Generating Empathetic Responses with a Large Scale Dialog Dataset

Yubo Xie and Pearl Pu
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne, Switzerland
{yubo.xie, pearl.pu}@epfl.ch

Abstract

The task of empathetic response generation aims at generating syntactically correct and, more importantly, emotionally appropriate responses following previous dialog turns. Existing models either directly incorporate pre-defined emotion information to guide the response generation, or use deterministic rules to decide the response emotion, ignoring the subtle emotion interactions captured in human conversations. With the advent of advanced language models, it is possible to learn the nuanced emotional exchanges captured in natural language dialogs. To fully explore the range of emotions and dialog intents, it is important to curate a dataset large enough to shed light on the general understanding of human emotional interactions in our conversations. In this paper, we describe in detail the curation process of a large-scale dialog dataset where each utterance is labeled with one of 32 emotions and 9 intent categories. We then show how to build a multi-turn empathetic dialog model that performs well compared to its baselines over 6,000 human evaluated instances.

1 Introduction

Endowing chatbots with empathy is an important step towards building human-like dialog systems. The task requires dialog models to generate responses that are emotionally appropriate following the conversation context. Some of the existing work (Zhou et al., 2018; Huang et al., 2018; Zhou and Wang, 2018; Colombo et al., 2019; Santhanam and Shaikh, 2019; Song et al., 2019; Shen and Feng, 2020) proposes various models for emotion-controllable dialog generation, which requires a manually defined emotion label as input to guide the generation of the response. However this might be impractical when deploying the chatbots in reality. Other work (Asghar et al., 2018; Li and Sun, 2018; Zhong et al., 2019; Lin et al., 2019; Li et al., 2020) adopts some manually defined rules, either explicitly or implicitly, to decide the emotion state for the response to be generated, e.g., following/reversing the speaker's emotion, or just maximizing the emotion content in the response. However, psychology literature does not confirm such deterministic rules for emotion interactions in human conversations. On the other hand, such emotional interactions can be observed in natural language dialogs and if a dataset of empathetic interaction existed a data-driven neural approach can learn these emotional exchanges. To fully explore the range of emotions and dialog intents, it is important to curate a dataset large enough to shed light on the general understanding of human emotional interactions in our conversations. For this reason, the previously curated datasets (Busso et al., 2008; Li et al., 2017; Hsu et al., 2018; Poria et al., 2019; Chatterjee et al., 2019) were too small and the emotion categories were too limited, leaving much room for improvement. Another limitation of current empathetic chatbots is the insufficient treatment of neutral categories, either ignoring the neutral category or grouping them into a single category (e.g., *other* or *neutral*).

Evaluation of empathetic chatbots also remains a challenging problem in the dialog generation community. Earlier work (Liu et al., 2016) has shown that automatic metrics designed for machine translation and automatic summarization have weak or no correlation with human judgement. Therefore, human evaluation for chatbots has been widely adopted. However, we found all the existing work that leveraged human evaluation on empathetic chatbots only included a rather limited amount of test samples, usually hundreds of them. While few of them conducted the experiment on crowdsourcing platforms, the majority just recruited three to five raters to do the evaluation tasks. We believe the limited number of test samples and human evaluators could both potentially lead to biased results.

To fill the aforementioned gaps, we describe the curation process of a large-scale multi-turn dialog dataset from the OpenSubtitles2018 corpus (Lison et al., 2019). To obtain a more fine-grained treatment of emotions, we trained a RoBERTa (Liu et al., 2019) based emotion classifier capable of predicting one of the 32 emotion categories from the EmpatheticDialogues dataset (Rashkin et al., 2019) and 8 additional dialog intents that we have manually delineated (plus one neutral category). This classifier was used to select out emotional dialogs from the curated OpenSubtitles dialog dataset, forming a total number of 1M multi-turn empathetic dialogs. We also designed an empathetic dialog model that learns the emotion exchanges in dialog data at a more fine-grained level and generates empathetic responses accordingly. The model consists of three parts: (1) an encoder that encodes the input dialog context into vector representations; (2) a response emotion/intent predictor that decides the appropriate emotion/intent for the response to be generated; (3) a decoder that generates the response based on the predicted response emotion/intent, while constantly attending to the encoder outputs. Finally, we carefully designed a human evaluation experiment on the crowdsourcing platform that enabled the workers to finish the evaluation tasks more easily, and at the same time kept them engaged by introducing bonus checkpoints. We tested a total number of 6,000 dialogs, which, to the best of our knowledge, has never been attempted before for the evaluation of empathetic chatbots.

2 Related Work

Emotional dialog datasets Most of the existing emotional dialog datasets are small in size and have limited number of emotion categories. Busso et al. (2008) developed the IEMOCAP by recording data from actors in dyadic sessions who performed selected scripts to elicit 5 emotion types. Li et al. (2017) created the DailyDialog dataset from English learning websites, consisting of 13K multi-turn dialogs manually labeled with 7 emotions. The EmotionLines dataset (Hsu et al., 2018) contains 2,000 dialogs collected from Friends TV scripts and EmotionPush chat logs, labeled with 7 emotions. Poria et al. (2019) extended the EmotionLines dataset to a multimodal setting, containing 1,433 dialogs from Friends TV scripts. Chatterjee et al. (2019) proposed the EmoContext dataset collected from users' interaction with a conversational agent, which contains 38K dialogs labeled with 4 emotions. Rashkin et al. (2019) curated the EmpatheticDialogues dataset containing 25K dialogs collected from a crowdsourcing platform by letting workers communicate with each other based on 32 emotion categories.

Neural response generation Vinyals and Le (2015) trained the seq2seq network on IT helpdesk dialogs and OpenSubtitles data. Shang et al. (2015) applied an attention mechanism to the seq2seq network and trained it on short-text social media dialogs. To adapt the seq2seq model to a multi-turn setting, Serban et al. (2016) designed a hierarchical encoder-decoder structure, based on which Xing et al. (2018) devised a hierarchical attention mechanism so that the model could pay attention at both token-level and utterance-level.

