Modelling Context Emotions using Multi-task Learning for Emotion Controlled Dialog Generation

Deeksha Varshney, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Patna, India
Email: {1821cs13, asif, pb}@iitp.ac.in

Abstract

A recent topic of research in natural language generation has been the development of automatic response generation modules that can automatically respond to a user's utterance in an empathetic manner. Previous research has tackled this task using neural generative methods by augmenting emotion classes with the input sequences. However, the outputs by these models may be inconsistent. We employ multi-task learning to predict the emotion label and to generate a viable response for a given utterance using a common encoder with multiple decoders. Our proposed encoder-decoder model consists of a self-attention based encoder and a decoder with dot product attention mechanism to generate response with a specified emotion. We use the focal loss to handle imbalanced data distribution, and utilize the consistency loss to allow coherent decoding by the decoders. Human evaluation reveals that our model produces more emotionally pertinent responses. In addition, our model outperforms multiple strong baselines on automatic evaluation measures such as F1 and BLEU scores, thus resulting in more fluent and adequate responses.

1 Introduction

One of the key skills for dialogue agents in a dialog system is to acknowledge the feelings of the user and respond accordingly. It is quite instinctive for humans to identify and understand other people's emotions but is quite hard for Artificial Intelligence (AI) systems due to the lack of representative publicly available data sets for training and evaluating an intelligent and robust dialog management system. Table 1 shows an example of emotion labelled conversation from the dataset. The example shows how two different emotionally inclined responses can lead a conversation in two different directions. An engaging conversation usually involves empathetic responses by conversing partners which can have varied emotion labels.

Agent 1	**Do you like wearing hats? It has so many functions.**	Curious
Agent 2	I don't like them on myself but I know a lot of people that can pull them off.	Neutral
Agent 1	Yes me as well. In the military hats denote a nationality, branch of service, rank or regiment.	Curious
Agent 2	Yes. I love hats! I have a wide variety of hats and wear them for different reasons.	Happy
Agent 1	Yes... Even I like it too !! Specially I am on vacation, roaming around I do carry 2–3 hats. And I wear it according to my dressing style.	Happy

Table 1: A snippet of two different emotionally inclined conversations with a common query.

It is important to capture user's affective information by any dialog agent to build an intelligent and socially engaging open-domain chatbot. For learning new tasks, we often apply the knowledge we have acquired by learning similar tasks. For instance, in Table 2 the context history has several utterances with Happy, Fearful, Disgusted and Curious to dive deeper emotion and the target responses are labelled with Fearful and Happy emotion. Context emotion can play an important role in transferring the target style while predicting the responses. The words *terrified, scary, afraid* and *like* can help in generating responses with the given target emotion label respectively. An auxiliary task of emotion classification can help in improving the main task of text generation.

In prior research, neural network based models handled the emotion controlled generation by either appending the target emotion label (Zhou et al., 2018; Zhou and Wang, 2017; Wang and Wan, 2018; Hu et al., 2017; Huang et al., 2018; Logeswaran et al., 2018; Song et al., 2019) or by using emotion embeddings (Asghar et al., 2018) in addition to

Agent 1	Are you **afraid** of snakes?	Curious
Agent 2	Hi, I am a little! but I was surprised there are none in New Zealand!	Happy
Agent 1	Sounds like a perfect place for me lol, I'm **terrified** of them	Fearful
Agent 2	Wow! I can understand , I am more **terrified** of crocodiles but it seems they are closer to birds than to snakes!	Fearful
Agent 1	Some snakes can even fly to catch their prey so thats **scary**	Curious
Agent 2	Wow, I would **like** to see that! And did you know its head is designed to swallow prays larger than them	Happy
Agent 1	Yeah I did know that, thats actually a bit disgusting, watching them eat prey	Disgusted
Agent 2	It looks like monkeys are terrified of snakes too!	Happy
Agent 1	They are? monkey are smart, they should stay as far as they can of snakes, dangerous animals	Fearful
Agent 2	Maybe you are terrified of snakes! But do you like dancing?	Happy

Table 2: Example conversations from the topical chat dataset showing different context emotion labels.

the input sentence representation. Although label information is effective, still it seems to be underutilized for effective response generation. Wang and Wan (2018) showed sentiment transfer using discriminator networks.

We hypothesize that emotional construct in a conversation can be formed by focusing on specific words in a dialog. To acknowledge the presence of annotated emotion labels in a multi-turn conversation, we perform emotion analysis of user utterances as an auxiliary task for open-domain dialogue generation. Our objective is to generate responses according to the target emotion style. Specifically, if we want to choose words that can provide information about the emotion of a sentence, we exploit an emotion classification model to govern the selection strategy. We train a self-attention (Vaswani et al., 2017) based encoder to compute the context features in a dialog. Words with higher attention weights are selected to be in the set of selections while decoding the response.

In this work, we propose to apply multi-task learning to leverage emotion information for open-domain response generation. Multi-task learning allows the encoder to learn common and prominent features in the input sequence. Our emotion-incorporated weights achieve a good balance between language fluency and emotion quality in model responses. We utilize focal loss (Lin et al., 2017) for emotion classification to address the imbalanced structure of the emotion distribution in the dataset. Furthermore, to attain better attention scores, we compute consistency loss in order to preserve the attention performance of individual tasks. Our empirical study does not show performance degradation in language fluency while classifying emotion-rich sequences.

We evaluate our proposed model on the Topical Chat dataset (Gopalakrishnan et al., 2019). We design human evaluation to score the following three metrics, *viz.* fluency, adequacy and emotional accuracy of the generated response. The human evaluation results indicate that our model improves not only the fluency and adequacy scores but also the emotional accuracy scores. In addition, we conduct automatic evaluation on the topical chat dataset. The automatic evaluation results show that our method improves significantly on the F1 and BLEU metrics.

