mmfp2437

Reflecting on Experiences for Response Generation

Chenchen Ye*#, Lizi Liao$, Suyu Liu$, Tat−Seng Chua*#

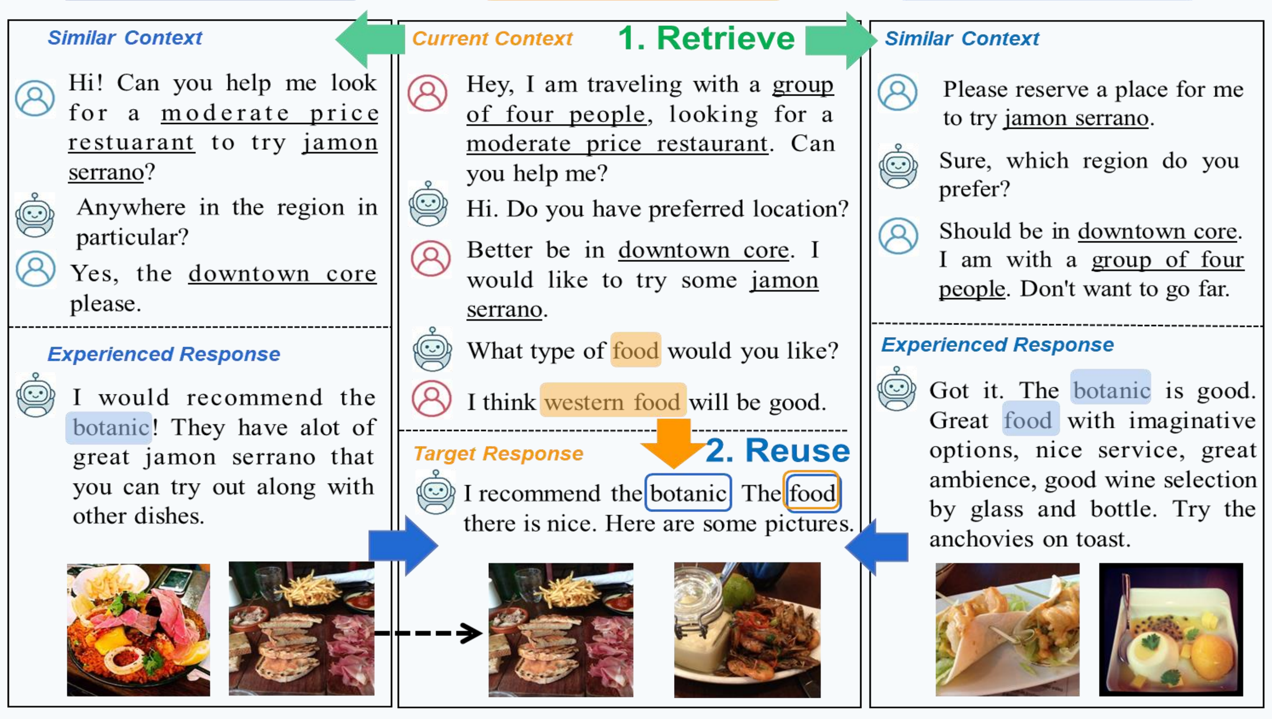
* National University of Singapore; # Sea-NExT Joint Lab; $ Singapore Management University

chenchenye.ccye@gmail.com, lzliao@smu.edu.sg, suyuliu2022@phdcs.smu.edu.sg, dcscts@nus.edu.sg

Abstract

- Multimodal dialogue systems face the following challenges:
  1. Automatically generate **context-specific responses** instead of safe but general responses;
  2. Naturally coordinate between **different information modalities**;
  3. Intuitively **explain the reasons** for generated responses and improve a specific response **without re-training whole model**.
- We propose a neural **case-based reasoning** framework to **reflect on experiences for multimodal response generation (RERG)**, which consists of two modules:
  1. A multimodal contrastive learning enhanced **retrieval model** for soliciting similar dialogue instances;
  2. A cross copy based **reuse model** to explore the current dialogue context (*vertical*) and similar dialogue instances' responses (*horizontal*) for response generation simultaneously.
- Extensive experiments validate the superiority of RERG on the mentioned challenges.

Method



- **Retrieval Module**
  - **Textual Contrastive Learning: follows SimCSE [1]**

$$L_{textual} = -log \frac{exp(s_i^{z_i} \cdot s_i^{z_i'}/\tau)}{\sum_{j=0}^{N'} exp(s_i^{z_i} \cdot s_j^{z_j'}/\tau)}$$

s_i, s_j are the encoded text contexts, z_i, z_i', z_j are different dropout masks.

  - **Visual Contrastive Learning: follows MoCo-v2 [2]**

$$L_{visual} = -log \frac{exp(q_i \cdot k_i^{+}/\tau)}{\sum_{j=0}^{M} exp(q_i \cdot k_j^{i}/\tau)}$$

q_i, k_i^j are the encoded image contexts from query and key encoders; q_i, k_i^{+} augment from the same image.

