# A Multi-Turn Emotionally Engaging Dialog Model

**Yubo Xie**
**Ekaterina Svikhnushina**
**Pearl Pu**
yubo.xie@epfl.ch
ekaterina.svikhnushina@epfl.ch
pearl.pu@epfl.ch
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

## ABSTRACT

Open-domain dialog systems (also known as chatbots) have increasingly drawn attention in natural language processing. Some of the recent work aims at incorporating affect information into sequence-to-sequence neural dialog modeling, making the response emotionally richer, while others use hand-crafted rules to determine the desired emotion response. However, they do not explicitly learn the subtle emotional interactions captured in human dialogs. In this paper, we propose a multi-turn dialog system aimed at learning and generating emotional responses that so far only humans know how to do. Compared with two baseline models, offline experiments show that our method performs the best in perplexity scores. Further human evaluations confirm that our chatbot can keep track of the conversation context and generate emotionally more appropriate responses while performing equally well on grammar.

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Natural language interfaces**.

## KEYWORDS

chatbots, affective computing, deep learning, natural language processing

**ACM Reference Format:**
Yubo Xie, Ekaterina Svikhnushina, and Pearl Pu. 2020. A Multi-Turn Emotionally Engaging Dialog Model. In *IUI '20 Wokrshops, March 17, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn



## 1 INTRODUCTION

Many application areas show significant benefits of integrating affect information in natural language dialogs. In earlier work on human computer interaction, Klein et al. [16] found user's frustration caused by a computer system can be alleviated by computer-initiated emotional support, by providing feedback on emotional content along with sympathy and empathy. Recently, Hu et al. [14] developed a customer support neural chatbot, capable of generating dialogs similar to the humans in terms of empathic and passionate tones, potentially serving as proxy customer support agents on social media platforms. In a qualitative study [47], participants expressed an interest in chatbots capable of serving as an attentive listener and providing motivational support, thus fulfilling users' emotional needs. Several participants even noted a chatbot is ideal for sensitive content that is too embarrassing to ask another human. Finally Bickmore and Picard [3] showed a relational agent with deliberate social-emotional skills was respected more, liked more, and trusted more, even after four weeks of interaction, compared to an equivalent task-oriented agent.

Recent development in neural language modeling has generated significant excitement in the open-domain dialog generation community. The success of sequence-to-sequence (seq2seq) learning [5, 37] in the field of neural machine translation has inspired researchers to apply the recurrent neural network (RNN) encoder-decoder structure to response generation [42]. Following the standard seq2seq structure, various improvements have been made on the neural conversation model. For example, Shang et al. [34] applied attention mechanism [2] to the same structure on Twitter-style microblogging data. Li et al. [17] found the original version tend to favor short and dull responses. They fixed this problem by increasing the diversity of the response. Li et al. [18] modeled the personalities of the speakers, and Xing et al. [44] developed a topic aware dialog system. We call work in this area globally neural dialog generation. For a comprehensive survey, please refer to [4].

More recently, researchers started incorporating affect information into neural dialog models. While a central theme

seems to be making the responses emotionally richer, existing approaches mainly follow two directions. In one, an emotion label is explicitly required as input so that the machine can generate sentences of that particular emotion label or type [49]. In another group of work, the main idea is to develop handcrafted rules to direct the machines to generated responses of the desired emotions [1, 48]. Both approaches require an emotion label as input (either given or handcrafted), which might be unpractical in real dialog scenarios.

Furthermore, to the best of our knowledge, the psychology and social science literature does not provide clear rules for emotional interaction. It seems such social and emotional intelligence is captured in our conversations. This is why we decided to take the automatic and data-driven approach. In this paper, we describe an end-to-end Multi-turn Emotionally Engaging Dialog model (MEED), capable of recognizing emotions and generating emotionally appropriate and human-like responses with the ultimate goal of reproducing social behaviors that are habitual in human-human conversations. We chose the multi-turn setting because a model suitable for single-turn dialogs cannot effectively track earlier context in multi-turn dialogs, both semantically and emotionally. Since being able to track several turns is really important, we made this design decision from the beginning, in contrast to most related work where models are only trained and tested on single-turn dialogs. While using a hierarchical mechanism to track the conversation history in multi-turn dialogs is not new (e.g., HRAN by Xing et al. [45]), to combine it with an additional emotion RNN to process the emotional information in each history utterance has never been attempted before.

Our contributions are threefold. (1) We describe in detail a novel emotion-tracking dialog generation model that learns the emotional interactions directly from the data. This approach is free of human-defined heuristic rules, and hence, is more robust and fundamental than those described in existing work. (2) We compare our model, MEED, with the generic seq2seq model and the hierarchical model of multi-turn dialogs (HRAN). Offline experiments show that our model outperforms both seq2seq and HRAN by a significant amount. Further experiments with human evaluation show our model produces emotionally more appropriate responses than both baselines, while also improving the language fluency. (3) We illustrate a human-evaluation procedure for judging machine produced emotional dialogs. We consider factors such as the balance of positive and negative emotions in test dialogs, a well-chosen range of topics, and dialogs that our human evaluators can relate. It is the first time such an approach is designed with consideration for human judges. Our main goal is to increase the objectivity of the results and reduce judges' mistakes due to out-of-context dialogs they have to evaluate.

## 2 RELATED WORK

### Neural Dialog Generation

Vinyals and Le [42] were one of the first to model dialog generation using neural networks. Their seq2seq framework was trained on an IT Helpdesk Troubleshooting dataset and the OpenSubtitles dataset [21]. Shang et al. [34] further trained the seq2seq model with attention mechanism on a self-crawled Weibo (a popular Twitter-like social media website in China) dataset. Meanwhile, Xu et al. [46] built a customer service chatbot by training the seq2seq model on a dataset collected with conversations between customers and customer service accounts from 62 brands on Twitter.

The standard seq2seq framework is applied to single-turn response generation. In multi-turn settings, where a context with multiple history utterances is given, the same structure often ignores the hierarchical characteristic of the context. Some recent work addresses this problem by adopting a hierarchical recurrent encoder-decoder (HRED) structure [32, 33, 35]. To give attention to different parts of the context while generating responses, Xing et al. [45] proposed the hierarchical recurrent attention network (HRAN), using a hierarchical attention mechanism. However, these multi-turn dialog models do not take into account the turn-taking emotional changes of the dialog.

### Neural Dialog Models with Affect Information

Recent work on incorporating affect information into natural language processing tasks has inspired our current work. They can be mainly described as affect language models and emotional dialog systems.

