Persona aware Response Generation with Emotions

Mauajama Firdaus*, Naveen Thangavelu†, Asif Ekbal‡ and Pushpak Bhattacharyya§
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Bihar, India 801103
Email: *mauajama.pcs16@iitp.ac.in, †naveent.cs16@iitp.ac.in, ‡asif@iitp.ac.in, §pb@iitp.ac.in

***Abstract*—Conversational systems are the perfect examples of human-machine interactions. The conversational agents while interacting with humans lack the ability to express emotions and behave inconsistently, making the conversations boring and non-interactive. In this work, we propose the task of persona aware emotional response generation in which the system can generate specific and consistent responses in accordance to the provided personality information and the conversational history. To make the responses interactive and interesting we intend to infuse the emotions in the responses that help in making the responses more human-like. We propose a persona aware attention framework employing an encoder-decoder approach. We investigate different ways to include the desired emotions in the responses. Experimental results on the PersonaChat dataset shows that our proposed framework outperforms the baseline models and can generate interactive and emotional responses.**

***Index Terms*—Response generation, Persona, Emotions, Attention, Encoder-decoder**

I. Introduction

Conversational agents are the best examples of human-machine interactions. With the progress in Artificial Intelligence (AI), Natural Language Processing (NLP) and Machine Learning, conversational agents have shown remarkable growth since the last few years. Conversational agents commonly known as dialogue systems have gained immense importance with its widespread applications in our day to day lives. The personal assistants like Apple's Siri, Amazon's Alexa, Microsoft's Cortana are being extensively used for assisting humans in their everyday works. Dialogue systems are majorly divided into two types. Open-domain dialogue systems [1], [2] are the type of conversational agents that deal with open-ended conversations with no fixed goal or objective. These systems consist of conversations having various topics. On the other hand, there are goal-oriented or task-oriented dialogue systems [3], [4] which comprise of dialogues having a certain objective to be fulfilled for the user. Dialogue systems comprise of different modules that help in satisfying the user's goals. One of the key components of every dialogue system is response generation. Though the dialog manager decides on "what to say" to the user, the task of "how to say" the information to the user is handled by the response generation module. Hence, the generation module must respond in such a manner that it makes the conversation interactive and interesting leading to customer and user satisfaction.

Table I: An example from the PersonaChat dataset

Persona 1	**Persona 2**
As a child , I won a national spelling bee.	*I'm very athletic.*
I've been published in the new yorker magazine.	*I have brown hair.*
I am a gourmet cook.	*I love bicycling.*
I've perfect pitch.	*I hate carrots.*
[Person 1] Hi! i work as a gourmet cook.	
[Person 2] I don't like carrots. I throw them away.	
[Person 1] Really. But, I can sing pitch perfect .	
[Person 2] I also cook, and I ride my bike to work.	

One of the long-standing goals of AI is to infuse human-like behavior in machines. Every individual has a personality and is driven by emotions. The ability to converse with a consistent personality helps in bringing consistency and specificity in responses. Recently, researchers are focusing on incorporating personality information on chit-chat [5], [6] and goal-oriented [7], [8] conversational systems. Due to the unavailability of persona aware datasets, the authors in [5] introduced a PersonaChat dataset where the individual state information about their personality is expressed in a few texts for open-domain chit-chat conversational systems. In Table I, we present an example from the dataset, from which it is evident that the speakers while conversing with one another is capable of maintaining the persona information. This helps in making the conversation interactive and also facilitates building user's trust and confidence [9]. The capability to maintain a consistent persona is imperative for conversational systems for proper interaction with the user in a coherent and natural manner.

Conversational agents in the form of personal assistants not only assist human in completing their desired goals, but also behaves as a companion to them. Therefore, it is essential to empower the conversational agents with the ability to perceive and express emotions to make them capable of interacting with the user at the human level. These agents help in enhancing user satisfaction [10], while reducing the breakdowns in conversations [11] and providing user retention. Hence, dialogue systems capable

* Corresponding Author

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

of generating replies while considering the emotional state of the user is the most desirable advancement in Artificial Intelligence (AI).

