Use of a Taxonomy of Empathetic Response Intents to Control and Interpret Empathy in Neural Chatbots

Anonymous ACL submission

Abstract

A recent trend in the domain of open-domain conversational agents is enabling them to converse empathetically to emotional prompts. Current approaches either follow an end-to-end approach or condition the responses on similar emotion labels to generate empathetic responses. But empathy is a broad concept that refers to the cognitive and emotional reactions of an individual to the observed experiences of another and it is more complex than mere mimicry of emotion. Hence, it requires identifying complex human conversational strategies and dynamics in addition to generic emotions to control and interpret empathetic responding capabilities of chatbots. In this work, we make use of a taxonomy of eight empathetic response intents in addition to generic emotion categories in building a dialogue response generation model capable of generating empathetic responses in a controllable and interpretable manner. It consists of two modules: 1) a response emotion/intent prediction module; and 2) a response generation module. We propose several rule-based and neural approaches to predict the next response's emotion/intent and generate responses conditioned on these predicted emotions/intents. Automatic and human evaluation results emphasize the importance of the use of the taxonomy of empathetic response intents in producing more diverse and empathetically more appropriate responses than end-to-end models.

1 Introduction

End-to-end neural dialogue response generation has revolutionized the design of open-domain conversational agents or chatbots due to requiring little or no manual intervention and its ability to largely generalize (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015). It overcomes many limitations of traditional rule-based response generation techniques such as the cost of domain expertise and predictability of responses. But due to the black-box nature of these end-to-end models, they offer very little controllability to the developer and generate responses that are difficult to interpret (Wu et al., 2018, 2020; Gupta et al., 2020), making these approaches less reliable and fail-safe (d'Avila Garcez and Lamb, 2020). A recent example is Microsoft's Taybot that started producing unintended, and offensive tweets denying the Holocaust as a result of learning from racist and offensive information on Twitter (Lee, 2016). Having control over the generated responses would have enabled the chatbot to avoid malicious intentions and carefully choose how to converse. Thus, it is important to look at ways how developers can gain control over the responses generated by end-to-end neural response generation models and how they can be made interpretable.

Dialogue context:	
Speaker:	*I think that the girl of my dreams likes somebody else. I feel very sad about it.*
Listener:	*Ooh, am so sorry about that. Have you tried to talk to her?*
Speaker:	*It's tough as she has been out of the country for a month, so I will likely discuss it when she returns.*
Possible responses:	
(No control)	*Have you talked to her about it yet?* (Repetitive)
(No control)	*I don't think that's a good idea.* (Not encouraging to the speaker)
(Conditioned on: **Encouraging**)	*I hope everything works out for you.* (Empathetically appropriate)

Table 1: An example dialogue showing how controllability affects response generation.

Recent research has taken efforts to induce controllability and interpretability into end-to-end models. For example, Xu et al. (2018) explore how the flow of human-machine interactions can be managed by introducing dialogue acts as policies to the dialogue generation model. Sankar and Ravi (2019) show that conditioning the response generation process on interpretable dialogue attributes such as dialogue acts and sentiment helps to eliminate repetitive responses and makes the model more interesting and engaging.

In contrast to task-oriented dialogue systems designed to help people complete specific tasks, open-domain chatbots are designed to engage users in human-machine conversation for entertainment and emotional companionship (Wu and Yan, 2018). Hence, in open-domain conversations, controllability should also be studied with respect to aspects such as humor, personality, emotions, and empathy, which cannot be achieved using generic dialogue acts. In this study, our focus is on controlling empathy in open-domain chatbot responses, which requires understanding conversational strategies used in human-human empathetic conversations.

Earlier studies gain control in this aspect by conditioning the response on either manually specified (Zhou et al., 2018; Zhou and Wang, 2018; Hu et al., 2018; Song et al., 2019) or automatically predicted (Chen et al., 2019) sentiment or emotion labels. However, an analysis by Welivita and Pu (2020) on human-human conversations of the EmpatheticDialogues dataset (Rashkin et al., 2018) reveals, listeners are much more likely to respond to positive or negative emotions with specific empathetic intents such as *acknowledgment*, *consolation* and *encouragement*, rather than expressing similar or opposite emotions. They introduce a taxonomy of eight response intents that can better describe empathetic human responses to emotional dialogue prompts. In this paper, we explore how end-to-end response generation can be combined with more advanced control of empathy by utilizing the above taxonomy of empathetic response intents in addition to existing emotion categories. To provide a glimpse of what we aim to achieve, in Table 1 we show how conditioning the response on an empathetic response intent chosen based on the dialogue history can serve in producing a more empathetically appropriate response. It avoids repetitive or sub-optimal responses generated by end-to-end approaches without any control.