Empathetic response generation Lubis et al. (2018) designed a hierarchical encoder-decoder model that captures the user's emotion state and takes it into account when generating the response, by incorporating an emotion encoder that predicts the emotion label of the current dialog turn. Xie et al. (2019) proposed a multi-turn emotionally engaging dialog model by modeling the emotion states in the dialog history. Shin et al. (2020) adopted a reinforcement learning framework that provides a higher reward to the generative model if it promotes the user's future emotion state. Li et al. (2019) adopted an adversarial learning framework and proposed two discriminators to evaluate if the generated response is empathetic and elicits more positive emotions by considering the emotion words in the gold response and the next reply.

3 Data Curation

Existing empathetic dialog corpora are usually limited in size and training solely on these datasets could not give us a chatbot with desirable performance. Therefore, we would like to take advantage of transfer learning and pre-train the dialog model on a huge amount of dialog data (not necessarily empathetic), and then fine-tune it on a possibly much smaller empathetic dialog dataset.

3.1 Extracting Dialogs from Movie Subtitles

To obtain a large-scale dialog dataset, we relied on the OpenSubtitles2018 corpus (Lison et al.,

Total # dialogs	4,010,009
Total # turns	18,849,440
Total # tokens	312,574,468
Average # turns per dialog	4.70
Average # tokens per turn	16.58
Average # tokens per dialog	77.95

Table 1: OpenSubtitles dialogs after cleaning.

Total # dialogs	1,000,000
Total # turns	3,488,300
Total # tokens	66,447,274
Average # turns per dialog	3.49
Average # tokens per turn	19.05
Average # tokens per dialog	66.45

Table 2: Empathetic OpenSubtitles dialogs.

2019), which contains text collected from movie subtitles spread over 60 languages, and is a good source of human conversations written by professional screenwriters. We only used the English part, which has 447K subtitle files, 441M sentences and 3.2B tokens. Due to the lack of speaker information in the OpenSubtitles corpus, before extracting the dialogs, we followed the same procedure proposed by Lison and Meena (2016) and built an SVM classifier to determine whether two consecutive lines in one subtitle file are actually spoken by the same character and should be in the same dialog turn. As a result, we obtained a turn segmentation accuracy of 76.69%.

We then separated these turns into dialogs by adopting a heuristic rule based on timestamps: for each subtitle file, we calculate the gap between the starting time of each turn and the ending time of its previous turn. If this time gap is greater than 5 seconds, we cut off at this position and regard these two turns as belonging to different dialogs. An exception is when the timestamp information is missing for one of the two turns. In this case, we just regard them as belonging to one dialog. In this way, we obtained 9M dialogs from the whole English OpenSubtitles corpus. To further clean the dataset, we applied a sequence of steps to remove undesirable utterances. The detailed procedure is described in Appendix A. As a result, we obtained 4M cleaned OpenSubtitles dialogs. Table 1 summarizes some statistics of this curated dataset.

3.2 Empathetic OpenSubtitles Dialogs

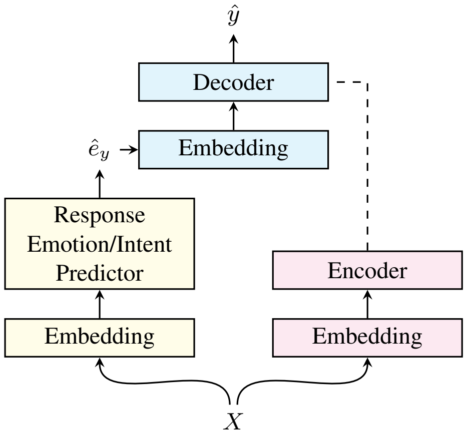
Many existing empathetic dialog datasets are small in size due to the expensive procedure of data collection, usually done manually by human. In this paper, we created a large-scale empathetic dialog dataset by first training a sentence-level fine-grained emotion classifier and then selecting out emotional dialogs from the cleaned OpenSubtitles dataset aforementioned.

To build the emotion classifier, we fine-tuned RoBERTa (Liu et al., 2019) on the situation sentences from the EmpatheticDialogues (Rashkin et al., 2019) training set (labeled with 32 fine-grained emotions), and 7K listener utterances labeled with 8 empathetic intents (*questioning*, *agreeing*, *acknowledging*, *sympathizing*, *encouraging*, *consoling*, *suggesting*, and *wishing*) plus one *neutral* category (all other not mentioned intents). The 7K intent-labeled utterances were obtained by first manually labeling 521 sentences and then expanding through searching most frequent n-grams for each intent. The classifier achieved an accuracy of 65.88% on the EmpatheticDialogues test set. We applied the obtained classifier on all cleaned OpenSubtitles dialogs, and calculated a probability distribution over the 41 categories for each utterance. We then define the emotionality of each utterance as the sum of the probability values of the 32 emotion categories, and the emotionality of each dialog as the averaged emotionality values of its utterances. We selected the top 1M dialogs with highest emotionality values to form the empathetic OpenSubtitles dialog dataset, whose statistics are listed in Table 2. The distribution of emotions/intents in this dataset can be found in Appendix A. Some samples of the OpenSubtitles dialogs can be found in Appendix D. We will make this dataset as well as all the OpenSubtitles dialogs publicly available.

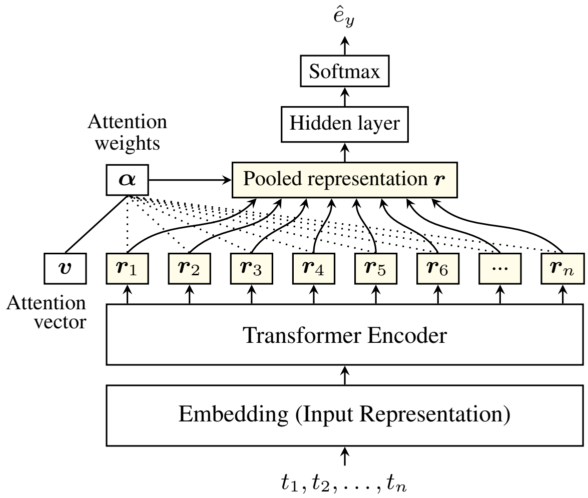
We combined the 32 emotion categories with the extra 8 intents because we think emotions in dialogs can also be treated as intents. For example, we label an utterance "angry" because the speaker is trying to communicate his/her feeling of anger. Emotional experience is primarily caused by an external event. In dialogs, this emotional experience is caused or shared by the interlocutors, and thus can also be regarded as dialog intents.

