Emotion-aware and Intent-controlled Empathetic Response Generation using Hierarchical Transformer Network

Tulika Saha, Sophia Ananiadou
National Centre for Text Mining, The University of Manchester, United Kingdom
Email: sahatulika15@gmail.com, sophia.ananiadou@manchester.ac.uk

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892592

***Abstract*—Enriching any dialogue systems to exhibit empathy is fundamental for delivering human-like conversations. Empathetic interactions in the form of empathetic dialogue generation has been studied widely in recent times. Existing models either incorporate emotion as a feature at the encoding side or as a latent variable explicitly at the decoder to condition their response on. While understanding speaker emotion is integral to expressing empathy, another aspect of being empathetic necessitates responding with an appropriate emotion (also known as emotional regulating intents) to speaker's mental state. To integrate these multiple aspects, in this paper, we propose a Hierarchical Transformer Network (HTN), an amalgamation of the recently introduced Transformer model and Hierarchical Encoder Decoder (HRED) architecture to capture the speaker emotion and dialogic context. For generating intent controlled empathetic responses, we draw insights from Reinforcement Learning (RL) to optimize rewards implicitly. The proposed approach is demonstrated on two benchmark open-domain empathetic datasets. The empirical evaluation (both automated and manual) demonstrates the system capability by way of outperforming several baselines and the state of the art models.**

***Index Terms*—Empathy, Generation, Emotion, Intent, Transformer**

I. Introduction

The advancements of Artificial Intelligence (AI) and Natural Language Processing (NLP) have engaged vivid interest of researchers in creating automated Dialogue Systems [1]. Effective communication between humans and Virtual Assistants (VAs), popularly known as Natural Language Generation (NLG) is fundamental for any conversational agent. In recent years, two major approaches relevant to NLG have emerged. The first type covers open domain conversations [2] which are often casual chit-chatting. The second type is goal-oriented dialogues [3], in which the VA must engage with the user to complete a specific task of a domain. Advanced neural models, for example, Transformer Networks [4], Memory Networks [5] etc., have only recently gained considerable ability and access to large-scale datasets to generate useful responses in a chit-chat context.

For any human-facing VA, it is important and desirable for the VA to respond appropriately by acknowledging the feelings and emotions of the user, a trait commonly referred to as empathetic responding [6]. Recently, empathetic chatbots [7] have revolutionised dialogue systems as it has shown to create stronger human-VA interaction [8] since humans naturally express and perceive emotion in natural language in order to strengthen their feeling of social bonding [9], [10]. The goal of empathetic response generation task is to produce contextually relevant responses which are syntactically correct and, most importantly, are emotionally aligned. The two primary empathetic components [11] are : *(i)* the capability to understand human emotions and feelings, i.e., being emotion-aware and; *(ii)* the capability to respond with an appropriate emotion to another person's mental states. It is to be noted that in the context of dialogues, when a listener wants to for e.g., acknowledge or console the speaker, it is referred to as expressing an emotional intent rather than simply emotion[1]. For the former aspect, there have been various efforts recently devoted to increasing dialogue models' ability to grasp the feelings of interlocutors, which makes the responses somewhat empathetic [12], [13]. For the latter case, many existing works [14] employ a pre-specified emotion/intent label, typically a latent variable explicitly at the time of decoding to generate emotional responses. However, because an additional label is required as input (at the decoder), this may be impractical for implementing the chatbots in practise. Recent works [15], [16] have also seemed to rely on ruled based generation either explicitly or implicitly for predicting the emotion/intent of the response to be generated, for e.g., following/reversing the speaker's emotion, or just maximizing the emotion content in the response. Such deterministic criteria, however, are not supported by psychological literature, and they neglect the complex exchanges observed in human conversations, where the listener frequently demonstrates more neutral empathetic intents.

Precisely, common challenges with these chit-chat oriented empathetic VAs are : *(i)* normally these VAs are trained by gradient descent using maximum-likelihood estimation (MLE) objective function. However, these models often tend to produce generic responses for e.g., *"I don't know"* irrespective of the input and get caught in an infinite loop by generating repetitive responses. This is because with solely MLE based models, it is not very clear how well the VAs perceive the goal of chatbot creation, i.e., training the VA to engage the humans in conversation by producing diverse, empathetic and informative responses; *(ii)* inadequacy of the VAs to

[1]Thus, emotion/intent is used synonymously in the context of the listener

assimilate long-term memory explicitly as they are trained to generate responses conditioned on the recent dialogue context or history; *(iii)* providing emotion/intent label of the responses explicitly at the time of generation to capture empathetic-emotional interactions which make the VA less realistic. These problems or issues combined together results in an overall inadequate and unsatisfiable responses and experience for the users to converse with the VA.