Our empathetic response generation model consists of two modules: 1) a response emotion or intent prediction module; and 2) a response generation module. We experiment with both rule-based and neural approaches for predicting the next response's emotion or intent. For the rule-based approaches for predicting the response emotion/intent, we develop two decision tree-based response emotion and intent prediction methods. For the neural approach for predicting the response emotion/intent, we develop a classifier based on the BERT transformer architecture (Vaswani et al., 2017; Devlin et al., 2019). The reason why we evaluate the performance of rule-based approaches is that they are much simpler than neural models and save a lot of training time and resources. Thus, if considerable performance can still be achieved through rule-based approaches compared to the baselines, it is worth considering the use of such simpler approaches over sophisticated neural approaches, especially in resource-limited environments. The emotions and intents predicted by these methods are then used to condition the responses generated by the response generation module. For training and evaluating these models, we use two state-of-the-art dialogue datasets containing empathetic conversations: 1) the EmpatheticDialogues dataset (Rashkin et al., 2018); and 2) the EDOS (Emotional Dialogues in OpenSubtitles) dataset (Welivita et al., 2021). The automatic and human evaluation results confirm the importance of the use of the taxonomy in generating more diverse and empathetically more appropriate responses than end-to-end models.

Our contributions in this paper are three folds. 1) We explore the ability of a taxonomy of empathetic response intents in controlling and interpreting the responses generated by open-domain conversational agents for emotional prompts. 2) We propose an empathetic response generation model consisting of a response emotion/intent prediction module and a response generation module to generate empathetic responses in a controllable and interpretable manner. 3) We experiment with both rule-based and neural approaches in predicting the next response's emotion or intent and evaluate their performance in conditional generation of empathetic responses using automatic and human evaluation metrics against standard baselines.

2 Literature Review

Existing conversational agents are designed for either open-domain or specific task completion (Gao et al., 2018). Regarding the former, a common practice is to generate dialogue in an end-to-end fashion (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015). Often responses generated by these methods are unpredictable and not fail-safe (d'Avila Garcez and Lamb, 2020). Hence, recent research has focused on methods to control and interpret the responses generated by open-domain neural conversational agents. Mainly we

find three methods they use to control the generated response: 1) by a manually specified value (Zhou et al., 2018; Zhou and Wang, 2018; Hu et al., 2018; Song et al., 2019); 3) by rules that are predefined or derived from the training data (Hedayatnia et al., 2020); 3) by an automatically predicted value from a neural network model (Xie and Pu, 2021; Wu et al., 2018; Sankar and Ravi, 2019; Santhanam et al., 2020; Ke et al., 2018; Lee et al., 2020).

Specially in studies addressing emotional response generation, a manually specified sentiment, emotion (Zhou et al., 2018) or an emoji (Zhou and Wang, 2018) was used to control the sentiment or emotionality of the responses generated. Later, more and more research focused on automatically predicting values or deriving rules such that they could be used to control the generated response without manual intervention. For example, Sankar and Ravi (2019) used an RNN based policy network to predict the next dialogue act given previous dialogue turns and dialogue attributes. Hedayatnia et al. (2020) used rules designed as a set of dialogue act transitions from common examples in the Topical-Chat corpus (Gopalakrishnan et al., 2019) to plan the content and style of target responses.

But all the above work focused on achieving controllability using generic dialogue acts or generating controlled emotional responses conditioned on similar or opposite emotions, emojis, or sentiment tags. These labels do not suffice the controlled generation of meaningful empathetic responses because humans demonstrate a wide range of emotions and intents when regulating empathy (Welivita and Pu, 2020). Previous work also lacks comparisons between rule-based and automatic conditioning methods used to control response generation. In this work, we address the above gaps by investigating how empathy in neural responses can be controlled using a taxonomy of eight empathetic response intents (Welivita and Pu, 2020), in addition to 32 emotion categories, while evaluating the applicability of both rule-based and automatic control mechanisms for this task.

