PARL: A Dialog System Framework with Prompts as Actions for Reinforcement Learning

Tao Xiang[a], Yangzhe Li, Monika Wintergerst[b], Ana Pecini, Dominika Młynarczyk and Georg Groh[c]

Department of Informatics, Technical University of Munich, Munich, Germany

Keywords: Open-Domain Dialog Systems, Prompting, Reinforcement Learning, Conversational AI.

Abstract: The performance of most current open-domain dialog systems is limited by the (training) dialog corpora due to either generation-based or retrieval-based learning patterns. To circumvent this limitation, we propose PARL, an open-domain dialog system framework using **P**rompts as **A**ctions for **R**einforcement **L**earning. This framework requires a (fixed) open-domain dialog system as the backbone and trains a behavior policy using reinforcement learning to guide the backbone system to respond appropriately with respect to a given conversation. The action space is defined as a finite set of behaviors in the form of natural language prompts. Preliminary results show that with the guidance of the behavior policy, the backbone system could generate more engaging and empathetic responses.

1 INTRODUCTION

Open-domain dialog systems are a popular natural language processing (NLP) task because of their potential in real-life applications, such as Google Meena (Adiwardana et al., 2020) or Facebook Blenderbot (Roller et al., 2021). Current methods for open-domain dialog systems can be generally categorized into retrieval-based and generation-based methods, where both require high-quality dialog corpora: Retrieval-based systems need a pre-collected paired conversation dataset for retrieving responses, while generation-based systems need a large amount of training data for supervised learning (Ni et al., 2022). Therefore, the performance of dialog systems depends heavily on the quality of the dialog corpus and in theory, it is difficult for dialog systems trained with these methods to exceed the quality of the training set.

Inspired by reinforcement learning (RL) applications surpassing human performance, such as AlphaGo Zero (Silver et al., 2017), our research objective in this work is to explore whether RL can further improve the performance of dialog systems in order to outperform training set level or even reach a human-like quality. To this end, we train a behavior policy that decides which system action to perform according to the current dialog history with RL. The system actions are defined as general human behaviors during one-on-one conversations in the form of natural language prompts, such as "greeting the other" or "comforting the other". After the system action is confirmed, it is fed together with the dialog history to a fixed dialog system, which then generates a response.

In this position paper, we introduce PARL, an open-domain dialog system framework using **P**rompts as **A**ctions for **R**einforcement **L**earning [1]. It is considered a framework because the definition of the action space, the backbone dialog system, and the training of the policy network can all be modified in future work. Its general pipeline is given as follows:

1. Define the actions as natural language prompts. An example is "comfort me".
2. Train or use a pre-trained open-domain dialog system as the fixed backbone.
3. Train a policy network that maps dialog history to actions with reinforcement learning.
4. Feed the dialog history and action prompt to the backbone to generate responses.

[a] https://orcid.org/0000-0001-6217-6560
[b] https://orcid.org/0000-0002-9244-5431
[c] https://orcid.org/0000-0002-5942-2297

[1] Code and models: https://github.com/TUM-NLPLab-2022/PARL-A-Dialog-System-Framework-with-Prompts-as-Actions-for-Reinforcement-Learning

Xiang, T., Li, Y., Wintergerst, M., Pecini, A., Młynarczyk, D. and Groh, G.
PARL: A Dialog System Framework with Prompts as Actions for Reinforcement Learning.
DOI: 10.5220/0011725200003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 633-640
ISBN: 978-989-758-623-1; ISSN: 2184-433X

The remainder of this paper is organized as follows: In section 2 we review related work in recent years, in section 3 we dive into details of our framework, in section 4 we introduce the evaluation methods used in this work, in section 5 we demonstrate and analyze the evaluation results, in section 6 we highlight advantages and limitations of the proposed framework, and in section 7 we conclude our work and discuss directions for future work.

2 RELATED WORK

2.1 Open-Domain Dialog Systems

In the past few years, the area of open-domain dialog systems has achieved significant progress with the development of deep learning. Typically, deep learning methods for open-domain dialog systems can be categorized into retrieval-based and generation-based approaches.

