Conditional Variational Autoencoders for Emotionally-aware Chatbot Based on Transformer

Zhechen Su, Yubo Xie
École Polytechnique Fédérale de Lausanne
Lansanne, Switzerland
{zhechen.su, yubo.xie}@epfl.ch

***Abstract*—Rising demand for artificial intelligence-powered chatbots with sentiment analysis is creating new growth opportunities for numerous areas. Thus, building empathetic natural language processing agents becomes an interdisciplinary field of natural language processing. Some researchers presented the seq2seq model to make responses more emotional, while others tried the generative model for more variation. However, it is arduous to control the emotion and sentiment of generated sentences. In this paper, we focus on the need for the variational empathetic chatbot. The model combines the plain Transformer chatbot model and Conditional Variational Autoencoders (CVAE). With the help of neural emotional classifiers and pre-trained weights from RoBERTa, our model achieves the best score in automatic and human evaluation. Experimentally, we show in the quantitative and qualitative analyses that the proposed models can successfully generate high-quality abstractive conversation responses following designated emotions.**

I. Introduction

In recent years, conversational agents or dialogue systems development are gaining more attention from both industry and academia [5]. Chatbot or multi-turn textual conversational model is a crucial research direction of dialogue systems, and it can conduct communication with a human by using natural language processing skills. One of the standard strategies to build agents is using simple rule-based approaches [1], [14]. With the development of machine learning techniques, more complex neural network models like sequence-to-sequence model [12], reinforcement model [15] were presented and got the outstanding performance.

Nowadays, textual conversational agents are used in many areas of our life. Apart from accomplishing tasks with specific domains like customer services and shopping assistance, the agents can also act as conversational partners. One challenging problem that arises in this domain is how to provide users with better engagement, which leads to higher satisfaction. In other words, the main development focus of building chatbots is to have a humanizing machine that has a better user-engagement when communicating with humans [6]. Some works were proposed to improve chatbot's user-engagement, such as building a context-aware chatbot [16] and injecting personality into the machine [19]. Other works also try to incorporate affective computing to build emotionally-aware chatbots [7], [13], [22].

Most of the recent emotionally-aware chatbots were built by using an encoder-decoder architecture with sequence-to-sequence model. Some studies also tried to model this task as a reinforcement learning task [10], in order to get more generic responses and let the chatbot able to achieve long-term conversation. However, few studies have yielded dialogues with emotions and diversity.

To make chatbots' responses more diverse and sentimental, we propose a new model named CVAE-Transformer. This model enjoys the advantage of Conditional Variational Autoencoders (Conditional VAE) [4], which can produce different but coherent responses under the same context. Also, replacing traditional encoders and decoder by pre-trained model - RoBERTa [11] enables the model to detect more subtle linguistic information, and deliver reliable performance. Then, we conducted a human evaluation to assess the quality of the generated emotional text. The results suggest that our method is capable of generating state-of-the-art emotional text at scale.

The main contributions of this work are as follows:

- The model explicitly inject emotion into responses. Furthermore, our trained emotion classifier shows that this injection mechanism performs well.
- Thanks to the attribution of Conditional VAE, our model enjoys significantly greater diversity than traditional seq2seq models.
- We apply several state-of-the-art generative models to train an emotional response generation system, and analysis confirms that our models deliver strong performance.

In the next section, we outline related work on building emotionally engaging chatbots, as well as neural generative models. Then, we will introduce the emotional dataset we use and explain why we use it. Next, we will describe the neural models we applied for the task. Finally, we will show automatic evaluation and human evaluation results, and some generated examples.

II. Related Works

Recently, many of emotionally engaging chatbots are built by encoder-decoder architecture with sequence-to-sequence learning. Through maximizing the likelihood of response, these seq2seq learning models are trained to incorporate rich data and generate an appropriate answer. Underlying seq2seq architecture is composed of two recurrent neural networks (RNNs), one as an encoder processing the input and one as a decoder generating the response. Long short term memory (LSTM) or gated recurrent unit (GRU) was the most dominant variant of RNNs, which used to learn the conversational

dataset in these models. Zhong et al. [21] extended the seq2seq model and imported the Attention mechanism. This mechanism enables the decoder to focus only on some significant parts in the input at every decoding step. Li's group [10] implemented a reinforcement learning approach to solve this task in order to get more universal responses and let the chatbot able to achieve long-term conversation.