The challenges faced by these empathetic conversational agents (discussed above) hint the need of a conversational framework with the ability to (i) model long-term memory of speaker utterances in the ongoing conversation; (ii) integrate emotion of the speaker to generate emotion-aware responses; (iii) guide the generation process implicitly to acknowledge the feelings of the speaker. To achieve these objectives, we draw on insights from Reinforcement Learning (RL), which has been applied widely in Dialogue Systems in a variety of contexts [17], [18]. In this paper, we present an emotion-aware and emotion/intent-controlled empathetic response generation framework capable of optimizing long-term rewards. The proposed model is a Hierarchical Transformer Network (HTN) which is an amalgamation of the recently introduced Transformer model [4] and Hierarchical Encoder Decoder (HRED) architecture [19], [20] to capture the speaker's emotion and dialogic context. We implicitly supervise the HTN model to learn empathetic-emotional interactions and generate emotion/intent-controlled empathetic responses using the RL objective. The proposed model thus, aims to leverage the strength of the HTN model to learn and ensure compositional semantic meaning of utterances while also benefitting from the power of RL in order to optimize long-term rewards. The proposed model is demonstrated on two benchmark empathetic conversational datasets.

The key contributions of this paper are as follows :

- *We propose an emotion-aware and emotion/intent-controlled empathetic response generation framework that utilizes an amalgamation of the recently introduced Transformer model and Hierarchical Encoder Decoder (HRED) architecture*;
- *The utility of RL helps optimize long-term rewards to implicitly supervise the HTN model to precisely learn the empathetic-emotional interactions present in the dialogue dataset*;
- *Empirical results indicate that the proposed approach outperforms several of its counterparts and state of the art models*.

II. Related Works

a) Neural Response Generation: The area of chit-chat based response generation has been studied widely over the years in different aspects. In [17], authors developed a NLG framework that optimize long-term rewards in a RL set-up by utilizing the traditional SEQ2SEQ model. The authors of [21] developed an adversarial framework consisting of a discriminative model to discriminate between actual and false chats on a simple chatbot for evaluating its performance. Authors of [20], developed an alternative to the HRED model for response generation which typically models hierarchy of sequences using two RNNs, one for word level encoding and the other for utterance level encoding of a dialogue. The recent advent in deep learning owing to the introduction of robust and context-sensitive pre-trained language models [22] have revolutionized several downstream tasks including response generation frameworks.

b) Empathetic Response Generation: In [23], authors created a hierarchical encoder-decoder model that identified the user's emotional state and utilized it to generate the response. In [7], authors proposed a empathetic conversational dataset with 32 fine-grained emotion intents manually tagged for the speaker utterances. However, due to expensive human effort, this dataset is limited in size containing 25k conversations. In [24], authors analyzed the conversations in [7] dataset and proposed nine more categories of fine-grained emotion/intents for the listener. Subsequently, in [25], authors introduced a large-scale, silver-standard empathetic conversational dataset with 41 fine-grained categories [24] of emotion/intent of the speaker-listener. In [26], authors publicly released a conversational AI based empathetic response generation model utilizing the traditional Seq2Seq model with transformer encoder structure. Authors of [12], proposed an emotion aware empathy generation model by utilizing the traditional Seq2Seq model with transformer encoder structure.

III. Proposed Methodology

A. Problem Definition

In this manuscript, our aim is to solve the task of empathetic response generation in dialogues between two speakers based on the textual information available. The empathetic response generated should be conditioned on multiple aspects of the conversation (considered in this paper) as : **(i)** the conversational history, also known as context; **(ii)** emotion of the speaker; and **(iii)** emotion of the listener for whom the response is being generated. Formally, for a given speaker (S_1) utterance, $U_t^1 = (u_{t,1}, u_{t,2}, ..., u_{t,n})$, a conversational context/history, $C = (c_1, c_2, ..., c_{t-1})$, emotion of the speaker, e_u^1 and emotion of the listener, e_y^2, the task is to generate next empathetic textual response of the listener (typically the next speaker (S_2)), $Y_t^2 = (y_{t,1}, y_{t,2}, ..., y_{t,n''})$, where n and n'' are the length of the input and response utterances, respectively. Thus, we focus on generating the next empathetic response for the listener which should be coherent and in accordance with the several aspects of communication for an enhanced human-human or human-machine interactions.

B. Methodology

In this section, we will discuss each of the components of the proposed framework.

1) Hierarchical Transformer Network: The architectural framework of our proposed method is shown in Figure 1, which is an amalgamation of the Transfomer [27] and Hierarchical Encoder Decoder (HRED) networks [19], [20]. In contrast to the commonly used sequence to sequence