A retrieval-based open-domain dialog system matches user utterances with present queries in a pre-collected human conversation dataset and retrieves responses of similar queries as candidate responses. Then a scoring algorithm scores these candidates and the response with the highest score is selected. Recent work on retrieval-based systems includes (Zhou et al., 2016), (Zhou et al., 2018), and (Gu et al., 2020). One drawback of retrieval-based systems is their dependence on the pre-collected dataset, which is difficult to construct. Additionally, the pre-existing responses can only cover a limited scope of conversations. This poses limits to real-world open-domain conversations, which include an arbitrarily wide range of topics.

Generation-based dialog systems, on the other hand, possess the potential of generating unseen responses. Recent work on generation-based systems focuses on fine-tuning pre-trained language models on dialog datasets (Saleh et al., 2020; Wolf et al., 2019; Zhang et al., 2020; Adiwardana et al., 2020). While generation-based dialog systems alleviate the limited scope problem of retrieval-based systems, their performance still heavily depends on the quality of training corpora.

2.2 RL for Dialog Systems

Compared to supervised learning, reinforcement learning in the area of open-domain dialog systems is still in the exploratory stage. One popular RL research direction is to optimize a dialog system pre-trained with supervised learning. For example, (Jaques et al., 2019) optimize for sentiment and several other conversation metrics by learning from a static batch of human-bot conversations using Batch RL. (Saleh et al., 2020) propose using RL to reduce toxicity in an open-domain dialog setting in order to ensure the model produces more appropriate and safe conversations. In these settings, the action space is usually infinite with actions being system responses of various lengths. In contrast, (Xu et al., 2018) explicitly define an action space consisting of dialog acts that represent human behaviors during conversation and train a policy model that decides appropriate dialog acts with respect to dialog history.

2.3 Prompting Language Models

Prompting, which can for instance be used to steer multi-task generalist agents (Reed et al., 2022), has recently also been explored as a way to enhance the performance of language models. (Radford et al., 2019) employ prompts to guide zero-shot generation for tasks such as translation. (Raffel et al., 2020) use task-specific prefixes as prompts in their text-to-text framework for various NLP tasks. (Lee et al., 2021) use natural language descriptions for requested domains and slots as prompts to guide the generation of a slot value for the requested domain and slot in the dialog state tracking task.

Inspired by this body of work, we propose to use natural language prompts as actions in RL to guide the backbone dialog system to behave accordingly. We explicitly define an action space similarly to (Xu et al., 2018), but with actions as natural language prompts. To the best of our knowledge, we are the first to propose using natural language prompts as actions in RL for optimizing dialog systems.

3 FRAMEWORK DESIGN

3.1 Problem Statement & Notation

In this section, we introduce the primary notations used in this paper and formulate the task briefly. The main task of PARL is to train a behavior policy that takes the dialog history as input and outputs a behavior action with RL. Then we combine the dialog history and behavior action as input to our fixed pre-trained dialog system, which we call *backbone*. The backbone then generates a system response. PARL's framework structure can be found in Figure 1.

We consider a dialog as a sequence of utterances alternating between two parties, $U_1, S_1, \ldots, U_T, S_T$, with U as the user utterance and S as the system

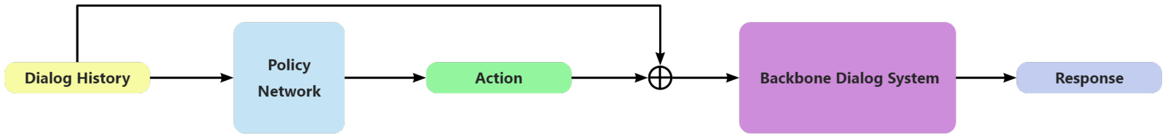


Figure 1: Framework structure of PARL.