Though maintaining a consistent personality is important to gain user's trust, at the same time it is essential to respond emotionally to build a connection with the user. From Table I it is visible that the agent can maintain a unique personality while conversing with the user but it lacks the emotional connection with the user. Therefore, the conversation is more like stating facts rather than a real conversation. Hence in this work, we propose the task of infusing emotional content in the responses while preserving a consistent persona. From the Table, the response to *Person 1* could be more empathetic like *That's a great job, but I don't like carrots and throw them away.* This response has a happy undertone than the ground-truth response which is neutral and contains only facts about *Person 2*. Emotional responses are interesting and provide a medium for a better conversation. From the example, it is clear that just having a persona in a response is not enough for generating engaging responses. The emotional aspect should also be introduced in the responses to make it more human-like and natural. Our present work is one of the first work that handles both the persona and emotion in responses.

The key contributions of this work are:

- We propose the task of generating emotional responses while considering the persona information also in the responses.
- We propose a novel persona aware attention approach with the ability to infuse the emotion information in the responses.
- We adopt a semi-supervised approach to annotate the PersonaChat dataset with emotions.
- Experimental results show that our proposed framework is capable of maintaining a consistent persona while generating emotional responses.

II. Related Work

Natural language generation (NLG) has become increasingly important in large applications, such as the dialog systems [1], [2], [12]–[14] and many other natural language interfaces. The response generation offers the medium by which a conversational agent can interact with its user to help the users accomplish their intended goals. In [13] a sequence to sequence framework was proposed for generating responses. The reinforcement learning paradigm was explored in [1] for generating diverse responses. Our work differs from these primary response generation framework in the sense that we intend to design a system that is capable of maintaining a consistent persona while generating emotional responses.

Persona information is an important aspect of response generation. Earlier works on persona-based conversational models [15] incorporated speaker embeddings to infuse persona information in the responses. To incorporate persona in chit-chat models the authors in [5], [6] introduced a PersonaChat dataset that includes personal information of the speakers. This dataset has been extensively used to build persona-based dialogue systems [16]–[19]. The authors in [16] used a meta-learning framework to include persona information in the generated responses. Similarly, the authors in [17] employed a hierarchical pointer network for generating persona-based responses. The authors in [18] used persona information to generate diverse responses by employing conditional variational autoencoder. Our present work differs from the existing works on the PersonaChat dataset as we intend to use the persona information while generating emotional responses. Persona information is also being exploited in goal-oriented dialogue systems [7], [8], [20]. The authors in [7] introduced persona information in babI dialog dataset for creating better responses. The authors in [20] introduced persona information by employing persona and position information in the responses. As personalization has been considered in responses we intend to take a step ahead by inculcating the desired emotions in the responses.

Lately, emotional text generation has gained immense popularity [21]–[27]. In [28], an emotional chatting machine (ECM) was proposed that was built upon seq2seq framework for generating emotional responses. ECM employs an internal and external memory that regulates the implicit change in emotional state and models the explicit emotional expression by selecting the emotion or generic words at every time-step respectively. Recently, a lexicon-based attention framework was employed to generate responses with a specific emotion [19]. Emotional embedding along with affective sampling and regularizer was employed to generate affect driven dialogues in [29]. Our present research differs from these existing works as we propose a novel task of generating responses with emotions having a consistent persona. To the best of our knowledge, this is the very first attempt to provide a benchmark setup for persona aware emotional response generation in a dialogue setting.

III. Methodology

In the present section, we first discuss the problem statement followed by the proposed methodology for generating emotional responses in a dialogue system having persona information of the speakers. The architectural diagram of the proposed framework is depicted in Figure 1.

A. Problem Definition

Our current work addresses the task of generating persona aware dialogue generation with desired emotions in accordance to the conversational history. The dialogue consists of utterances along with the persona information (in a couple of sentences) of the speakers and given a context of k turns the goal is to generate the next response with the desired emotion e. More precisely,

for a given dialogue context having k utterances $D = U_1, U_2, \ldots, U_k$ as input where each utterance comprises of $U_k = w_{k,1}, w_{k,2}, \ldots, w_{k,n}$ words and a set of persona information $P = P_1, P_2, \ldots, P_m$ along with the emotion embedding e_v, the task is to generate the emotional response $Y' = y_1, y_2, \ldots, y_{n'}$ for the desired emotion e based upon the input dialogue context and the persona information.