Ghosh et al. [11] made the first attempt to augment the original LSTM language model with affect treatment in what they called Affect-LM. At training time, Affect-LM can be considered as an energy based model where the added energy term captures the degree of correlation between the next word and the affect information of the preceeding text. At text generation time, affect information is also used to increase the appropriate selection of the next word. A key component in Affect-LM is the use of a well established text analysis program, LIWC (Linguistic Inquiry and Word Count) [28]. For every sentence, for example, "I unfortunately did not pass my exam", the model generates five emotion features denoting (*sad*: 1, *angry*: 1, *anxiety*: 1, *negative emotion*: 1, *positive emotion*: 0). This makes Affect-LM both capable of distinguishing affect information conveyed by each word in the language modeling part and aware of the preceeding text's emotion in each generation step. In a similar vein, Asghar et al. [1] appended the original word embeddings with a VAD affect model [43]. VAD is a vector model, as opposed to a categorical model (LIWC), representing a given emotion in each of the valence, arousal, and dominance axes. In

contrast to Affect-LM, Asghar's neural affect dialog model aims at generating explicit responses given a particular utterance. To do so, the authors designed three affect-related loss functions, namely minimizing affect dissonance, maximizing a affective dissonance, and maximizing affective content. The paper also proposed the affectively diverse beam search during decoding, so that the generated candidate responses are as affectively diverse as possible. However, literature in affective science does not necessarily validate such rules. In fact, the best strategy to speak to an angry customer is the de-escalation strategy (using neutral words to validate anger) rather than employing equally emotional words (minimizing affect dissonance) or words that convey happiness (maximizing affect dissonance).

The Emotional Chatting Machine (ECM) [49] takes a post and generates a response in a predefined emotion category. The main idea is to use an internal memory module to capture the emotion dynamics during decoding, and an external memory module to model emotional expressions explicitly by assigning different probability values to emotional words as opposed to regular words. Zhou and Wang [50] extended the standard seq2seq model to a conditional variational autoencoder combined with policy gradient techniques. The model takes a post and an emoji as input, and generates the response with target emotion specified by the emoji. Hu et al. [14] built a tone-aware chatbot for customer care on social media, by deploying extra meta information of the conversations in the seq2seq model. Specifically, a tone indicator is added to each step of the decoder during the training phase.

In parallel to these developments, Zhong et al. [48] proposed an affect-rich dialog model using biased attention mechanism on emotional words in the input message, by taking advantage of the VAD embeddings. The model is trained with a weighted cross-entropy loss function, which encourages the generation of emotional words.

### Summary

As much as these work in the above section inspired our work, our approach in generating affect dialogs is significantly different. Most of related work focused on integrating affect information into the transduction vector space using either VAD or LIWC, we aim at modeling and generating the affect exchanges in human dialogs using a dedicated embedding layer. The approach is also completely data-driven, thus absent of hand-crafted rules. To avoid learning obscene and callous exchanges often found in social media data like tweets and Reddit threads [29], we opted to train our model on movie subtitles, whose dialogs were carefully created by professional writers. We believe the quality of this dataset can be better than those curated by crowdsource platforms. For modeling the affect information, we chose to use LIWC because it is a well-established emotion lexical resource, covering the whole English dictionary whereas VAD only contains 13K lemmatized terms.

## 3 MODEL

We describe our model one element at a time, from the basic structure, to the hierarchical component, and finally the emotion embedding layer.

We first consider the problem of generating response $\boldsymbol{y}$ given a context $\boldsymbol{X}$ consisting of multiple previous utterances by estimating the probability distribution $p(\boldsymbol{y} \mid \boldsymbol{X})$ from a data set $\mathcal{D} = \{(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N}$ containing $N$ context-response pairs. Here

$$\boldsymbol{X}^{(i)} = \left(\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \ldots, \boldsymbol{x}_{m_i}^{(i)}\right) \tag{1}$$

is a sequence of $m_i$ utterances, and

$$\boldsymbol{x}_j^{(i)} = \left(x_{j,1}^{(i)}, x_{j,2}^{(i)}, \ldots, x_{j,n_{ij}}^{(i)}\right) \tag{2}$$

is a sequence of $n_{ij}$ words. Similarly,

$$\boldsymbol{y}^{(i)} = \left(y_1^{(i)}, y_2^{(i)}, \ldots, y_{T_i}^{(i)}\right) \tag{3}$$

is the response with $T_i$ words.

Usually the probability distribution $p(\boldsymbol{y} \mid \boldsymbol{X})$ can be modeled by an RNN language model conditioned on $\boldsymbol{X}$. When generating the word $y_t$ at time step $t$, the context $\boldsymbol{X}$ is encoded into a fixed-sized dialog context vector $\boldsymbol{c}_t$ by following the hierarchical attention structure in HRAN [45]. Additionally, we extract the emotion information from the utterances in $\boldsymbol{X}$ by leveraging an external text analysis program, and use an RNN to encode it into an emotion context vector $\boldsymbol{e}$, which is combined with $\boldsymbol{c}_t$ to produce the distribution. The overall architecture of the model is depicted in Figure 1. We are going to elaborate on how to obtain $\boldsymbol{c}_t$ and $\boldsymbol{e}$, and how they are combined in the decoding part.

### Hierarchical Attention

The hierarchical attention structure involves two encoders to produce the dialog context vector $\boldsymbol{c}_t$, namely the word-level encoder and the utterance-level encoder. The word-level encoder is essentially a bidirectional RNN with gated recurrent units (GRU) [5]. For utterance $\boldsymbol{x}_j$ in $\boldsymbol{X}$ ($j = 1, 2, \ldots, m$), the bidirectional encoder produces two hidden states at each word position $k$, the forward hidden state $\boldsymbol{h}_{jk}^{\mathrm{f}}$ and the backward hidden state $\boldsymbol{h}_{jk}^{\mathrm{b}}$. The final hidden state $\boldsymbol{h}_{jk}$ is then obtained by concatenating the two,

$$\boldsymbol{h}_{jk} = \mathrm{concat}\left(\boldsymbol{h}_{jk}^{\mathrm{f}}, \boldsymbol{h}_{jk}^{\mathrm{b}}\right). \tag{4}$$

The utterance-level encoder is a unidirectional RNN with GRU that goes from the last utterance in the context to the first, with its input at each step as the summary of the corresponding utterance, which is obtained by applying a Bahdanau-style attention mechanism [2] on the word-level