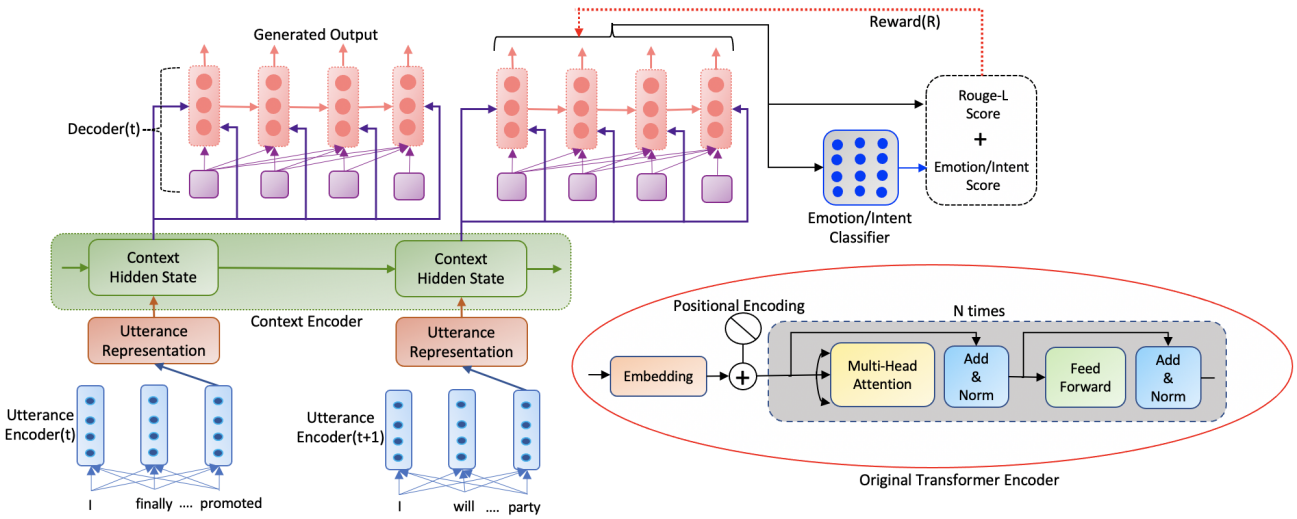


Fig. 1: Overall architectural diagram of the proposed empathetic response generation framework

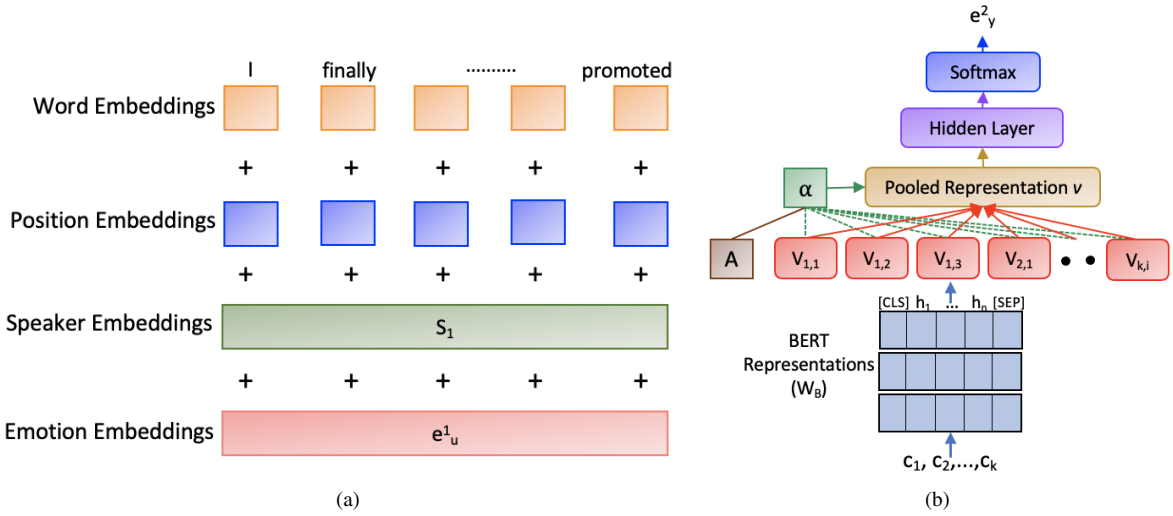


Fig. 2: (a) Input Representation of the HTN, (b) Architecture of the emotion/intent classifier

(SEQ2SEQ) model [28], [29], in HRED, a distinct RNN over the encoder RNN is used to simulate the conversational context or dialogue history, resulting in a hierarchical structure known as the hierarchical encoder. However, the main issue with RNN based models is its inability to provide parallelization while processing. The RNN based models provide sequential processing, i.e., the value of the next time-step cannot be computed unless we have the output of the current. This makes RNN based approaches slow. Also, since the sequences are encoded in the word level for RNN based models, they are incapable of preserving context at the encoder side for longer sequences. In order to counter this, we make use of the Transformer network [27], popular for its robustness over its predecessors as it process sentences as a whole using attention mechanisms and positional embeddings. Thus, our proposed methodology is an amalgamation of HRED architecture with the Transformer model, known as Hierarchical Transformer Network (HTN). The main components of the HTN are *utterance encoder*, *context encoder* and *decoder* which are discussed as follows :

a) Utterance Encoder: To encode the utterance of a speaker, U_k into vector representations, we employ the transformer encoder structure [27] as our utterance encoder. The transformer encoder typically comprises of multiple layers of multi-head self-attention module followed by a feed-forward layer with residual connections [30] and layer normalization [31]. The utterance in consideration, U_k is marked by its corresponding speaker label, $S_{U_k} \in \{0, 1\}$ (for two speakers). Along with the speaker label, we incorporate the speaker's emotion, $e^{U_k} \in [0, g)$ (g represents the number of emotion/intent categories) in the utterance encoder for a richer semantic representation of the input. Assuming that for a token, $u_{a,b}$, its corresponding token id and position id are represented as $w_{u_{a,b}} \in [0, |v|)$ (v is the vocabulary of the dataset) and $p_{u_{a,b}} \in [0, l)$ (l is the maximum input length of the encoder), respectively, each token is represented as the