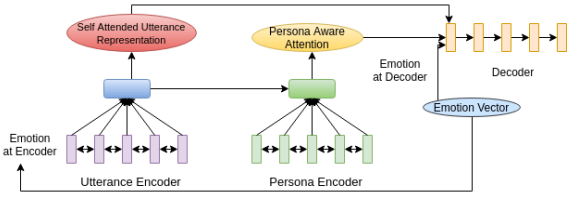

Figure 1: Architectural diagram of the proposed framework

B. Proposed Framework

We construct a response generation model based upon the encoder-decoder framework. Our proposed framework comprises of utterance and persona encoders followed by a decoder for generating the desired emotional responses as shown in Figure 1.

1) Utterance Encoder: Given an utterance U_k, a bidirectional GRU (BiGRU) is employed to encode each word $w_{k,i}$, $i \in (1, ..., n)$ represented by d-dimensional embeddings. We concatenate the last hidden representation from both unidirectional GRUs to form the final hidden representation of a given utterance. The final hidden state of the utterance GRU serves as the initial state of the decoder GRU.

$$h_{U,k,i} = BiGRU_u(w_{k,i}, h_{U,k,i-1}) \tag{1}$$

2) Persona Encoder: The persona encoder encodes the persona texts $P = P_1, P_2, \ldots, P_m$ into fixed dimensional vectors. Given a persona text P_m, a bidirectional GRU is employed to encode each word $w_{m,j}$, $j \in (1, \ldots, n')$ in the persona text by a d-dimensional embedding. We concatenate the last hidden representation from both unidirectional GRUs to form the final hidden representation of a given persona text represented as follows:

$$h_{P,m,j} = BiGRU_p(w_{m,j}, h_{P,m,j-1}) \tag{2}$$

The final persona representation is the concatenation of all the persona text given by $h_P = [h_{P,m,1}] \odot [h_{P,m,2} \odot \ldots \odot [h_{P,m,j}]$. Here $\odot$ represents concatenation.

3) Persona Aware Attention: In the baseline sequence to sequence model, we incorporate persona information of the speakers to improve the performance of the system. To focus on different persona information mentioned in the text, we employ persona-aware attention.

$$\alpha_a = softmax({W_a}^T h_{U,k,i}), U_a = \alpha_a {h_{U,k,i}}^T \tag{3}$$

The self-attended utterance embedding is used as a query vector U_a to compute the attention distribution over the persona features represented by h_P.

$$\beta_a = softmax({U_a}^T W_{a'} h_P), P_a = \beta_a {h_P}^T \tag{4}$$

where, ${W_a}^T$ and $W_{a'}$ are trainable parameters.

4) Decoder: In the decoding stage, we employ unidirectional GRU that generates words sequentially conditioned on the self-attended utterance vector U_a, the attended hidden representation of the persona P_a and the previously decoded words. We use randomly initialized embedding to represent the desired emotion labels. Global Attention mechanism [30] is incorporated to enhance the performance of the decoder GRU. The attention layer is applied to the hidden state of utterance encoder using decoder state d_t as the query vector. The concatenation of the utterance vector and the decoder state is used to compute the final probability distribution over the output tokens.

$$s_{d,t} = GRU_d(y_{t-1}, [s_{d,t-1}, U_a, P_a]) \tag{5}$$

$$c_t = \sum_{i=1}^{k} \alpha_{t,i} h_{U,k,i}, \tag{6}$$

$$\alpha_{t,i} = softmax(h_{U,k,i}^T W s_{d,t-1}) \tag{7}$$

5) Emotion: To include the desired emotions in the responses we use two approaches to include the emotion vector in the responses. In the first method we prepend the desired emotion at the beginning of the utterance representation as follows:

$$U_{k'} = e_v, w_{k,1}, w_{k,2}, \ldots, w_{k,n} \tag{8}$$

Here, e_v represents the emotion vector. The emotion appended utterance representation is encoded using the utterance encoder and the final representation (including the emotion vector) is fed as input to the decoder for generating the responses with the desired emotions.

