Reflecting on Experiences for Response Generation

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

Suyu Liu  
Singapore Management University  
liusuyu4@gmail.com

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

Tat-Seng Chua  
Sea-NExT Joint Lab, National University of Singapore  
dcscets@nus.edu.sg
Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

I recommend the botanic. The food there is nice. Here are some pictures.
Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

That sounds good.

Challenges

- **Context-specific responses**
- **Towards user-specific requests**
- Emphasize the context-response mapping over the whole training corpus
- Tend to assign high probabilities to safe but universal responses (Li et al., 2016)
Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

I recommend the botanic. The food there is nice. Here are some pictures.

• **Context-specific responses**

• **Coordination between the different modalities**

• **Coordinate** multimodal response components

• **Treat different modalities separately**
Multimodal Dialogue Systems - Challenges

Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

I recommend the botanic. The food there is nice. Here are some pictures.

Challenges

• Context-specific responses
• Coordination between the different modalities
• Explainability
• Generalizability

• Why?

• If we want to improve the responses for certain dialogue situation

-> Retraining & Catastrophic Forgetting
Case-based Reasoning (CBR)

Observation:
Humans often solve a new problem by recollecting and adapting the solutions to multiple related problems that they encountered in the past. (Ross, 1984)

A typical sketch of a CBR system:
- Retrieve
- Reuse
- Revise (if needed)
CBR – recent application in research

- **Knowledge-based question answering**
  - A Simple Approach to Case-Based Reasoning in Knowledge Base (Das et al., 2020)
  - CBR-KBQA (Das et al, 2021, EMNLP)

- **Natural language modeling**
  - kNN-LM (Khandelwal, 2020, ICLR)

- **Machine translation**
  - kNN-MT (Khandelwal, 2021, ICLR)
  - Adaptive kNN-MT (Zheng, 2021, ACL-IJCNLP)

- **Multimodal dialogue response generation**
  - structured triple queries / complex dialogue queries / multimodal cases
  - pure textual sequences
Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

I recommend the botanic. The food there is nice. Here are some pictures.

Please reserve a place for me to try jamon serrano.

Sure, which region do you prefer?

Should be in downtown core. I am with a group of four people. Don't want to go far.

Got it. The botanic is good. Great food with imaginative options, nice service, great ambiance, good wine selection by glass and bottle. Try the anchovies on toast.
RERG

• A neural components based CBR framework to Reflect on Experiences for Response Generation

**Retrieval Module**

• Intra-modality Contrastive Learning
• Case-level Triplet Ranking

**Reuse Module**

• Reuse for Text Response – Cross Copy
• Reuse for Image Response

Hi, I am traveling with a group of four people in downtown. We would like to try some jamon serrano like this...

I recommend the botanic. Great food there. Try the anchovies on toast also.
**Textual Contrastive Learning** for text context embedding $s_i$:

- Encode text context $C_i^S$ with BERT as $s_i^{z_i} = f_s(C_i^S, z_i; \theta_s)$, where $z^i$ is a random mask for dropout.

- Similar to SimCSE (Gao et al., 2021), the key here is to get a positive embedding pair with different dropout masks $z^i$ and $z^j$.

- Training objective inside a minibatch of size $N'$ is:

$$L_{textual} = -\log \frac{\exp(s_i^{z_i} \cdot s_i^{z_i'}/\tau)}{\sum_{j=0}^{N'} \exp(s_i^{z_i} \cdot s_j^{'z_j}/\tau)}$$
Visual Contrastive Learning for image context embedding $q_i$:

- Augment the context image $C_i^l$ to $C_i^{l'}$ and $C_i^{l''}$.