4 An Empathetic Dialog Model

The problem could be defined as follows: given a dialog context X consisting of one or more utter-



(a) Overall architecture showing how the model works in inference mode. Dashed line denotes multi-head attention.



(b) A detailed illustration of the response emotion/intent predictor. Dotted lines denote attention mechanism.

Figure 1: Illustrations of our dialog model.

ances $u_1, u_2, \ldots, u_m$, spoken between two people, try to generate a response $\hat{y}$ that not only follows the dialog context but also is emotionally appropriate. Our model consists of three modules: (1) an encoder responsible for encoding the input X into vector representations; (2) a response emotion/intent predictor which takes X as input and decides in which emotion/intent the model should respond; (3) a decoder responsible for generating the actual response. We use Transformer (Vaswani et al., 2017) encoder structure for our encoder and emotion/intent predictor, and Transformer decoder structure for our decoder. Figure 1a gives an overall depiction of the whole model architecture. All the three modules have the same input representation (i.e., embedding layers), which we describe in detail next.

4.1 Input Representation

The input representation is illustrated in Figure 2. We use the RoBERTa tokenizer to tokenize the utterances $u_1, u_2, \ldots, u_m$ in the input dialog context X, and concatenate them by two special tokens: <s> and </s>, as shown in the figure. For our model to have a better understanding of the input dialog context, in addition to the word embeddings and position embeddings in the original Transformer architecture, we also have emotion embeddings. Specifically, for each utterance u_i, we use the same emotion classifier described in Section 3.2 to obtain an emotion representation in the form of a probability distribution on 41 emotions/intents. The label with maximum probability value is denoted as e_{u_i}, representing the emotion/intent expressed by utterance u_i. Similar to word embeddings, we embed this emotion/intent e_{u_i} into a vector space with the same dimensionality as other embeddings, so that they could add up. The same emotion embedding is used for all the tokens in the same utterance. To further differentiate between the speakers, we augment the input representation with segment embeddings. Utterances spoken by the same person would have the same segment embedding. The encoder and decoder share the same embedding tables.

4.2 Response Emotion/Intent Predictor

We relied on a data-driven approach to decide the emotion/intent of the response to be generated, by designing an emotion/intent classifier to predict the emotion/intent of the ground-truth response y, based on the context X. As shown in Figure 1b, we use a Transformer encoder to get a context-dependent vector representation $\boldsymbol{r}_i$ for each of the input token t_i. To pool these high-level representations into a single vector, we use a simple attention mechanism and incorporate a trainable vector $\boldsymbol{v}$ to obtain an attention weight $\boldsymbol{\alpha}_i$ for $\boldsymbol{r}_i$,

$$\boldsymbol{\alpha}_i = \frac{\exp(\boldsymbol{v}^T \boldsymbol{r}_i)}{\sum_{j=1}^{N} \exp(\boldsymbol{v}^T \boldsymbol{r}_j)}. \tag{1}$$

The aggregate representation $\boldsymbol{r}$ is then

$$\boldsymbol{r} = \sum_{i=1}^{N} \boldsymbol{\alpha}_i \boldsymbol{r}_i. \tag{2}$$

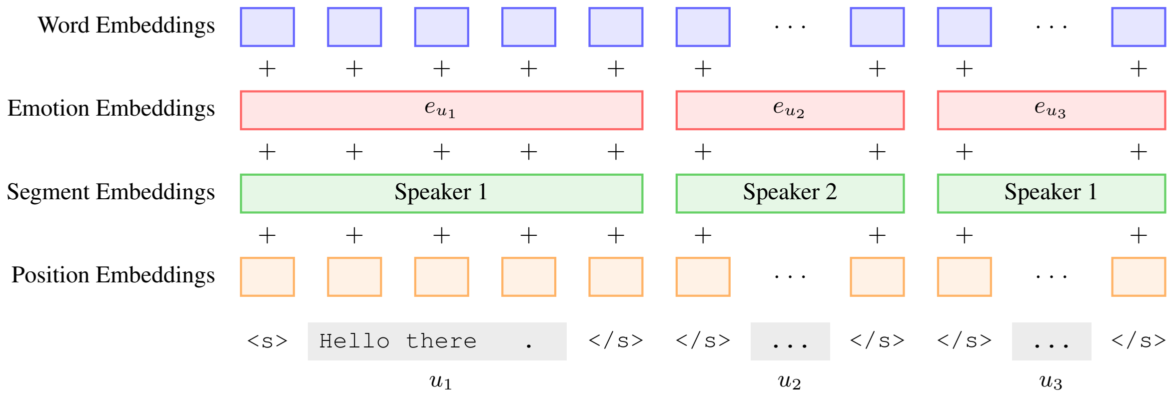


Figure 2: Input representation of our dialog model.

$\boldsymbol{r}$ is fed into a hidden layer followed by a softmax layer to produce $\hat{e}_y$, denoting the predicted emotion/intent of the response to be generated.

4.3 Training

The response emotion/intent predictor is trained separately from the encoder/decoder, which means the training phase is a bit different from what is illustrated in Figure 1a. In particular, the response emotion/intent predictor is independently trained to minimize the cross entropy loss of $\hat{e}_y$ with respect to e_y (true emotion/intent of y). While training the encoder and decoder simultaneously, we just feed e_y into the embedding layers of the decoder, and try to minimize the cross entropy loss of $\hat{y}$ with respect to y.

We also experimented with jointly training the response emotion/intent predictor and the encoder/decoder, by combining two loss functions like in a multi-task setting. However, we found the generated responses quite generic compared with training the two components separately, plus joint training also introduces more hyperparameters to be tuned. Moreover, having them trained separately endows the decoder with more controllability—the decoder is able to generate responses according to a specified emotion/intent label.

5 Evaluation

We trained our empathetic dialog model and the baselines on three datasets and evaluated them in *held-out* setting (meaning the test data comes from the same domain as the training data) and *zero-shot* setting (meaning the test data comes from a different domain than the training data), using both automatic metrics and human judgement via crowdsourcing.

Total # dialogs	24,850
Total # turns	107,220
Total # tokens	1,722,602
Average # turns per dialog	4.31
Average # tokens per turn	16.07
Average # tokens per dialog	69.32

Table 3: EmpatheticDialogues dataset.