The key contributions and/or attributes of our current work are summarized as follows:

1. We propose an effective deep multi-task framework that performs emotion classification and response generation.
2. To handle the imbalanced data distribution, we use Focal Loss (Lin et al., 2017) instead of regular cross entropy loss for emotion classification of utterances.
3. To maintain uniformity between the attention weights of different tasks, we utilise consistency loss (Nishino et al., 2019) in addition to the original task-specific losses.

2 Related Work

Early representative works were mostly based on the manually hand-crafted rules (Skowron, 2010; Polzin and Waibel, 2000), for generating responses with a specific emotion. Although rule-based approaches show high accuracy they often fail to handle complex emotions, especially for large corpora. In (Prendinger and Ishizuka, 2005), computational experiments established that empathetic agents ensure good communication. Ochs et al. (2008) designed an empathetic virtual agent that can express emotions based on cognitive appraisal theories which require an extensive hand-crafted rule base.

In recent years, there is an emerging research trend in an end-to-end neural network based generative conversational systems (Vinyals and Le, 2015;

Shang et al., 2015). To improve the content quality of neural conversational models, many techniques have been proposed, such as improving response diversity using Conditional Variational Autoencoders (CVAE) (Zhao et al., 2017) and encoding commonsense knowledge using external facts corpus (Ghazvininejad et al., 2018).

By expressing emotions, people show their mutual respect, empathy and understanding to each other, and thus improve the relationship between them. Emotional chatting machine (ECM) (Zhou et al., 2018) extended the basic encoder-decoder architecture using three mechanisms, *viz.* emotion category embedding, internal emotion memory, and external memory in order to generate sequence with a particular emotion label. Affect transfer in text using Recurrent Neural Networks (RNNs) (Ghosh et al., 2017) and text generation using emojis as the target labels (Zhou and Wang, 2017) was proposed for controlled generation of text. The research reported in (Niu and Bansal, 2018; Golchha et al., 2019) introduced state-of-the-art techniques for stylistic transfer of user behaviour, such as courteousness (e.g. polite, rude or neutral). Li et al. (2019) proposed an empathetic dialogue system (EmpGAN) based on adversarial learning comprising of a multi-resolution empathetic generator along with two interactive discriminators.

Song et al. (2019) presented an attention framework based on emotion-lexicons. Colombo et al. (2019) generated affect driven dialogues using emotion embeddings and affective sampling methods. Various techniques that can capture user's emotional state empathetic response generation were developed in (Asghar et al., 2018; Lubis et al., 2018) . An affective attention based model coupled with weighted cross-entropy loss was proposed by Zhong et al. (2019) for affective dialogue generation. Lin et al. (2020) built an empathetic chatbot which fine-tunes a Generative Pre-trained Transformer (GPT) with multiple objectives: response language modeling, response prediction, and dialogue emotion detection.

Multi-task learning, with deep neural networks which learn from different related-tasks has achieved remarkable success in improving the performance of many natural language processing (NLP) tasks (Luong et al., 2015a; Hashimoto et al., 2016; Liu et al., 2019). A multi-task learning framework usually consists of an encoder which is shared across multiple tasks to learn a common set of shared features. Moreover, the encoder learns to focus more on important and desirable features, and ignores redundant and noisy features (Ruder, 2017). Rashkin et al. (2018) proposed a new dataset with $\sim$ 25k conversations empathetic dialogue generation. The conversations in the dataset are prepared for a given emotion label. As opposed to this, our model handles dataset which has different emotion labels for every utterance in a dialog. As per our knowledge there is no existing work that has proposed the multi-task learning architecture for heterogeneous emotions in a conversation.

In our current work, we propose a multi-task framework with a shared multi-head self-attention based hierarchical encoder for response generation and emotion classification. We also utilize focal loss for emotion classification. Additionally, we incorporate a consistency based loss to enable persistent output generation for our multi-task architecture. The experiments are performed on the knowledge and emotion grounded Topical Chat dataset (Gopalakrishnan et al., 2019) containing a significant amount of human-human conversations in open-domain setting. Our approach tends to produce adequate responses.

3 Methodology

3.1 Problem Statement

In this work, we aim to produce emotion controlled responses for multi-turn conversations using relevant context knowledge and emotion labels. Let $U = u^{(1)}, ..., u^{(k)}, ..., u^{(K)}$ denote the set of K utterances of our multi-turn conversation. We represent I words of the k-th utterance as $u^{(k)} = w_1^{(k)}, ..., w_i^{(k)}, ..., w_I^{(k)}$. Each utterance $u^{(k)}$ is tagged with an emotion label $e^{(k)}$ i.e $E = e^{(1)}, ..., e^{(k)}, ..., e^{(K)}$. Hence, our task is to generate a response $y = y_1, y_2, ..., y_m$ with m words given the set of previous k context utterances and emotion labels.