3 Methodology

Our controllable and interpretable empathetic response generation architecture consists of two modules: 1) the response emotion/intent prediction module; and 2) the response generation module. The emotion or intent predicted by the first module is input into the second to condition the response generated by the second module. In the following sections we discuss the datasets used for our experiments, the different rule-based and automatic emotion/intent prediction methods we propose, how the emotions and intents predicted by these modules are used to generate responses that are both controllable and interpretable, and the different evaluation methods we utilize to compare the performance of these approaches on two state-of-the-art dialogue datasets containing emotional dialogue prompts.

3.1 Datasets

We utilized the EmpatheticDialogues dataset proposed by Rashkin et al. (2018), and the OS (OpenSubtitles) and EDOS (Emotional Dialogues in OpenSubtitles) dialogue datasets proposed by Welivita et al. (2021) to train and evaluate our models. The EmpatheticDialogues (ED) dataset contains ≈25K open-domain human-human conversations carried out between a speaker and a listener. Each conversation is conditioned on one of 32 emotions selected from multiple annotation schemes. The OS and EDOS datasets are curated by applying a series of preprocessing and turn segmentation steps on the movie and TV subtitles in the OpenSubtitles 2018 corpus (Lison et al., 2019). The EDOS dataset contains 1M highly emotional dialogues filtered from the rest of the OS dialogues. Even though the speaker and listener turns in the OS and EDOS datasets are not clearly defined, we assumed the odd-numbered turns (1, 3, 5, ...) as speaker turns and even-numbered turns (2, 4, 6, ...) as listener turns for our experiments. We used the OS dialogues dataset containing ≈3M dialogues for pre-training and the ED and EDOS datasets to separately fine-tune the models. The statistics of these datasets are denoted in Table 2. From each dataset, 80% of the data was used for training, 10% for validation, and the remaining 10% for testing.

Dataset	Dialogues	Turns	Turns/dialogue
OS	2,989,774	11,511,060	3.85
ED	24,847	107,217	4.32
EDOS	1,000,000	2,940,629	2.94

Table 2: Statistics of the datasets used for training and evaluating the models.

We used a BERT (Devlin et al., 2019) transformer-based dialogue emotion classifier proposed by Welivita et al. (2020) to automatically annotate all dialogue turns in the above datasets. This classifier is trained on 25K situation descrip-

Empathetic intent	Example response
1. Questioning	*What's the matter? What's wrong?*
2. Agreeing	*Exactly, I get that entirely!*
3. Acknowledging	*Sounds awesome!*
4. Encouraging	*Just give it a trial.*
5. Consoling	*I hope everything works out for you.*
6. Sympathizing	*I am sorry to hear that.*
7. Wishing	*Congrats, that's a step forward.*
8. Suggesting	*Maybe you should talk to her.*

Table 3: The taxonomy of listener specific empathetic response intents used to achieve controllability and interpretability in the responses generated.

tions from EmpatheticDialogues labeled with 32 emotion classes, 7K EmpatheticDialogues listener turns labeled with eight empathetic response intents and *Neutral*, and 14K emotion and intent annotated dialogue turns from the OSED dataset. It has a final annotation accuracy of 65.88% over 41 labels, which is significant compared to the other state-of-the-art dialogue emotion classifiers (Welivita et al., 2020). We use the emotion and intent labels suggested by the above classifier as ground-truth labels for our experiments.

3.2 Response Emotion/Intent Prediction

To generate controlled and interpretable empathetic responses, we utilized 32 fine-grained emotions and a taxonomy of listener-specific empathetic response intents. The 32 emotions are emotion categories on which dialogues in the EmpatheticDialogues dataset are conditioned on (Rashkin et al., 2018). They range from basic emotions derived from biological responses (Ekman, 1992; Plutchik, 1984) to larger sets of subtle emotions derived from contextual situations (Skerry and Saxe, 2015). We further utilized the taxonomy of empathetic response intents proposed by Welivita and Pu (2020), which is derived by analysing the listener responses in the EmpatheticDialogues dataset. These intents are denoted in Table 3 along with corresponding examples. To predict the emotion or intent of the next response, we propose several rule-based and neural response emotion/intent prediction methods, which are described in the following subsections.