In the second approach, instead of providing the emotion information at the encoder side, we feed the emotion embeddings e_v during decoding at every decoder time-step. To include the emotion vector in the decoder, there is a slight change in Equation (5) and the new equation is as follows:

$$s_{d,t} = GRU_d(y_{t-1}, [s_{d,t-1}, U_a, P_a, e_v]) \tag{9}$$

6) Training and Inference: We employ commonly used teacher forcing [31] algorithm at every decoding step to minimize the negative log-likelihood on the model distribution. We define $y^* = \{y_1^*, y_2^*, \ldots, y_m^*\}$ as the ground-truth output sequence for a given input

$$\mathcal{L}_{ml} = -\sum_{t=1}^{m} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*) \tag{10}$$

We apply uniform label smoothing [32] to alleviate the common issue of low diversity in dialogue systems, as suggested in [33].

7) Baseline Models: As mentioned before, this is one of the very first attempts that considers persona information for generating emotional responses in a dialogue setting, and hence we did not find any closely related baselines in the literature. The existing works on persona [5], [16], [18] are not suitable baselines because they do not deal with incorporating emotions in the responses. Similarly, ECM [28], EMOTICONS [29] and EmoDS [19] are also not appropriate baselines as there is no provision of persona in them. Hence, for our baselines, we implement the following models:

- Seq2Seq: For this baseline only the input utterance is considered and no persona information is provided to the model.
- Seq2Seq + Attn: In this baseline, we add global attention [30] to enhance the performance of the decoder with no persona information.
- Seq2Seq + Attn + PAA: In this baseline we employ persona-aware attention for incorporating persona information in the generated response.
- Seq2Seq + Attn + EE: The emotion information in this baseline is incorporated at the encoder side along with the utterance for the generation of emotional responses. Here, persona information is not included in the generation.
- Seq2Seq + Attn + ED: The emotion vector is provided directly to the decoder for generating the responses with desired emotions. Again in this baseline we do not consider the persona information.

IV. Dataset

Due to the unavailability of an emotion-labeled persona aware dataset, we employ a semi-supervised approach to label the PersonaChat dataset.

A. Dataset Description

We perform experiments on the recently released ConvAI2 benchmark dataset, which is an extended version (with a new test set) of the persona-chat dataset [5]. The conversations are obtained from crowd workers who were randomly paired and asked to act the part of a given persona. This dataset contains 164,356 utterances in over 10,981 dialogues and has a set of 1,155 personas, each consisting of at least four profile texts. The testing set contains 1,016 dialogues and 200 never seen before personas. As the dataset is not labeled with emotions we annotate the utterances with the emotion classifier described below to achieve our goal of incorporating the emotions along with persona.

B. Emotion Classifier

We apply a semi-supervised approach for annotating the PersonaChat dataset. To label the dataset we make use of the EmpatheticDialogues(EmpD) dataset [27] which has 25k conversations grounded in emotional situations. The dataset is crowdsourced and has 32 fine-grained emotions, covering a wide range of positive and negative emotions. For training the emotion classifier, we used the situation description and the label, as input-label pair of the EmpatheticDialogues dataset. We trained several classifiers such as CNN, LSTM, Bi-LSTM on the EmpD dataset for predicting emotions from the given set of 32 classes. We also employed the transformer-based architecture used for building the emotion classifier. We used the BERT for Sequence classification model proposed in [36]. Evaluation results of the various classifiers for emotion are demonstrated in Table II. As stated in [37], the highest classification accuracy achieved was 48%, using the situation description and the label in the DeepMoji chain-thaw model proposed in [38]. By using the BERT based architecture, we were able to get an improvement of more than 10% for emotion classification.

Table II: Classification scores of Emotion on Empathetic-Dialogue data. E-F1 denote the weighted average F1 score of emotion

Model	E-F1
LSTM	37.06
CNN	34.90
Bi-LSTM	39.87
BERT [34]	61.74
RoBERTa [35]	59.89

Table III: Dataset Statistics

Dataset Statistics	Train	Valid	Test
No. of Dialogues	7686	1640	1655
No. of Utterances	124816	19680	19860
Avg. turns per Dialogue	12.51	12.73	12.74
Avg. words in a Response	11.89	9.57	10.75
No. of emotions per dialogue	7.4	6.5	5.1
No. of unique words	20322	13415	15781

C. Dataset Preparation

Finally for labeling the entire PersonaChat [5] dataset, we use the best-performing classifier *viz.* BERT. While the emotion labels of the PersonaChat dataset are not completely gold due to automatic annotation, we believe this dataset is good enough to be used for the generation, which is similar to what was found in [19], [28]. Detailed statistics of the PersonaChat dataset are provided in Table III.