In dialog generation, our work is in line with the recent progress of the application of Variational Autoencoder (VAE) [9]. The encoder of VAE represents textual input as a probability distribution, and then samples from the distribution to generate responses. Nevertheless, the original frameworks do not support end-to-end generation. In order to have condition options for results, Conditional VAE (CVAE) [4] has additional information in the training distribution process. Zhao's research in dialog generation [20] shows that dialog generated by VAE models enjoy significantly higher diversity than traditional seq2seq models, which is a preferable property toward building true-to-life dialog agents.

MojiTalk was introduced by Zhou et al. [23] in 2017, which is based on Conditional VAE and got impressive results. They trained the model by Twitter data labeled with emojis naturally. Being inspired by their work, we also use Conditional VAE when building a chatbot. To improve the model's linguistic comprehension capacity, a pre-trained transformer [11] replaces encoders and decoders of Conditional VAE.

III. Dataset

Like other artificial intelligent agents, building chatbot also needs a dataset to produce a meaningful conversation as a human-like agent. Therefore, some studies proposed datasets that contain textual conversation annotated by different emotion categories. To make our model performs appropriately, we choose EMPATHETICDIALOGUES [13] being our dataset. This novel dataset contains 25k conversations conducted by 810 different participants. Particularly, this dataset including three parts, namely context, emotional labels, and prompts. The speaker and listener generate prompts under a given context and its emotion. We employ emotional labels and context texts to facilitate training and evaluate the textual conversational system. We take the predicted response's emotion as a condition of the CVAE-Transformer model and textual contents as one of the inputs of the model. Comparing to other datasets like Twitter conversation [7], EMPATHETIC-DIALOGUES provides more reliability and precision because it is generated manually. This dataset offers more balanced coverage of emotions than would appear in public social media content.

IV. Our Approach

Our model, CVAE-Transformer, is adapted by Transformer and Conditional VAE. In this section, we will give a precise problem formulation of emotionally engaging chatbots at first. Then, our Transformer and Conditional VAE will be introduced from components to hierarchical structure independently. Finally, we will describe the method to combine two models as well as our emotion classifier, which provides emotional condition inputs.

A. Problem Formulation

The emotional response task can be formulated as follows: given a context $\boldsymbol{X}$ consisting of several previous utterances, the model should generate a response y by maximizing the probability distribution $p(\boldsymbol{y}|\boldsymbol{X})$ from a data set

$$\mathcal{D} = \left\{ \left(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{e}^{(i)} \right) \right\}_{i=1}^{N} \tag{1}$$

where N is the number of context-response pairs, $\boldsymbol{X}^{(i)} = (\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \ldots, \boldsymbol{x}_{m_i}^{(i)})$ is a sequence of utterances, and $\boldsymbol{y}^{(i)} = (y_1^{(i)}, y_2^{(i)}, \ldots, y_{T_i}^{(i)})$ is response with T_i words. $\boldsymbol{e}^{(i)}$ is the emotion condition predicted from $\boldsymbol{X}^{(i)}$.

B. Our Transformer

Transformer plays a crucial part in our model. We choose RoBERTA [11], an enhanced pre-trained model using Transformer [17]. Like most seq2seq models, the structure of the transformer is also composed of an encoder and a decoder. Besides, the model includes a stack of multi-head attention layers and point-wise, fully connected feed-forward networks for the encoder and the decoder. Using techniques like Attention, Transformer acquires a strong capability to understand inputs and apply the formation to many NLP areas.

1) Input Representation: Our input representation is different from the original Transformer since the input text in our task is not continuous. We use a similar approach proposed in BERT [3], where the input representation of a given word is constructed by concatenating the word, segment and position embeddings:

$$IR_{w_j^i} = WE_{w_j^i} \oplus SE_i \oplus PE_j \tag{2}$$

where $IR_{w_j^i}$ is the input representation of j-th word in i-th sentence, $WE_{w_j^i}$ is the word embedding of w_j^i, SE_i is the segment embedding of i-th sentence and PE_j is the position embedding of j-th word. For convenience, we denote the packages of a set of input representations for encoder and decoder as IR^E and IR^D respectively. Transformer takes advantage of the positional embedding to encode order within a conversation.