5.1 Datasets

Three datasets were involved in the evaluation:

- **OpenSubtitles dialogs**. As described in Section 3.1, these dialogs were obtained by segmenting the movie subtitles. See Table 1 for the statistics. Note that for the purpose of pre-training, we excluded the empathetic OpenSubtitles dialogs (Table 2), resulting in around 3M dialogs. We denote this dataset as OS.

- **Empathetic OpenSubtitles dialogs**. The curation process is described in Section 3.2. The total number of dialogs is 1M (statistics listed in Table 2). We denote this dataset as OSED.

- **EmpatheticDialogues dataset**. This dataset is created by Rashkin et al. (2019) and contains around 25K dialogs collected from the crowdsourcing platform. See Table 3 for the statistics. We denote this dataset as ED.

We split each of the three datasets into training set (80%), validation set (10%), and test set (10%). Among the dialogs of each test set, we further selected out 2,000 to form a combined test set of 6,000 dialogs, for the purpose of evaluating the models on automatic metrics and human judgement via crowdsourcing.

Model	OS				OSED				ED			
	PPL	D1	D2	SES	PPL	D1	D2	SES	PPL	D1	D2	SES
Pre-trained (OS)	24.8	.046	.159	.172	37.8	.046	.154	.126	564.6	.044	.167	.178
Fine-tuned (OSED)	26.9	.044	.139	.162	32.3	.056	.165	.137	452.6	.031	.107	.176
Fine-tuned (ED)	88.9	.030	.109	**.174**	140.8	.028	.096	.130	19.3	.026	.091	**.316**
Raw (ED)	793.9	.009	.032	.144	1615.0	.008	.027	.098	35.8	.008	.029	.278
Ours (OS)	**22.0**	**.064**	**.210**	.168	31.9	.061	.197	.130	487.3	.046	.171	.174
Ours (OS → OSED)	22.8	.057	.196	.168	**28.5**	**.070**	**.225**	**.171**	391.7	**.051**	**.199**	.207
Ours (OS → ED)	84.3	.038	.153	.165	125.7	.036	.138	.116	**17.2**	.036	.140	.299

Table 4: Automatic evaluation results. Here PPL denotes perplexity, D1 and D2 denote Distinct-1 and -2, and SES denotes the sentence embedding similarity. X → Y means pre-training on X and then fine-tuning on Y.

Model	OS			OSED			ED		
	P	R	F-1	P	R	F-1	P	R	F-1
Random	.1484	.0240	.0285	.0382	.0250	.0266	.0989	.0165	.0215
Ours (OS)	**.2210**	**.3960**	.2312	.0109	.1040	.0198	.0942	.3070	.1442
Ours (OS → OSED)	.2012	.1480	.1537	**.1029**	**.1495**	**.0917**	.1288	.2630	.1674
Ours (OS → ED)	.2166	.3265	**.2502**	.0253	.0870	.0239	**.2660**	**.3530**	**.2864**

Table 5: Weighted precision, recall and F-1 scores of the response emotion/intent predictor in our model on the three datasets. X → Y means pre-training on X and then fine-tuning on Y.

5.2 Baselines

Similar to the work of Rashkin et al. (2019), we adopted the full Transformer model as our baseline, and based on the training strategies, we have the following variants:

- **Pre-trained**. To take advantage of transfer learning, we pre-trained the full Transformer model on the curated OS dataset, which contains around 3M dialogs. The large scale of this training set is expected to provide a good starting point for fine-tuning.
- **Fine-tuned**. We took the pre-trained full Transformer, and then fine-tuned it on two empathetic dialog datasets: our curated OSED dataset, and the ED dataset, respectively.
- **Raw**. To test the effectiveness of pre-training, we directly trained the full Transformer on the ED dataset, and then compared it with the fine-tuned models.

Note that we did not include the EmoPrepend-1 model by Rashkin et al. (2019) as our baseline, because in their paper, its human evaluation performance is actually reported to be worse than the fine-tuned Transformer. All the models have a hidden size of 300, and were trained until the minimum validation loss was reached. For inference we used beam search with beam size 32 and 4-gram repeats blocking. Further details regarding the implementation parameters can be found in Appendix B.

5.3 Automatic Evaluation

Most of the existing automatic metrics directly compare the generated response with the ground-truth provided by human, often in a simple way. Due to the inherent diversity of human conversations, this is not suitable for dialog models, since for the same prompt, there could exist many responses that are equally good. In fact, Liu et al. (2016) has shown that word-overlap-based metrics (specifically BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004)) and word embedding metrics all exhibit weak or no correlation with human judgements. To this end, we did not adopt these metrics in our experiment, but instead considered the following:

- **Perplexity**. Perplexity is a model-dependent metric that measures how well a probability model predicts a given sample. In our case, a lower perplexity score indicates better capability of generating the ground-truth response.
- **Distinct-1 and -2**. The Distinct-1 and -2 metrics (Li et al., 2016) measure the diversity of the generated responses by calculating the ratio of unique unigrams or bigrams over the total number of unigrams or bigrams in the

Model	OS			OSED			ED		
	Good	Okay	Bad	Good	Okay	Bad	Good	Okay	Bad
Pre-trained (OS)	.3097	.2878	.4025	.2975	.2933	.4091	.1799	.3037	.5164
Ours (OS)	.3166	.3158	.3676	.3073	.3288	.3639	.1863	.3088	.5049
Ours (OS $\rightarrow$ OSED)	.3175	.3036	.3789	.2926	.3034	.4040	.2097	.2891	.5012
Ours (OS $\rightarrow$ ED)	**.3513**	.3125	.3362	**.3535**	.3093	.3372	**.4890**	.3033	.2077

Table 6: Human evaluation results on each of the three test sets. Numbers have been normalized across the three quality categories on each test set. X $\rightarrow$ Y means pre-training on X and then fine-tuning on Y.

generated responses.

- **Sentence Embedding Similarity**. For this metric, we use Sentence-BERT (Reimers and Gurevych, 2019) to obtain an embedding for the generated response as well as the ground-truth, and then calculate the cosine similarity between the two embeddings.