summation of the following embeddings :

$$E_{u_{a,b}} = M_w[w_{u_{a,b}}] + M_p[p_{u_{a,b}}] + M_S[S_{U_a}] + M_e[e_{U_a}] \quad (1)$$

where $M_w \in \mathbb{R}^{|v|\times d}$, $M_p \in \mathbb{R}^{l\times d}$, $M_S \in \mathbb{R}^{2\times d}$ and $M_e \in \mathbb{R}^{g\times d}$ represents the embedding matrices of word, position, speaker and emotion of the speaker, respectively. Thus, this transformer based utterance encoder converts the embedding of the utterance, E_{U_k} into a list of hidden representations, $h_{k,1}, h_{k,2}, ..., h_{k,n}$. We use the hidden state at the last position, $h_{k,n}$ as the final textual representation of the utterance, H_{U_k}. A detailed representation of the input at the encoder is depicted in Figure 2a.

b) Context Encoder: The context encoder is yet another transformer encoder albeit applied on the utterance level instead of word level (as in utterance encoder) to capture context across a dialogue. This is achieved by applying the transformer on a sequence of utterance representations, $H_{U_1}, H_{U_2}, ..., H_{U_k}$ and subsequently obtaining a context-sensitive dialogue representation, $H_C = (h_{c_1}, h_{c_2}, ..., h_{c_k})$ occurred so far. In this manner, a hierarchical encoder is created on top of the utterance, U_k to model the conversational context.

c) Decoder: The context encoder's final hidden state representation is used as the decoder's initial state representation. Similarly, we employ the transformer decoder structure [27] as our decoder. In the current work, we slightly modify the original transformer decoder structure which comprises of two multi-head attention modules followed by a feed-forward layer to incorporate conversational context in both the encoder and decoder. However, in our case, the model only needs one multi-head attention module to learn the decoder context as the context at the encoder is available in the form of a vector (H_C) followed by the point-wise feed-forward layer as in the original transformer decoder. To generate the next empathetic textual response, $Y_k = (y_{k,1}, y_{k,2}, ..., y_{k,n''})$, the decoder then generates words consecutively at each time step (say j) based on the previously decoded words, $y_{k,1}, y_{k,2}, ..., y_{k,j-1}$ and final state representation of the context encoder, H_C. At each time-step j, decoder produces the probability of output tokens w using softmax.

$$P(y_{k,j}|y_{j-1}, ..., y_1) = exp(w_j H_d) / \sum_{j=1}^{w} exp(w_j H_d) \quad (2)$$

where H_d represents hidden state representation of the decoder after generating token at $j-1^{th}$ step. Initially, the HTN model is supervisedly trained with the negative log likelihood, i.e., the MLE objective function. Assuming, $Y_k^* = y_{k,1}^*, y_{k,2}^*, ..., y_{k,n''}^*$ is the ground-truth response or output sequence, then the MLE loss is defined as :

$$L_{MLE} = -\sum_{j=1}^{n''} \log P(y_j|y_{j-1}, ..., y_1) \quad (3)$$

This HTN model trained with the MLE objective function is utilized below for initialization to produce intent (emotion regulating intents) controlled empathetic responses.

2) Emotion/Intent Classifier: We train an emotion/intent classifier to predict the emotion/intent of the generated empathetic response, Y_k by the HTN model based on the context, $C = (c_1, c_2, ..., c_k)$, i.e., including both the speaker utterance and the available conversational history. We use the pre-trained BERT model [32] to obtain context-dependent vector representation, $v_{k,i}$ for each of the input token, $c_{k,i}$. In order to obtain a single vector representation, we pool these token-wise representations using a simple attention mechanism. A trainable attention vector, A is introduced to obtain an attention weight, $\alpha_{k,i}$ for each of the tokens, $v_{k,i}$ as,

$$\alpha_{k,i} = \exp(A^T v_{k,i}) / \sum_{l=1}^{k} \sum_{m=1}^{i} \exp(A^T v_{l,m}) \quad (4)$$

The aggregated representation, v is obtained as,

$$v = \sum_{l=1}^{k} \sum_{m=1}^{i} \alpha_{l,m} v_{l,m} \quad (5)$$

This representation, v is then fed into feed-forward layer followed by a softmax layer to predict the emotion/intent, e_y^2 of the ground-truth response, Y_k^*. We use this trained classifier as a pre-trained module to predict the emotion/intent of the generated response, Y_k by the HTN model (explained below). The architectural diagram of the emotion/intent classifier is presented in Figure 2b.

3) Reinforcement Learning (RL): The sequence of utterances in a dialogue can be considered as actions chosen by the HTN model based on a policy learned. The model is then tweaked using the MLE parameters to learn a policy that maximises long-term future rewards [17]. The elements of the RL-based training are explored below.

a) State and Action: The state is similar to the input of the HTN model, i.e., context, H_C comprising of history and current utterances (explained above), $[S(H_C = h_{c_1}, h_{c_2}, ..., h_{c_k})]$. The action a generates the utterance, Y_k at the next time step. Because the generated sequence might be of any length, the action space is unlimited. As a result, the policy, $\Pi(Y_k|S(H_C))$ is defined by its parameters and is learning the mapping from states to actions.

b) Reward: Here, we discuss the task-specific reward functions, r, used to evaluate the predicted output, Y_k against the true output, Y_k^*.