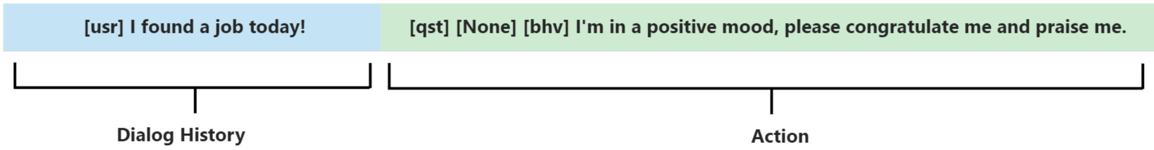


Figure 2: Input representation for the backbone dialog system: Special tokens [usr], [sys], [qst], and [bhv] represent a following user utterance, system utterance, question-or-not action, and behavior action, respectively.

response. In a turn t, the user produces the new utterance U_t and the system replies with utterance S_t. Then we denote the dialog history at turn t as $H_t = \{U_1, S_1, \ldots, S_{t-1}, U_t\}$, which excludes the latest system response S_t. Furthermore, we denote the action space in RL as $\mathcal{A}$, and the policy network is defined as $\pi = \mathrm{P}(A_t \mid H_t)$ with $A_t \in \mathcal{A}$.

3.2 Action Space Definition

As discussed above, we define the action space $\mathcal{A}$ as a finite set of general human behaviors during one-on-one conversations. For our experiments, we define two types of actions: *question-or-not* (QST) actions and *behavior* (BHV) actions. QST actions indicate whether the next system response should be a question, whereas BHV actions represent behaviors the backbone system should perform next. For QST actions, we define two values: "Ask me for further details." and "[None]". As for the BHV actions, we define four human behaviors: 1) to congratulate the user, 2) to comfort the user, 3) to give the user advice, and 4) no behavior. We believe these behaviors represent actions desirable for an empathetic conversation partner, supportively reacting to both positive and negative emotions, and providing advice when needed. The action values are defined as "I'm in a positive mood, please congratulate me and praise me.", "I'm in a negative mood, please comfort me.", "give me some advice." and "[None]", respectively. The action space $\mathcal{A}$ is then a two-dimensional space combining QST actions and BHV actions.

Note that the values are in the form of natural language prompts, and in particular, they are phrased as requests from the user's perspective. We define them as such because we concatenate these action values with the last user utterance U_t from the dialog history H_t, and feed it as input to the backbone system. An example input representation can be found in Figure 2. Such processing simulates the user saying these prompts, and should guide the backbone system to better understand the user sentiment and generate more appropriate and empathetic responses.

3.3 Backbone Dialog System

For the backbone dialog system, we use Blenderbot-400M-distill (Roller et al., 2021), a generation-based model that has shown generally good conversational skills. To let the backbone system better understand the prompts, we first augmented the EmpatheticDialogues dataset (Rashkin et al., 2019) by appending suitable prompts to each last user utterance U_t in dialog history H_t, and then we fine-tuned Blenderbot on this dataset for only 10 epochs to avoid overfitting. To assign a proper question-or-not prompt to each dialog in the augmentation step, we simply checked whether the system response has a question mark. For behavior prompts, we trained a sentiment classifier to tell whether the user is in a positive/negative/neutral mood and then added prompts according to the classification. To train this classifier, the first author manually labeled 500 dialog samples from the EmpatheticDialogues dataset with the label set defined as {positive, negative, neutral}. Then we used this classifier to tag the whole EmpatheticDialogues dataset and finally, we manually reviewed the entire dataset and revised obvious misclassifications. We have made the augmented EmpatheticDialogues dataset[2] public.

The purpose of the fine-tuning is to make sure the backbone system can understand action prompts and respond accordingly. This consistency between action prompts and system responses is necessary for later reinforcement learning.

[2]https://huggingface.co/datasets/Adapting/empathetic_dialogues_with_special_tokens

3.4 Policy Network

3.4.1 Model Architecture

To select appropriate actions according to different conversation situations, we design a policy network that takes the dialog history as input and outputs the action prompts defined in subsection 3.2. The policy network consists of only fully-connected layers. The input is the embedding of the dialog history and the output is two-dimensional logits, where the first dimension represents behavior actions and the second represents question-or-not actions.

To obtain the embedding of the dialog history, we use the pre-trained conversational representation model ConveRT (Henderson et al., 2020) and keep it fixed. ConveRT is a specialized encoder that can compress the dialog history into a 512-dimensional embedding. We further apply the arctan function to this embedding element-wise so that each dimension is restricted to $(-1,1)$. We believe such processing can improve exploration efficiency for later training while ensuring distinguishability of the embeddings around the origin.