The results of automatic evaluation are shown in Table 4. Our model achieves lower perplexity scores than the corresponding full Transformer on all the three datasets. Here we have an extra model configuration, Raw (ED), to compare with Fine-tuned (ED), in order to see the effects brought by pre-training. As we can see, without pre-training on OS, the model gets much worse performance on the perplexity scores. This indicates that pre-training and then fine-tuning is preferred to directly training on a target dataset. On Distinct-1 and -2, our model always has a higher score than the corresponding full Transformer model, suggesting that by injecting additional emotion information, the dialog system could be guided to generate more diverse responses. We also observe that on the ED dataset, our model fine-tuned on OSED actually has the highest Distinct scores, even though it has never seen the ED data. We conjecture that this is because the OSED dataset is much bigger than the ED dataset, and contains text that is more diverse. Table 5 lists the weighted precision, recall, and F-1 scores of the response emotion/intent predictor for different model configurations.

If we consider a zero-shot setting, meaning the model is evaluated on data from a different domain than its training data, we see from Table 4 that all models achieves higher perplexity scores on zero-shot test data. In particular, models trained on the OS (OSED) dataset achieves lower perplexity on the OSED (OS) dataset, compared with the results on the ED dataset. This is because OS and OSED dialogs are actually curated from the same source, while the source of ED data is quite different. Moreover, models trained on OSED has better perplexity scores on OS dataset, due to the performance boost brought by fine-tuning.

5.4 Human Evaluation via Crowdsourcing

Human evaluation for dialog models has been widely adopted due to the limitations of automatic metrics. However, the experiment should be carefully designed so that the raters clearly understand the instructions and are constantly engaged in the evaluation tasks. Moreover, most of the existing work only recruited a limited number of raters to evaluate a test set of small size, therefore leading to possibly biased results. In this paper, we carefully designed a human evaluation experiment that enables the raters to work on the evaluation tasks more easily and at the same time keeps them engaged by incorporating bonus checkpoints.

We conducted our human evaluation experiment on Amazon Mechanical Turk (MTurk). The 6,000 test dialogs were randomly shuffled and then split into 600 Human Intellligence Tasks (HITs), with each HIT containing 10 dialogs to be evaluated. For each test dialog, we included the generated responses from four candidate models, i.e., Pre-trained (OS), Ours (OS), Ours (OS $\rightarrow$ OSED), and Ours (OS $\rightarrow$ ED). Unlike the existing work, we adopted an evaluation strategy combining scoring and ranking where the workers can drag candidate responses to one of the three pre-defined areas: *good*, *okay*, and *bad*, according to whether the response is emotionally appropriate following the given dialog context. In this way, it is easier for the workers to finish the tasks, and we also benefit from the accurate scoring results. In order to make the workers more engaged in the evaluation, and also encourage those providing high-quality answers, for each HIT we attached a bonus task to three ED dialogs, by adding the ground-truth response as a candidate. If the worker successfully put the ground-truth into the *good* or *okay* category, he/she will receive a bonus point. We gave a bonus

Context (from OS):
- Well, now, you know, most of your average people... won 't know a venomous reptile from a large garden worm.

Pre-trained (OS): I don 't know what you're talking about.
Ours (OS) (questioning): What do you mean, you won 't know?
Ours (OS $\rightarrow$ OSED) (terrified): I'm not afraid of snakes.
Ours (OS $\rightarrow$ ED) (terrified): Oh my gosh, I hate snakes. What kind of snake was it?
Ground-truth (afraid): Oh, no. It has to be an absolute, authentic, poisonous snake... or he won 't work with it.

Context (from OSED):
- Yes, it's been a long wait. The second half of King of Comedy will now begin!
- Hyo-jin. Yes. I'm excited to see what stars the second half will bring.
- Yes. Welcome our first contestants. Who are they?

Pre-trained (OS): The second half will bring the second half will be the second half.
Ours (OS) (questioning): Who are they? What do they look like?
Ours (OS $\rightarrow$ OSED) (excited): Let's begin!
Ours (OS $\rightarrow$ ED) (excited): I can't wait to see who they are.
Ground-truth (wishing): Warriors. Welcome!

Context (from ED):
- after she got her 1st job and new car, she started doing girls night out. Out until 3am one morning. I checked her chat history and sure enough she had kissed another guy >:(
- Wow, you must have felt jealous
- jealous isn't the one I felt. I wanted to rip her apart lol

Pre-trained (OS): what are you doing here?
Ours (OS) (questioning): Whoa whoa whoa, whoa. whoa. what are you talking about?
Ours (OS $\rightarrow$ OSED) (jealous): You don 't know what you're talking about
Ours (OS $\rightarrow$ ED) (agreeing): I know what you mean. I would have done the same thing.
Ground-truth (acknowledging): Hahahaa, its not easy though, I would have done same

Table 7: Some samples of the generated responses. X $\rightarrow$ Y means pre-training on X and then fine-tuning on Y.

of $0.1 to those workers who obtained all the three bonus points. More details of the human evaluation setup, including a screenshot of the interface, can be found in Appendix C.

In total we received 24,000 answers from the MTurk experiment (4 answers for each of the 6,000 dialogs). We discarded answers from low-quality workers, i.e., those who provided the same answer for almost all dialogs, and those who completed the tasks in less than five minutes and failed to obtain at least two bonus points. Then, to calculate the human evaluation scores, we further selected out those assignments with at least two bonus points, and obtained a total number of 21,630 answers. The human evaluation results on the three individual test sets are shown in Table 6. From the table we see that our model outperforms the full Transformer on all three datasets (Pre-trained (OS) v.s. Ours (OS)), and of all the four model configurations, our model trained on ED achieves the highest percentage of good response on all three datasets, meaning training on ED enables the model to gain both good held-out performance and good zero-shot performance. Compared with our model only pre-trained on OS, it achieves better performance on OS and ED if fine-tuned on OSED, but not on OSED itself, meaning this model has a good zero-shot performance but the held-out performance is somehow lower. This could be explained by the unbalanced emotion/intent distribution in the OS dataset. As discussed in Section 5.3, for our model trained on OS, the response emotion/intent predictor would usually predict the dominating "questioning" category. For OSED dialogs, since the response emotion/intent is more difficult to predict, responding in questions is probably safer.