2) Encoder: Encoder is composed of N identical layers which is shown in Fig 1. Each Layer consists of two sub-layers, namely a multi-head self-attention mechanism and a fully connected feed-forward network. Each of these sub-layers adds a residual connection and normalisation. The output of the sub-layer can be expressed as:

$$SubLayerOutput = LayerNorm(x + (SubLayer(x))) \tag{3}$$

The two kinds of SubLayers are introduced below:

- Multi-head Attention: The multi-head attention mechanism computes attention weights, i.e., a softmax distribution, for each word within a sentence, including the word

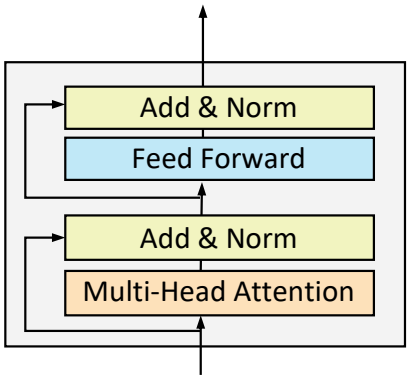

Fig. 1: The structure of Transformer Encoder

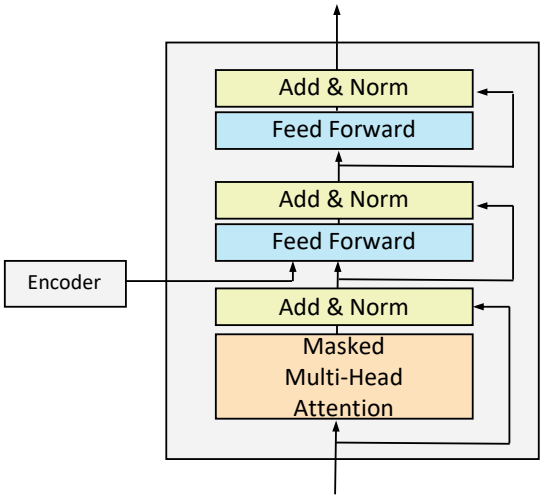

Fig. 2: The structure of Transformer Decoder

itself. Specifically:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where the input consists of queries Q and keys K of dimension d_k, and values V of dimension d_v. The queries, keys and values are linearly projected h times, to allow the model to jointly attend to information from different representation, concatenating the result. For multi-head:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \quad (5)$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, $W^O \in \mathbb{R}^{d_{model} \times hd_v}$.

- Position-wise feed-forward networks: On top of the multi-head attention, there is a feed-forward network that consists of two layers with a ReLU activation in between. The network projects the vector obtained by Multi-Head Attention to a larger space (the space is enlarged by 4 times in the paper). In that large space, it is easier to extract the required information (using the ReLU activation function) And finally project back to the original space of the token vector. The feed-forward networks FFN is denoted as:

$$FFN(x) = ReLU\left(xW_1 + b_1\right)W_2 + b_2 \quad (6)$$

where W is weights and b is bias.

Each encoder layer takes as input the output of the previous layer, allowing it to attend to all positions of the previous layer.

3) Decoder: The decoder has a similar architecture as the encoder, stacking N identical layers of multi-head attention with feed-forward networks. From Fig 2, we can see that the difference between them is decoder has two multi-head attention sub-layers: 1) a decoder self-attention and 2) encoder-decoder attention. The decoder self-attention attends on the previous predictions, masked one position by one position while encoder-decoder attention performs attention between the final encoder representation and the decoder representation.

C. Conditional VAE

Conditional Variational Autoencoder (CVAE) also makes use of encoder-decoder structures. Besides encoders and decoders, CVAE also has a recognition network and a prior network to represent sentences as distribution. Then the model takes a sample from generated distribution each time and decodes the sample to prediction. Mathematically, CVAE is trained by maximizing a variational lower bound on the conditional likelihood of X given c, according to:

$$p(X|c) = \int p(X|z, c)p(z|c)dz \quad (7)$$

where X represents the encoded response, c is the condition: $c = Emb(v_e) \oplus Enc(v_{inp})$ and z is a latent variable capturing the distribution of responses.