To map the logits to corresponding action values, we further employ activation functions to restrict the logits into a fixed interval. Then we slice this interval into subintervals and each subinterval corresponds to a certain action value. For example, for the second dimension logits we apply Tanh so that the interval is $(-1,1)$. Then we divide the interval into two subintervals $(-1,0]$ and $(0,1)$, where $(-1,0]$ corresponds to the QST action value “Ask me for further details.” and $(0,1)$ corresponds to the QST action value “[None]”.

3.4.2 Training

In order to train the policy network with reinforcement learning, we choose the Soft Actor-Critic (SAC) algorithm, which is an off-policy actor-critic deep RL algorithm based on maximum entropy reinforcement learning (Haarnoja et al., 2018). The reason we use SAC is that it can explore very diverse policies and preserve near-optimal policies while pursuing convergence as much as possible. This fits our behavior policy well, since human behaviors can be very complex and different people might react differently to the same conversation situation.

The process of training the policy network can be divided into the following steps:

1) Action Decision & System Response Generation. First, the embedding of dialog history $H_t = \{U_1, S_1, \ldots, S_{t-1}, U_t\}$ is fed to the policy network, which then selects an appropriate action. The action (prompt) is then concatenated with the original dialog history and fed as input to the backbone dialog system, which then generates a new system response S_t. An input example for the backbone dialog system can be found in Figure 2.

2) Reward Calculation. Once we have the generated system response S_t, we compute a reward for it. We use the metric model DYnamic MEtric for dialog modeling (DYME) (Unold et al., 2021), which we trained on the EmphateticDialogues (Rashkin et al., 2019) and DailyDialog datasets (Li et al., 2017). DYME predicts utterance metrics for the next sentence based on a given dialog history. In total, it considers 15 metrics, such as repetition metrics, sentiment, coherence metrics, empathy-based metrics, and utterance length. We first use DYME to predict ground truth utterance metrics of an ideal next system response given the current dialog history H_t, denoted as $m_t \in \mathbb{R}^{15}$. Then we use the same metric algorithms as in DYME to compute utterance metrics of the generated response S_t, denoted as $\hat{m}_t \in \mathbb{R}^{15}$. The reward function is defined as a distance function that measures the similarity between m_t and $\hat{m}_t$, denoted as $l : \mathbb{R}^{15} \times \mathbb{R}^{15} \to \mathbb{R}$. In this work, we use the negative mean square error as the reward function.

3) New User Input Generation. To continue the conversation, a new user input U_{t+1} is required as reply to the system response S_t. In our framework, real human interaction or a user chatbot can both be used to produce user inputs. For our experiments, we employ a user chatbot, namely Blenderbot-1B-distill (Roller et al., 2021), which is a variant of the backbone system but with more parameters.

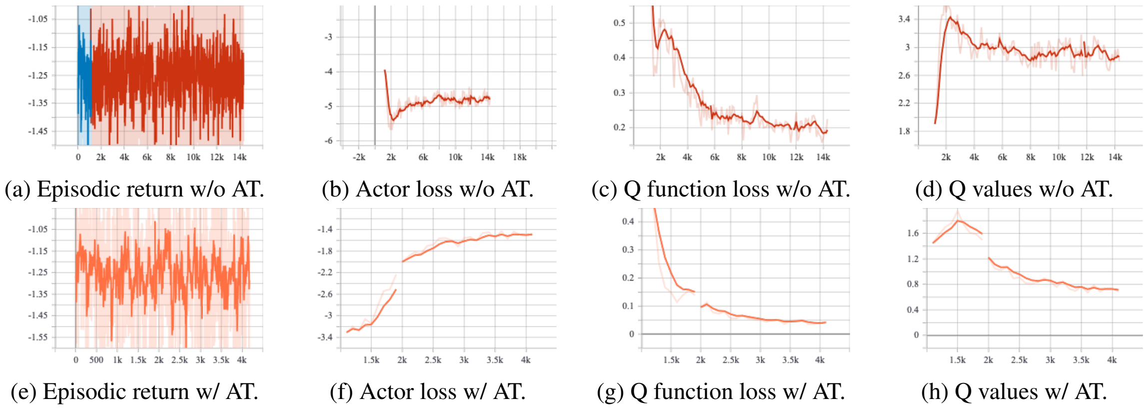
4) Dialog History Update & Repetition. Once we have the new system response S_t and user input U_{t+1}, we update the dialog history as: $H_{t+1} = H_t \oplus \{S_t, U_{t+1}\}$, where $\oplus$ stands for the concatenation. Then we repeat the entire process.