3.2 Encoder-Decoder Model

3.2.1 Encoder

The encoder is used to transform the input utterance into a hidden representation $q^{(k)}$. The embedding, e, of the current word, $e(w_i^{(k)})$) and the positional embedding *PE(i)* is fed as input to the encoder. The combined embedding representation is subsequently passed into the Gated Recurrent Unit (GRU) model (Cho et al., 2014) which encodes the input utterance and yields relevant features. We

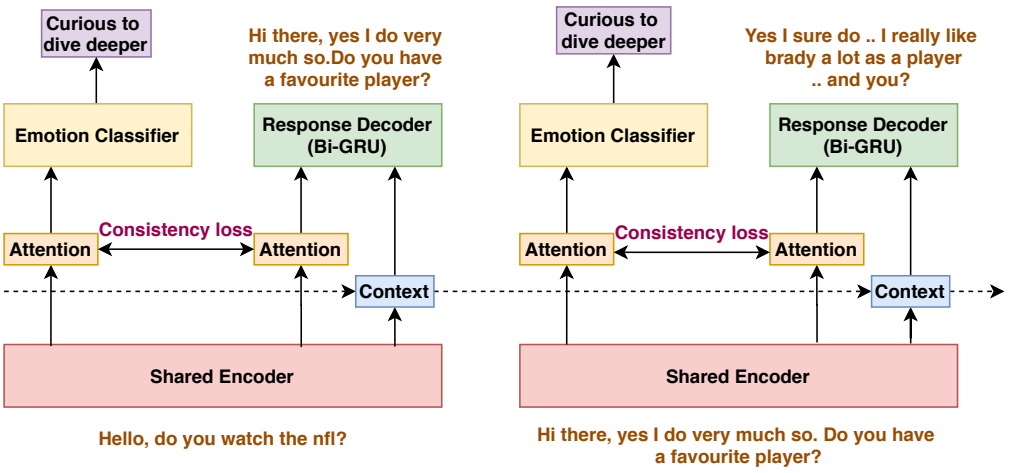

Figure 1: Proposed model architecture

apply self-attention (Vaswani et al., 2017) on the input features. Here, 'n' is the number of layers.

$$Ip_u^{(k)} = [w_1, ..., w_I] \quad (1)$$

$$w_i^{(k)} = e(w_i^{(k)}) + PE(i) \quad (2)$$

$$h_{e,t}^{(k)} = GRU_e(w_t^{(k)}, h_{e,t-1}^{(k)}) \quad (3)$$

$$(D^{(k)})^n = \text{MultiHead}(h^{(k)}, h^{(k)}, h^{(k)}) \quad (4)$$

$$(E^{(k)})^n = \text{FFN}((D^{(k)})^n) \quad (5)$$

$$q^{(k)} = (E^{(k)})^n \quad (6)$$

3.2.2 Context-level Encoder:

We use a GRU network to address the previous context of utterances in a multi-turn conversation. The initial state of the decoder GRU is initialised with the final hidden state of the context GRU.

$$h_c^{(k)} = GRU_c(q^{(k)}, h_c^{(k-1)}) \quad (7)$$

3.2.3 Decoder:

Intuitively, this layer takes what we have decoded so far, $h_{d,t-1}^{(k)}$, and all of what we have encoded, $q^{(k)}$, to produce a vector, $a_t^{(k)}$, that represents attention weights which signifies most important words in the source sentence in order to correctly decode, $\hat{y}_{t+1}$. We then calculate the energy, $e_{e,ij}^{(k)}$, between them by concatenating them together and passing them through a linear layer (attn) and a $\tanh$ activation function. The desired conditioning on previous utterances (context history) is obtained by initializing the hidden state of the GRU decoder with the final hidden state from the context GRU, $h_c^{(k)}$ i.e $h_{d,0}^{(k)} = h_c^{(k)}$.

$$h_{d,t}^{(k)} = GRU_d(e(y_t^{(k)}), w_t^{(k)}, h_{d,t-1}^{(k)}) \quad (8)$$

$$e_{e,ij}^{(k)} = v^T \tanh(\text{attn}(h_{d,t-1}^{(k)}), q_j^{(k)})) \quad (9)$$

$$a_t^{(k)} = \text{softmax}(e_{e,ij}^{(k)}) \quad (10)$$

$$g_t^{(k)} = a_t^{(k)} q_j^{(k)} \quad (11)$$

$$P(\hat{y}_{t+1}/y_{<t}) = softmax(e(y_t^{(k)}), g_t^{(k)}, h_{d,t}^{(k)}) \quad (12)$$

3.3 Multitasking Dialog Generation and Emotion Recognition

We perform multi-tasking using a shared encoder layer for encoding input sequences and two decoder layers for utterance prediction and classification. Figure 1 gives an overview of our proposed model.

Shared encoder: We use the encoder from Section 3.2.1 which converts the input sequence into hidden vectors ($q^{(k)}$) which is used across multiple tasks.

Classifier: The classifier transforms the shared representation from the encoder into the emotion class probability $p_c^{(k)}$.

$$p_c^{(k)} = \text{softmax}(Wq^{(k)} + b) \quad (13)$$

Decoder: We employ a GRU based decoder which takes the hidden representation from the shared encoder and generates a response $y = y_1, y_2, ..., y_m$ comprising of m words.

3.4 Focal Loss

Focal Loss (Lin et al., 2017) is employed to address imbalance between the emotion classes during training. We use focal loss as a replacement

of cross entropy loss for emotion recognition. It is defined in Eq 14, where γ is a focusing parameter.

$$L_1 = -(1 - p_c^{(k)})^{\gamma} log(p_c^{(k)}) \quad (14)$$

3.5 Consistency Loss

We use the "consistency loss" (Nishino et al., 2019) to reduce the difference between the attention weights from different tasks. Attention agreement favours emotional words while decoding the responses. The consistency loss between two different tasks is defined as follows:

$$L_{cl} = \sum_{i=1}^{I} |\max_j e_{p,ij}^{(k)} - \max_j e_{q,ij}^{(k)}|+ \quad (15)$$

where $e_{p,ij}^{(k)}$ is the attention weight for every k-th utterance for the p-th task. To compare the two attention weights, a ramp function $|x|+$ is used.