V. Experiments

In this section, we present the implementation details along with the evaluation metrics used to evaluate the model's output (both automatic and human evaluation).

A. Implementation Details

All the implementations are done using the PyTorch[1] framework. For all the models including baselines, the batch size is set to 32. The utterance encoder is a bidirectional GRU with 600 hidden units in each direction. We use the dropout [39] with probability 0.45. During

[1]https://pytorch.org/

decoding, we use a beam search with beam size 10. We initialize the model parameters randomly using a Gaussian distribution with the Xavier scheme [40]. The hidden size for all the layers is 512. We employ AMSGrad [41] as the optimizer for model training to mitigate the slow convergence issues. We use uniform label smoothing with $\epsilon = 0.1$ and perform gradient clipping when the gradient norm is over 5. To reduce data sparsity all the numbers and names are replaced with <number> and <person>. All the out-of-vocabulary (OOV) words are replaced with the <UNK> token. We use 300-dimensional word-embedding initialized with Glove [42] embedding pre-trained on Twitter. Previous 3 turns are considered for dialogue history and maximum utterance length is set to 50. The variance σ^2 of Gaussian Kernel Layer is set as 1. We ran 15 epochs, and the proposed model took about 3 days on a Titan X GPU machine.

B. Evaluation Metrics

For proper evaluation of our model, we employ both automatic and human evaluation methods.

1) Automatic Evaluation Metrics: To evaluate the model at emotion and grammatical level, we report the results using the standard automatic metrics. To evaluate our proposed framework at the content level we report Perplexity [43]. Lesser perplexity scores signify that the generated responses are grammatically correct and fluent. We also report the results using standard metrics like BLEU-4 [44] and Rouge-L [45] to measure the ability of the generated response for capturing the correct information. We report Distinct-1 and Distinct-2 metrics that measure the distinct n-grams in the generated responses and are scaled with respect to the total number of generated tokens in order to avoid repetitive and boring responses [1]. To measure the emotional content in the generated responses we calculate the emotion accuracy using the pre-trained classifier (BERT) on the responses generated by the baseline and proposed models.

2) Human Evaluation Metrics: To analyze the response quality of the generated responses we use human evaluation to study the efficiency of the different baseline and proposed models. From the generated responses we randomly take 700 responses from the test dataset for qualitative evaluation. For a given input along with persona information, three annotators with post-graduate exposure were assigned to evaluate the correctness, emotion and persona consistency of the generated responses by the different approaches for the following three metrics:

1) Fluency (F): This metric is used to measure the grammatical correctness of the generated response. It checks that the response is fluent and does not contain any errors.
2) Emotion (E): It is used to judge whether the generated response is in accordance with the desired emotions.
3) Persona Consistency (PC): For this metric, we take care of the fact that the response generated is in accordance with the persona information of the speaker provided in the form of texts and is also coherent with the conversational history.

The scoring scheme for the human evaluation metrics in case of fluency is measured as follows:0- incomplete response or else incorrect response, 1- moderately correct response, and 2- correct response. The scoring scheme for emotion and persona consistency is 0: for the absence of emotion in the reply and the reply is inconsistent to the specified persona and 1: for the presence of emotion in the response along with the consistency of the response with the persona information. For the human evaluation metrics, we calculate the Fleiss' kappa [46] to determine the inter-rater consistency. For fluency, the kappa score is 0.75, and for emotion and persona consistency it is 0.77, indicating "substantial agreement".

VI. Results and Discussion

In this section, we present the experimental results (both automatic and human) along with the necessary analysis of the generated response.

