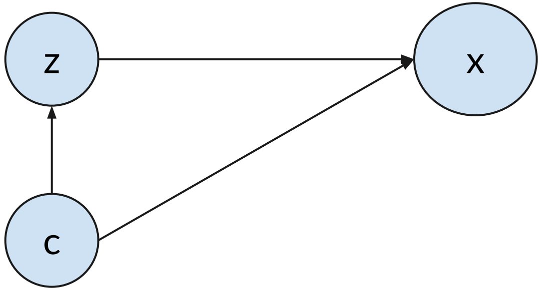
Bowman et al. [2015] described two methods to alleviate the problem :

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.

$$\mathcal{L}'(\theta, \phi; x, c) = \mathcal{L}(\theta, \phi; x, c) + \mathbf{E}_{q_\phi(z|c,x,y)}[\log p(x_{bow}|z, c)] \quad (6)$$



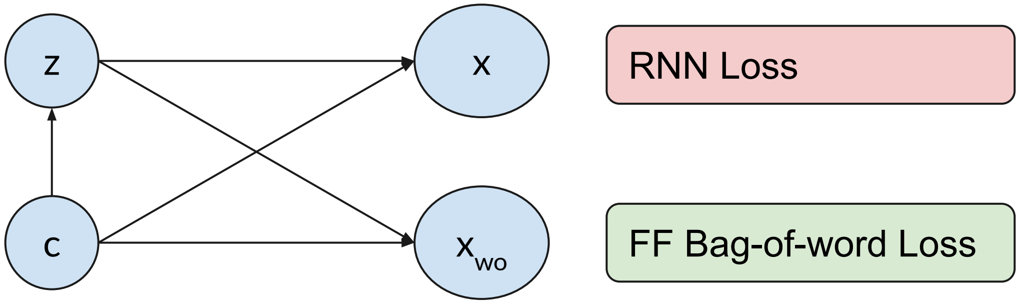
RNN Loss

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.

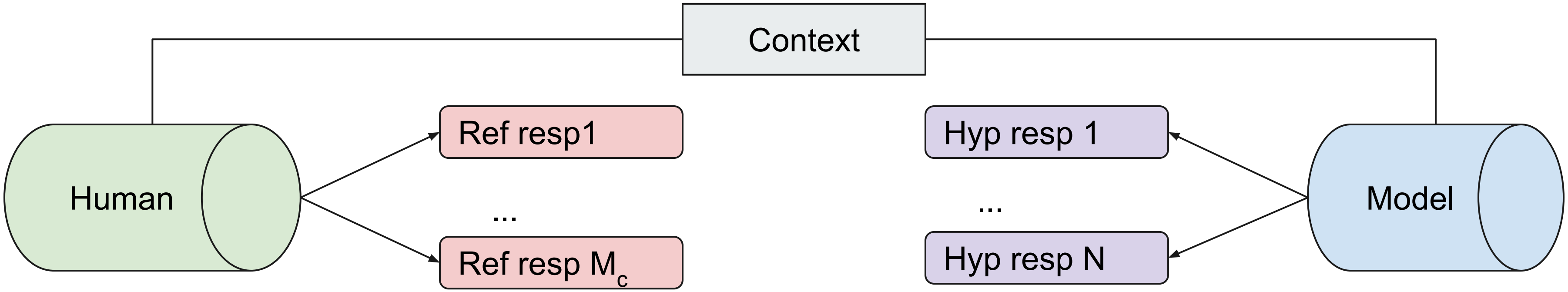
$$\mathcal{L}'(\theta, \phi; x, c) = \mathcal{L}(\theta, \phi; x, c) + \mathbf{E}_{q_\phi(z|c,x,y)}[\log p(x_{bow}|z, c)] \quad (6)$$



