Reinforcement Learning Based Personalized Neural Dialogue Generation

Tulika Saha(✉), Saraansh Chopra, Sriparna Saha, and Pushpak Bhattacharyya

Indian Institute of Technology Patna, Bihta, India
sahatulika15@gmail.com, saraansh.chopra@gmail.com,
sriparna.saha@gmail.com

Abstract. In this paper, we present a persona aware neural reinforcement learning response generation framework capable of optimizing long-term rewards carefully devised by system developers. The proposed model utilizes an extension of the recently introduced Hierarchical Encoder Decoder (HRED) architecture. We leverage insights from Reinforcement Learning (RL) and employ policy gradient methods to optimize rewards which are defined as simple heuristic approximations that indicate good conversation to a human mind. The proposed model is demonstrated on two benchmark datasets. Empirical results indicate that the proposed approach outperforms their counterparts that do not optimize long-term rewards, have no access to personas, standard models trained using solely maximum-likelihood estimation objective.

Keywords: Natural language generation · Dialogue · Persona · Hierarchical encoder-decoder · Reinforcement learning

1 Introduction

Efficient communication between human and Virtual Agent (VA) in the form of Natural Language Generation (NLG) has been a long-standing goal for any conversational agent [2]. In recent times, two notable paradigms of research have emerged pertaining to NLG. The first category includes open domain conversations typically nonchalant chit-chatting [4]. The second ones are goal-oriented dialogue generation where the VA is required to interact with the user in natural language to solve a particular task of a domain [7]. Neural models such as Bi-LSTMs or Memory Networks [10] recently had enough capacity and access to large-scale datasets and seemed to produce meaningful responses in a chit-chat environment. However, conversing with such generic chit-chat models for a while exposes its weaknesses quickly. Also, to ensure that the VA provides a more natural, human-like and coherent conversational experience, it is imperative for the VA to exhibit a persona and reciprocatively understand users persona to increase users' engagement level and to gain its trust and confidence. Lately, researches involving chit-chat VAs are classified into two broad categories: (i) first one is implicit model where a user's persona is learnt implicitly from the dialogue data

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1332, pp. 709–716, 2020.
https://doi.org/10.1007/978-3-030-63820-7_81

and is depicted as the user's spoken utterance embedding [5]; (ii) second is the explicit model where the user's persona is available explicitly, i.e., the produced responses from the VA are explicitly conditioned either on a given profile with various attributes [15] or on a text-described persona [13].

The challenges faced by these chit-chatting VAs hint the need of a conversational framework with the ability to (i) depict a consistent persona of the speaker (say A) and incorporate persona of the speaker (B) while generating responses in a dialogue and vice-versa; (ii) integrate carefully engineered rewards that mimic a human-like conversational experience as closely as possible; (iii) model long-term memory of generated utterances in the ongoing conversation. To realize these goals, we gather insights from Reinforcement Learning (RL) which has been used extensively in Dialogue Systems in various aspects [7]. In this paper, we propose a persona aware neural reinforcement learning response generation framework capable of optimizing long-term rewards carefully devised by system developers. The proposed model utilizes an extension of the recently introduced Hierarchical Encoder Decoder (HRED) architecture [9] and models conversation between two speakers conditioned on their respective personas to explore and examine the space of possible actions while simultaneously learning to maximize expected rewards. The simulated speakers need to learn an efficient persona aware dialogue policy from the ongoing dialogue simulations using well-known RL algorithm namely policy gradient methods [12] instead of employing the MLE (maximum likelihood estimate) objective used in traditional HRED models. The proposed model is demonstrated on two different datasets where personas are viewed in different ways, i.e., explicit and implicit (as stated above). Empirical results indicate that the proposed approach outperforms several strong baselines.

The key contributions of this paper are as follows : ***i.*** *Propose a persona aware neural reinforcement learning response generation framework that utilizes an extension of the recently introduced HRED architecture tuned in accordance to the personas of the speaker;* ***ii.*** *The utility of RL helps optimize long-term rewards in the on-going dialogue resembling approximations that characterize a good conversation to a human mind;* ***iii.*** *Empirical results indicate that the proposed approach outperforms their counterparts that do not optimize long-term rewards, have no access to personas, standard models trained using solely MLE objective.*