3.2.1 Baselines

As a baseline, we sample a response emotion or intent from the set of eight empathetic response intents plus the most recent emotion encountered in the last $k(k = 3)$ dialogue turns. This is based on the observations by Welivita and Pu (2020) on the EmpatheticDialogues dataset (Rashkin et al., 2018). They state that in human empathetic conversations, the listener's response to emotional prompts mostly contain an empathetic response intent identified by their taxonomy or a statement with similar emotion. This baseline is inspired by the work of Hedayatnia et al. (2020), in which the response dialogue act is chosen among the most frequently seen dialogue acts based on an equal probability distribution.

The other baseline we used when generating responses is the plain end-to-end transformer model proposed by Vinyals et al. (2017), in which no conditioning is used when generating the response.

3.2.2 Rule-based Decision Tree Approaches

We propose two non-neural, decision tree-based response intent prediction methods that leverage the knowledge of the emotion-intent flow of the dialogues in the training dataset. The basic idea of a decision tree for this context is denoted along with an example in Figure 1. The probabilities of emotions and intents in the branches in the decision tree are learned from the training data itself by traversing through dialogues using a window of size k, where k is the maximum depth of the decision tree. The window is moved forward two dialogue turns at a time capturing the probability of speaker-listener emotion-intent exchanges in the training dataset.

Here, we used a window of size 4 mainly because most dialogues contained in the ED, OS, and EDOS datasets were limited to four dialogue turns. During inference, an emotion or an intent is sampled based on the sequence of emotions and intents in the previous $(k-1)$ dialogue turns. We used two different methods: 1) argmax; and 2) probabilistic sampling, to sample the response emotion or intent from the decision tree. In the argmax method, we chose the emotion or intent with the highest probability in the decision tree based on the sequence of emotions and intents in the previous $(k-1)$ dialogue turns. In the probabilistic sampling method, we sampled an emotion or an intent based on the distribution of probabilities in the decision tree given the sequence of emotions and intents in the previous $(k-1)$ dialogue turns. We refer to these two decision tree-based methods as *DT (argmax)*, and *DT (prob. sampled)*.

We have more control over the above methods than neural response intent prediction methods since we can foresee where the dialogue will be directed by visualizing the decision trees beforehand. For example, the decision trees generated using the EmpatheticDialogues and EDOS training

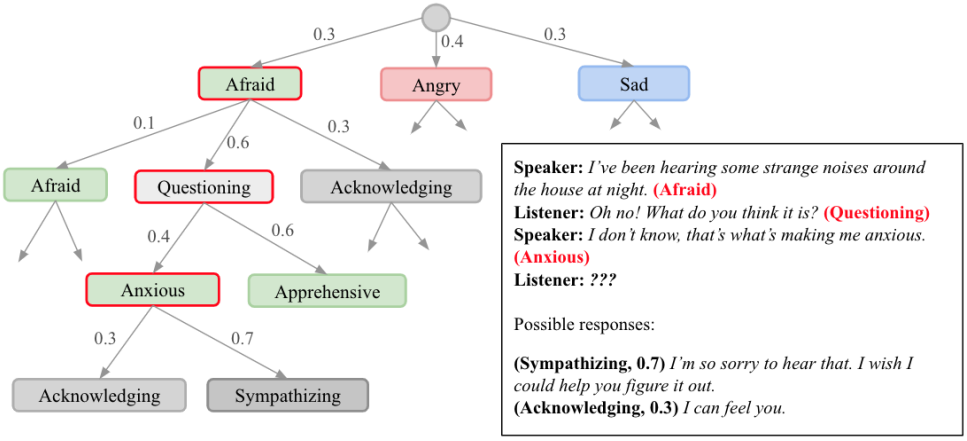
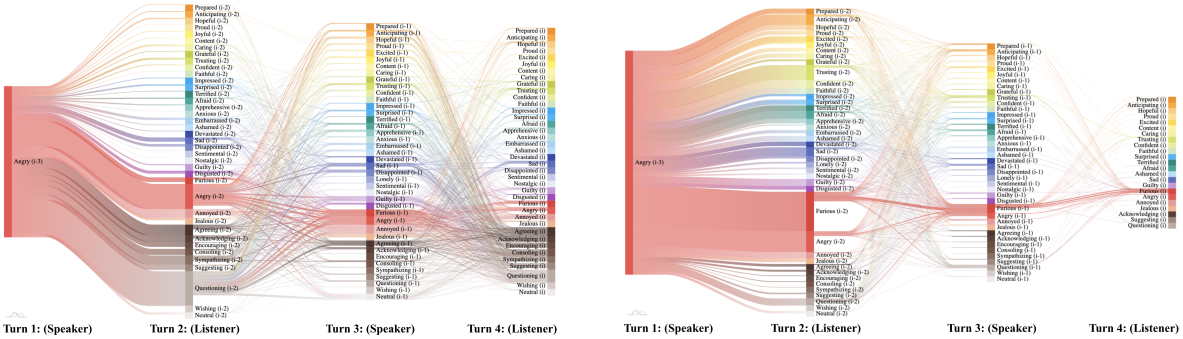