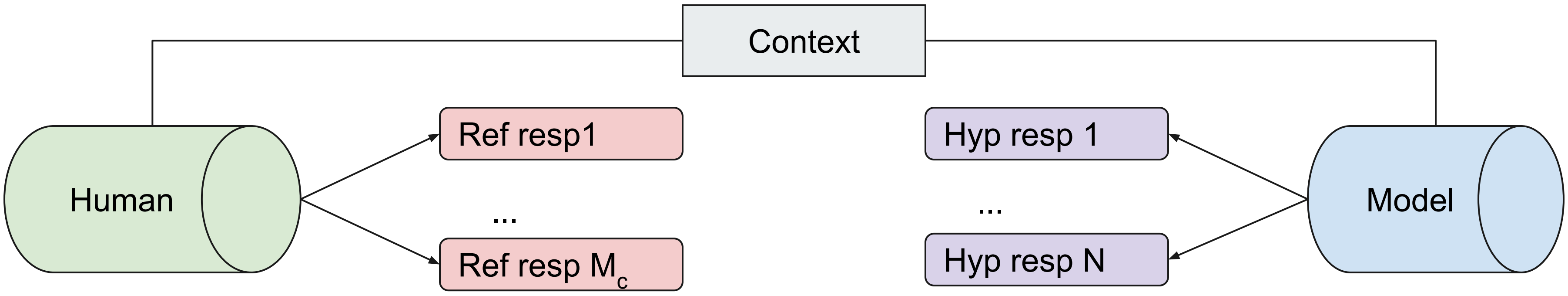
Dataset

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



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.