2 Related Works

There exist numerous works in the literature that have addressed the task of chit-chat based Dialogue Generation in different aspects. In [4], authors proposed the first-ever NLG framework that utilizes RL to optimize long-term rewards using the traditional SEQ2SEQ model. In [8], authors proposed an extension of the HRED model for dialogue generation that models hierarchy of sequences using two RNNs, i.e., one at word level and the other at the utterance level of a dialogue. However, none of these works models persona of the speaker while addressing the task of NLG. There are plenty of works that use persona of the speaker in implicit and explicit ways in the NLG framework. In [5], authors proposed a large-scale REDDIT dataset to model personas of the speaker by utilizing the SEQ2SEQ model for the task of dialogue generation. In [13], authors proposed a PERSONA-CHAT dataset with text-described persona for each speaker. In [14], authors proposed a PersonalDialogue dataset, where personas are available as a given profile with various attributes such as age, gender, location etc. In [15], authors proposed a personalized dialogue generation model that utilizes the encoder-decoder framework along with attribute embeddings to capture and incorporate rich persona. However, none of these works utilizes RL for optimizing long-term rewards as an approximations to different features of the human mind.

3 Proposed Approach

In this section, we will describe the different components of the proposed methodology.

Hierarchical Encoder Decoder. The overall architecture of the proposed framework is shown in Fig. 1, which is an extension of the recently introduced Hierarchical Encoder Decoder architecture [8,9]. Here, the conversational context or dialogue history is modelled using a separate RNN over the encoder RNN, thus forming a hierarchical structure known as the hierarchical encoder. The proposed network is built upon the HRED to include personas of the speaker involved in the conversation along with the other textual modalities. The main components of the HRED (in this paper) are *utterance encoder*, *context encoder*, *persona encoder* and *decoder* which are discussed as follows:

Utterance Encoder. Given a speaker utterance X_k, a Bidirectional Gated Recurrent Unit (Bi-GRU) [1] has been used to encode each of the words of the utterance, $x_{k,i}, i \in (1, n)$ depicted by a w dimensional word vector/embeddings.

Persona Encoder. For a speaker utterance, each persona sentence of the speaker is encoded separately. Each of the persona utterances is represented as a w dimensional word embedding with a linear layer on top of them. Finally, the representations across all the persona sentences are summed up to obtain the persona encoding. The encoded persona is passed through a 1-hop memory network [10] with a residual connection [13], using all the spoken utterances of the speaker as context and the persona representation as memory.

Context Encoder. The obtained representation from both the utterance and persona encoders are concatenated in every turn and fed as input to the context encoder which is a context GRU.

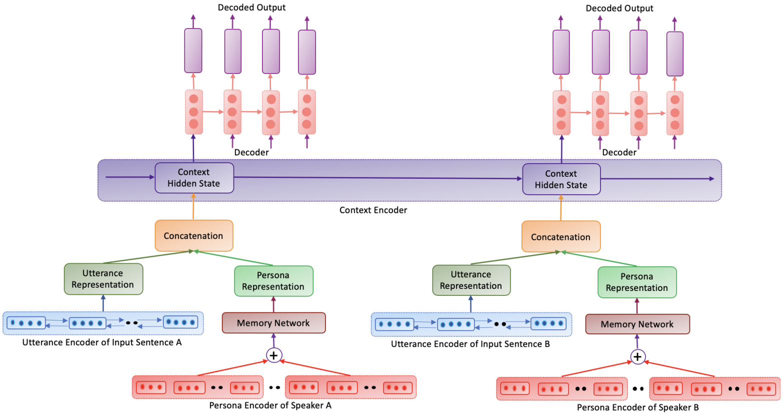

Fig. 1. The proposed persona aware HRED framework for a conversation with 2 turns

Decoder. The final hidden state representation of the context encoder is used as the initial state representation of the decoder. Similarly, a GRU unit has been employed for the decoder part. The decoder then generates word sequentially at every time step (say t) conditioned on the decoded words prior to t and final state representation of the context encoder. At each time-step t, decoder produces the probability of output token w using softmax. The HRED model explained above is initially trained with the negative log likelihood, i.e., the MLE objective function in a supervised manner to generate semantically plausible responses which will be used below for initialization.

3.1 Reinforcement Learning

A dialogue is typically represented as an alternating sequence between two speakers as $X_{a,1}, X_{b,1}, X_{a,2}, X_{b,2}, ..., X_{a,i}, X_{b,i}$. This sequence of utterances can be viewed as actions that are taken according to a policy learnt by the HRED model. Next, with the help of the MLE parameters, the model is tuned to learn a policy that maximizes long-term future rewards [4]. The components of the RL based training are discussed below.

State and Action. The state is denoted by the output of the hierarchical encoder (explained above) which represents the encoding of the current utterance conditioned on the persona of the speaker along with the dialogue history, $[S(H_X, H_{P,i}, H_C)]$. The action a, is the utterance to generate in the next time-step i.e., Y_t. The action space is infinite as the sequence generated can be of arbitrary length. So, the policy $\Pi(Y_t|S(H_X, H_{P,i}, H_C))$ learns the mapping from states to actions and is defined by its parameters.