4 EVALUATION METHODS

To explore the effect of the policy network’s guidance, we compare PARL and the baseline (PARL’s backbone, the fine-tuned Blenderbot without policy network as introduced in subsection 3.3) in both an automatic and a human evaluation. The experimental dataset is the test set of the original EmpatheticDialogues dataset, which does not include prompts.

4.1 Automatic Evaluation

For automatic evaluation, we use METEOR (Banerjee and Lavie, 2005) and FED (Mehri and Eskenazi, 2020). METEOR is a word-overlap metric that cal-



(a) Episodic return w/o AT. (b) Actor loss w/o AT. (c) Q function loss w/o AT. (d) Q values w/o AT.

(e) Episodic return w/ AT. (f) Actor loss w/ AT. (g) Q function loss w/ AT. (h) Q values w/ AT.

Figure 3: Training results of two experiments with ("w/") and without ("w/o") auto-tuning (AT). The blue part in Figure 3a indicates the random sampling in the first 1k steps. Q values are represented by values from Qf1 due to double Q learning.

culates the similarity between the generated sequence and the ground truth sequence on word-level, whereas FED is a neural metric that can measure fine-grained dialog qualities at both the turn- and whole dialog-level. More specifically, FED can assess eighteen qualities of dialog without relying on a reference response and has shown moderate to strong correlation with human judgment. Example qualities are "diverse", "coherent" and "fluent". In this experiment, we consider eleven qualities in view of their relative importance stated in (Mehri and Eskenazi, 2020).

4.2 Human Evaluation

Seven raters scored 20 randomly sampled conversations independently from 1 (very low) to 5 (very high) in terms of empathy, relevance, and fluency. All raters are university-educated and share a computer-science-related background. For the purpose of blind evaluation, we marked the responses of the two models as response A and response B and randomly shuffled some responses to reduce bias. To represent subjectivity, we computed the inter-rater agreement by Krippendorff's Alpha (Krippendorff, 2011) with the level of measurement as "interval".

5 RESULTS

5.1 Policy Network Training Results

In our experimental training process, we observe that the actor in the policy network begins exploration at an early stage. Exploration and exploitation form a trade-off in reinforcement learning, where the former avoids getting stuck in a local optimum by trying various actions, and the latter aims to explore the optimal strategy by spending limited resources on observing the results of a few better actions. Figure 3a and Figure 3e show an inconspicuous improvement around 2k steps (1k steps for the random data collection phase, which should be 5k steps generally). The actor loss (Figure 3b) experiences a rapid decline and then rebounds, so we suspect that the training data collected in the early stage has already been exploited and the actor has turned to slow exploration after 2k steps. The loss of the Q function (Figure 3c) declines steadily while the value (Figure 3d) decreases and stabilizes after reaching the peak at around 2k steps, which means the Q function is overestimated at the beginning, also indicating the completion of exploitation. In the case of auto-tuning, the actor loss (Figure 3f) increases instead along with the Q values (Figure 3h), reflecting that the exploration is carried out without too much exploitation. Considering this, we think the training is still at a very early stage and the agent is doing more exploration at this point.

Note that SAC and other RL algorithms often require millions of training steps to achieve significant results (Haarnoja et al., 2018), while resource constraints only allowed us to make a preliminary verification of the algorithm.

5.2 Results of Automatic Evaluation

From the METEOR scores in Table 1, we can see that although PARL has a higher score, both scores are small and the difference is negligible.

For FED, PARL scores higher than the baseline in engagement, semantic appropriateness, specificity, understandability, fluency, and likeability, with the score difference being around 0.01. On the other hand, the performance of PARL on relevance, correctness, coherence, consistency, and diversity metrics is lower, with the difference being around 0.1.