3.6 Training: Dialog generation

We denote the negative log-likelihood loss for dialog generation using L_2.

$$L_2 = -\sum_{t=1}^{m} logP(\hat{y}_{t+1}/y_{<t}) \quad (16)$$

The overall loss function for our proposed model is calculated as the total sum of losses from the two tasks and the consistency loss:

$$L_{all} = L_1 + L_2 + L_{cl} \quad (17)$$

where L_1 and L_2 signify the loss of the emotion classification and dialog generation task. L_{cl} indicates the consistency loss.

4 Datasets and Experiments

4.1 Dataset

We perform our experiments on the knowledge and emotion grounded Topical Chat dataset (Gopalakrishnan et al., 2019) with ∼11K dialogues. It is a multi-turn conversational dataset in which every utterance is annotated with an emotion label. There are a total of eight emotions (angry, disgusted, fearful, sad, happy, surprised, curious to dive deeper, and neutral) in the dataset. The data is split into 5 distinct groups: Train, Valid Frequent, Valid Rare, Test Frequent, and Test Rare. The frequent set contains conversations on entities frequently seen in the training set. The rare set contains conversations on entities infrequently seen in the training set. Table 3 provides the details of the dataset.

	#Conversation	#Utterances
Train	8628	188378
Valid Frequent	539	11681
Valid Rare	539	11692
Test Frequent	539	11760
Test Rare	539	11770

Table 3: Dataset details

Emotion Classes	Original Count
Curious to dive deeper	101162
Surprised	38254
Disgusted	1848
Sad	3070
Neutral	51796
Happy	36845
Angry	1133
Fearful	1174

Table 4: Distribution of emotion classes in topical chat dataset

4.2 Baselines

In order to prove the usefulness of our model, we compare it with the following baselines:

1. **HRED:** This baseline is defined based on the hierarchical encoder-decoder model by Serban et al. (2015, 2016). In this, the encoder RNN encodes the words of the utterances, and the context RNN encodes the dialog history.
2. **HRED-A:** We apply word-level attention (Luong et al., 2015b) to the encoder of the HRED model to capture important words of the input sequence.
3. **HRED-SA:** Another extension to the generative hierarchical Seq2Seq model with self-attention mechanism on the encoder which takes the dialog conversations as input.
4. **EmoHRED-A-FL-CL:** We extend the HRED-A model to EmoHRED-A-FL-CL, a deep multi-task learning framework that jointly performs the task of both response generation and emotion analysis. We add focal loss and consistency loss to the existing task specific losses.

To prove the effectiveness of our consistency loss in EmoHRED-SA-FL-CL, we conduct ablation study by removing the consistency loss from the EmoHRED-SA-FL-CL model. We name the model as EmoHRED-SA-FL. We also show the strength of the focal loss by eliminating FL from EmoHRED-SA-FL model. The resulting model is named as EmoHRED-SA.

Models	PPL *(Freq/Rare)*	BLEU% *(Freq/Rare)*	F1% *(Freq/Rare)*	Div.(n=1) *(Freq/Rare)*	Div.(n=2) *(Freq/Rare)*	Fluency *(Freq/Rare)*	Adequacy *(Freq/Rare)*	EA *(Freq/Rare)*
HRED	45.61 / 70.30	2.4 / 1.9	0.14 / 0.10	0.88 / 0.87	0.89 / 0.88	1.65 / 1.60	0.85 / 0.70	0.50 / 0.45
HRED-A	41.42 / 71.31	2.3 / 1.8	0.15 / 0.11	**0.91 / 0.90**	**0.90 / 0.90**	1.70 / 1.65	0.90 / 0.84	0.52 / 0.54
HRED-SA	36.63 / 54.87	2.1 / 1.8	0.21 / 0.15	0.83 / 0.82	0.84 / 0.84	1.70 / 1.65	0.98 / 0.88	0.60 / 0.55
EmoHRED-A-FL-CL	36.08 / 51.06	2.1 / 1.7	0.23 / 0.12	0.87 / 0.87	0.87 / 0.88	1.85 / 1.80	1.45 / 1.35	0.74 / 0.64
EmoHRED-SA-FL-CL	**35.45 / 50.45**	**2.6 / 2.1**	0.23 / **0.19**	0.88 / 0.87	0.89 / 0.88	**1.95 / 1.90**	**1.50 / 1.45**	**0.80** / 0.60
EmoHRED-SA-FL	36.34 / 54.82	2.3 / 1.9	**0.25** / 0.13	0.86 / 0.82	0.86 / 0.84	1.80 / 1.80	1.01 / 0.95	0.64 / **0.65**
EmoHRED-SA	36.04 / 52.98	2.3 / 1.8	0.24 / 0.13	0.88 / 0.83	0.83 / 0.84	1.83 / 1.81	0.93 / 0.81	0.53 / 0.51

Table 5: Evaluation results using automatic and human evaluation metrics for baseline, ablation, and our proposed model. Bold face indicates leading results for each metric.

4.3 Experimental Setup:

For the HRED model, we use a single layer bi-directional GRU (Cho et al., 2014). We extend the HRED model to HRED-A using the global attention mechanism (Luong et al., 2015b) at the encoder. For our proposed self-attention-based model, the number of encoder and decoder layers is set to 2 and the number of attention heads is 8 with the filter size equal to 2048. Word embedding dimension is chosen as 300, hidden dimension is set to 300. For the generator, we use the ADAM optimizer whose learning rate is fixed to 0.0001. While decoding the responses we use beam search with beam size set to 4.