Reward. Here, we will discuss few key factors that are attributed to the success of any dialogue and discuss how the approximations to these factors can be induced in computable reward function, r. Below are the reward functions used.

Cosine Triplet Loss. At each turn of the conversation, we expect speakers involved in the dialogue to keep adding new information and refrain from generating repetitive responses, i.e., we expect $X_{a,t}$ and $X_{a,t-1}$ to be diverse, thus, we penalize the semantic similarity between consecutive turns of the same speaker (say A here). Similarly, we also need to counter situations where the generated responses between the two speakers are highly rewarded but are not coherent in the sense of the topic being discussed, i.e., we expect that the conversation at a time-step t between two speakers, $X_{a,t}$ and $X_{b,t}$ be consistent and appropriate. Intuitively, this idea represents the triplet loss function [3] where the anchor is compared to two different inputs. We extend this idea to the semantic similarity between utterances. Let $X_{b,t}$ be the anchor (say). So, the goal is to minimize the semantic similarity between $X_{a,t}$ and $X_{a,t-1}$ and to maximize the semantic similarity between $X_{a,t}$ and $X_{b,t}$ and vice-versa with the speakers as the dialogue progresses. Let $H_{X,a,t}$, $H_{X,b,t}$ be hidden state encoder representations of two different speakers A and B at time-step t, respectively, and $H_{X,a,t-1}$ be the representation of speaker A at time-step $t-1$. Then the reward function is:

$$r_1 = (cos(H_{X,a,t}, H_{X,b,t}) - cos(H_{X,a,t}, H_{X,a,t-1}) - \gamma) \tag{1}$$

Negative Log Likelihood. The generated utterances by the speakers must pave way for the communication further, i.e., speakers should respond in a way that makes it easier for the other speaker to respond too. So, we penalize the generated response with the negative log likelihood if it is a dull response such as "*I don't know*", "*I have no idea*" and so on. A list L with such dull responses is manually created with 8 turns. So, the reward function is:

$$r_2 = -(1/N_L) \sum_{s \in L} (1/N_s) log_{X,t}(s|a) \tag{2}$$

where N_L represents the cardinality of L and N_s represents the number of tokens in the dull response s.

Thus, the final reward for an action a is the weighted sum of the rewards explained above:

$$r(a|S(H_X, H_{P,i}, H_C)) = \alpha r_1 + (1-\alpha) r_2 \tag{3}$$

where α is 0.75. This reward is obtained after the end of each generated utterance. Policy Gradient algorithm [12] is used to optimize these rewards. The policy model Π is initialized using the pre-trained HRED model (using the MLE objective function).

4 Implementation Details

In this section, we first describe the details of the dataset used followed by the experimentation details.

4.1 Dataset

We perform experiments on two different datasets in which personas are represented in different ways i.e., explicit and implicit. For the explicit representation of the personas, we use the PERSONA-CHAT dataset [13]. This dataset contains 10,907 dialogues amounting to 1,62,064 utterances. For the implicit case, we use the REDDIT dataset [5] which contains 1.7 billion comments/utterances. This is a large-scale dataset with the utterances covering the persona profile of 4.6 million users. To extract persona utterances from this dataset, we follow the same method as used in the source paper.

4.2 Hyperparameters and Evaluation Metric

FastText embeddings of dimension 300 have been used to represent words of an utterance. The hidden size for all the GRU units is 512. The model is initialized randomly with a Gaussian distribution. For decoding, beam search with beam size of 5 is used. Adam optimizer is used to train the model. A learning rate of 0.0001 was found to be optimum. The models are automatically evaluated using standard metrics namely BLEU-1 score [6], perplexity and embedding based metric [9].

5 Results and Analysis

We compare our proposed approach with several baselines to highlight the contributions and gains of different components of the proposed approach in terms of performance. The different baselines are: **i.** ***SEQ2SEQ***: This is the traditional sequence to sequence, i.e., encoder-decoder framework without the use of persona or RL based training; **ii.** ***HRED-1***: This is the standard HRED model with a context of one without the usage of persona and RL; **iii.** ***HRED-2***: HRED model with a context of two previous turns without the usage of persona and RL; **iv.** ***HRED-3***: HRED model with a context of three previous turns without the usage of persona and RL; **v.** ***HRED-3 + MN***: This model includes HRED-3 and Memory Network to model persona without the usage of RL based training; **vi.** ***HRED-3 + MN + RL***: This is the proposed model that includes both the persona of the speakers along with RL based training to optimize long-term rewards.