- Following MoCo-v2 (Chen et al., 2020), embed them into query and key feature vectors by two Resnet network:

  $$q_i = f_q(C_i^{l'}; \theta_q)$$
  $$k_i = f_k(C_i^{l''}; \theta_k).$$

- Training objective with the momentum queue is:

  $$L_{visual} = -\log \frac{\exp(q_i \cdot k_i^{+}/\tau)}{\sum_{j=0}^M \exp(q_i \cdot k_i^j/\tau)}$$
RERG – Retrieval Module – Case-level

- **Triplet Ranking** for multimodal context representation $c_i$:
  - **Anchor**: Embed current dialogue context to $c_i$ as $c_i = f_{MLP}([s_i; q_i])$
  - **Positive**: Encode similar dialogue case’s context to $c_i^+$
  - **Negative**: Select the **batch-hardest case** embedding $c_i^-$
  - Triplet ranking loss with margin $\epsilon$ and **dot product** as the similarity function is:
    $$L_{triplet} = \max(0, \epsilon - \text{sim}(c_i, c_i^+) + \text{sim}(c_i, c_i^-)).$$
RERG – Retrieval Module

- Intra-modality Contrastive Learning: Capture **intrinsic patterns** of text and image contexts
- Case-level Triplet Ranking: learn semantic information from **higher-level similarity**

- Thus, the **retrieval module** is trained via the **total loss** as:

\[
L_{\text{retrieval}} = L_{\text{textual}} + \lambda_1 \cdot L_{\text{visual}} + \lambda_2 \cdot L_{\text{triplet}}
\]

---

Hi, I am traveling with a group of four people in downtown. We would like to try some jamon serrano ...
Current Dialogue

Hi! Can you help me look for a moderate price restaurant to try jamon serrano?

Anywhere in the region in particular?

Yes, the downtown core please.

I would recommend the botanic! They have a lot of great jamon serrano that you can try out along with other dishes.

Experienced Response

Current Context

Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

Target Response

I recommend the botanic. The food there is nice. Here are some pictures.

Experienced Response

Please reserve a place for me to try jamon serrano.

Sure, which region do you prefer?

Should be in downtown core. I am with a group of four people. Don't want to go far.

Got it. The botanic is good. Great food with imaginative options, nice service, great ambience, good wine selection by glass and bottle. Try the anchovies on toast.

Cross Copy

- Vallina Generation
- Vertical Copy
- Horizontal Copy
RERG – Reuse Module – Text Response

- Reuse for Text Response – Cross Copy
  - Vallina generation process: a typical encoder-decoder network from context to response
  - Vertical Copy from current dialogue context
  - Horizontal Copy from the similar contexts’ responses

\[ \alpha = \text{sigmoid}(W_2 \cdot [h_t^{\text{dec}}; w_{t-1}; P_t^{\text{vertical}} \cdot H^{CS}_R]) \]

\[ \beta_k = \text{sigmoid}(W_3 \cdot [h_t^{\text{dec}}; w_{t-1}; P_t^{\text{horizontal}} \cdot H^{R_k}_S]) \]

Hi, I am traveling with a group of four people in downtown. We would like to try some jamon serrano ...

I recommend the botanic. Great food there. Try the anchovies on toast also.

\[ H^{CS} = [v_1, \ldots, v_{|CS|}] \]
RERG – Reuse Module – Image Response

• Reuse for Image Response
  o Weighted query vector $q_{sum}$ by text context similarity $\gamma_k$
  o Rank all images in datastore by the cosine similarity with $q_{sum}$, select the top-ranked image

Hi, I am traveling with a group of four people in downtown. We would like to try some jamon serrano ...

Bi-GRU  \[ H^c_s = [v_1, ..., v_{|c_s|}] \]

Attention  

Retrieval

Top-k $\text{sim}(c_i, c_j)$

$\gamma_k$

$h^c_s$

$h^c_k$

$R^I_k$

Query Image Encoder

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$q_1$

$q_j$

Rank using cosine similarity $q_{sum} = \sum_{k=0}^{K} \gamma_k \cdot q_k$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$

$\cdot$
Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

I recommend the botanic! They have a lot of great jamon serrano that you can try out along with other dishes.

Got it. The botanic is good. Great food with imaginative options, nice service, great ambience, good wine selection by glass and bottle. Try the anchovies on toast.
RERG – Explainability

Ground-truth Response

You can try going to Pizza Express.

Generated Response

You can try going to Pizza Express. Someone has said: Great pizza and outdoor seating!