4.4 Evaluation Metrics

Automatic Evaluation: We utilise the most well-known metrics for evaluating a sequence such as BLEU (Papineni et al., 2002), F1, perplexity (PPL) (Vinyals and Le, 2015) and n-gram diversity (Div.) (Gopalakrishnan et al., 2019).

1. **Perplexity:** We define perplexity in Equation 18. It is a measurement of how well a model can predict human responses. We report perplexity values on our frequent and rare test. N is the total number of samples in the test set and N_w is the total number of tokens in the entire test set.

$$PPL = \exp\{-\frac{1}{N}\sum_{i=1}^{N_w} log(P(y|U))\} \quad (18)$$

2. **BLEU:** To evaluate the predicted responses we compute BLEU score, a word-based metric which performs n-gram matching with the ground truth responses.
3. **F1:** We compute *unigram* F1-score[1] between the model prediction and the ground truth responses.
4. **N-gram diversity:** We evaluate the informativeness and diversity of sentences using N-gram diversity. It is defined in Eq 19. M is the total number of samples in the test set. The result is shown under the column - Div. (n=1) and Div. (n=2) in Table 5 on the frequent and rare test set.

$$Div = \frac{1}{M}\left[\frac{\text{\# unique n-grams}}{\text{\# words in predicted response}}\right] \quad (19)$$

Human Evaluation: To measure the quality of the generated text from a human perspective, we randomly sample 100 conversations from each model and with the help of two experts with post-graduate exposure we evaluated the predicted responses using the following metrics:

(i) Fluency: It is used to measure the grammatical correctness. (ii) Adequacy: It is used to measure contextual relevancy of the predicted response. (iii) Emotional Accuracy (EA): It checks how accurately one can infer the target emotion in the predicted response.

We assign a scores in {0,1,2} (representing "wrong", "acceptable" and "perfect") for indicating the level of fluency and adequacy of responses. We measure the emotional accuracy on a scale of 0-1 with '0' indicating the incorrect emotion and '1' the correct emotion. We compute the Fleiss' kappa (Fleiss, 1971) score, to measure the inter-annotator agreement. We obtain a kappa score of 0.90, 0.75, 0.76 for fluency, adequacy, and emotional content respectively denoting "good agreement".

5 Results and Analysis

We present the results for all our experiments in this section. Detailed results using both the automatic and human evaluation methods are shown in Table 5.

[1] https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py

5.1 Automatic evaluation results

In Table 5, we observe that the proposed model has high uni-gram and bi-gram diversities, demonstrating that the models learn to decode fluent and informative responses with great diversity. We observe relatively fewer repeated segments in the responses generated by our proposed model owing to a good Div.(n=1) and Div.(n=2) score. We observe significant improvement in BLEU and F1-scores when compared with the baseline models which support our multi-task learning architecture. Our proposed model seems to utilize the multi-task learning phenomenon and effectively utilize emotion labels associated with each utterances.

We also perform an ablation study for better understanding the contributions of the attributes of our model. As shown in Table 5, after we remove the consistency loss, both the emotion accuracy and perplexity performance become obviously worse, indicating that to generate persistent outputs, consistency between attention weights is critical for emotion understanding and model generation quality. We also test the importance of focal loss using the EmoHRED-SA model. As shown in Table 5, after we eliminate the focal loss, there is significant drop in EA and F1 which justifies our use of focal loss. We perform statistical significance test between our proposed and the baselines models using t-test at 5% (0.05) significance level, and showed that the improvement in our model is statistically significant.

5.2 Human evaluation results

Table 5 illustrates that our proposed model outperforms the other baseline models in terms of fluency, adequacy and emotion quality. Owing to a good fluency score of our proposed model, we observed fewer copying of sentences from the input utterance in the predicted response. The increment in the adequacy scores w.r.t baseline models verifies that the response generated by the proposed model comes out as more relevant. The emotional content score determines that the generated responses are more in line with the emotional sensitivity of the sentences.

In Table 6, we present few examples of the responses generated by one of the baseline model (HRED) and our proposed model given the desired emotion. As shown in the table, the responses predicted by the *EmoHRED-SA-FL-CL* model has mostly predicted adequate and emotionally relevant responses as compared to the baseline HRED model. For the fourth utterance, even though the HRED model gives an emotionally relevant reply but it seems highly inadequate with respect to the context where as the *EmoHRED-SA-FL-CL* model responds with an emotionally as well as contextually relevant reply. Detailed examples with outputs from all of our baseline and proposed model with the required emotion label can be found in the appendix.

5.3 Error Analysis:

In this section, we report the most commonly occurring errors that our proposed and baseline models encounter.

1. Common phrases: Some common phrases are repeated in the generated response. For instance *'i don't think i've ever heard about it though'*, *'i don't know much about it so i don't know much about it either.'* and *'i 'm not sure either. i've never been there'*. Due to data scarcity and less diversity in the data, the models may only have learned to predict the most frequent utterances. Since the dialogues are inherently ambiguous, predicting them accurately would require more data.

2. Repetition: The proposed model (EmoHRED-SA-FL-CL), in a few cases, go on repeating the information present in the predicted response. Predicted Response: *that's terrible. i'll have to check that out. i'll have to check it out!*. This lowers the count of unique uni-gram words in the generated response i.e the F1-score.

3. Emotional inconsistencies: In some cases, the proposed model (EmoHRED-SA-FL-CL) is unable to produce responses of particular emotion labels due to less occurrence of instances from those classes (angry, sad, fearful and disgusted). The less frequent emotions classes like anger, sad, fearful and disgusted get confused with the recurring classes like curious to dive deeper and surprised. Also, instances from 'Happy' and 'Surprised' emotion classes gets mixed up with each other. For example, in Table 6, the target response for Utterance 5 should have the emotion 'Happy' but it gets confused with the emotion 'Surprised' and generates an irrelevant response. Table 4 shows the distribution of emotion classes present in the dataset. More detailed examples can be found in the appendix.