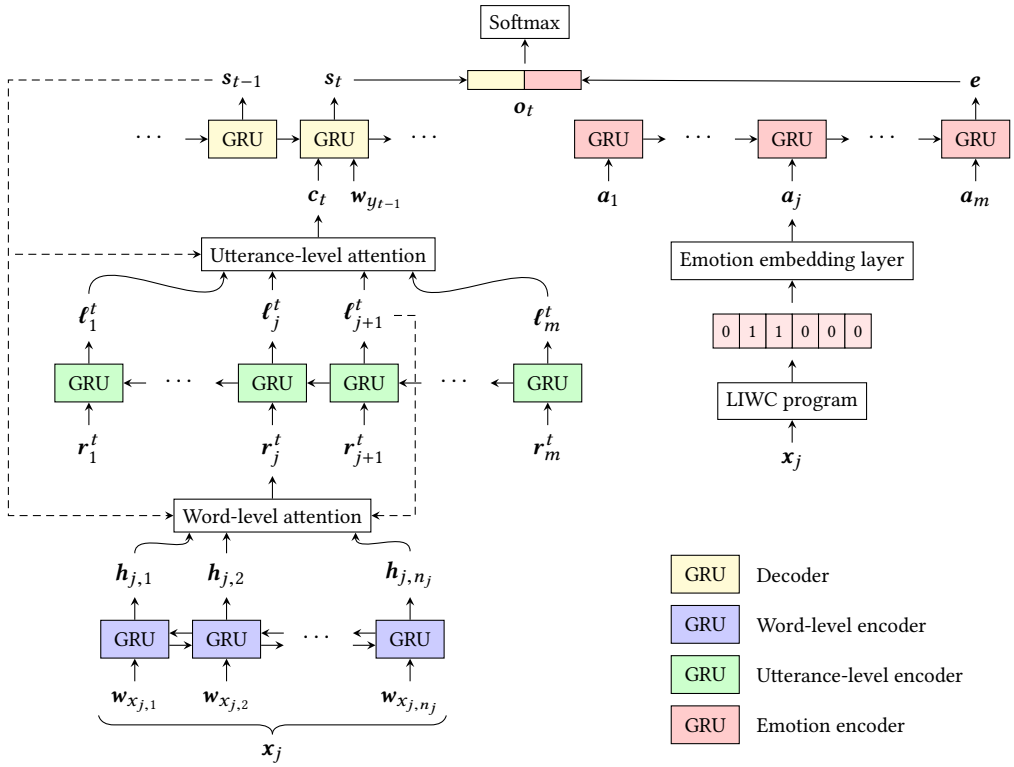


**Figure 1: The overall architecture of our model.**

encoder output. More specifically, at decoding step $t$, the summary of utterance $\boldsymbol{x}_j$ is a linear combination of $\boldsymbol{h}_{jk}$, for $k = 1, 2, \ldots, n_j$,

$$\boldsymbol{r}_j^t = \sum_{k=1}^{n_j} \alpha_{jk}^t \boldsymbol{h}_{jk}. \tag{5}$$

Here $\alpha_{jk}^t$ is the word-level attention score placed on $\boldsymbol{h}_{jk}$, and can be calculated as

$$a_{jk}^t = \boldsymbol{v}_a^T \tanh(\boldsymbol{U}_a \boldsymbol{s}_{t-1} + \boldsymbol{V}_a \boldsymbol{\ell}_{j+1}^t + \boldsymbol{W}_a \boldsymbol{h}_{jk}), \tag{6}$$

$$\alpha_{jk}^t = \frac{\exp(a_{jk}^t)}{\sum_{k'=1}^{n_j} \exp(a_{jk'}^t)}, \tag{7}$$

where $\boldsymbol{s}_{t-1}$ is the previous hidden state of the decoder, $\boldsymbol{\ell}_{j+1}^t$ is the previous hidden state of the utterance-level encoder, and $\boldsymbol{v}_a$, $\boldsymbol{U}_a$, $\boldsymbol{V}_a$ and $\boldsymbol{W}_a$ are word-level attention parameters. The final dialog context vector $\boldsymbol{c}_t$ is then obtained as another linear combination of the outputs of the utterance-level encoder $\boldsymbol{\ell}_j^t$, for $j = 1, 2, \ldots, m$,

$$\boldsymbol{c}_t = \sum_{j=1}^{m} \beta_j^t \boldsymbol{\ell}_j^t. \tag{8}$$

Here $\beta_j^t$ is the utterance-level attention score placed on $\boldsymbol{\ell}_j^t$, and can be calculated as

$$b_j^t = \boldsymbol{v}_b^T \tanh(\boldsymbol{U}_b \boldsymbol{s}_{t-1} + \boldsymbol{W}_b \boldsymbol{\ell}_j^t), \tag{9}$$

$$\beta_j^t = \frac{\exp(b_j^t)}{\sum_{j'=1}^{m} \exp(b_{j'}^t)}, \tag{10}$$

where $\boldsymbol{s}_{t-1}$ is the previous hidden state of the decoder, and $\boldsymbol{v}_b$, $\boldsymbol{U}_b$ and $\boldsymbol{W}_b$ are utterance-level attention parameters.

### Emotion Encoder

The main objective of the emotion embedding layer is to recognize the affect information in the given utterances so that the model can respond with emotionally appropriate replies. To achieve this, we need an encoder to distinguish the affect information in the context, in addition to its semantic meaning. Equally we need a decoder capable of selecting the best and most human-like answers.