- **ROUGE-L Metric Score** (r_1) : This metric assures the matching of the longest common sub-sequence between the predicted and the true output. This is done to make the predicted response as closer as possible to the ground-truth response.
- **Emotion/Intent Score** (r_2) : With the availability of robust language models, it is now possible to learn delicate emotional interactions directly from the dataset itself. To ensure that the empathetic-emotional interactions are

learnt precisely from a given dataset distribution, we implicitly supervise the HTN model to generate emotion/intent controlled empathetic response, Y_k aligned with the conversational context, C and emotion of the speaker, e_u^1. This emotion/intent score metric will ensure that the emotion/intent of the generated output, Y_k by the HTN model is consistent with the emotion/intent of the true output, Y_k^*. Thus, the reward is :

$$r_2 = \begin{cases} +1, & \text{if } EIC(Y_k) == EIC(Y_k^*) \text{ and } eis \geq 0.5 \\ +eis, & \text{if } EIC(Y_k) == EIC(Y_k^*) \text{ and } eis \leq 0.5 \\ -eis, & \text{if } EIC(Y_k) \neq EIC(Y_k^*) \end{cases}$$

where EIC is the pre-trained emotion/intent classifier (explained above). eis is the emotion/intent score (i.e., softmax score) obtained from the classifier.

Thus, the final reward (R) is the weighted average of all the above terms as,

$$R = (r_1 * \beta + r_2 * (1 - \beta))/2$$

where β is the parameter of the model. These rewards are optimised using the Policy Gradient algorithm [33]. The policy model Π is initialized using the trained HTN model (using the MLE objective function). So, the final loss back-propagated to the HTN model is a combined objective function as :

$$L_{comb} = \eta L_{RL} + (1 - \eta) L_{MLE}$$

where η is the parameter of the model and L_{RL} and L_{MLE} are the losses calculated from the RL and MLE objectives, respectively.

C. Experimental Details

a) Dataset: We evaluated our model on two benchmark empathetic conversational datasets :

- **EmpatheticDialogues (ED)** dataset [7] : ED dataset is a gold dataset collected from real participants/workers comprising of 24,850 conversations which are grounded on 32 fine-grained emotion labels. The dataset is split into 80%-10%-10% for training, validation and testing sets comprising of 19533, 2770 and 2547 dialogues , respectively, for the generation framework. We randomly sample 10k conversations from the dataset to train the emotion/intent classifier.
- **Emotional Dialogs in OpenSubtitles (EDOS)** dataset [25] : EDOS dataset is a large-scale, silver dataset collected from movie and TV subtitles comprising of 1M conversations. The EDOS dataset is a subset of a larger corpus OpenSubtitles Dialogs (OS) dataset [34] containing approx. 4M conversations. The EDOS dataset is grounded on 32 emotion labels (similar to ED dataset) as well as eight empathetic response intents and one Neutral tag resulting in a total of 41 fine-grained categories of emotion/intent. We split the dataset into 80%-10%-10% for training, validation and testing sets for the generation framework. For training the emotion/intent classifier, we make use of the dataset released by the authors [25] specifically for the classifier.

TABLE I: Results of the emotion/intent classifier on different datasets

Model	EDOS		ED	
	Accuracy	F1-score	Accuracy	F1-score
Bi-LSTM	69.52	0.6147	72.74	0.6890
Bi-LSTM+Attention	71.05	0.6206	73.84	0.6933
BERT	76.38	0.7215	80.44	0.7731
BERT+Attention	**77.82**	**0.7302**	**82.04**	**0.7896**

b) Hyper-parameters: The RoBERTa [35] tokenizer is used to tokenize the speaker utterances with a vocabulary size of 50,265. The encoder, context encoder and decoder has 4 layers each with 6 heads in the multi-head self-attention module. The hidden dimension size is 768. We use Adam optimizer to train the model with a learning rate of 0.00004 and dropout rate of 0.1. The model parameters are initialized using the xavier distribution. The model captures conversational history of previous 3 dialogue turns. Beam search with beam size of 10 is employed at the time of decoding. For the baseline models using GPT and GPT-2, we decode using default temperature and top-k values. For the emotion/intent classifier, Categorical Crossentropy is used as the loss function. Here, as well Adam optimizer is used to train the model with a learning rate of 0.001.

c) Evaluation Metrics: The generative models are automatically evaluated using metrics such as BLEU-1 score [36], perplexity, ROUGE-L score [37] and embedding based metric [19]. A human evaluator from the authors' affiliation was asked to rate the quality of the responses generated given the speaker utterance and the context. During testing, the evaluator was shown 50 randomly selected simulated dialogues and asked to score the quality of the generated responses on a scale of 1 (worst) to 5 (best) based on the following metrics :

- ***Fluency*** : The generated response should be correct in terms of syntax and grammar;
- ***Coherency*** : The model should generate responses based on the current trajectory of the conversation, i.e. what is now being discussed;
- ***Empathetic*** : The response generated by the model should be empathetic in essence.

Finally, the average of the scores for different metrics are computed and reported below.