We use networks to capture distribution. $p_P(z|c)$ represents prior network distribution, which is trained to approach prior $p(z|c)$. $q_R(z|X, c)$ represents recognition network distribution, which is trained to approach true posterior $p(z|X, c)$. In addition, decoder is used to approximate $p(X|z, c)$.By assuming that the latent variable z has a multivariate Gaussian distribution with a diagonal covariance matrix, the lower bound to $\log p(x|c)$ can be written by:

$$\begin{aligned} -\mathcal{L}\left(\theta_D, \theta_P, \theta_R; X, c\right) &= \text{KL}\left(q_R(z|X, c)\|p_P(z|c)\right) \\ &\quad - \mathbb{E}_{q_R}(z|X, c)\left(\log p_D(X|z, c)\right) \end{aligned} \quad (8)$$

where θ_D, θ_P, θ_R are parameters of Decoder, Prior network and Recognition network respectively. And $D_{KL}(||)$ denotes KL-divergence.

Through minimizing KL-divergence between prior network and recognition network, the model has the ability to make similar prediction from samples of two distribution. Specially, during testing, the recognition network is absent, and we sample from prior network.

D. CVAE-Transformer

As shown in Fig 3, our CVAE-Transformer inherits the encoder and decoder of the Transformer as well as the same attention mechanism. Upon the Transformer, CVAE-Transformer also utilizes latent variables, z in our case, for learning the distribution of coherent responses conditioned on given emotions. Then, based on the attention memory, condition, and latent variables, a response is finally generated from the decoder. This kind of structure makes our model achieved better than previous works of CVAE on textual data.

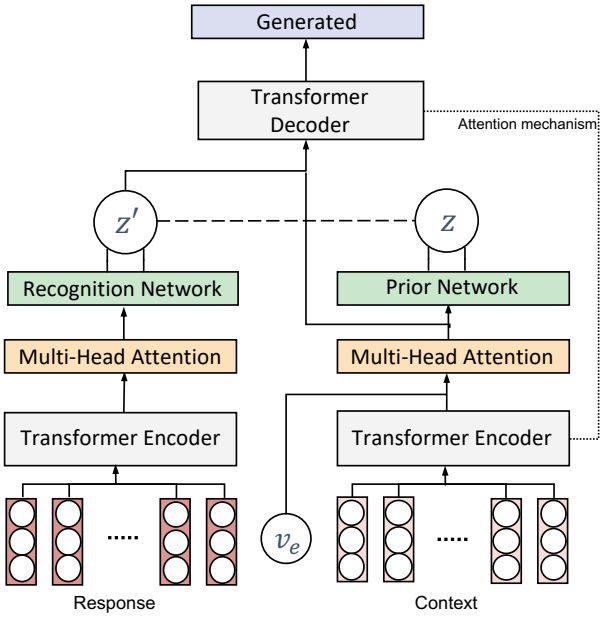

Fig. 3: The structure of CVAE-Transformer

The model takes context (the concatenation of previous utterances), response, and v_e as inputs. Response and context are embedded as Tranformer's input and sent to their encoder. v_e, the emotional condition, is the predicted result of pre-trained classifiers given context information. We train the emotional classifier by situation descriptions and their emotion labels in EMPATHETICDIALOGUES. The model can reach an accuracy of 58.6% at last. By providing a true response to the emotion classifier, we can get a 32-dimension vector v_e to denotes emotional probability distribution, or condition in our case. Then, after one-hot representation and embedding, the vector is concatenated with output of context Transformer encoder and sent to Multi-head Attention layer. The motivation of Multi-head Attention layer is to filter out unnecessary information and decrease vector dimensions, so that the training process is accelerated. Using the outputs of Attention layers, Prior Network and Recognition Network is capable of approaching probability distribution through an MLP to get mean and log variance of z's (z''s) distribution. After that, we run a reparameterization trick [8] to sample latent variables. The variables concatenating with the input of Prior Network are passed to Transformer decoder. Finally, we employ the beam search algorithm [18] for the optimal response with dictionary probability. Especially, while during testing, the target response is absent, and z by the prior network is passed to the decoder.

Model	Perplexity	Emotion Accuracy (Top1)
Plain Transformer	8.0	12.3%
MEED	6.2	11.1%
CVAE-Transformer	**6.1**	**58.8%**

TABLE I: Generation perplexity and emoji accuracuy (Top 1) of the three models.

$$\mathcal{L}' = \mathcal{L} + \mathcal{L}_{KL} + \mathcal{L}_{bow} \tag{9}$$

The equation is the loss function we use to train CVAE-Transformer model. $\mathcal{L}$ is the reconstruction loss, known as the mean-squared error or cross-entropy between the actual response and prediction. $\mathcal{L}_{KL}$ is the KL-divergence loss between z and z'. When processing text data, the VAE models with recurrent neural networks may first learn to ignore the latent variable, and explain the data with the more easily optimized decoder [2]. Thus, the latent variables lose their functionality. In order to alleviate this problem and keep a balance between KL loss and reconstruction loss. We use techniques of KL annealing, early stopping [2] and bag-of-word loss $\mathcal{L}_{bow}$ [20].