We are able to achieve this goal, i.e., capturing the emotion information carried in the context $\boldsymbol{X}$, in the encoder, thanks to LIWC. We make use of the five emotion-related categories, namely *positive emotion*, *negative emotion*, *anxious*, *angry*, and *sad*. This set can be expanded to include more categories if we desire a richer distinction. See the discussion section for more details on how to do this. Using the newest version of

the program LIWC2015,[1] we are able to map each utterance $\boldsymbol{x}_j$ in the context to a six-dimensional indicator vector $\mathbf{1}(\boldsymbol{x}_j)$, with the first five entries corresponding to the five emotion categories, and the last one corresponding to *neutral*. If any word in $\boldsymbol{x}_j$ belongs to one of the five categories, then the corresponding entry in $\mathbf{1}(\boldsymbol{x}_j)$ is set to 1; otherwise, $\boldsymbol{x}_j$ is treated as neutral, with the last entry of $\mathbf{1}(\boldsymbol{x}_j)$ set to 1. For example, assuming $\boldsymbol{x}_j$ = "he is worried about me", then

$$\mathbf{1}(\boldsymbol{x}_j) = [0, 1, 1, 0, 0, 0], \tag{11}$$

since the word "worried" is assigned to both *negative emotion* and *anxious*. We apply a dense layer with sigmoid activation function on top of $\mathbf{1}(\boldsymbol{x}_j)$ to embed the emotion indicator vector into a continuous space,

$$\boldsymbol{a}_j = \sigma(\boldsymbol{W}_e \mathbf{1}(\boldsymbol{x}_j) + \boldsymbol{b}_e), \tag{12}$$

where $\boldsymbol{W}_e$ and $\boldsymbol{b}_e$ are trainable parameters. The emotion flow of the context $\boldsymbol{X}$ is then modeled by an unidirectional RNN with GRU going from the first utterance in the context to the last, with its input being $\boldsymbol{a}_j$ at each step. The final emotion context vector $\boldsymbol{e}$ is obtained as the last hidden state of this emotion encoding RNN.

### Decoding

The probability distribution $p(\boldsymbol{y} \mid \boldsymbol{X})$ can be written as

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{X}) &= p(y_1, y_2, \ldots, y_T \mid \boldsymbol{X}) \\ &= p(y_1 \mid \boldsymbol{c}_1, \boldsymbol{e}) \prod_{t=2}^{T} p(y_t \mid y_1, \ldots, y_{t-1}, \boldsymbol{c}_t, \boldsymbol{e}). \end{aligned} \tag{13}$$

We model the probability distribution using an RNN language model along with the emotion context vector $\boldsymbol{e}$. Specifically, at time step $t$, the hidden state of the decoder $\boldsymbol{s}_t$ is obtained by applying the GRU function,

$$\boldsymbol{s}_t = \mathrm{GRU}(\boldsymbol{s}_{t-1}, \mathrm{concat}(\boldsymbol{c}_t, \boldsymbol{w}_{y_{t-1}})), \tag{14}$$

where $\boldsymbol{w}_{y_{t-1}}$ is the word embedding of $y_{t-1}$. Similar to Affect-LM [11], we then define a new feature vector $\boldsymbol{o}_t$ by concatenating $\boldsymbol{s}_t$ (which we refer to as the language context vector) with the emotion context vector $\boldsymbol{e}$,

$$\boldsymbol{o}_t = \mathrm{concat}(\boldsymbol{s}_t, \boldsymbol{e}), \tag{15}$$

on which we apply a softmax layer to obtain a probability distribution over the vocabulary,

$$\boldsymbol{p}_t = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{o}_t + \boldsymbol{b}), \tag{16}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters. Each term in Equation (13) is then given by

$$p(y_t \mid y_1, \ldots, y_{t-1}, \boldsymbol{c}_t, \boldsymbol{e}) = \boldsymbol{p}_{t, y_t}. \tag{17}$$

[1] https://liwc.wpengine.com/

**Table 1: Statistics of the two datasets.**

| | Cornell | DailyDialog |
|---|---|---|
| # dialogs | 83,097 | 13,118 |
| # utterances | 304,713 | 102,977 |
| Average # turns | 3.7 | 7.9 |
| Average # words / utterance | 12.5 | 14.6 |
| Training set size | 142,450 | 46,797 |
| Validation set size | 10,240 | 10,240 |

We use the cross-entropy loss as our objective function

$$\mathcal{L} = -\frac{1}{\sum_{i=1}^{N} T_i} \sum_{i=1}^{N} \log p\big(\boldsymbol{y}^{(i)} \mid \boldsymbol{X}^{(i)}\big). \tag{18}$$

## 4 EVALUATION

We trained our model using two different datasets and compared its performance with HRAN as well as the basic seq2seq model by performing both offline and online testings.

### Datasets

We used two different dialog corpora to train our model—the Cornell Movie Dialogs Corpus [6] and the DailyDialog dataset [20].

- **Cornell Movie Dialogs Corpus**. The dataset contains 83,097 dialogs (220,579 conversational exchanges) extracted from raw movie scripts. In total there are 304,713 utterances.
- **DailyDialog**. The dataset is developed by crawling raw data from websites used for language learners to learn English dialogs in daily life. It contains 13,118 dialogs in total.

We summarize some of the basic information regarding the two datasets in Table 1.

In our experiments, the models were first trained on the Cornell Movie Dialogs Corpus, and then fine-tuned on the DailyDialog dataset. We adopted this training pattern because the Cornell dataset is bigger but noisier, while DailyDialog is smaller but more daily-based. To create a training set and a validation set for each of the two datasets, we took segments of each dialog with number of turns no more than six,[2] to serve as the training/validation examples. Specifically, for each dialog $\boldsymbol{D} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$, we created $M - 1$ context-response pairs, namely $\boldsymbol{U}_i = (\boldsymbol{x}_{s_i}, \ldots, \boldsymbol{x}_i)$ and $\boldsymbol{y}_i = \boldsymbol{x}_{i+1}$, for $i = 1, 2, \ldots, M-1$, where $s_i = \max(1, i - 4)$. We filtered out those pairs that have at least one utterance with length greater than 30. We also reduced the frequency of those pairs

[2] We chose the maximum number of turns to be six because we would like to have a longer context for each dialog while at the same time keeping the training procedure computationally efficient.

whose responses appear too many times (the threshold is set to 10 for Cornell, and 5 for DailyDialog), to prevent them from dominating the learning procedure. See Table 1 for the sizes of the training and validation sets. The test set consists of 100 dialogs with four turns. We give more detailed description of how we created the test set in the section of human evaluation.

### Baselines and Implementation

Our choice of including S2S is rather obvious. Including HRAN instead of other neural dialog models with affect information was not an easy decision. As mentioned in the related work, Asghar's affective dialog model, the affect-rich conversation model, and the Emotional Chatting Machine do not learn the emotional exchanges in the dialogs. This leaves us wondering whether using a multi-turn neural model can be as effective in learning emotional exchanges as MEED. In addition, comparing S2S and HRAN also gives us an idea of how much the hierarchical mechansim is improving upon the basic model. This is why our final comparision is based on three multi-turn dialog generation models: the standard seq2seq model (denoted as S2S), HRAN, and our proposed model, MEED. In order to adapt S2S to the multi-turn setting, we concatenate all the history utterances in the context into one.