IV. Results and Analysis

A series of experiments were conducted for evaluating the proposed framework.

A. Comparison with the Baselines

Here, we compare our proposed approach against different baseline models.

a) Emotion-Intent Classifier: We compare the BERT-attention based emotion/intent classifier against its non-attention baseline as well as simpler baselines such as Bi-LSTM based models utilizing GloVe embeddings [38]

TABLE II: Automatic evaluation results on the EDOS dataset

Model	Embedding Metrics			BLEU	ROUGE-L	PPL
	Average	Extrema	Greedy			
Seq2Seq(Bi-LSTM)	0.504	0.285	0.332	3.20	3.95	56.84
HRED(Bi-LSTM)	0.527	0.314	0.357	4.29	5.55	60.16
GPT	0.567	0.344	0.382	5.94	9.07	62.48
GPT-2	0.626	0.395	0.435	8.05	14.38	63.22
Seq2Seq(Transformer) (OS-EDOS)	0.580	0.352	0.400	6.01	9.21	59.97
Seq2Seq(Transformer)	0.572	0.348	0.395	5.98	9.14	63.85
HTN (OS-EDOS)	0.629	0.401	0.452	8.12	14.77	62.33
HTN	0.622	0.390	0.429	8.02	14.36	66.85
HTN+RL(r_1) (OS-EDOS)	0.635	0.400	0.446	8.56	14.78	64.18
HTN+RL(r_2) (OS-EDOS)	0.641	0.403	0.457	8.82	15.01	63.90
HTN+RL(r1+r2) (OS-EDOS)	**0.650**	**0.416**	**0.502**	**8.95**	**15.36**	64.04

TABLE III: Automatic evaluation results on the ED dataset

Model	Embedding Metrics			BLEU	ROUGE-L	PPL
	Average	Extrema	Greedy			
Seq2Seq(Bi-LSTM)	0.296	0.103	0.230	1.48	2.22	65.72
HRED(Bi-LSTM)	0.298	0.107	0.231	1.52	2.24	68.46
GPT	0.396	0.285	0.304	3.03	5.82	63.09
GPT-2	0.398	0.281	0.307	3.56	6.21	63.81
Seq2Seq(Transformer) (OS-ED)	0.352	0.171	0.290	2.57	3.71	73.81
Seq2Seq(Transformer)	0.311	0.127	0.242	2.14	3.02	69.34
HTN (OS-ED)	0.495	0.283	0.308	3.81	6.64	74.41
HTN	0.377	0.210	0.297	3.04	4.08	69.60
HTN+RL(r_1) (OS-ED)	0.496	0.280	0.311	3.88	6.72	74.06
HTN+RL(r_2) (OS-ED)	0.506	0.293	0.324	4.06	7.53	74.58
HTN+RL(r1+r2) (OS-ED)	**0.512**	**0.301**	**0.336**	**4.19**	**7.62**	74.45

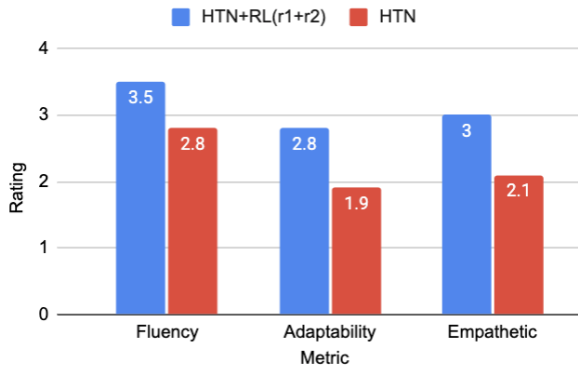
with/without integrating the attention module. The performance of the emotion/intent classifier and its baselines are reported in Table I. As evident the BERT based models produced better results as opposed to simpler models utilizing GloVe embeddings consistent with the literature. This is because BERT learns context-sensitive embeddings which makes it easier for the classifier to learn context-oriented emotional interactions. Additionally, the usage of the attention module boosts the performance of the classifier in both the set-up as it allows to identify relevant segments that helps to distinguish amongst emotion/intent(s). We also observe that the emotion/intent classifier produces better results on the ED dataset as opposed to the EDOS dataset. This can be attributed to the presence of much more fine-grained emotion/intents present in the EDOS dataset (about 41) compared to the 32 emotions in the ED dataset. Since the BERT+Attention based model gave better results on both the dataset, we chose this classification model for further experimentation on our RL based generation framework.

b) Empathetic Response Generation Framework: We compare our proposed *HTN+RL* generation framework with several baselines as follows : **(i)** Seq2Seq and HRED with Bi-LSTMs : these models are trained on traditional Seq2Seq and HRED models with Bi-LSTMs by utilizing GloVe based embeddings on individual datasets; **(ii)** GPT and GPT-2 : these models are fine-tuned on individual datasets using the pre-trained language models, GPT-1 (base) and GPT-2 (base); **(iii)** Seq2Seq (Transformer) : these are based on traditional Seq2Seq model but utilizing transformer encoder-decoder structures. OS-EDOS/OS-ED implies that the models are first pre-trained on the OS dataset and then further fine-tuned on the empathetic EDOS and ED dataset, respectively; **(iv)** HTN (Transformer) : these are the HTN model (proposed in the paper) without RL based fine-tuning.