5.5 Case Study

In this section, we give some sample responses generated by the models in Table 7. We took one dialog from each test set (OS, OSED and ED). We can observe that most of the generated responses are syntactically correct (exceptional cases are from Pre-trained (OS)). The models could understand the dialog context and generate appropriate responses. For example, in the first dialog, our models fine-tuned on OSED and ED recognize and understand the word "reptile" in the context, and then as response, generate the word "snakes". We can also observe from the table that the response emotions predicted by our models (fine-tuned on OSED and ED) are reasonable and follow the emotions embedded in the dialog context. Moreover, the generated responses are indeed consistent with the predicted emotions. Note that our model trained on OS has a big chance of predicting the "questioning" category, which is due to the unbalanced distribution in the training set. More samples of the generated responses can be found in Appendix D.

6 Conclusion

In this paper, we curated a large-scale dialog dataset from the OpenSubtitles corpus. Pre-training dialog models on this dataset could largely boost the performance of down-stream empathetic response generation. We proposed an empathetic dialog model capable of learning the dialog emotion interactions at a more fine-grained level, and producing emotionally appropriate responses accordingly. The model was evaluated through a carefully designed human evaluation experiment on the crowdsourcing platform, on a large-scale test set never attempted before. As future work, we would like to improve the accuracy of the response emotion/intent predictor, which we found plays a vital role in generating empathetic responses.

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *Proceedings of ECIR 2018*, volume 10772, pages 154–166.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005*, pages 65–72.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 Task 3: EmoContext contextual emotion detection in text. In *Proceedings of SemEval@NAACL-HLT 2019*, pages 39–48.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of NAACL-HLT 2019*, pages 3734–3743.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *CoRR*, abs/1606.08415.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of LREC 2018*.

Chenyang Huang, Osmar R. Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of NAACL-HLT 2018*, pages 49–54.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Jingyuan Li and Xiao Sun. 2018. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. In *Proceedings of EMNLP 2018*, pages 678–683.

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*, pages 110–119.

Qintong Li, Hongshen Chen, Zhaochun Ren, Zhumin Chen, Zhaopeng Tu, and Jun Ma. 2019. EmpGAN: Multi-resolution interactive empathetic dialogue generation. *CoRR*, abs/1911.08698.

Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. 2020. Empathetic dialogue generation via knowledge enhancing and emotion dependency modeling. *CoRR*, abs/2009.09708.

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*, pages 986–995.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out@ACL 2004*, pages 74–81, Barcelona, Spain.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of EMNLP-IJCNLP 2019*, pages 121–132.

Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation for movie & TV subtitles. In *2016 IEEE Spoken Language Technology Workshop, SLT 2016*, pages 245–252.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of LREC 2018*.

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*, pages 2122–2132.

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.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of AAAI 2018*, pages 5293–5300.

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*, pages 311–318.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL 2019*, pages 527–536.

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*, pages 5370–5381.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3980–3990.

Sashank Santhanam and Samira Shaikh. 2019. Emotional neural language generation grounded in situational contexts. *CoRR*, abs/1911.11161.

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*, pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of ACL 2015*, pages 1577–1586.

Lei Shen and Yang Feng. 2020. CDL: curriculum dual learning for emotion-controllable response generation. In *Proceedings of ACL 2020*, pages 556–566.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *Proceedings of ICASSP 2020*, pages 7989–7993.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of ACL 2019*, pages 3685–3695.

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 NeurIPS 2017*, pages 5998–6008.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Yubo Xie, Ekaterina Svikhnushina, and Pearl Pu. 2019. A multi-turn emotionally engaging dialog model. *CoRR*, abs/1908.07816.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Proceedings of AAAI 2018*, pages 5610–5617.

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*, pages 7492–7500.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of AAAI 2018*, pages 730–739.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of ACL 2018*, pages 1128–1137.

A The Cleaning Procedure of the OpenSubtitles Dialogs

After segmenting the subtitle files in the OpenSubtitles corpus into dialogs, we further clean the dataset with the following steps:

- Remove redundant spaces in the utterances (e.g., spaces at the beginning and the end, and unnecessary spaces between the tokens);
- Remove utterances starting with "previously on ..." (narration at the beginning of TV episodes);
- Remove utterances that simply repeat previous turns;
- Remove utterances that do not start with alphabet, digit, "`'`" (single quote), or "`"`" (double quote);
- For utterances in the form of "character : ...", remove the character information and keep the remaining part;
- Remove utterances with length (number of tokens) less than 2 or greater than 100;
- Remove utterances with percentage of alphabet letters less than 60%;
- Remove utterances with percentage of distinct tokens less than $2/3$;
- Reduce frequency of any utterance to 100.

Whenever we remove an utterance, we discard all the following utterances in the same dialog.

The distribution of emotions/intents of the last utterance in the OSED dataset is plotted in Figure 3.

B Implementation Parameters

Here we summarize some of the parameters of the model implementation:

- We use the RoBERTa tokenizer to tokenize the input utterances, and the vocabulary size is 50,265. We allow a maximum number of 100 tokens as the input to the model.

- We use 4 sub-layers in the encoder and decoder, with 6 heads in the multi-head attention. The dimension of the hidden units is 300, and the dimension of the pointwise feed-forward layers is 1200. We use a dropout rate of 0.1, and the GELU (Hendrycks and Gimpel, 2016) activation function for the hidden layers.

- The loss function is optimized with the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 5×10^{-5}.

- For inference, we use beam search with a beam size of 32. To prevent the models from generating repetitive tokens or n-grams, we modified the beam search algorithm so that at each time step, if any of the branches contains repetitive 4-grams, we set the log probability of this branch to infinitely negative, to stop it from being further expanded.

All the models were trained with a batch size of 512, on machines with 4 Nvidia Titan X Pascal GPUs, 2 Intel Xeon E5-2680 v3 CPUs, and 256GB RAM. Table 8 lists the training details as well as the validation performance for all the models.

C Human Evaluation Setup

The 6,000 test dialogs were split into 600 HITs, with each HIT containing 10 dialogs to be evaluated. We allowed a maximum of 4 workers working on the same HIT, and gave \$0.4 for completing a HIT. When launching the experiment, we only included workers from English speaking countries, i.e., US, AU, NZ, GB, and CA. We also required the workers to have at least 100 approved assignments, and the approval rate is at least 95%. To avoid having the same worker working on too many HITs, we ran a custom script at the backend that constantly checked the worker statistics and blocked the worker if he/she had already finished 50 HITs.