For all the models, the vocabulary consists of 20,000 most frequent words in the Cornell and DailyDialog datasets, plus three extra tokens: <unk> for words that do not exist in the vocabulary, <go> indicating the begin of an utterance, and <eos> indicating the end of an utterance. Here we summarize the configurations and parameters of our experiments:

- We set the word embedding size to 256. We initialized the word embeddings in the models with word2vec [26] vectors first trained on Cornell and then fine-tuned on DailyDialog, consistent with the training procedure of the models.
- We set the number of hidden units of each RNN to 256, the word-level attention depth to 256, and utterance-level 128. The output size of the emotion embedding layer is 256.
- We optimized the objective function using the Adam optimizer [15] with an initial learning rate of 0.001.
- For prediction, we used beam search [39] with a beam width of 256.

We have made the source code publicly available.[3]

### Evaluation Metrics

The evaluation of chatbots remains an open problem in the field. Recent work [22] has shown that the automatic evaluation metrics borrowed from machine translation such as BLEU score [27] tend to align poorly with human judgement. Therefore, in this paper, we mainly adopt human evaluation, along with perplexity and BLEU score, following the existing work.

*Automatic Evaluation.* Perplexity is a measurement of how a probability model predicts a sample. It is a popular method used in language modeling. In neural dialog generation community, many researchers have adopted this method, especially in the beginning of this field [32, 42, 45, 48–50]. It measures how well a dialog model predicts the target response. Given a target response $\boldsymbol{y} = \{y_1, y_2, \ldots, y_T\}$, the perplexity is calculated as

$$\begin{aligned} \text{ppl}(\boldsymbol{y}) &= p(y_1, y_2, \ldots, y_T)^{-1/T} \\ &= \exp\left[ -\frac{1}{T} \sum_{t=1}^{T} \log p(y_t \mid y_1, \ldots, y_{t-1}) \right]. \end{aligned} \tag{19}$$

Thus a lower perplexity score indicates that the model has better capability of predicting the target sentence, i.e., the humans' response. Some researchers [19, 34, 48] argue that perplexity score is not the ideal measurement because for a given context history, one should allow many responses. This is especially true if we want our conversational agents to speak more diversely. However, for our purpose, which is to speak emotionally appropriately and as human-like as possible, we believe this is a good measure. We do recognize that it is not the only way to measure chatbots' performance. This is why we also conducted human evaluation experiment.

BLEU score is often used to measure the quality of machine-translated text. Some earlier work of dialog response generation [17, 18] adopted this metric to measure the performance of chatbots. However, recent study [22] suggests that it does not align well with human evaluation. Nevertheless, we still include BLEU scores in this paper, to get a sense of comparison with perplexity and human evaluation results.

*Human Evaluation.* Human evaluation has been widely used to evaluate open-domain dialog generation tasks. This approach can include any criterion as we judge appropriate. Most commonly, researchers have included the model's ability to generate grammatically correct, contextually coherent, and emotionally appropriate responses, of which the latter two properties cannot be reliably evaluated using automatic metrics. Recent work [1, 48, 49] on affect-rich conversational chatbots turned to human opinion to evaluate both fluency and emotionality of their models. But such human experiments are sensitive to risk factors if the experiment is not carefully designed. They include whether the intructions are clear, whether they have been tested with users before hand, and whether there is a good balance of the human judgement tasks. Further, if a test set for human evaluation is prepared

[3] https://github.com/yuboxie/meed

by randomly sampling the dialogs from the dataset, it may include out-of-context dialogs, causing confusion and ambiguity for human evaluators. Unbalanced emotional distribution of the test dialogs may also lead to biased conclusions since the chatbot's abilities are evaluated on the unrepresentative sample.

To take into account the above issues, we took several iterations to prepare the instructions and the test set before conducting the human evaluation experiment. Part of our test set comes from the DailyDialog dataset, which consists of meaningful complete dialogs. To compensate for the inbalance, we further curated more negative emotion dialogs so that the final set has equal emotion distributions. We provide the details about the test data preparation process and the evaluation experiment below.

*Preparation of Natural Dialog Test Set.* We first selected the emotionally colored dialogs with exactly four turns from the DailyDialog dataset. In the dataset each dialog turn is annotated with a corresponding emotional category, including the neutral one. For our purposes we filtered out only those dialogs where more than a half of utterances have non-neutral emotional labels, resulting in 78 emotionally positive dialogs and 14 emotionally negative dialogs. We recruited two human workers to augment the data to produce more emotionally negative dialogs. Both of them were PhD students from our university (males, aged 24 and 25), fluent in English, and not related to the authors' lab. We found them via email and messaging platforms, and offered 80 CHF (or roughly US $80) gift coupons as incentive for each participant. The workers fulfilled the tasks in Google form[4] following the instructions and created five negative dialogs with four turns, as if they were interacting with another human, in each of the following topics: *relationships*, *entertainment*, *service*, *work and study*, and *everyday situations*. The Google form was released on 31 January 2019, and the workers finished their tasks by 4 February 2019. Subsequently, to form the final test set, we randomly selected 50 emotionally positive and 50 emotionally negative dialogs from the two pools of dialogs described above.

*Human Evaluation Experiment Design.* In the final human evaluation of the model, we recruited four more PhD students from our university (1 female and 3 males, aged 22–25). Three of them are fluent English speakers and one is a native speaker. The recruitment proceeded in the same manner as described above; the raters were offered 80 CHF (or roughly US $80) per participant gift coupons for fulfilling the task, and extra 20 CHF (or roughly US $20) coupon was promised as a bonus to the rater judged to be the most serious. For the evaluation survey, we also leveraged Google form. Specifically, we randomly shuffled the 100 dialogs in the test set, then we used the first three utterances of each dialog as the input to the three models being compared (S2S, HRAN, and MEED), and obtain the respective responses. Dialog contexts and three models' responses were included into Google form. According to the context given, the raters were instructed to evaluate the quality of the responses based on three criteria:

(1) *Grammatical correctness*—whether or not the response is fluent and free of grammatical mistakes;
(2) *Contextual coherence*—whether or not the response is context sensitive to the previous dialog history;
(3) *Emotional appropriateness*—whether or not the response conveys the right emotion and feels as if it had been produced by a human.

For each criterion, the raters gave scores of either 0, 1 or 2, where 0 means bad, 2 means good, and 1 indicates neutral. For this survey, the Google form was launched on 12 February 2019, and all the submissions from our raters were collected by 14 February 2019.