Figure 1: Visualization of a simpler version of our decision tree approach to predict the response emotion or intent.



(a) Dataset: EmpatheticDialogues (b) Dataset: EDOS

Figure 2: Decision trees generated using the EmpatheticDialogues and EDOS training datasets when the emotion of the beginning dialogue prompt is *Angry*.

datasets when the emotion of the beginning dialogue prompts is *Angry* are denoted in Figure 2. As it could be observed, in the ED dataset, the listeners mostly respond to speakers' emotions with one of the intents from the taxonomy of empathetic response intents. The EDOS dataset by nature is more dramatic, in which both the speaker and the listener become emotional. This phenomenon is called "emotional contagion" in the psychological literature (Hatfield et al., 1993). For example in EDOS, if the speaker is angry, the listener also tends to reply back with anger. These communication patterns could clearly be visualized with the decision trees created and the developer can predict beforehand how the chatbots whose responses are conditioned on these emotion-intent patterns would behave for a given emotional prompt.

3.2.3 Neural Response Emotion and Intent Predictor

An automatic method for predicting the next response's emotion or intent is using a neural network-based response emotion/intent predictor. An advantage of using neural approaches to determine the emotion or intent of the next response is that they can leverage clues from the semantic content of the previous dialogue turns in addition to the flow of emotions and intents when predicting the response emotion or intent. Our neural response emotion/intent predictor consists of a BERT transformer-based encoder architecture (representation network) followed by an attention layer for aggregating individual token representations, a hidden layer, and a softmax as depicted in Figure 3. The BERT-base architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters is used as the representation network. It is initialized with weights from the pre-trained language model RoBERTa (Liu et al., 2019).

We concatenate the previous k dialogue turns as depicted in Figure 3 and they are input to the encoder of the model. The emotions and intents corresponding to these k dialogue turns are added to the word embeddings and positional embeddings

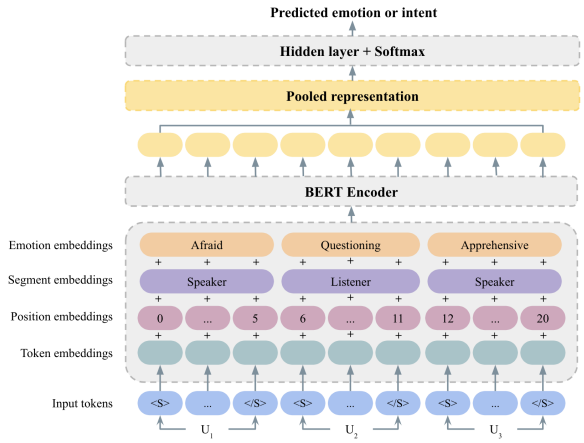


Figure 3: Architecture of the neural response emotion/intent predictor.