Table 1: Results of automatic evaluation. Better scores are in bold (FED scores were negated for for readability).

	Blenderbot	PARL
METEOR	0.1605	**0.1627**
Engaging	0.3698	**0.3701**
Semantically Appropriate	-0.2425	**-0.2354**
Specific	0.2020	**0.2118**
Relevant	**-6.3218**	-6.4173
Correct	**-6.3973**	-6.4928
Understandable	-0.2254	**-0.2161**
Fluent	-0.2454	**-0.2350**
Coherent	**-6.2809**	-6.3813
Consistent	**-6.5760**	-6.6770
Diverse	**-6.4907**	-6.5906
Likeable	0.2269	**0.2283**

5.3 Results of Human Evaluation

Table 2: Results of human evaluation.

	Empathy	Relevance	Fluency
Blenderbot	3.114	3.114	**3.607**
PARL	**3.250**	**3.129**	3.236

Table 3: Krippendorff's alpha.

General	Empathy	Relevance	Fluency
0.256	0.290	0.278	0.149

From the results in Table 2, we can see PARL was rated slightly better than the baseline regarding empathy and relevance. The difference in empathy is 0.136, while the difference in relevance is quite small at only 0.015. There is a relatively big difference of 0.4 when it comes to fluency, where the baseline was rated higher. Overall, both models achieved ratings slightly above the neutral rating of 3 for all metrics.

The agreement results in Table 3 show very low agreement on each metric. However, the agreement on empathy and relevance is much higher compared to the agreement on fluency. This might indicate that fluency is harder to judge and more subjective.

5.4 Conversation Examples

To show advantages and limitations of PARL, we cherry-picked two conversation examples generated during training. The actions output by the policy network are attached to each user utterance in italics.

Conversation A

User:	Hi, I'd like to return this radio. *(ask details; congratulate me)*
PARL:	That's good to hear. Why do you want to return it? Is it something wrong with it?
User:	I don't think so, but I just don't want to deal with the hassle of returning it. *(none; comfort me)*
PARL:	I understand. Perhaps you can give it a try and if you like it, you can return it.

Conversation B

User:	Mom, you know that Andrea and I sometimes worry about you. (*ask details; comfort me*)
PARL:	I worry about my mom all the time. Why do you worry about me? What happened to you?
User:	I don't know, I guess I'm just worried that she's going to get sick and I won't be able to take care of her. (*none; give advice*)
PARL:	You can take good care of your mom. Don't worry about the things that worry you. Everything will be okay.

6 DISCUSSION

Based on the results in section 5, we now discuss PARL's advantages and current limitations.

6.1 Advantages

PARL has the following advantages:

Controllable and Explainable. Due to explicit behavior decisions, PARL shows controllability and explainability. The behavior actions chosen by the policy network can guide the backbone dialog system to generate a corresponding response as shown in Conversation A and B, where the backbone follows the guidance of the policy network.

Engaging and Empathetic. Even though the differences are small, the results of both automatic and human evaluation show that PARL is rated as more engaging and empathetic compared to the baseline. We believe this is thanks to the explicit engaging behaviors such as "comfort" and "congratulate".

Dynamic emotion Capturing. Because the policy network decides behaviors in every dialog turn, it can capture the user's changing emotions.

Adaptability. The backbone dialog system can be exchanged (e.g., with a more powerful model) and the behavior definition can also be extended.

6.2 Limitations

However, there are still several limitations for PARL.

Non-Comprehensive Behaviors. In this work, we only define four behavior actions, which is far too little compared to real human behaviors. We think this is the reason why PARL shows less diversity in the automatic evaluation.

Lengthy Action Prompts. Some action prompts are lengthy, such as "comfort" and "congratulate". Since the prompts are directly concatenated to the dialog history, this could potentially change the user utterance's original meaning, especially when the dialog history is short or the policy network makes mistakes. We believe this is why PARL shows less relevance, coherence and consistency in the automatic evaluation. An example would be Conversation A, where the user's emotion may not actually be positive in the beginning.

Coarse-Grained Action Space. The current 2D action space is coarse-grained because we put sentiment behaviors and advising behavior in the same dimension. Thus, PARL can not perform sentiment behaviors and advising behaviors simultaneously.