### Results and Analysis

In this subsection, we present the experimental results of the automatic evaluation metric as well as human judgement, followed by some analysis.

*Automatic Evaluation Results.* Table 2 gives the perplexity and BLEU scores obtained by the three models on the two validation sets and the test set. As shown in the table, MEED achieves the lowest perplexity and the highest BLEU score on all three sets. We conducted $t$-test on the perplexity obtained, and results show significant improvements of MEED over S2S and HRAN on the two validation sets (with $p$-value $< 0.05$).

*Human Evaluation Results.* Table 3, 4 and 5 summarize the human evaluation results on the responses' grammatical correctness, contextual coherence, and emotional appropriateness, respectively. In the tables, we give the percentage of votes each model received for the three scores, the average score obtained, and the agreement score among the raters. Note that we report Fleiss' $\kappa$ score [10] for contextual coherence and emotional appropriateness, and Finn's $r$ score [9] for grammatical correctness. We did not use Fleiss' $\kappa$ score for grammatical correctness. As agreement is extremely high, this can make Fleiss' $\kappa$ very sensitive to prevalence [13]. On the contrary, we did not use Finn's $r$ score for contextual coherence and emotional appropriateness because it is only reasonable when the observed variance is significantly less than the chance variance [40], which did not apply to these two criteria. As shown in the tables, we got high agreement among the raters for grammatical correctness, and fair

[4]We provide the link to the form used for creating the dialogs: https://forms.gle/rPagMZYuYJ3M3Sq8A, hoping to help other researchers reproduce the same procedure. However, due to privacy concerns, we do not plan to release this dataset.

**Table 2: Perplexity and average BLEU scores achieved by the models. Avg. BLEU: average of BLEU-1, -2, -3, and -4. Validation set 1 comes from the Cornell dataset, and validation set 2 comes from the DailyDialog dataset.**

| | Perplexity | | | Avg. BLEU | | |
|---|---|---|---|---|---|---|
| | Validation Set 1 | Validation Set 2 | Test Set | Validation Set 1 | Validation Set 2 | Test Set |
| S2S | 43.136 | 25.418 | 19.913 | 1.639 | 2.427 | 3.720 |
| HRAN | 46.225 | 26.338 | 20.355 | 1.701 | 2.368 | 2.390 |
| MEED | **41.862** | **24.341** | **19.795** | **1.829** | **2.635** | **4.281** |

**Table 3: Human evaluation results on grammatical correctness.**

| | +2 | +1 | 0 | Avg. Score | $r$ |
|---|---|---|---|---|---|
| S2S | 98.0 | 0.8 | 1.2 | 1.968 | 0.915 |
| HRAN | 98.5 | 1.3 | 0.2 | 1.982 | 0.967 |
| MEED | 99.5 | 0.3 | 0.2 | **1.992** | 0.981 |

**Table 4: Human evaluation results on contextual coherence.**

| | +2 | +1 | 0 | Avg. Score | $\kappa$ |
|---|---|---|---|---|---|
| S2S | 25.8 | 19.7 | 54.5 | 0.713 | 0.389 |
| HRAN | 37.3 | 21.2 | 41.5 | 0.958 | 0.327 |
| MEED | 38.5 | 22.0 | 39.5 | **0.990** | 0.356 |

**Table 5: Human evaluation results on emotional appropriateness.**

| | +2 | +1 | 0 | Avg. Score | $\kappa$ |
|---|---|---|---|---|---|
| S2S | 21.8 | 25.2 | 53.0 | 0.688 | 0.361 |
| HRAN | 30.5 | 28.5 | 41.0 | 0.895 | 0.387 |
| MEED | 32.0 | 27.8 | 40.2 | **0.917** | 0.337 |

agreement among the raters for contextual coherence and emotional appropriateness.[5] For grammatical correctness, all three models achieved high scores, which means all models are capable of generating fluent utterances that make sense. For contextual coherence and emotional appropriateness, MEED achieved higher average scores than S2S and HRAN, which means MEED keeps better track of the context and can generate responses that are emotionally more appropriate and natural. We first conducted Friedman test [12] and then $t$-test on the human evaluation results (contextual coherence and emotional appropriateness), showing the improvements of MEED over S2S are significant (with $p$-value $< 0.01$).

The comparison between perplexity scores and human evaluation results further confirms the fact that in the context of dialog response generation, perplexity does not align with human judgement. In Table 2, for all the three sets, HRAN performs worse than S2S in terms of perplexity. However, for all of the three criteria in human evaluation, HRAN actually outperforms S2S. Based on this, we conclude that perplexity alone is not enough for evaluating a dialog system.

*Visualization of Output Layer Weights.* We may wonder how HRAN and MEED differ in terms of the distributional representations of their respective vocabularies (words in the language model, and affect words). We decided to visualize the output layer weights as word embedding representations using dimensionality reduction technique for the various models.

In the decoding phase, Equation (16) takes $\boldsymbol{o}_t$, the concatenation of the language context vector $\boldsymbol{s}_t$ and the emotion context vector $\boldsymbol{e}$, and generates a probability distribution over the vocabulary words by applying a softmax layer. The weight matrix of this softmax layer is denoted as $\boldsymbol{W}$, whose shape is $|V| \times 2d$, where $|V|$ is the vocabulary size and $d = 256$ is the hidden state size of the RNNs. Thus the $i$th row of the weight matrix $\boldsymbol{W}_i$ can be regarded as a vector representation of the $i$th word in the vocabulary. Since we concatenate the language context vector and the emotion context vector as the input to the softmax layer, the first half of the weight vector $\boldsymbol{W}_i$ corresponds to the language context vector, and the second half corresponds to the emotion context vector. We refer to them as language model weights and emotion weights, respectively. If the emotion embedding layer is learning and distinguishing affect states correctly, we will see clear differences in the visualization.