Table 1 shows the results of all the baselines and the proposed models for both the datasets. As evident from the tables, all the HRED based models perform better than the traditional SEQ2SEQ model. Also, amongst the HRED models, conversational context of three turns gave the best results in terms of all the

Table 1. Results of all the baseline and proposed models. MN represents Memory Network, PPL represents perplexity.

Models	Dataset									
	PERSONA-CHAT					REDDIT				
	Embedding Metrics			PPL	BLEU-1	Embedding Metrics			PPL	BLEU-1
	Average	Extrema	Greedy			Average	Extrema	Greedy		
SEQ2SEQ	0.6084	0.3593	0.3999	84.779	0.0952	0.5918	0.2896	0.3013	106.371	0.0891
HRED-1 (context-1)	0.6042	0.3364	0.4014	47.633	0.1026	0.621	0.319	0.34	95.261	0.1047
HRED-2	0.6057	0.3367	0.4024	49.797	0.1014	0.588	0.299	0.315	103.186	0.1052
HRED-3	0.6060	0.3345	0.4043	47.739	0.1025	0.616	0.331	0.356	78.263	0.1059
HRED-3 + MN	0.6073	0.3387	0.4052	216.232	0.1096	0.621	0.337	0.369	111.150	0.1089
HRED-3 + MN + RL	**0.6200**	**0.3512**	**0.4162**	**96.474**	**0.1183**	**0.6311**	**0.325**	**0.384**	**150.851**	**0.1167**

metrics. This shows that long-term assimilation of memory across the dialogue helps produce better responses. In *HRED-3* model, memory network is incorporated to model the personas of the speaker to demonstrate a consistent persona in the conversation. Finally, in this model, we have incorporated the RL based training. As seen from the table, this model produces the best results amongst all the baseline models. Thus, each of these components aids the performance of the proposed model. All the reported results are statistically significant as we have performed Welch's t-test [11] at 5% significance level.

6 Conclusion and Future Work

In this paper, we propose a persona aware neural reinforcement learning response generation framework capable of optimizing long-term rewards carefully devised by system developers. The proposed model utilizes an extension of the HRED architecture and models conversation between two speakers conditioned on their respective personas to explore and examine the space of possible actions while simultaneously learning to maximize expected rewards. A thorough evaluation is carried out on two benchmark datasets with automated metrics as well as human evaluation. Empirical results indicate that the proposed approach outperforms several strong baselines. Future works include developing a more sophisticated and efficient end-to-end dialogue generation framework along with extending these approaches for creating end-to-end NLG models for low-resource language.

Acknowledgement. Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations. ICLR (2015). http://arxiv.org/abs/1409.0473

2. Dušek, O., Novikova, J., Rieser, V.: Evaluating the state-of-the-art of end-to-end natural language generation: the e2e nlg challenge. Comput. Speech Lang. **59**, 123–156 (2020)
3. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. CoRR abs/1703.07737 (2017). http://arxiv.org/abs/1703.07737
4. Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., Gao, J.: Deep reinforcement learning for dialogue generation. In: Su, J., Carreras, X., Duh, K. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP (2016). https://doi.org/10.18653/v1/d16-1127
5. Mazaré, P., Humeau, S., Raison, M., Bordes, A.: Training millions of personalized dialogue agents. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018). https://doi.org/10.18653/v1/d18-1298
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002). https://www.aclweb.org/anthology/P02-1040/
7. Saha, T., Gupta, D., Saha, S., Bhattacharyya, P.: Reinforcement learning based dialogue management strategy. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11303, pp. 359–372. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04182-3_32
8. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: Schuurmans, D., Wellman, M.P. (eds.) Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (2016). http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957
9. Serban, I.V., et al.: A hierarchical latent variable encoder-decoder model for generating dialogues. In: Singh, S.P., Markovitch, S. (eds.) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, USA (2017). http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567
10. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems (2015). http://papers.nips.cc/paper/5846-end-to-end-memory-networks
11. Welch, B.L.: The generalization ofstudent's' problem when several different population variances are involved. Biometrika (1947)
12. Zaremba, W., Sutskever, I.: Reinforcement learning neural turing machines-revised. arXiv preprint arXiv:1505.00521 (2015)
13. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: i have a dog, do you have pets too? In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL (2018). https://www.aclweb.org/anthology/P18-1205/
14. Zheng, Y., Chen, G., Huang, M., Liu, S., Zhu, X.: Personalized dialogue generation with diversified traits. arXiv preprint arXiv:1901.09672 (2019)
15. Zheng, Y., Zhang, R., Huang, M., Mao, X.: A pre-training based personalized dialogue generation model with persona-sparse data. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI (2020). https://aaai.org/ojs/index.php/AAAI/article/view/6518