Utterance 1	**do you enjoy sports?**	**Neutral**
EmoHRED-SA-FL-CL	i do! it's very interesting. i used to play video games when i was a kid i don't have any time.	Happy
HRED	i do. do you ?	Curious
Utterance 2	**i enjoy the getting together with friends part of it but i could take or leave the actual sports events except when i am watching my children play. how about you what's your favorite sport?**	**Happy**
EmoHRED-SA-FL-CL	i do too.	Neutral
HRED	i enjoy watching football games as well.	Happy
Utterance 3	**i would say football**	**Neutral**
EmoHRED-SA-FL-CL	nice! speaking of football, did you know women were not allowed to play at the same time ?	Curious
HRED	i don't really like football, but i don't think i've ever heard of it, what about you?	Curious
Utterance 4	**that is the favorite sport in my household as well! tom brady is quite the dedicated player both on and off the field, don't you think?**	**Curious**
EmoHRED-SA-FL-CL	i don't know much about him.	Neutral
HRED	i do not have a favorite team but i don't really follow it.	Neutral
Utterance 5	**yes, he is arguably the greatest steal in nfl history**	**Neutral**
EmoHRED-SA-FL-CL	i can't believe he has never eaten a strawberry ha ha. have you ever heard of him ?	Surprised
HRED	i think it's cool that the nfl has no written rule against female players. i don't know how that's possible.	Surprised

Table 6: Generated examples are from a continuous conversation from the frequent test set. EmoHRED-SA-FL-CL and HRED predict responses using the previous set of utterances and emotion labels.

5.4 Comparisons to the state-of-the-art

The original paper (Gopalakrishnan et al., 2019) proposed a transformer-based model, and observed a F1-score on frequent and rare test set of 0.16 / 0.16 and a Div.(n=1) and Div.(n=2) of 0.85 / 0.84 and 0.86 / 0.86, respectively. Following the diversity scores, we notice a significant improvement with Div.(n=1) and Div.(n=2) of 0.88 / 0.87 and 0.89 / 0.88, respectively, for our proposed model. Similarly, using CAiRE (Lin et al., 2020) we obtained a F1-score on frequent and rare test set of 0.13 / 0.13 and a Div.(n=1) and Div.(n=2) of 0.87 / 0.83 and 0.86 / 0.85, respectively. However, it is to be noted that like us, (Gopalakrishnan et al., 2019) and (Lin et al., 2020) did not focus on taking into consideration the context utterance and instead simply concatenated the context utterances and passed them as a single sequence into the transformer model. We also observe a significant improvement in the F1-score for our proposed model. We achieve a score of 0.23 / 0.19 for our task of emotion-controlled dialog generation. We adopt ECM (Zhou et al., 2018) for comparision, a Seq2Seq model that first proposed to generate emotional response using emotion category embeddings, internal and external memory mechanisms. We concatenate the dialog history into a long sequence and feed as input to the model. We compute automatic evaluation metric - F1-score of 0.14 / 0.13 and BLEU score of 1.9 / 1.6. Our model clearly outperforms the baselines with a huge margin.

6 Conclusion and Future Work

In this paper, we propose a new deep learning framework for modeling emotion-grounded conversations using emotion labels as the guiding attributes. Building an emotion-aware conversational agent is crucial in enhancing user interactions with long, engaging conversations.

Extensive experiments show that the predicted responses expressed high levels of emotional accuracy and content adequacy. We have also provided details of different kinds of errors found in section 5.3. In general, we show how a related task of emotion recognition along with appropriate loss functions can ensure emotional relevancy of the generated response and improves user engagement.

In the future, we intend to use pre-trained language models for the task of dialog generation using emotion labels. We also aim to extend our model to handle knowledge-grounded conversations.

7 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). Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya Ph.D. scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation

(formerly Media Lab Asia).

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *NAACL*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. *arXiv preprint arXiv:1704.06851*.

Hitesh Golchha, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Courteously yours: Inducing courteous behavior in customer care responses using reinforced pointer generator network. 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)*, pages 851–860.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. 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, Volume 2 (Short Papers)*, pages 49–54.

Qintong Li, Hongshen Chen, Zhaochun Ren, Zhumin Chen, Zhaopeng Tu, and Jun Ma. 2019. Empgan: Multi-resolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13622–13623.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015a. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015b. Effective approaches to attention-based neural machine translation. *EMNLP*.

Toru Nishino, Shotaro Misawa, Ryuji Kano, Tomoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2019. Keeping consistency of sentence generation and document classification with multi-task learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3186–3196.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Magalie Ochs, Catherine Pelachaud, and David Sadek. 2008. An empathic virtual dialog agent to improve human-machine interaction. In *Proceedings of the 7th international joint conference on Autonomous*

agents and multiagent systems-Volume 1, pages 89–96.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Thomas S Polzin and Alexander Waibel. 2000. Emotion-sensitive human-computer interfaces. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses user's affective states. *Applied Artificial Intelligence*, 19(3-4):267–285.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8).

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Marcin Skowron. 2010. Affect listeners: Acquisition of affective states by means of conversational systems. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 169–181. Springer.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuan-Jing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7492–7500.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xianda Zhou and William Yang Wang. 2017. Mojitalk: Generating emotional responses at scale. *arXiv preprint arXiv:1711.04090*.

A Samples Generated by the Proposed Model

A.1 Predicted responses when we have different emotion labels for every utterance.