With t-SNE [25], we are able to reduce the dimensionality of the weights to two, and visualize them in a straightforward way. For better illustration, we selected 100 most frequent (emotionally) positive words and 100 most frequent negative words from the vocabulary, and used t-SNE to project the corresponding language model weights and emotion weights to two dimensions. Figure 2 gives the results in three subplots. Since HRAN does not have the emotion context vector, we just visualized the whole output layer weight vector, which does a similar job as the language model weights in

[5] https://en.wikipedia.org/wiki/Fleiss%27_kappa#Interpretation

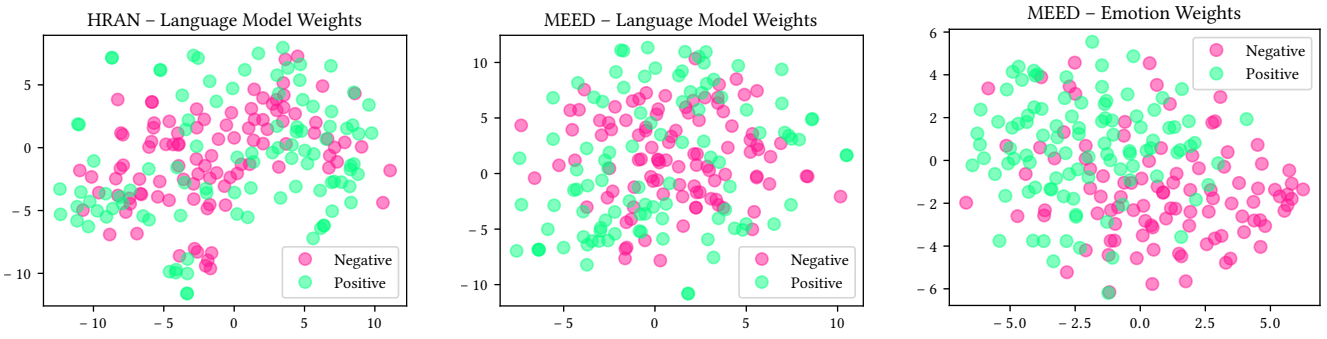


**Figure 2: t-SNE visualization of the output layer weights in HRAN and MEED. 100 most frequent positive words and 100 most frequent negative words are shown. The weight vectors in MEED are separated into two parts and visualized individually.**

MEED. We can observe from the first two plots that positive words (green dots) and negative words (red dots) are scattered around and mixed with each other in the language model weights for HRAN and MEED respectively, which means no emotion information is captured in these weights. On the contrary, the emotion weights in MEED, in the last plot, have a clearer clustering effect, i.e., positive words are mainly grouped on the top-left, while negative words are mainly grouped at the bottom-right. This gives the hint that the emotion encoder in MEED is capable of tracking the emotion states in the conversation history.

*Case Study.* We present four sample dialogs in Table 6, along with the responses generated by the three models. Dialog 1 and 2 are emotionally positive and dialog 3 and 4 are negative. For the first two examples, we can see that MEED is able to generate more emotional content (like "fun" and "congratulations") that is appropriate according to the context. For dialog 4, MEED responds in sympathy to the other speaker, which is consistent with the second utterance in the context. On the contrary, HRAN poses a question in reply, contradicting the dialog history.

## 5 DISCUSSION

In this section, we briefly discuss how our framework can incorporate other components, as well as several directions to extend it.

### Emotion Recognition

To extract the affect information contained in the utterances, we used the LIWC text analysis program. We believe this emotion recognition step is vital for a dialog model to produce emotionally appropriate responses. However, the choice of emotion classifier is not strictly limited to LIWC. It could be replaced by other well-established affect recognizer or one that is more appropriate to the target domain. For example, we can consider using more fine-grained emotion categories from GALC [31], or using DeepMoji [8], which was trained on millions of tweets with emoji labels and is more suitable for tweet-like conversations. However, for DeepMoji, the 64 categories of emojis do not have a clear and exact correspondence with standardized emotion categories, nor to the VAD vectors.

### Training Data

We pre-trained our model on the Cornell movie subtitles and then fine-tuned it with the DailyDialog dataset. We adopted this particular training order because we would like our chatbot to talk more like human chit-chats, and the DailyDialog dataset, compared with the bigger Cornell dataset, is more daily-based. Since our model learns how to respond properly in a data-driven way, we believe having a training dataset with good quality while being large enough plays an important role in developing an engaging and user-friendly chatbot. Thus, in the future, we plan to train our model on the multi-turn conversations that we have already extracted from the much bigger OpenSubtitles corpus and the EmpatheticDialogues dataset.[6]

### Evaluation

Evaluation of dialog models remains an open problem in the response generation field. Early work [18, 30, 36] on response generation used automatic evaluation metrics borrowed from the machine translation field, such as the BLEU score, to evaluate dialog systems. Later on, Liu et al. [22] showed that these metrics correlate poorly with human judgement. Recently, a number of researchers began developing automatic and data-driven evaluation methods [24, 38], with the ultimate goal of replacing human evaluation. However they are still in an early stage. In this paper, we used both perplexity measures and human judgement in our experiments to finalize our model. In other words, using the perplexity measures, we were able to determine when to stop training our model. But this condition does not gurantee the optimal results until

[6]https://github.com/facebookresearch/EmpatheticDialogues

**Table 6: Sample model responses. For each dialog, the ground truth is included in a pair of parentheses.**

| | Context | Model Responses |
|---|---|---|
| 1 | A: I'm happy to see you again.<br>B: Mee too.<br>A: We should do this more often.<br>(B: Okay, I'll give you a ring next week.) | S2S: I hope so.<br>HRAN: Thanks a lot.<br>MEED: That sounds like fun. |
| 2 | A: Thank god! I am finished writing that service guide! It took me forever!<br>B: When did you finish?<br>A: This morning! No more overtime, and no more headaches!<br>(B: Well, I'm glad to hear it. Have a cup of coffee!) | S2S: When?<br>HRAN: Why?<br>MEED: Congratulations! |
| 3 | A: I think that's settled.<br>B: I'm tired of your cut-and-dried opinions. Who do you think you are!<br>A: How dare you speak to me like this.<br>(B: Why not?) | S2S: What is it?<br>HRAN: What is it?<br>MEED: Are you serious? |
| 4 | A: This concert was awful.<br>B: Agreed, the musicians were not in harmony.<br>A: It was too painful. Never again.<br>(B: That's for sure!) | S2S: Congratulations!<br>HRAN: Why not?<br>MEED: That's true. |

human judgement test can validate them. We thus highly recommend this combination, which is also a common practice in the research community [45, 48–50].

### Model Extensions

Our model uses RNNs to encode the input sequences, and GRU cells to capture long-term dependency among different positions in the sequences. Recent advances in natural language understanding have proposed new network architectures to process text input. Specifically, the Transformer [41] uses pure attention mechanisms without any recurrence structures. Compared with RNNs, the Transformer can capture better long-term dependency due to the self-attention mechanism, which is free of locality biases, and is more efficient to train because of better parallelization capability. Following the Transformer architecture, researchers found that pre-training language models on huge amounts of data could largely boost the performance of downstream tasks, and published many pre-trained language models such as BERT [7] and RoBERTa [23]. As future work, we would like to adopt the Transformer architecture to replace the RNNs in our model, and initialize our encoder with pre-trained language models. We hope to increase the performance of response generation.