V. Experiments and Results

In this section, we first analyze the overall results of our models, including perplexity, loss, and emotion accuracy. Then we take a closer look at the generation quality as well as our models' capability of expressing emotion.

A. Settings

We split the conversations into approximately 80% train, 10% validation, and 10% test partitions. In other words, the model learns from 19533 conversations, and we have 2523 conversations as test set. For each conversation, we concatenate all utterances except the last response as context, whereas the last response is the ground truth.

For hyper-parameters, we assign each Transformer encoder and decoder 6 layers. We embed the emotion vector to 768 dimensions, which is identical to the model dimension of RoBERTa. The size of the latent variables is 128. MLPs in recognition/prior network are 3 layered with $tanh$ activation function. The model is trained for 40000 steps of 32 batch size.

B. Baselines

We compare our models with the following baselines:

Plain Transformer: The original Transformer [17] is compared. The model contains Transformer encoder and decoder only. The same input representation as our model is fed to the encoder.

MEED: The model is presented by Yubo Xie base on Transformer. MEED is a multi-task model, which training the

1. Context	
- i hate when i fart in public it is so embarrassing - You're unable to hold them in? - i can hold them but some of them are just accidental and it is the worst, depends on what i eat haha	
Model	**Response**
Plain Transformer	Oh, I'm sorry to hear that.
MEED	How many times have you had this happen?
CVAE	I know how that feels.
GoundTruth	I'm sorry, that must be pretty rough on you!
2. Context	
- When my daughter was born doctors said she wouldn't ever go to school and basically be mentally handicapped. - And how is she doing now - She is a normal A-B student and just started 8th grade	
Model	**Response**
Plain Transformer	I am sure she will be fine
MEED	That is a tough situation to be in.
CVAE	That is great to hear. You must be very proud of her.
GoundTruth	That is awesome. Sometimes a second opinion is all you need
3. Context	
- My children are getting old. They are starting to spread the wings. So, I would have to let go and let them fly. - That can't be easy. How old are your children? Sometimes I wish we could keep our kids small forever! - I have 23, 20, 17 and 14. They are doing stuff that I question but I don't want them to not tell me stuff.	
Model	**Response**
Plain Transformer	I'm sorry you have to deal with that. I'm sorry you have to deal with that.
MEED	I'm sorry you have to deal with that.
CVAE	Aw, that's so sad. I'm sorry you're having to go through that.
GoundTruth	Yeah you want them to learn, but it's definitely hard to see them doing things that you might not agree with.

TABLE II: Cases: Generated Samples from CVAE-Transformer

Transformer encoder not only as encoder itself, but also as an emotion predictor. Using a softmax layer, the model generates a response emotion, and passes the vector to decoder.

C. General Results

After training 20 epochs, the KL loss is converged to 0.0125, Bag of Words loss is 5.24, and reconstruction loss is 2.07. The losses prove that the models managed a balance between the two items of loss, and it confirms that they have successfully learned a meaningful latent variable.

Table I gives the perplexity scores of three models and TOP 1 emotion accuracy on the test set. As shown in the table, CVAE-Transformer reaches the best perplexity and significantly increases the emoji accuracy over baseline models. These results confirm that proposed methods are effective toward the generation of emotional responses.

To evaluate the quality of generated prediction, we choose perplexity, a popular metric for language model. Perplexity indicates how much difference the generated response comparing to ground truth. The lower perplexity score indicates that the model has a higher capability of predicting the target sentence. Although our model surpasses others, the score is not strong evidence for the model's ability. On the one hand, the scores' differences between models are tiny, and any factors like initialization and training duration could affect the results. On the other hand, the task is for dialogue, and it is known that a conversation has various proper responses. Thus, the score is not convincing due to lacking golden standards.