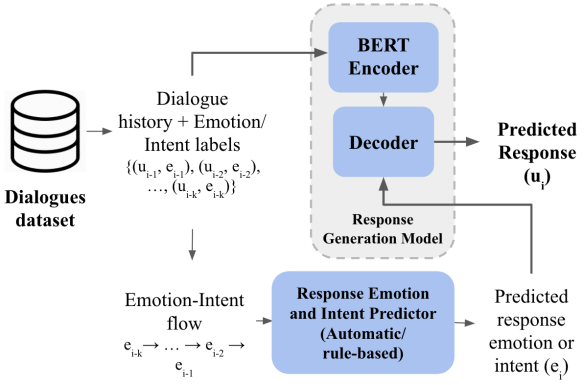


Figure 4: Overall architecture of the controllable and interpretable empathetic response generation model.

in the original transformer architecture. This additional knowledge helps the model to get a better understanding of the flow of emotions and intents in the previous dialogue turns. The emotions and intents are embedded into a vector space having the same dimensionality as the word and position embeddings so they can add up. In addition, we also incorporate segment embeddings that differentiate between speaker and listener turns. We pre-trained the model on the OS dialogues dataset and fine-tuned it separately on ED and EDOS datasets. The hyper-parameters used during training and other training details are described in the appendices.

3.3 Response Generation

For response generation, we used a plain transformer-based encoder-decoder architecture (end-to-end model) as a baseline (Vaswani et al., 2017). To generate controlled empathetic responses, we incorporated the different response emotion/intent prediction methods described above as input to the decoder. Figure 4 shows the overall architecture of our models.

The input representation for the encoder of the generation model is the same as the input representation used for the neural response emotion/intent predictor described in section 3.2.3. The vector representation generated by the encoder is input into the decoder along with the embedding of the emotion or intent predicted by the response emotion/intent predictor. During training, instead of the predicted emotion or intent, we used the ground-truth emotion or intent. The generation model is first pre-trained on OS dialogues and then fine-tuned on ED and EDOS datasets separately.

4 Evaluation and Results

4.1 Automatic Evaluation Results

Evaluation by means of automatic metrics was carried out separately for response emotion/intent prediction and conditional response generation. The following subsections describe the results obtained in these evaluations.

4.1.1 Prediction Performance

The weighted precision, recall, F1, and balanced accuracy scores computed for different response emotion/intent prediction methods across ED and EDOS testing datasets are indicated in Table 4.

According to the weighted precision, recall, F1, and accuracy scores, the neural emotion/intent predictor performed the best compared to other prediction methods. Among rule-based approaches for response emotion/intent prediction, the DT (argmax) method performed the best. The DT (argmax) method had considerable improvement in recall, F1, and accuracy scores over the equally sampled baseline.

4.1.2 Generation Performance

To evaluate the performance of response generation, we computed the perplexity, diversity metrics (distinct unigram and distinct bigram scores), and vector extrema cosine similarity on ED and EDOS testing datasets. They are denoted in Table 5. We also evaluated the responses generated by a model conditioned on the ground-truth emotion or intent of the next response to see how well the taxonomy of empathetic response intents alone contributes to better empathetic response generation performance.

According to the results, the models whose response was conditioned on the ground-truth re-

Model	Trained on: OS + ED Tested on: ED				Trained on: OS + EDOS Tested on: EDOS			
	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.
Equally sampled	0.1138	0.0667	0.0638	0.0410	0.0981	0.0221	0.0232	0.0285
DT (argmax)	0.0959	0.0883	0.0883	0.0692	0.0755	0.1016	0.0799	0.0419
DT (prob. sampled)	0.0715	0.0663	0.0680	0.0480	0.0627	0.0616	0.0619	0.0345
Neural predictor	0.1634	0.1636	0.1472	0.1163	0.1306	0.1712	0.1181	0.0679

Table 4: Weighted precision, recall, F1 and accuracy scores computed for ED and EDOS test datasets. The cells in dark green indicate the best scores and the cells in light green indicate the second best scores.

