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.
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]

User: I am feeling quite happy today.
... (previous utterances)
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 keeping 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]

\[
\mathcal{L}(\theta, \phi; x, c) = -KL(q_\phi(z|x, c) || p_\theta(z|c)) + \mathbb{E}_{q_\phi(z|c, x)}[\log p_\theta(x|z, c)] 
\leq \log p(x|c)
\]
Knowledge-Guided CVAE (kgCVAE)

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

\[
L(\theta, \phi; x, c, y) = -KL(q_\phi(z|x, c, y) || P_\theta(z|c)) + \mathbb{E}_{q_\phi(z|x, c, y)}[\log p(x|z, c, y)] + \mathbb{E}_{q_\phi(z|x, c, y)}[\log p(y|z, c)] \quad (4)
\]
Training of (kg)CVAE

Reconstruction loss

KL-divergence loss
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.

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

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

(6)
## Dataset

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

Human

Ref resp1

Ref resp M_c

Context

... 

Hyp resp 1

Hyp resp N

Model
Quantitative Metrics

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

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

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

\(d(r, h)\) 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)

- **Encoder** → **Sampling Decoder** → Baseline
- **Encoder** → **Greedy Decoder** → CVAE
- **Encoder** → **Greedy Decoder** → kgCVAE

Note: The diagram shows the flow of data through the models, with sampling and greedy decoding steps indicated.
<table>
<thead>
<tr>
<th>Metrics</th>
<th>Perplexity (KL)</th>
<th>BLEU-1 (p/r)</th>
<th>BLEU-2 (p/r)</th>
<th>BLEU-3 (p/r)</th>
<th>BLEU-4 (p/r)</th>
<th>A-bow (p/r)</th>
<th>E-bow (p/r)</th>
<th>DA (p/r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>35.4 (n/a)</td>
<td>0.405/0.336</td>
<td>0.3/0.281</td>
<td>0.272/0.254</td>
<td>0.226/0.215</td>
<td>0.387/0.337</td>
<td>0.701/0.684</td>
<td>0.736/0.514</td>
</tr>
<tr>
<td>CVAE</td>
<td>20.2 (11.36)</td>
<td>0.372/0.381</td>
<td>0.295/0.322</td>
<td>0.265/0.292</td>
<td>0.223/0.248</td>
<td>0.389/0.361</td>
<td>0.705/0.709</td>
<td>0.704/0.604</td>
</tr>
<tr>
<td>kgCVAE</td>
<td>16.02 (13.08)</td>
<td><strong>0.412/0.411</strong></td>
<td><strong>0.350/0.356</strong></td>
<td><strong>0.310/0.318</strong></td>
<td><strong>0.262/0.272</strong></td>
<td>0.373/0.336</td>
<td><strong>0.711/0.712</strong></td>
<td>0.721/0.598</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
<th>KL Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>122.0</td>
<td>0.05</td>
</tr>
<tr>
<td>KLA</td>
<td>111.5</td>
<td>2.02</td>
</tr>
<tr>
<td>BOW</td>
<td>97.72</td>
<td>7.41</td>
</tr>
<tr>
<td>BOW+KLA</td>
<td>73.04</td>
<td>15.94</td>
</tr>
</tbody>
</table>
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?


## Training Details

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Embedding</td>
<td>200 Glove pre-trained on Twitter</td>
</tr>
<tr>
<td>Utterance Encoder Hidden Size</td>
<td>300</td>
</tr>
<tr>
<td>Context Encoder Hidden Size</td>
<td>600</td>
</tr>
<tr>
<td>Response Decoder Hidden Size</td>
<td>400</td>
</tr>
<tr>
<td>Latent Z Size</td>
<td>200</td>
</tr>
<tr>
<td>Context Window Size</td>
<td>10 utterances</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam learning rate=0.001</td>
</tr>
</tbody>
</table>
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