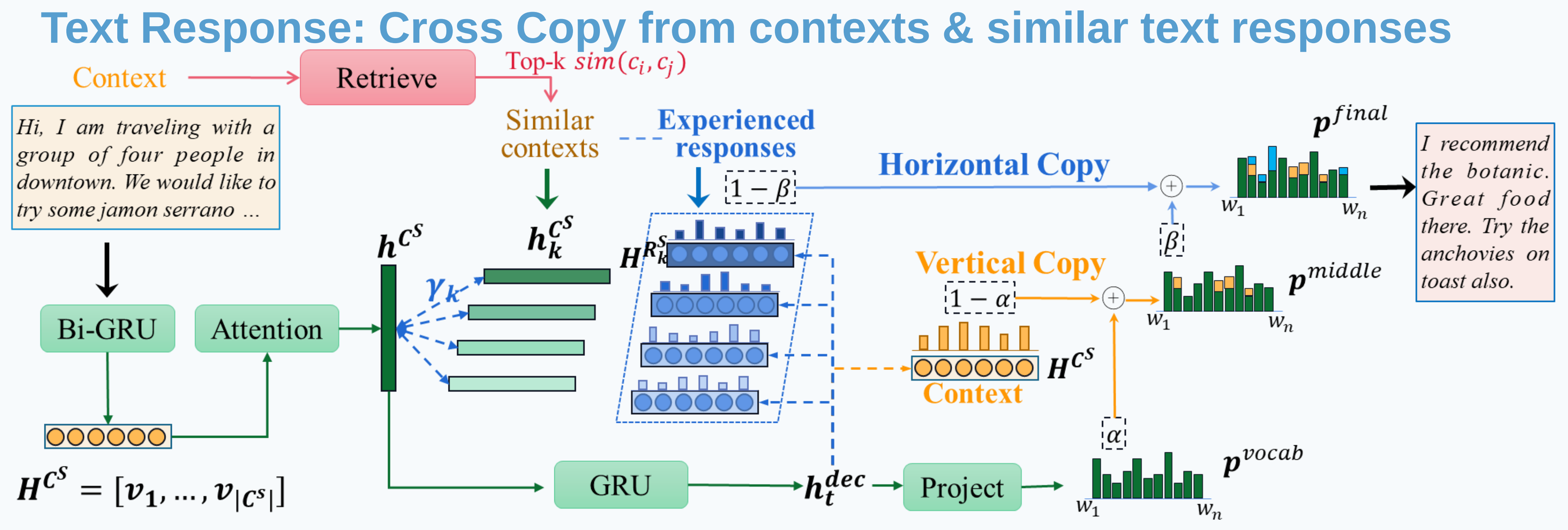
  - **Case-level Triplet Ranking**

$$c_i = f_{MLP}([s_i; q_i])$$

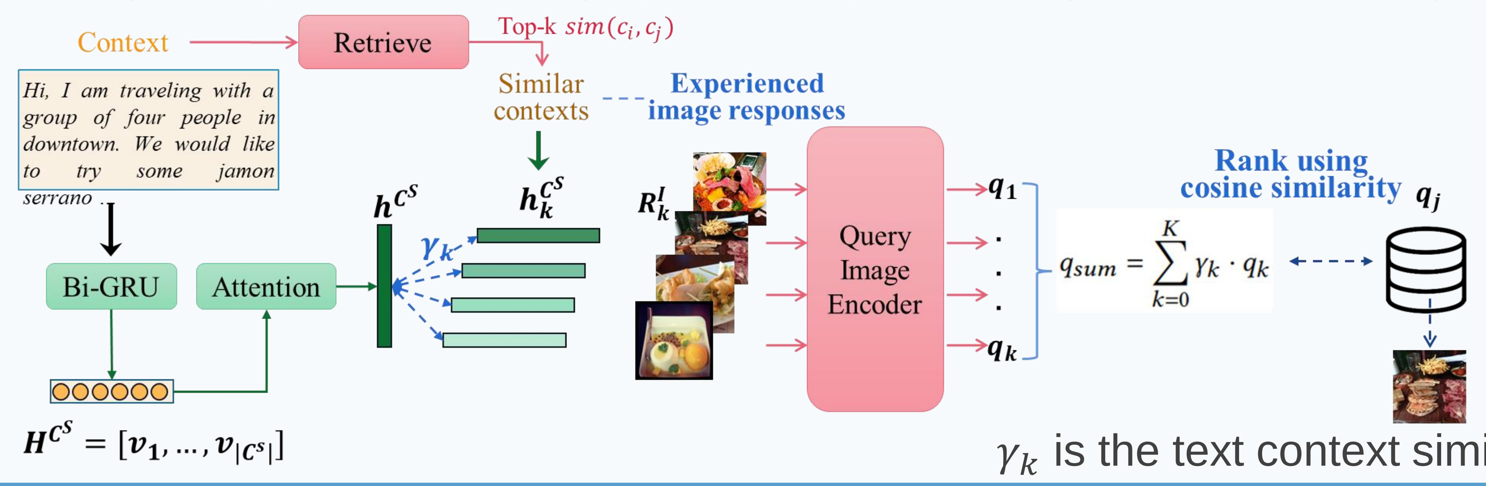
$$L_{triplet} = max(0, \epsilon - sim(c_i, c_i^{+}) + sim(c_i, c_i^{-})).$$

c_i^{+} is the ground-truth similar cases; c_i^{-} is the batch-hardest cases; sim function is dot product.