The other goal of our model is emotional injection. According to results, the emotion accuracy of MEED is close to Plain Transformer. In other words, MEED hardly uses emotional information, although it holds emotional input. Surprisingly, due to the properties of CVAE, the accuracy of our model reaches 58.8%, which shows the emotion is embedded successfully. Additionally, we can control the emotion of responses explicitly, but we still use the emotion vector from classifier directly instead of assigning by ourselves. The reason is the models must have a similar emotion vector for models, or we cannot compare their results because of different precondition. As the meaning of different emotions may overlap, it is good to try multi-emotion condition instead of one hot in the future.

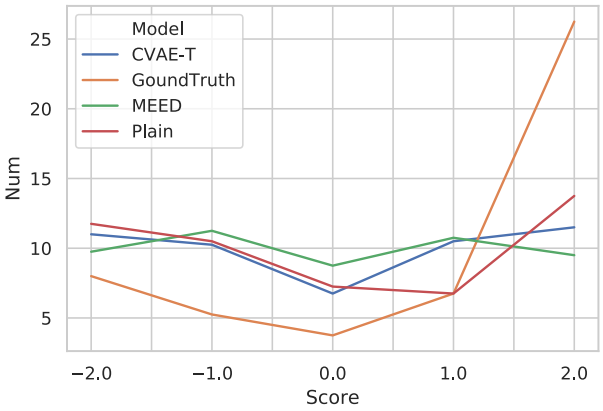

Fig. 4: Human Evaluation Score Distribution

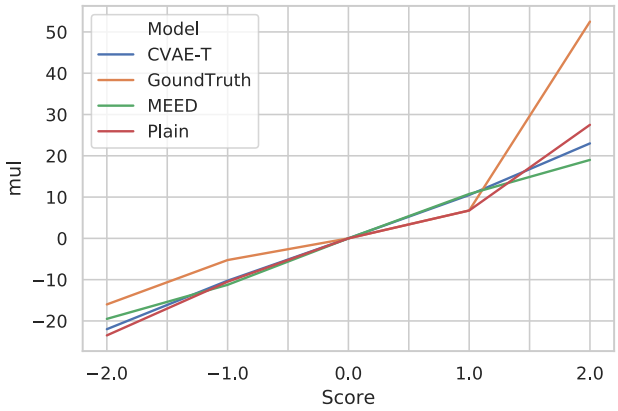

Fig. 5: Weighted Score Distribution

And the machine may learn that similarity and give multiple possible labels as the answer.

VI. Case Study

We take some generated responses from Plain Transformer, MEED, and CVAE-Transformer, and list these samples in Table II. From the cases, we find that more coherent and reasonable responses are generated from CVAE-Transformer, which confirms the high score. However, compared with GroundTruth, which is written by a human, our model still has some drawbacks about informativeness and coherence.

From these three examples, Plain Transformer always generates responses starting with "I am", one of the most frequent expressions in the corpus. This fact indicates that the model tends to choose the safest answer and form a predictable pattern without emotional instruction. Moreover, due to the decoding method, Plain Transformer is unstable and has a problem of duplication as Example 3. For MEED, emotional injection and optimized structure enable the model to generate more diverse sentences than baselines. However, as shown in Example 1, MEED's response is embarrassing and hard to be accepted. The fact means emotional information does not engage, and its coherence should be improved.

Comparing to two other models, CVAE-Transformer makes the most human-like and natural predictions, which is in line with previous quantitative analysis. In Example 3, we can see that our model can use an informal expression like "Aw" and produce a longer answer instead of duplication. These features make the model performs best in these models. Nevertheless, there still are some parts to be improved. As a listener, the model might be good at generating responses to the last sentences, but they do not provide or ask for new information to help the conversation continue.

VII. Human Evaluation

We find four volunteers who are willing to taking the survey, which aims at evaluating CVAE-Transformer's performers. The survey includes 50 random cases choosing from test set. In each case, there are context and responses generated from 3 models: Plain Transformer, MEED, and CVAE-Transformer. Plus, Ground Truth is added in order to indicate the upper bound. All responses are anonymous and random ordered to promise fairness and correctness of the survey. For each response, the volunteer is asked to assign a review mapping to integral score from -2 to 2. In this way, the average score for 50 cases ranges from -100 to 100, corresponding to the model's predictions. For each response, we take the average score of volunteers as the final score.