Table II and III shows the performance of the proposed model and its different counterparts on the EDOS and ED dataset, respectively. As evident, the traditional SEQ2SEQ and HRED models were unable to produce better results compared to its transformer based counterparts (Seq2Seq and HTN) on individual datasets. This shows that long-term assimilation of memory for a particular utterance and across the dialogue was not appropriately learnt by the traditional SEQ2SEQ and HRED models. However, the transformer based models leverages from its ability to capture context-sensitive information across longer sequences. The transformer based Seq2Seq and HTN models pre-trained on the OS dataset and further fine-tuned on the individual empathetic datasets produced better results compared to its counterparts which weren't pre-trained. Supposedly, the pre-training on the OS dataset provided a stronger base for these neural models to be aligned with the context while further fine-tuning on empathetic datasets were focused on learning stronger empathetic-emotional interactions. Furthermore, the models trained to optimize long-term rewards produced better results in comparison to the base HTN models. This suggests that the RL objective did, in fact, aid in the generation of empathetic responses congruent with the context, rather than simply learning to be accurate at the token level. Out of the two RL objectives, emotion/intent score based rewards was rather more beneficial to boost the performance of the generation framework. However, the best result was obtained for the model trained with both the RL objectives. This is because both the rewards worked together towards ensuring that the generated responses are as close to the ground truth in terms of emotion/intent polarity and semantics. All the reported results are statistically significant as we have performed Welch's t-test [39] at 5% significance level. In Figure 3a, we report the results of the proposed (HTN+RL) and the basic HTN models pre-trained on the OS dataset during the human evaluation phase. As evident, the proposed framework attained the highest average fluency, adaptability and empathetic scores of 3.5, 2.8 and 3.0 respectively. However, both these models generated moderate replies consistent with the context, thus, demonstrating the need to address longer context/sequences more effectively. We present few examples of generated responses from the proposed and baseline models in Fig 3b. As evident, the baseline HTN model without any RL training generated generic responses, devoid of empathetic-emotional interactions. Whereas the proposed framework learnt a fair trade-off between being consistent with the speaker's emotional state and providing empathetic response.

B. Comparison with the State of the Art (SOTA)

We also evaluate the performance of our proposed model compared against the state of the art models. Table IV shows



(a)

Example-1
S_1 : It 's pretty new . Please take a look ... I don 't want it ...
Ground Truth : Do you want me to take a look?
HTN+RL(r_1+r_2) : Do you? You want me to take a look?
HTN : look! twhere?

Example-2
S_1 : Shit ! How could I forget to turn off the chiller ?
Ground Truth : Goddamnit , what have I done ? What ?
HTN+RL(r_1+r_2) : I'm the one who got them
HTN : i f*** up. i screwed up

(b)

Fig. 3: (a) Human evaluation scores of the models, (b) Sample response generated by the baseline and the proposed models

TABLE IV: Performance analysis against SOTA

Model	EDOS		ED	
	BLEU	ROUGE-L	BLEU	ROUGE-L
CAIRE (GPT) [26]	5.94	9.07	3.03	5.82
Seq2Seq (Transformer) + e_u^1 [12]	5.96	9.10	2.06	2.98
Seq2Seq (Transformer) [7]	5.98	9.14	2.14	3.02
Seq2Seq (Transformer) + Pre-trained-OS [25]	6.01	9.21	2.57	3.71
HTN+RL(r1+r2)	**8.95**	**15.36**	**4.19**	**7.62**

the performance of our proposed approach and several recent state of the art models. We re-implement the state of the art models following the information detailed in their respective work. This was done due to the absence of a standard train-test set for the EDOS dataset or unavailability of SOTA results on these specific dataset. As evident, the proposed model outperformed SOTA models on both the dataset.

C. Error Analysis

We thoroughly examined the performance of the best performing model (HTN+RL) and identified various cases in which the model failed, some of which are detailed below : **(i)** ***Incorrect Information*** : In some cases, the model outputs inaccurate utterances when contrasted to the truth, such as out of context responses. For example, Ground truth: "*Yes you can try that. It's always advisable to ask for help*", Predicted: "*I hope you get some rest.*"; **(ii)** ***Generic Response*** : When compared to the truth, some of the generated utterances are found to be general, i.e., lacking in empathic imparting expressions. For example, Ground truth: "*i've heard that your brother francisco was a hero.*", Predicted: "*i've heard about your brother*"; and **(iii)** ***Repetition*** : In a few instances, the model continues to repeat phrases from the ground truth. For example, Ground truth: "*Im so glad you are doing better and you found something to keep you busy!!!*", Predicted: "*glad that you are busy keep busy keep busy and do better*".