A. Automatic Evaluation Results

The experimental results of the baseline, as well as the proposed framework, are presented in Table IV. It is evident from the table that our proposed model outperforms the baselines for all the metrics and the improvement is statistically significant [2]. The model with emotion given directly to the decoder performs better than the model in which the emotion is provided at the beginning of the utterance at the encoder side. By applying persona aware attention (PAA) in the seq2seq framework it is visible that the model performs better in case of BLEU metric with an improvement of 4.6% from the baseline Seq2Seq model. This clearly shows that by adding the persona information the model is able to generate informative responses that are more like the ground-truth response. In case of the proposed framework with emotion at the decoder (ED) along with PAA shows an improvement of 2 and 6.6 BLEU points from the baseline seq2seq model with PAA and the baseline seq2seq model, respectively.

It is obvious from the results that the generated responses by the proposed framework are better than the baseline seq2seq model as there is a drop in the perplexity scores of about 6% and 3% in the case of ED and EE, respectively. For Rouge-L, the proposed framework outperforms the baseline methods having at least 1% improvement.

We also report the emotion accuracy of the generated response. It is quite obvious that the responses conditioned with emotions have higher accuracy than the models with no emotion information. Also, through experimental

[2]we perform statistical significance t-test [**?**] and it is conducted at 5% (0.05) significance level

Table IV: Experimental results of different models. Here PAA represents Persona-Aware Attention, EE represents Emotion at Encoder, ED represents Emotion at Decoder

Model Description		Perplexity	BLEU	Rouge-L	Emotion Accuracy	Distinct-1	Distinct-2
Baseline Approaches	Seq2Seq	59.11	0.042	0.149	0.35	0.0125	0.0464
	Seq2Seq + Attn	58.23	0.047	0.151	0.38	0.0131	0.0472
	Seq2Seq + Attn + PAA	57.60	0.088	0.154	0.42	0.0163	0.0581
	Seq2Seq + Attn + EE	56.87	0.092	0.157	0.58	0.0155	0.0534
	Seq2Seq + Attn + ED	56.39	0.096	0.158	0.61	0.0158	0.0562
Proposed Approaches	**Seq2Seq + Attn + PAA + EE**	**55.59**	**0.099**	**0.162**	**0.65**	**0.0189**	**0.0844**
	Seq2Seq + Attn + PAA + ED	**52.68**	**0.108**	**0.169**	**0.67**	**0.0210**	**0.0923**

Table V: Results of Human Evaluation

Model Description		Fluency			Emotion		Persona Consistency	
		0	1	2	0	1	0	1
Baseline Approaches	Seq2Seq	27.36	45.83	26.81	75.93	24.07	77.20	22.80
	Seq2Seq + Attn	26.11	44.71	29.18	74.56	25.44	76.14	23.86
	Seq2Seq + Attn + PAA	23.41	42.96	33.63	73.81	26.19	51.64	48.36
	Seq2Seq + Attn + EE	24.17	43.11	32.72	59.33	40.67	70.88	29.12
	Seq2Seq + Attn + ED	23.05	42.88	34.07	57.49	42.51	70.31	29.69
Proposed Approaches	**Seq2Seq + Attn + PAA + EE**	**19.64**	**38.65**	**41.71**	**55.72**	**44.28**	**49.85**	**50.15**
	Seq2Seq + Attn + PAA + ED	**18.15**	**37.32**	**44.53**	**53.91**	**46.09**	**48.11**	**51.89**

results, it is noticeable that there is an increase of 2% in emotion accuracy in the ED framework in comparison to the EE framework. The possible reason for the improvement is that the decoder directly gets the emotion information which helps in infusing the correct emotions in the responses. In contrast to the baseline seq2seq model the proposed framework (Seq2Seq+Attn+PAA+ED) gets a very high improvement in emotion accuracy. We also provide the distinct-1 and distinct-2 results to showcase that the generated responses are diverse. From evaluation, it is evident that the proposed framework along with generating emotion and persona aware responses is also capable of making the response diverse and interactive.

B. Human Evaluation Results

For a thorough evaluation of our proposed framework, we perform the human evaluation, the results of which are reported in Table V. From the table, it is clearly evident that the proposed method performs better than the baseline models with respect to all the defined metrics. As fluency measures the grammatical correctness of the generated response, hence it can be concluded that the proposed framework generates responses that are fluent. As opposed to the baseline Seq2Seq framework the final model shows an improvement of 17% in case of fluency. Similarly, the emotional content of the generated response from the proposed framework with emotion information provided to the decoder has an increased emotion score of 1.8% than the model with emotion information at the encoder. This proves the fact that the desired emotion is expressed better in responses when the emotion information is given at very decoder step. Also, it is noticeable that there is a huge improvement in the emotional content of the generated responses from the different baseline models.