Context

Usr: I am thinking of a pizza place that has outdoor seating. 
Sys: Is there anything else you would like?
Usr: I would like if they have moderate prices and if they accept credit cards.

Experienced Response

You can try going to Pizza Express. Someone has said: Great pizza, great service, tables to eat outside, perfect on a weekday evening after work. And I would recommend the pizza like in the picture. 

I would recommend their pizza like in the picture.

Noted! In this case, I would recommend Pizza Express.
**RERG – Performance**

**Dataset:** MMConv (Liao et al, 2021), a multi-turn multimodal conversation dataset.  
- 5,106 dialogues, 5 domains, 39.8K utterances  
- Grounded on a venue database with 1,771 venues and 113,953 associated images

**Main baseline model:** employ the large pretrained GPT-2 model

### Main multimodal response generation results

<table>
<thead>
<tr>
<th>Group</th>
<th>Method</th>
<th>Textual Response</th>
<th>Image Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>Text-based</td>
<td>DialoGPT [46]</td>
<td>18.32</td>
<td>3.160</td>
</tr>
<tr>
<td></td>
<td>LaRL [50]</td>
<td>13.33</td>
<td>2.496</td>
</tr>
<tr>
<td></td>
<td>HDNO [40]</td>
<td>14.79</td>
<td>2.745</td>
</tr>
<tr>
<td>Multimodal</td>
<td>MMD [37] MMConv [23]</td>
<td>16.60</td>
<td>3.062</td>
</tr>
<tr>
<td></td>
<td>RERG_5</td>
<td>32.33</td>
<td>5.758</td>
</tr>
<tr>
<td></td>
<td>RERG_{gtk=5}</td>
<td>30.75</td>
<td>5.616</td>
</tr>
<tr>
<td></td>
<td>RERG_2</td>
<td>31.17</td>
<td>5.529</td>
</tr>
<tr>
<td></td>
<td>RERG_{-10}</td>
<td>29.66</td>
<td>5.374</td>
</tr>
<tr>
<td></td>
<td>RERG_{-10}</td>
<td>27.72</td>
<td>5.345</td>
</tr>
</tbody>
</table>

**Natural pattern** | **Targeted response** | **Related images**
RERG – Generalizability

Study on Unseen situations

- Consider all dialogues happened under the user goal $\pi$: "You plan to do shopping in Jurong East. Thus seek about shopping malls there (Westgate, Jem, and IMM)".

Existing: Time-consuming re-training or finetuning process to handle unseen situations. Such costly process may also lead to catastrophic forgetting.

RERG: A computationally much cheaper way:
- add similar cases to the retrieve datastore
- let the reuse module to construct responses with new top-k cases.
## RERG – Generalizability

### Study on Unseen situations

- Consider all dialogues happened under the **user goal** $\pi$: “You plan to do shopping in Jurong East. Thus seek about shopping malls there (Westgate, Jem, and IMM)”.

### Entity F1 score: task completion

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario</th>
<th>Remaining</th>
<th>Held-out</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMConv</td>
<td>Train on original cases</td>
<td>49.07</td>
<td>11.54</td>
</tr>
<tr>
<td></td>
<td>+ Fine-tune on additional cases</td>
<td>44.06</td>
<td>69.23</td>
</tr>
<tr>
<td></td>
<td>+ Fine-tune on all cases</td>
<td>47.39</td>
<td>57.69</td>
</tr>
<tr>
<td>RERG</td>
<td>Train on original cases</td>
<td>49.55</td>
<td>11.54</td>
</tr>
<tr>
<td></td>
<td>+ Add back to retrieve datastore</td>
<td>49.55</td>
<td>65.38</td>
</tr>
</tbody>
</table>

- **Catastrophic forgetting** without retraining
**RERG**: A neural case-based reasoning framework to reflect on experiences for multimodal response generation

- Context-specific response to fulfill user requests
- Coordination between modalities
- Explainability
- Generalizability

Thank you for listening!

Email: chenchenye.ccye@gmail.com