A.1.1 When the emotions occur frequently

We observe that the predicted responses as shown in Table 7 tend to follow the target emotions accurately, however they sometimes may lack adequacy.

A.1.2 When the emotions occur rarely

We observe that the predicted responses as shown in Table 8 fails to generate adequate as well as emotionally relevant responses.

A.2 Predicted responses when we have the same emotion label for every utterance.

We observe that the predicted responses as shown in Table 9 are very close to the ground truth response and are also emotionally accurate.

Utterance 1	do you enjoy sports?	Neutral
EmoHRED-SA-FL-CL	i do! it's very interesting. i used to play video games when i was a kid i don't have any time.	Happy
HRED	i do. do you?	Curious
HRED-A	yes, i do. do you?	Curious
HRED-SA	i do. do you?	Curious
EmoHRED-A-FL-CL	i do enjoy football, but i don't really follow it.	Curious
EmoHRED-SA-FL	i do. do you?	Curious
Utterance 2	i enjoy the getting together with friends part of it but i could take or leave the actual sports events except when i am watching my children play . how about you what's your favorite sport?	Happy
EmoHRED-SA-FL-CL	i do too.	Neutral
HRED	i enjoy watching football games as well.	Happy
HRED-A	i tend to enjoy college football teams.	Happy
HRED-SA	i am a big fan. i don't really know much about football	Neutral
EmoHRED-A-FL-CL	i am not a big football fan but i would love to watch it.	Happy
EmoHRED-SA-FL	i'm a big football fan, do you?	Curious
Utterance 3	i would say football	Neutral
EmoHRED-SA-FL-CL	nice! speaking of football, did you know women were not allowed to play at the same time?	Curious
HRED	i don't really like football, but i don't think i've ever heard of it, what about you?	Curious
HRED-A	football is a good sport. i'm a huge fan of football. i was surprised to learn that the nfl uses tracking chips in the players shoulder pads.	Surprised
HRED-SA	i'm not a fan but i don't really know much about football. i don't think i've ever heard of it.	Neutral
EmoHRED-A-FL-CL	i like the nfl. i don't think i would ever play soccer because i don't think i've ever seen it though.	Neutral
EmoHRED-SA-FL	i don't think i've ever seen it but i don't think i've ever seen it but i don't think i've ever seen it but i don't think i've ever seen it . did you know there is only 11 minutes of live gameplay?	Curious
Utterance 4	that is the favorite sport in my household as well ! tom brady is quite the dedicated player both on and off the field, don't you think?	Curious
EmoHRED-SA-FL-CL	i don't know much about him.	Neutral
HRED	i do not have a favorite team but i don't really follow it	Neutral
HRED-A	he is a good player. i think he is a good player but his sister was always better than him	Neutral
HRED-SA	i don't really know much about him	Neutral
EmoHRED-A-FL-CL	i am not sure if i am a football fan. i would like to be able to retire throwing kicks.	Happy
EmoHRED-SA-FL	well i am not sure who he is. he is one of my favorite shows	Neutral
Utterance 5	yes, he is arguably the greatest steal in nfl history	Neutral
EmoHRED-SA-FL-CL	i can't believe he has never eaten a strawberry. have you ever heard of him ?	Surprised
HRED	i think it's cool that the nfl has no written rule against female players. i don't know how that's possible.	Surprised
HRED-A	that's so cool! i'm going to have to look it up. i can't believe that tom brady has never eaten a strawberry in his entire life!	Surprised
HRED-SA	i didn't know that. i don't think i've ever heard of reggie miller. i don't think i've ever heard of him?	Neutral
EmoHRED-A-FL-CL	i wonder if he has ever eaten a strawberry? i wonder if he was born in kenya?	Curious
EmoHRED-SA-FL	i think he is a great player, i think he is a great player, i think he is a great player	Neutral

Table 7: Generated examples are from a continuous conversation from the frequent test set. Each model predicts responses using the previous set of utterances and emotion labels.