We measure the capability of the models to maintain a consistent persona while generating the responses. From the human evaluation results presented in the table, we can see that the Persona aware attention model shows a vast improvement of 25% from the baseline Seq2Seq model in inducing the persona information while generating the responses. In the proposed framework there is also an improvement from the baseline models in case of persona consistency metric. Hence, it can be concluded through human evaluation that the proposed framework not only is capable of generating emotional responses but also has the ability to maintain a specific persona. In Table VI, we present a few examples of the generated responses from the baseline as well as the proposed framework.

C. Error Analysis

After doing a detailed quantitative and qualitative analysis of the generated responses, we came across some of the mistakes made by our proposed framework. Some of the commonly occurring errors are:

- Repetition: In both the baselines as well as the proposed framework, there are some cases, where the information in the input is observed to be repeated. For example, Gold: *if I have time outside of hunting and remodeling homes*; Predicted: *if I have have time time hunting...*
- Unknown tokens: Since the proposed and baseline models use basic Seq2Seq framework, hence at times it generates unknown token (<UNK>) in case of named entities that occur less number of times in the training set. For example, Gold: *I am packing to visit my dad in China.*; Predicted: *I am going to visit my dad <UNK> <UNK> <UNK>.*
- Persona Inconsistency: The responses in some cases generated by the proposed framework is inconsistent with the persona information and lacks the specific details present in the persona texts of the speaker.
- Emotion mismatch: The baseline, as well as the proposed framework, is unable to express the desired emotions in the generated responses, thereby making the responses generic.

VII. Conclusion and Future Work

Response generation is a key component in every dialogue system. The ability to respond in a human-like

Table VI: Examples of responses generated by different models

Dialog Input	Persona Information		Ground-Truth	Generated Response
[Person 1] Hi! i work as a gourmet cook. [Person 2] I don't like carrots. I throw them away. [Person 1] Really. But, I can sing pitch perfect .	Persona 1 As a child, I won a national spelling bee. I've been published in the new yorker magazine. I am a gourmet cook. I've perfect pitch.	Persona 2 I'm very athletic. I have brown hair. I love bicycling. I hate carrots.	I also cook, and I ride my bike to work.	**Seq2Seq:** I also cook. **Seq2Seq+PAA+EE:** Really! I like to cook and ride a bike to work. (surprise) **Seq2Seq+PAA+ED:** Wow that's nice, but I like to cook and ride bike to work. (surprise)
[Person 1] Hi! how are you today? [Person 2] I had the day off, you? [Person 1] I only worked half a day at the bank.	Persona 1 I am a bank teller. I've never been out of the country. My favorite phone is as Iphone. I love to go hiking.	Persona 2 I m from Texas. I like basketball. I work many hours. My favorite band is imagine dragons.	I work a lot.	**Seq2Seq+PAA:** I work alot. **Seq2Seq+PAA+EE:** That's nice to hear but I have a busy schedule. (disappointed) **Seq2Seq+PAA+ED:** Ohh good for you but I work alot for my basketball. (disappointed)

manner is the ultimate goal of every conversational agent. In this work, we have focused on addressing two major aspects of response generation i.e., infusing emotions in the response and maintaining a consistent persona. We proposed a novel persona aware attention mechanism to the responses to make the generated response more specific, interactive and consistent with the speaker. Simultaneously, we also incorporate emotions in the response by employing two different methods. In the first method, we have included the desired emotion at the beginning of the utterance encoder while in the second approach we have provided the desired emotion vector directly to the decoder. Due to the unavailability of the emotion labeled persona dataset, we employ a semi-supervised approach to label the dataset with emotion labels. Both qualitative and quantitative analyses show that our proposed framework is capable of maintaining the persona information of the speaker while responding emotionally. The proposed method outperforms the various baselines in both automatic and human evaluation.