	Plain	MEED	CVAE-T	GroundTruth
Std	5.0	**4.2**	4.6	4.5
Average	0.25	-1	**1.25**	38

TABLE III: Human Evaluation Statistic Results

Table III describes statistic results of the survey. Ground Truth from human gets the highest score of 38. This score can be interpreted as the best performance chatbots could reach. Our model, CVAE-Transformer, outperforms the baseline models, and it achieves 1.25 of average and 4.60 of standard deviation. The results indicate that our model can generate more coherence and reasonable responses than baselines. Also, it is more stable and more possible to generate acceptable predictions than others.

We also plot the score distribution in Fig 4 to capture the features of models. In the plot, the x-axis is the categorical score of predictions, and the y-axis is the number of times the model got that score. Overall, models tend to generate extreme responses making lines' heads higher than middle parts. Also, at x=2, we can find Plain Transformer performs best among baseline models. However, considering other scores, Plain Transformer is unstable: it makes many best predictions and worst predictions. CVAE-Transformer and MEED are of a similar wavy line. Nevertheless, CVAE-Transformer would more like to take a chance and get more extreme scores.

Moreover, we multiply the x-axis score and the number for

each model. The higher results mul are, the better responses the model can make. Through the mul, we can track how many each category (x-axis) contributes to average, as well as identify models' advantages. As shown in Fig 5, GroundTruth achieves best at -2, -1, 2, and performs much better than others. All other models are nearly the same when score below 0. And in positive part, Plain is unstable and sometimes makes a good guess. MEED and CVAE-Transformer are linear, and CVAE-T has better results at 2.

Comparing to the average score of GroundTruth, the difference between three is subtle. In other words, building chatbots still has a long way to go.

VIII. Conclusion and future work

In this report, we investigate the possibility of using manual annotated emotional data for building emotionally engaged chatbots. We applied Transformer, Conditional Variational Autoencoders, as well as other techniques to build a prediction generation system that is capable of giving diverse and reasonable responses by the supervision of emotion. With the help of an emotion classifier trained by the same data set, we can provide expected emotion as a condition to the model and generated responses to evaluate the emotion accuracy. Also, we performed automatic and human evaluations to understand the quality of the generated predictions. Experimentally, the results show that our model can generate high-quality emotional responses. Our work provides a new method to build intelligent dialogue agents.

For future work, we are looking forward to import reinforcement leaning to this kind of probability-based multi-turn dialog model. We can model speakers as agents with an attribution like persona and emotion. The states of agents will transfer along with the conversation. In that way, the model is capable of simulating real dialogue processing, and the agents' emotions can change dynamically.

References

[1] H. Al-Zubaide and A. A. Y. Issa. Ontbot: Ontology based chatbot. *International Symposium on Innovations in Information and Communications Technology*, pages 7–12, 2011.

[2] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating Sentences from a Continuous Space. *arXiv e-prints*, page arXiv:1511.06349, Nov 2015.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805, Oct 2018.

[4] C. Doersch. Tutorial on variational autoencoders, 2016.

[5] H. Fang, H. Cheng, M. Sap, E. Clark, A. Holtzman, Y. Choi, N. A. Smith, and M. Ostendorf. Sounding board: A user-centric and content-driven social chatbot. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[6] E. Go and S. Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, 8 2019.

[7] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. *arXiv e-prints*, page arXiv:1803.02952, Mar 2018.

[8] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2575–2583. Curran Associates, Inc., 2015.

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2013.

[10] J. Li, X. Sun, X. Wei, C. Li, and J. Tao. Reinforcement learning based emotional editing constraint conversation generation, 2019.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu. AliMe chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 498–503, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[13] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: a new benchmark and dataset, 2018.

[14] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[15] I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke, S. Rajeshwar, A. de Brebisson, J. M. R. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, J. Pineau, and Y. Bengio. A deep reinforcement learning chatbot, 2017.

[16] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *CoRR*, abs/1506.06714, 2015.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *arXiv e-prints*, page arXiv:1706.03762, Jun 2017.

[18] S. Wiseman and A. M. Rush. Sequence-to-Sequence Learning as Beam-Search Optimization. *arXiv e-prints*, page arXiv:1606.02960, Jun 2016.

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

[20] T. Zhao, R. Zhao, and M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders, 2017.

[21] P. Zhong, D. Wang, and C. Miao. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss, 2018.

[22] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory, 2017.

[23] X. Zhou and W. Y. Wang. Mojitalk: Generating emotional responses at scale, 2017.