V. Conclusion and Future Works

Empathetic response generation is fundamental for any human-faced VAs to provide human-like conversational experience. Towards this aim, in this paper, we draw insights from AI to propose an emotion-aware and intent-controlled empathetic response generation framework. The proposed model is an amalgamation of the recently introduced Transformer model and HRED architecture to capture the speaker emotion and dialogic context. For generating intent controlled empathetic responses, we leverage from RL to optimize rewards in order to learn empathetic-emotional interactions implicitly. The proposed approach is demonstrated on two different benchmark open-domain empathetic datasets. The empirical evaluation (both automated and manual) demonstrates the system capability by way of outperforming several baselines and state of the art models. In future, attempts will be made to extend the framework in a multi-modal setting to incorporate additional information from speaker's facial representation to capture emotions. Additionally, a more sophisticated model will be established to capture long-term dialogic context. All these aspects will be investigated in our future works.

References

[1] A. Tiwari, T. Saha, S. Saha, S. Sengupta, A. Maitra, R. Ramnani, and P. Bhattacharyya, "A dynamic goal adapted task oriented dialogue agent," *Plos one*, vol. 16, no. 4, p. e0249030, 2021.

[2] T. Saha, S. Chopra, S. Saha, and P. Bhattacharyya, "Reinforcement learning based personalized neural dialogue generation," in *International Conference on Neural Information Processing*. Springer, 2020, pp. 709–716.

[3] T. Saha, S. Saha, and P. Bhattacharyya, "Towards sentiment-aware multi-modal dialogue policy learning," *Cognitive Computation*, pp. 1–15, 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[5] S. Sukhbaatar, J. Weston, R. Fergus *et al.*, "End-to-end memory networks," *Advances in neural information processing systems*, vol. 28, 2015.

[6] R. Roth-Hanania, M. Davidov, and C. Zahn-Waxler, "Empathy development from 8 to 16 months: Early signs of concern for others," *Infant Behavior and Development*, vol. 34, no. 3, pp. 447–458, 2011.

[7] H. Rashkin, E. M. Smith, M. Li, and Y. 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*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019, pp. 5370–5381.

[8] K. Liu and R. W. Picard, "Embedded empathy in continuous, interactive health assessment," in *CHI Workshop on HCI Challenges in Health Assessment*, vol. 1, no. 2, 2005, p. 3.

[9] F. Valente, "Empathy and communication: A model of empathy development," *Journal of new media and mass communication*, vol. 3, no. 1, pp. 1–24, 2016.

[10] T. Saha, A. Patra, S. Saha, and P. Bhattacharyya, "Towards emotion-aided multi-modal dialogue act classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4361–4372.

[11] S. G. Shamay-Tsoory, J. Aharon-Peretz, and D. Perry, "Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions," *Brain*, vol. 132, no. 3, pp. 617–627, 2009.

[12] R. Goel, S. Susan, S. Vashisht, and A. Dhanda, "Emotion-aware transformer encoder for empathetic dialogue generation," in *2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021 - Workshops and Demos, Nara, Japan, September 28 - Oct. 1, 2021*. IEEE, 2021, pp. 1–6.

[13] Y. Xie, E. Svikhnushina, and P. Pu, "A multi-turn emotionally engaging dialog model," in *Joint Proceedings of the Workshops on Human-AI Co-Creation with Generative Models and User-Aware Conversational Agents co-located with 25th International Conference on Intelligent User Interfaces (IUI 2020), Cagliari, Italy, March 17, 2020*, ser. CEUR Workshop Proceedings, W. Geyer, Y. Khazaeni, and M. Shmueli-Scheuer, Eds., vol. 2848. CEUR-WS.org, 2020.

[14] 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 Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3685–3695.

[15] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, "Affective neural response generation," in *European Conference on Information Retrieval*. Springer, 2018, pp. 154–166.

[16] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "Moel: Mixture of empathetic listeners," *arXiv preprint arXiv:1908.07687*, 2019.

[17] 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*, J. Su, X. Carreras, and K. Duh, Eds. The Association for Computational Linguistics, 2016, pp. 1192–1202.

[18] T. Saha, D. Gupta, S. Saha, and P. Bhattacharyya, "Towards integrated dialogue policy learning for multiple domains and intents using hierarchical deep reinforcement learning," *Expert Systems with Applications*, vol. 162, p. 113650, 2020.

[19] 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*, S. P. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 3295–3301.

[20] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 3776–3784.

[21] E. Bruni and R. Fernandez, "Adversarial evaluation for open-domain dialogue generation," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 284–288.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[23] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach," 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*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 5293–5300.

[24] A. Welivita and P. Pu, "A taxonomy of empathetic response intents in human social conversations," in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, D. Scott, N. Bel, and C. Zong, Eds. International Committee on Computational Linguistics, 2020, pp. 4886–4899.

[25] A. Welivita, Y. Xie, and P. Pu, "A large-scale dataset for empathetic response generation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 1251–1264.

[26] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, "Caire: An end-to-end empathetic chatbot," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 13 622–13 623.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[28] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks

[29] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, A. Moschitti, B. Pang, and W. Daelemans, Eds., 2014. [Online]. Available: https://doi.org/10.3115/v1/d14-1179

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.

[31] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

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

[34] P. Lison, J. Tiedemann, and M. Kouylekov, "Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Ro{bert}a: A robustly optimized {bert} pretraining approach," 2020. [Online]. Available: https://openreview.net/forum?id=SyxS0T4tvS

[36] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002*, 2002. [Online]. Available: https://www.aclweb.org/anthology/P02-1040/

[37] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[38] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[39] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, 1947.