- **Reuse Module**
  - **Text Response: Cross Copy from contexts & similar text responses**



  - **Image Response: Ranking by similarity with weighted similar images**



γ_k is the text context similarity.

Experiments

- **Dataset**
  - ✓ MMConv [3]
- **Evaluation Metric**
  - ✓ Text response: BLEU, NIST, ROUGE-L, Entity F1 and Match Rate
  - ✓ Image response: Recall@k
- **Baselines**
  - ✓ DialoGPT [4]; LaRL [5]; HDNO [6]; MMD [7]; MMConv [3].
- **Results**

Group	Method	Textual Response					Image Response		
		BLEU	NIST	ROUGE-L	Entity-F1	Match Rate	Recall@1	Recall@3	Recall@5
Text-based	DialoGPT [46]	18.32	3.160	0.4419	18.89	24.7	–	–	–
	LaRL [50]	13.33	2.496	0.3214	5.36	1.5	–	–	–
	HDNO [40]	14.79	2.745	0.3663	8.23	2.3	–	–	–
Multimodal	MMD [37]	16.60	3.062	0.3728	11.08	5.1	4.69	8.33	11.98
	MMConv [23]	32.33	5.758	0.5402	49.01	69.2	17.85	–	–
	RERG_5	30.75	5.616	0.5585	52.55	79.3	22.83	24.88	26.33
	RERG_$gt_{k=5}$	31.17	5.529	0.5776	54.36	80.6	23.43	25.60	36.57
	RERG_2	29.66	5.374	0.5591	51.55	81.9	33.94	35.51	36.23
	RERG_10	27.72	5.345	0.5322	46.69	69.8	14.37	16.67	17.75

Observations:
1. RERG achieves leading ROUGH-L performance, indicating that it learns **useful natural language patterns** from contexts and similar responses.
2. RERG outperforms largely on Entity F1 & Match Rate, indicating richer and more accurate entity and venue information for **specific user requests**.
3. RERG also leads image recall, indicating that **more relevant images** are provided and a **better coordination between modalities** is achieved.

- **Explainability**

RERG could explain how each component (**vallina decoding** in green, **vertical copy from context** in orange, and **horizontal copy from similar cases** in blue) contributes to the prediction probability of each generated word.

- **Study on Unseen Situations**
  - **Experiment setting**: Split training & testing set to dialogues that happen under a user goal π (additional training cases & held-out test set) and those happen under other goals (new training set & remaining test set). Evaluate task completion by Entity F1.

Method	Scenario	Remaining	Held-out
MMConv	Train on original cases	49.07	11.54
	+ Fine-tune on additional cases	44.06	69.23
	+ Fine-tune on all cases	47.39	57.69
RERG	Train on original cases	49.55	11.54
	+ Add back to retrieve datastore	49.55	65.38

Existing models requires **time-consuming retraining** and suffer form the problem of **catastrophic forgetting**.

To handle unseen situations, RERG provides a **computationally much cheaper** way: just need **add few similar cases** into the **retrieve datastore**, and then let the **reuse** module to construct response **with the new top-ranked cases**.

Conclusion

- **Main contributions**
  - ✓ Propose a neural case based reasoning framework to reuse context and retrieved experiences for multimodal response generation.
  - ✓ Generate more context-specific responses to fulfill user requests.
  - ✓ Achieve better coordination between text and image modalities.
  - ✓ Show explainability and generalizability of proposed model.
- **Future work**
  - ✓ Explore avenues for end-to-end learning for case based reasoning.
  - ✓ Improve the strategy planning part in handling dialogue situations that require consecutive turns of actions.

Reference

[1] Gao *et al.* 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In EMNLP, 6894-6910.
[2] He *et al.* 2020. Momentum contrast for unsupervised visual representation learning. In CVPR, 9729-9738.
[3] Liao *et al.* 2021. MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. In SIGIR. 605-612.
[4] Zhang *et al.* 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In ACL,270-278.
[5] Zhao *et al.* 2019. Rethinking Action Spaces for Reinforcement Learning in End-to-end Agents with Latent Variable Models. In ACL, 1208-1218.
[6] Wang *et al.* 2020. Modelling Hierarchical Structure between Dialogue Policy and Natural Language Generator with Option Framework for Task-oriented Dialogue System. In ICLR.
[7] Shubham *et al.* 2018. A Knowledge Grounded Multimodal Search-Based Conversational Agent. In SCAI@EMNLP.

A Joint Research Collaboration between
NUS National University of Singapore · Tsinghua University · University of Southampton
Supported by
National Research Foundation
Prime Minister's Office
Singapore

Sea-NExT Joint Lab
NATIONAL UNIVERSITY OF SINGAPORE
13 Computing Drive, Singapore 117417