Figure 4 is a screenshot of the welcome page of our human evaluation experiment on the crowdsourcing platform. Figure 5a shows the instructions and explains to the worker how the tasks work, where the worker can also try an example task by dragging and dropping the candidate responses to one of the defined areas, and then validate the answer and get the feedback. Figure 5b is a screenshot of the task page. This task includes a bonus checkpoint, meaning one of the candidate responses is the ground-truth. The worker can click the “Bonus Validation” button to check if he/she has successfully obtained the bonus point.

D More Samples of Model Outputs

Table 9 lists some more samples of the generated responses, with dialog contexts taken from the OS, OSED, and ED datasets.

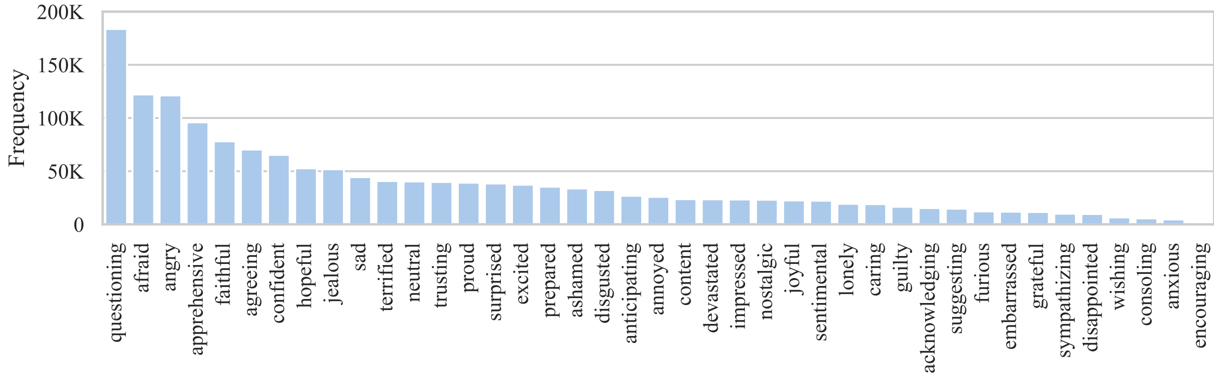


Figure 3: Distribution of emotions/intents of last utterances in the OSED dataset.

Model	# Parameters	# Training Epochs	Training Time	Validation PPL
Pre-trained (OS)	121M	50 epochs	171.00 hr	24.51
Fine-tuned (OSED)	121M	5 epochs	4.23 hr	31.78
Fine-tuned (ED)	121M	9 epochs	19.50 min	21.04
Raw (ED)	121M	55 epochs	1.87 hr	40.56
Ours (OS)	180M	50 epochs	181.38 hr	21.70
Ours (OS $\rightarrow$ OSED)	180M	6 epochs	4.88 hr	28.12
Ours (OS $\rightarrow$ ED)	180M	10 epochs	20.09 min	19.02

Table 8: Training details and validation performance of each model configuration.

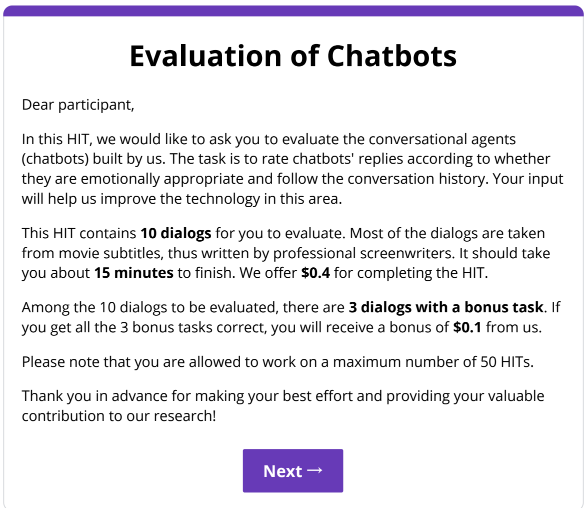
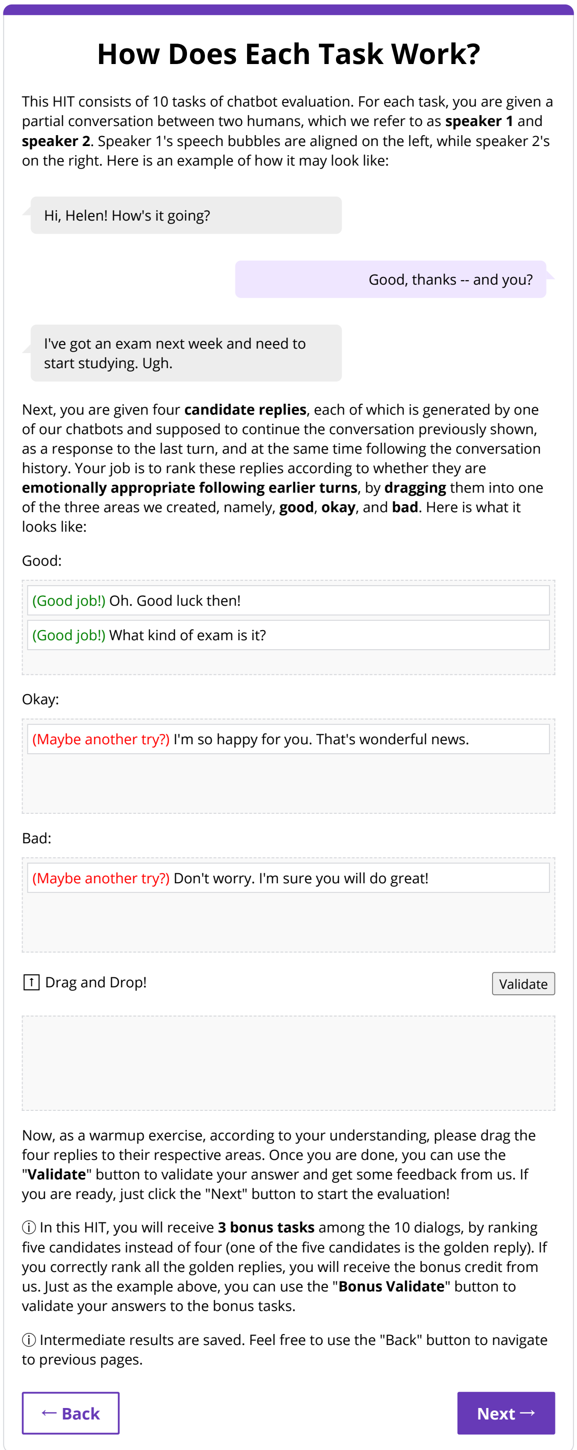
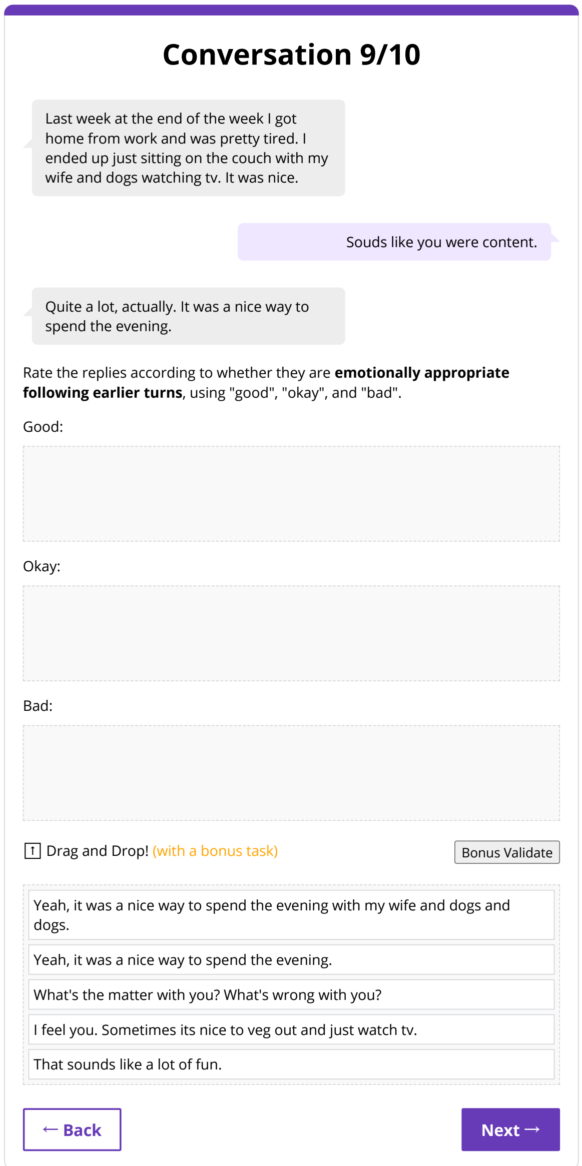