Utterance 1	hi, how are you?	Happy
EmoHRED-SA-FL-CL	hello there, i am well thank you. do you watch youtube?	Curious
HRED	i'm doing well, how about you?	Curious
HRED-A	i am good thank you. do you like basketball?	Curious
HRED-SA	i'm great! how are you?	Curious
EmoHRED-A-FL-CL	i 'm doing well thank you. do you like to travel?	Curious
EmoHRED-SA-FL	i'm good. do you like to travel?	Curious
Utterance 2	hi i m doing great how about you?	Curious
EmoHRED-SA-FL-CL	i am well, thanks. do you listen to country music ?	Curious
HRED	i'm great! do you like dogs ?	Curious
HRED-A	i am well thank you. have a great day!	Neutral
HRED-SA	i am well thank you. do you have a dog?	Curious
EmoHRED-A-FL-CL	i'm good do you have a favorite team?	Curious
EmoHRED-SA-FL	i'm good . do you like dogs ?	Curious
Utterance 3	i'm good, thanks for asking. do you like football ?	Curious
EmoHRED-SA-FL-CL	i do, what about you? do you have a favorite football team? do you have a favorite football team? do you have a favorite nfl team?	Curious
HRED	i do like football, do you?	Curious
HRED-A	i sure do. do you have a favorite team?	Curious
HRED-SA	i am not a huge fan of football. do you?	Curious.
EmoHRED-A-FL-CL	i do, i don't follow it much. do you?	Curious
EmoHRED-SA-FL	i do. do you?	Curious
Utterance 4	i like watching it sometimes, i guess it depends on who is playing lol how about u?	Curious
EmoHRED-SA-FL-CL	it's interesting that it has been around since 1989.	Curious
HRED	i don't watch football. i don't know much about football.	Neutral
HRED-A	yep, it is a very good game. do you have a favorite nfl team?	Curious
HRED-SA	i do too. did you know there is only 11 minutes of live gameplay?	Curious
EmoHRED-A-FL-CL	i am not a huge football fan but i don't think i've watched it a lot.	Neutral
EmoHRED-SA-FL	i do. i'm not a huge fan. do you?	Curious
Utterance 5	me too, the games can get kind of long to watch. the last super bowl i watched was over three hours long.	Disgusted
EmoHRED-SA-FL-CL	yea it's a classic game but i didn't think it started i see, i don't have a favorite team, i don't have a favorite team. i don't have a favorite team.	Neutral
HRED	that's cool i didn't know it was that long ago!	Surprised
HRED-A	wow that's really cool. i'm a steelers fan. i don't really have the time though.	Neutral
HRED-SA	i agree with you there, i didn't know there is a variation of soccer where 3 teams face each other at the same time.	Curious
EmoHRED-A-FL-CL	oh i see, i don't know much about football though, do you watch tv?	Curious
EmoHRED-SA-FL	i see, i don't have a favorite team, i don't have a favorite team. i don't have a favorite team.	Neutral
Utterance 6	that's so weird, especially since its only 11 min of game play	Curious
EmoHRED-SA-FL-CL	i agree with you. i know that soccer makes more money than football, basketball, baseball, golf, tennis and soccer.	Curious
HRED	yeah, i was not aware of that. i didn't know there is only 11 minutes of actual live gameplay!	Surprised
HRED-A	yes, i agree. did you know espn won an emmy for the superimposed yellow first down line?	Neutral
HRED-SA	i agree with you on that one. i don't think i've ever heard of it. i don't think i've ever heard of it.	Neutral
EmoHRED-A-FL-CL	i know . i don't know much about it though .	Neutral
EmoHRED-SA-FL	yes, i agree. i wonder if there is a variation of soccer where 3 teams face each other at the same time?	Curious
Utterance 7	i know, it makes me mad, sometimes to think about that. a lot of wasted time standing around. lol	Angry
EmoHRED-SA-FL-CL	for sure, do you have a favorite album?	Curious
HRED	yeah for sure. do you know who alex smith is?	Curious
HRED-A	that was back in 1916 when georgia tech beat cumberland 222 0!	Neutral
HRED-SA	i agree with you there, do you ever watch the world cup?.	Curious
EmoHRED-A-FL-CL	for sure lol i agree with you there. did you know 86 of nevada is owned by the government?	Curious
EmoHRED-SA-FL	i know right, i don't know much about football, i don't know much about it	Neutral

Table 8: Generated examples are from a continuous conversation from the frequent test set. Each model predicts responses using the previous set of utterances and emotion labels.

Utterance 1	do you watch football?	Curious
EmoHRED-A-FL-CL	i sure do. u of iowa painted the visitor locker room pink?	Curious
HRED	i do. do you have a favorite team?	Curious
HRED-A	i do watch it from time to time. what about you?	Curious
HRED-SA	i sure do. do you?	Curious
EmoHRED-A-FL-CL	i do, i am a patriots fan, what about you?	Curious
EmoHRED-SA-FL	i do. do you?	Curious
Utterance 2	i sure do. do you know the u of iowa painted the visitor locker room pink?	Curious
EmoHRED-A-FL-CL	i wonder if that is supposed to play?	Curious
HRED	yes i have heard of it. do you know who invented the sport?	Curious
HRED-A	hmm i have no idea. do you know why?	Curious
HRED-SA	yes i wonder why that is?	Curious
EmoHRED-A-FL-CL	i have no idea but i wonder if it was supposed to do it?	Curious
EmoHRED-SA-FL	yeah that is pretty cool, i wonder why that is	Curious
Utterance 3	i wonder if that is supposed to have some effect on the opposing team	Curious
EmoHRED-A-FL-CL	i wonder too. did you know bowlers made more than football players in the 60s?	Curious
HRED	yeah i wonder too. did you know espn won an emmy?	Curious
HRED-A	good point. did you know women can play too?	Curious
HRED-SA	i think so too. do you know why the managers wear uniforms?	Curious
EmoHRED-A-FL-CL	i am not sure either. do you know much about bowlers?	Curious
EmoHRED-SA-FL	i do not know. do you know who invented the sport?	Curious
Utterance 4	i wonder too . did you know bowlers made more than football players in the 60s?	Curious
EmoHRED-A-FL-CL	yes i did hear about that. do you know who alex smith is?	Curious
HRED	i guess that was back in 1916 when georgia tech beat cumberland 222 0	Neutral
HRED-A	i did not know that it was the highest grossing basketball movie of all time	Surprised
HRED-SA	i did not know that. do you know who invented the game?	Curious
EmoHRED-A-FL-CL	i am not sure but i guess it makes more sense then	Surprised
EmoHRED-SA-FL	wow that is a lot. did you know espn won an emmy?	Curious
Utterance 5	yes i guess football was not as popular back in the day as it is now.	Curious
EmoHRED-A-FL-CL	true. do you know who benjarvus green ellis is?	Curious
HRED	yeah me too. did you know espn won an emmy?	Curious
HRED-A	yeah true. did you know espn won an emmy once?	Curious
HRED-SA	yeah i agree. do you know who invented the game?	Curious
EmoHRED-A-FL-CL	yes i guess it makes sense since it makes more money than football players	Neutral
EmoHRED-SA-FL	true. do you know who invented the sport?	Curious

Table 9: Generated examples are from a continuous conversation from the frequent test set. Each model predicts responses using the previous set of utterances and emotion labels.