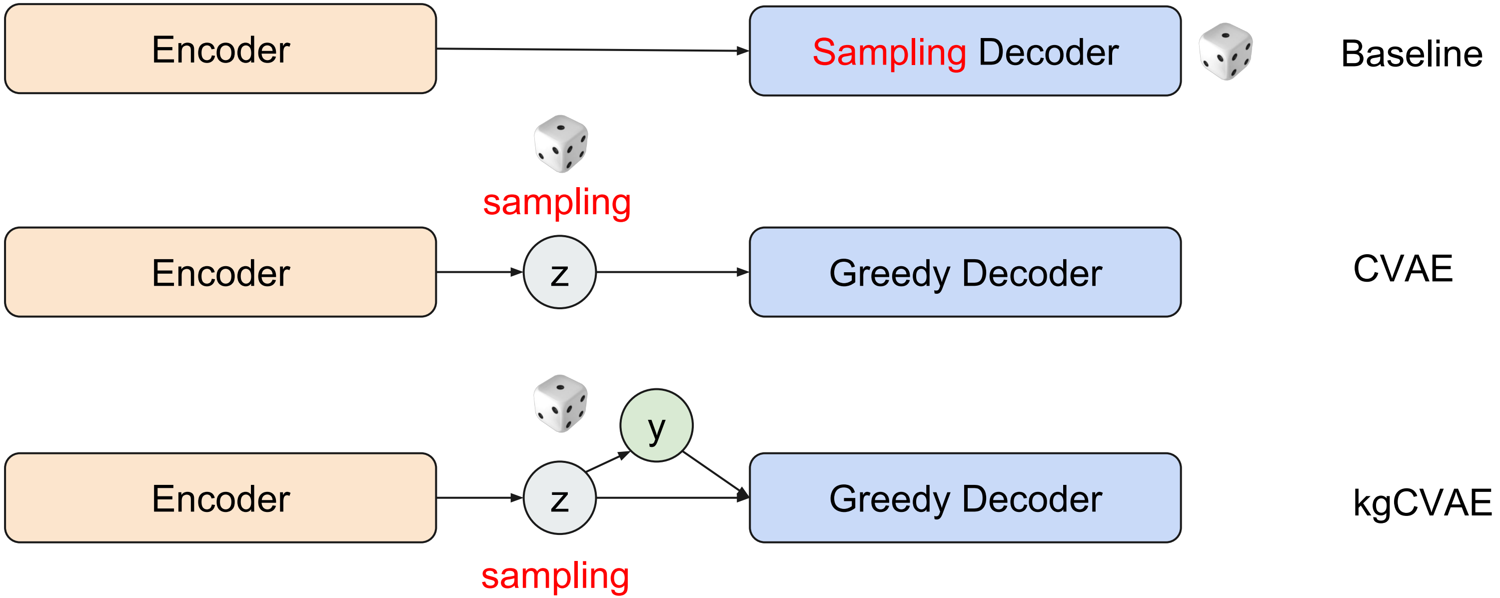
Distance Functions used for Evaluation

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

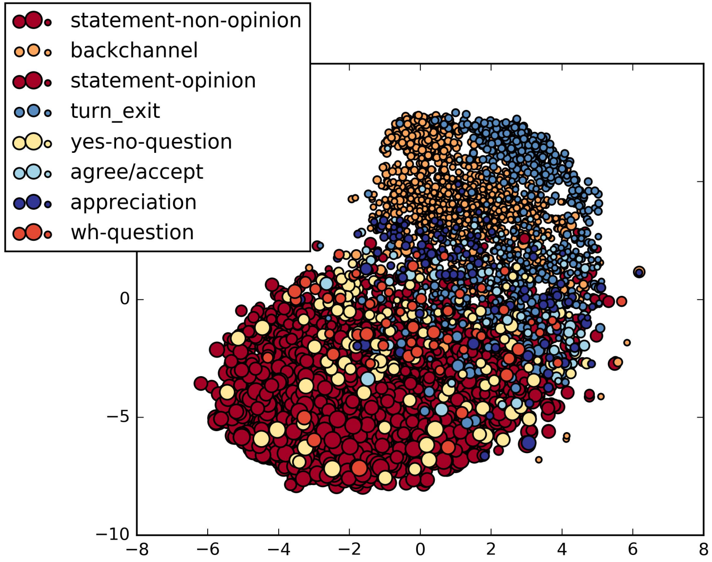
Qualitative Analysis

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



The Effect of BOW Loss

Same setup on PennTree Bank for LM [Bowman 2015]. Compare 4 setups:

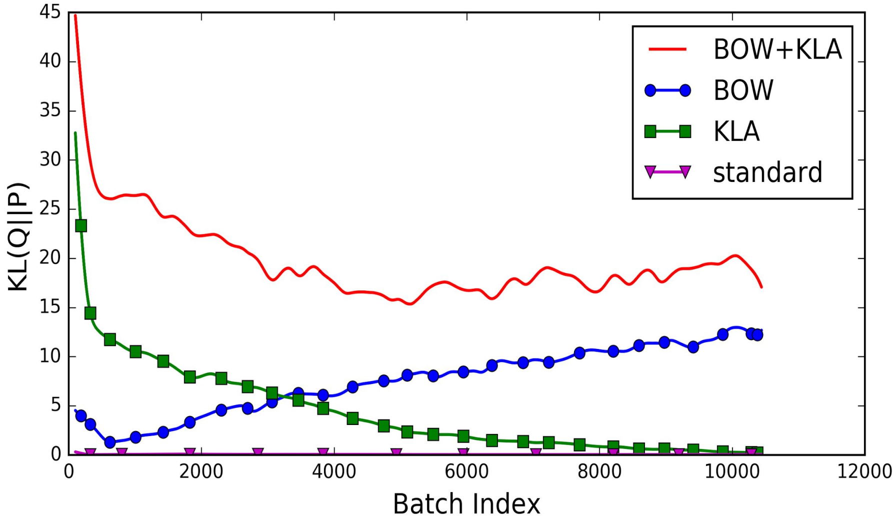
1. Standard VAE
2. KL Annealing (KLA)
3. BOW
4. BOW + KLA

Goal: low reconstruction loss + small but non-trivial KL cost

Model	Perplexity	KL Cost
Standard	122.0	0.05
KLA	111.5	2.02
BOW	97.72	7.41
BOW+KLA	73.04	15.94

KL Cost during Training

- Standard model suffers from *vanishing latent variable.*
- KLA requires *early stopping*.
- BOW leads to stable convergence with/without KLA.
- The same trend is observed on CVAE.



Conclusion and Future Work

- Identify the **ONE-TO-MANY** nature of open-domain dialog modeling
- Propose two novel models based on latent variables models for generating diverse yet appropriate responses.
- Explore further in the direction of leveraging both past linguistic findings and deep models for controllability and explainability.
- Utilize crowdsourcing to yield more robust evaluation.

Code available here! https://github.com/snakeztc/NeuralDialog-CVAE

Thank you!

Questions?

References

1. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*
2. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* .
3. Samuel R Bowman, Luke Vilnis, Oriol Vinyals, An- drew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* .
4. Diederik P Kingma and Max Welling. 2013. Auto- encoding variational bayes. *arXiv preprint arXiv:1312.6114* .
5. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*

Training Details

Word Embedding	200 Glove pre-trained on Twitter
Utterance Encoder Hidden Size	300
Context Encoder Hidden Size	600
Response Decoder Hidden Size	400
Latent Z Size	200
Context Window Size	10 utterances
Optimizer	Adam learning rate=0.001

Testset Creation

- Use 10-nearest neighbour to collect similar context in the training data
- Label a subset of the appropriateness of the 10 responses by 2 human annotators
- bootstrap via SVM on the whole test set (5481 context/response)
- Resulting 6.79 Avg references responses/context
- Distinct reference dialog acts 4.2