Figure 4: A screenshot of the welcome page of our human evaluation experiment.



(a) A screenshot of the instruction page of our human evaluation experiment.



(b) A screenshot of the task page of our human evaluation experiment. This task includes a bonus checkpoint.

Figure 5: Screenshots of our human evaluation experiment.

Context (OS): - What do you think happened to the lights, huh? **Pre-trained (OS)**: I don 't know. I've never seen anything like it. **Ours (OS) (questioning)**: What do you think happened to the lights? **Ours (OS → OSED) (afraid)**: I don 't know. I can 't see anything. **Ours (OS → ED) (afraid)**: I don't know what happened to the lights. **Ground-truth (agreeing)**: Our waiter told me that the power in Bogota is not very reliable. That's a lie. You and I both know it. They cut the lights on purpose! I know how you feel. And I want her back just as much as you do.
Context (OS): - Oh, this is really your day. - Having trouble with your little Brunhilde, dear? - All right then, tell me about your lady wrestler friend. **Pre-trained (OS)**: I don 't know what you're talking about. **Ours (OS) (questioning)**: What's she like? **Ours (OS → OSED) (jealous)**: I don 't know what she's talking about. I've never seen her before. **Ours (OS → ED) (questioning)**: What did she do to you? **Ground-truth (questioning)**: Who is she? An old flame, darling?
Context (OSED): - I'm telling you. The black kid's fixed the launch pad and we're outta here. Future's looking dim. Soon they won 't have money for soda. Let's go party somewhere else. - Party? Don 't you understand? If God had wanted us to have fun instead of the Ten Commandments he would have sent the Ten Suggestions. - Man, what have they done to you? **Pre-trained (OS)**: What have they done to me? **Ours (OS) (questioning)**: What have they done to you? **Ours (OS → OSED) (faithful)**: We're not going anywhere. We're staying here. **Ours (OS → ED) (neutral)**: I don't know what they have done to me, but I'm glad it worked out for you. **Ground-truth (content)**: Go in peace. I'm staying put.
Context (OSED): - Sinbad, what are you doing? It's just a bull up ahead. - Come on! I already took one of your eyes! I'm your monster, you're going! - Come get me! Come on! **Pre-trained (OS)**: I don 't know what you're talking about. **Ours (OS) (questioning)**: Sinbad, what are you doing? **Ours (OS → OSED) (angry)**: Come on, Sinbad! **Ours (OS → ED) (angry)**: I hate when that happens. **Ground-truth (furious)**: No! The last!
Context (ED): - Sometimes an old song runs through my head and I remember the times where I used to listen to songs so many times that I memorized the words - That sounds nice. It's always good to remember better times. What kind of music? - Like old kind of emo songs, full of angst and whatnot **Pre-trained (OS)**: Whatnot? **Ours (OS) (questioning)**: Don 't you think that's funny? **Ours (OS → OSED) (nostalgic)**: Like old kind of music **Ours (OS → ED) (nostalgic)**: Those are the good old days. **Ground-truth (nostalgic)**: Ha, I used to listen to that stuff back in high school. The My Chemical Romance gets me every time.
Context (ED): - Even though I could've gotten more money at another job, I've stayed with my current employer. It is a good place to work. - well that's good to hear I guess. what do you do? - I work in public relations. I create campaigns for companies. **Pre-trained (OS)**: what do you do? **Ours (OS) (questioning)**: what do they do in public relations? **Ours (OS → OSED) (questioning)**: What are you doing in public affairs? **Ours (OS → ED) (acknowledging)**: that sounds like a great job. **Ground-truth (questioning)**: what's your most successful campaign so far?

Table 9: More samples of the generated responses with dialog contexts taken from the OS, OSED, and ED datasets.