In future, along with the opportunity of extending the architectural designs and training methodologies to enhance the performance of our systems, we look forward to designing a specific component to enhance the natural language generation component of an end to end Chatbot, by including the appropriate mechanisms to interact with all its components (persona memory, emotion database, and the dialog manager). Moreover, we would focus on making the framework both persona and emotionally aware by using different techniques to incorporate emotions thereby generating better responses.

Acknowledgement

Authors duly acknowledge the support from the Project titled "Sevak-An Intelligent Indian Language Chatbot", Sponsored by SERB, Govt. of India (IMP/2018/002072).

References

[1] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 1192–1202.

[2] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2017, pp. 3295–3301.

[3] J. Juraska, P. Karagiannis, K. K. Bowden, and M. A. Walker, "A deep ensemble model with slot alignment for sequence-to-sequence natural language generation," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 152–162, 2018.

[4] D. Raghu and N. Gupta, "Hierarchical-pointer generator memory network for task oriented dialog."

[5] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2204–2213, 2018.

[6] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2775–2779, 2018.

[7] C. K. Joshi, F. Mi, and B. Faltings, "Personalization in goal-oriented dialog," *arXiv preprint arXiv:1706.07503*, 2017.

[8] L. Luo, W. Huang, Q. Zeng, Z. Nie, and X. Sun, "Learning personalized end-to-end goal-oriented dialog," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, vol. 33, 2019, pp. 6794–6801.

[9] H.-Y. Shum, X.-d. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, 2018.

[10] H. Prendinger, J. Mori, and M. Ishizuka, "Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game," *International journal of human-computer studies*, vol. 62, no. 2, pp. 231–245, 2005.

[11] B. Martinovski and D. Traum, "Breakdown in human-machine interaction: the error is the clue," in *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, 2003, pp. 11–16.

[12] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[13] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pp. 1577–1586, 2015.

[14] X. Wu, A. Martinez, and M. Klyen, "Dialog generation using multi-turn reasoning neural networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 2049–2059.

[15] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model,"

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, 2016.

[16] A. Madotto, Z. Lin, C.-S. Wu, and P. Fung, "Personalizing dialogue agents via meta-learning," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 2019, pp. 5454–5459.

[17] S. Yavuz, A. Rastogi, G.-L. Chao, and D. Hakkani-Tur, "Deepcopy: Grounded response generation with hierarchical pointer networks," *NeurIPS*, 2019.

[18] H. Song, W.-N. Zhang, Y. Cui, D. Wang, and T. Liu, "Exploiting persona information for diverse generation of conversational responses," *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 5190–5196, 2019.

[19] Z. Song, X. Zheng, L. Liu, M. Xu, and X.-J. Huang, "Generating responses with a specific emotion in dialog," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 2019, pp. 3685–3695.

[20] Q. Qian, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Assigning personality/identity to a chatting machine for coherent conversation generation," *IJCAI*, 2017.

[21] C. Huang, O. R. Zaiane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, 2018, pp. 49–54.

[22] J. Li and X. Sun, "A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 678–683, 2018.

[23] Z. Lin, P. Xu, G. I. Winata, Z. Liu, and P. Fung, "Caire: An end-to-end empathetic chatbot," *arXiv preprint arXiv:1907.12108*, 2019.

[24] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pp. 986–995, 2017.

[25] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "Affect-lm: A neural language model for customizable affective text generation," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 634–642, 2017.

[26] L. Kezar, "Mixed feelings: Natural text generation with variable, coexistent affective categories," in *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018,Student Research Workshop*, 2018, pp. 141–145.

[27] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 2019, pp. 5370–5381.

[28] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

[29] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia, "Affect-driven dialog generation," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3734–3743, 2019.

[30] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 1412–1421, 2015.

[31] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 2818–2826.

[33] S. Jiang and M. de Rijke, "Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots," *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pp. 81–86, 2018.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2018.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[37]

[38] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm."

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, 2010, pp. 249–256.

[41] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," *arXiv preprint arXiv:1904.09237*, 2019.

[42] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1532–1543.

[43] S. Chen, D. H. Beeferman, and R. Rosenfeld, "Evaluation metrics for language models," 1998.

[44] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[45] C.-Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, July 2004.

[46] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.