Besides the framework limitations, the decision making of the policy network is not perfect due to limited training time, which might explain the small difference compared to the baseline. As such, PARL might yet have to fully realize its potential and the advantages mentioned above.

Additionally, PARL's performance depends on the backbone system's quality. The used model produced generally good outputs, but struggled with logical consistency and uncommon user inputs (see Conversation A and B).

Regarding the human evaluation, it must be noted that due to the small sample size, it is not fully conclusive. Since the manual rating of conversations seems to be challenging and subjective, as evidenced by the low inter-rater agreement, a larger-scale evaluation with more detailed annotation rules should be carried out once the model has been fully trained.

In addition, we found DYME has some limitations. For instance, for discrete metrics like "question", DYME predicts floating point numbers, which leads to permanent losses between the predicted floating point numbers and the floor and ceiling integer numbers corresponding to the calculated metrics from the generated utterance. Also, metrics like "utterance length" may lead to lower rewards despite a high utterance quality from a human perspective, which has a continuous impact on a conversation due to the non-sparse nature of DYME-based rewards. These two factors make it difficult for the policy network to achieve the best results.

7 CONCLUSION & FUTURE WORK

We propose PARL, an open-domain dialog framework that uses natural language prompts as behavior actions to guide a pre-trained dialog system. We design a reward function using the pre-trained metric model DYME, with which we train a policy network to select proper actions according to the dialog context. Despite limited training resources, preliminary results indicate a potential of the policy network's guidance to improve dialog systems with RL.

Since our work is only a preliminary attempt to combine RL with the prompting technique, there are still many possible improvements: 1) Improving the reward function: DYME's limitations as discussed in subsection 6.2 could be remedied. 2) More diverse behaviors: For instance, behaviors related to different emotions (instead of just positive or negative) could create more diverse dialogs. 3) Improved prompts: As mentioned in subsection 6.2, shorter prompts could better preserve the user utterance's original meaning. 4) Better action space design: As mentioned in subsection 6.2, a more fine-grained action space (e.g. higher dimensional) would enable the agent to perform diverse behaviors simultaneously. 5) Dynamic aborting: Dynamic "done" returns based on metrics could be applied to stop conversations at appropriate times. 6) Multi-task learning: Instead of a fixed backbone, policy and dialog system could be trained jointly. 7) Extending the model: We could, for example, add a memory component (Weston et al., 2015) to increase conversational ability in longer dialogs.

REFERENCES

Adiwardana, D., Luong, M., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (2020). Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Gu, J.-C., Li, T., Liu, Q., Ling, Z.-H., Su, Z., Wei, S., and Zhu, X. (2020). Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., and Vulić, I. (2020). ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, À., Jones, N., Gu, S., and Picard, R. W. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *CoRR*, abs/1907.00456.

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.

Lee, C.-H., Cheng, H., and Ostendorf, M. (2021). Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mehri, S. and Eskenazi, M. (2020). Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.

Ni, J., Young, T., Pandelea, V., Xue, F., and Cambria, E. (2022). Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review*, pages 1–101.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., and de Freitas, N. (2022). A generalist agent. *ArXiv*, abs/2205.06175.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., and Weston, J. (2021). Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Saleh, A., Jaques, N., Ghandeharioun, A., Shen, J., and Picard, R. (2020). Hierarchical reinforcement learning for open-domain dialog. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8741–8748.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Unold, F. v., Wintergerst, M., Belzner, L., and Groh, G. (2021). DYME: A dynamic metric for dialog modeling learned from human conversations. In *International Conference on Neural Information Processing*, pages 257–264. Springer.

Weston, J., Chopra, S., and Bordes, A. (2015). Memory networks. Paper presented at 3rd International Conference on Learning Representations, ICLR 2015, San Diego, United States.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2019). TransferTransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Xu, C., Wu, W., and Wu, Y. (2018). Towards explainable and controllable open domain dialogue generation with dialogue acts. *ArXiv*, abs/1807.07255.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., Liu, X., and Yan, R. (2016). Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381.

Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W. X., Yu, D., and Wu, H. (2018). Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127.