## 6 CONCLUSION

We believe reproducing conversational and emotional intelligence will make social chatbots more believable and engaging. In this paper, we proposed a multi-turn dialog system capable of recognizing and generating emotionally appropriate responses, which is the first step toward such a goal. We have demonstrated how to do so by (1) modeling utterances with extra affect vectors, (2) creating an emotional encoding mechanism that learns emotion exchanges in the dataset, (3) curating a multi-turn and balanced dialog dataset, and (4) evaluating the model with offline and online experiments. For future directions, we would like to investigate the diversity issue of the responses generated, possibly by extending the mutual information objective function [17] to multi-turn settings. We would also like to adopt the Transformer architecture with pre-trained language model weights, and train our model on a much larger dataset, by extracting multi-turn dialogs from the OpenSubtitles corpus.

## REFERENCES

[1] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective Neural Response Generation. In *Proceedings of ECIR 2018*. 154–166. https://doi.org/10.1007/978-3-319-76941-7_12

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). arXiv:1409.0473 http://arxiv.org/abs/1409.0473

[3] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (2005), 293–327. https://doi.org/10.1145/1067860.1067867

[4] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explorations* 19, 2 (2017), 25–35. https://doi.org/10.1145/3166054.3166058

[5] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014.

Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of EMNLP 2014*. 1724–1734. http://aclweb.org/anthology/D/D14/D14-1179.pdf

[6] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of CMCL@ACL 2011*. 76–87. https://aclanthology.info/papers/W11-0609/w11-0609

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*. 4171–4186. https://aclweb.org/anthology/papers/N/N19/N19-1423/

[8] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using Millions of Emoji Occurrences to Learn Any-Domain Representations for Detecting Sentiment, Emotion and Sarcasm. In *Proceedings of EMNLP 2017*. 1615–1625. https://aclanthology.info/papers/D17-1169/d17-1169

[9] Robert H Finn. 1970. A Note on Estimating the Reliability of Categorical Data. *Educational and Psychological Measurement* 30, 1 (1970), 71–76.

[10] Joseph L Fleiss and Jacob Cohen. 1973. The Equivalence of Weighted kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[11] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *Proceedings of ACL 2017*. 634–642. https://doi.org/10.18653/v1/P17-1059

[12] David C Howell. 2016. *Fundamental Statistics for the Behavioral Sciences*. Nelson Education.

[13] George Hripcsak and Daniel F. Heitjan. 2002. Measuring Agreement in Medical Informatics Reliability Studies. *Journal of Biomedical Informatics* 35, 2 (2002), 99–110. https://doi.org/10.1016/S1532-0464(02)00500-2

[14] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. In *Proceedings of CHI 2018*. 415. https://doi.org/10.1145/3173574.3173989

[15] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[16] Jonathan Klein, Youngme Moon, and Rosalind W. Picard. 2001. This Computer Responds to User Frustration: Theory, Design, and Results. *Interacting with Computers* 14, 2 (2001), 119–140. https://doi.org/10.1016/S0953-5438(01)00053-4

[17] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of NAACL-HLT 2016*. 110–119. http://aclweb.org/anthology/N/N16/N16-1014.pdf

[18] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of ACL 2016*. http://aclweb.org/anthology/P/P16/P16-1094.pdf

[19] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of EMNLP 2016*. 1192–1202. http://aclweb.org/anthology/D/D16/D16-1127.pdf

[20] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of IJCNLP 2017*.

[21] Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of LREC 2016*. http://www.lrec-conf.org/proceedings/lrec2016/summaries/947.html

[22] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of EMNLP 2016*. 2122–2132. http://aclweb.org/anthology/D/D16/D16-1230.pdf

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[24] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings ACL 2017*. 1116–1126. https://doi.org/10.18653/v1/P17-1103

[25] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*. 311–318. http://www.aclweb.org/anthology/P02-1040.pdf

[28] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[29] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of ACL 2019*. 5370–5381. https://www.aclweb.org/anthology/P19-1534/

[30] Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of EMNLP 2011*. 583–593. http://www.aclweb.org/anthology/D11-1054

[31] Klaus R Scherer. 2005. What Are Emotions? And How Can They Be Measured? *Social science information* 44, 4 (2005), 695–729.

[32] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of AAAI 2016*. 3776–3784. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957

[33] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of AAAI 2017*. 3295–3301. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567

[34] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of ACL-IJCNLP 2015*. 1577–1586. http://aclweb.org/anthology/P/P15/P15-1152.pdf

[35] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of CIKM 2015*. 553–562. https://doi.org/10.1145/2806416.2806493

[36] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of NAACL-HLT 2015*. 196–205. http://aclweb.org/anthology/N/N15/N15-1020.pdf

[37] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS 2014*. 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-

learning-with-neural-networks

[38] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *Proceedings of AAAI 2018*. 722–729. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16179

[39] Christoph Tillmann and Hermann Ney. 2003. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics* 29, 1 (2003), 97–133. https://doi.org/10.1162/089120103321337458

[40] Howard E Tinsley and David J Weiss. 1975. Interrater Reliability and Agreement of Subjective Judgments. *Journal of Counseling Psychology* 22, 4 (1975), 358.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS 2017*. 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[42] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *CoRR* abs/1506.05869 (2015). arXiv:1506.05869 http://arxiv.org/abs/1506.05869

[43] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.

[44] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of AAAI 2017*. 3351–3357. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14563

[45] Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical Recurrent Attention Network for Response Generation. In *Proceedings of AAAI 2018*. 5610–5617. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16510

[46] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of CHI 2017*. 3506–3510. https://doi.org/10.1145/3025453.3025496

[47] Jennifer Zamora. 2017. I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations. In *Proceedings of HAI 2017*. 253–260. https://doi.org/10.1145/3125739.3125766

[48] 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 AAAI 2019*. 7492–7500. https://aaai.org/ojs/index.php/AAAI/article/view/4740

[49] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of AAAI 2018*. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16455

[50] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. In *Proceedings of ACL 2018*. 1128–1137. https://doi.org/10.18653/v1/P18-1104