Model	Trained on: OS + ED Tested on: ED				Trained on: OS + EDOS Tested on: EDOS			
	PPL	D-1	D-2	Embed. extrema	PPL	D-1	D-2	Embed. extrema
GT emotion/intent	11.74	0.0823	0.2812	0.5181	12.57	0.0846	0.2552	0.4539
End-to-end model	12.26	0.0544	0.1612	0.5015	13.13	0.0784	0.228	0.4365
Equally sampled	13.48	0.0761	0.2469	0.4824	14.20	0.0754	0.2229	0.433
DT (argmax)	13.23	0.0865	0.2977	0.4892	14.14	0.0727	0.2419	0.4458
DT (prob. sampled)	13.37	0.0795	0.2761	0.4828	14.23	0.0763	0.2418	0.436
Neural predictor	13.15	0.0835	0.2811	0.4851	13.97	0.0805	0.2415	0.4403

Table 5: Perplexity (PPL), diversity metrics (distinct unigrams: D-1; and distinct bigrams: D-2), and vector extrema cosine similarity (Embed. extrema) calculated on ED and EDOS testing datasets.

sponse emotion or intent performed the best in terms of perplexity and embedding extrema in both ED and EDOS datasets and in terms of diversity metrics in the EDOS dataset. These results emphasize the usefulness of the taxonomy of empathetic response intents and the 32 fine-grained emotion categories in generating controlled empathetic responses. The models incorporating the DT (argmax) approach scored the best in terms of diversity metrics in the ED test dataset.

4.2 Human Evaluation

In addition to the automatic metrics, we carefully designed a human evaluation experiment in Amazon Mechanical Turk (AMT) to evaluate responses' empathetic appropriateness. We selected a total of 1,000 dialogue cases: 500 ED and EDOS dialogues for testing. The AMT workers had to drag and drop responses generated by five models (end-to-end; models whose response was conditioned on the equally sampled baseline, DT argmax, DT prob. sampled and the neural predictor) into areas *Good*, *Okay*, and *Bad*, depending on their empathetic appropriateness. We neglected responses conditioned on the ground-truth emotion or intent since we are more interested in automatically predicted labels. We bundled 10 dialogues into a HIT (Human Intelligence Task) so that one worker works on at least 10 cases to avoid too much bias between answers. To evaluate the quality of the work generated, we included three quiz questions equally spaced in a HIT. In these, we included the ground-truth response among the other responses generated by the models. If a worker rated the ground-truth response either as *Good* or *Okay*, then a bonus point was added. To encourage attentiveness to the task, for those who obtained at least two out of three quiz questions correct, we gave a bonus of 0.1$. Three workers were allowed to work on a HIT and only the ratings that were agreed by at least two workers, both who have obtained bonuses, were taken to compute the final scores. As a result, 8.33% of the answers were disqualified. The results of the experiment are denoted in Table 6. The experiment yielded an inter-rater agreement (Fleiss' kappa) score of 0.2294 indicating fair agreement.

According to the results, the neural predictor scored the highest percentage of *Good* ratings in both ED and EDOS testing datasets. The models that use the equally sampled approach performed the worst producing the highest percentage of responses ranked *Bad*. An interesting observation is that the DT (argmax) method scored the most number of combined *Good* and *Okay* responses in ED and EDOS testing datasets confirming that rule-baled approaches such as the decision tree approach we propose could be used to control and interpret the responses without losing significant accuracy.

4.3 Case Study

In Table 7 we show some example responses generated by different models for a given dialogue context. It could be noticed that having no response control mechanism and having a response conditioned on an inappropriate intent both can result

Model	Trained on: OS + ED Tested on: ED				Trained on: OS + EDOS Tested on: EDOS			
	Good	Okay	(Good + Okay)	Bad	Good	Okay	(Good + Okay)	Bad
End-to-end model	46.94	15.72	62.66	37.34	27.48	27.02	54.50	45.50
Equally sampled	25.22	25.22	50.44	49.57	19.35	29.84	49.18	50.82
DT (argmax)	37.61	33.70	71.31	28.69	21.70	36.72	58.42	41.58
DT (prob. sampled)	24.17	21.68	45.85	42.13	21.96	31.78	53.74	46.26
Neural predictor	50.00	17.39	67.39	32.61	24.82	33.01	57.83	42.17

Table 6: Human evaluation results (as a percentage) corresponding to ED and EDOS testing datasets.

Dialogue context:

Speaker: *Last week when Toys R Us closed it really made me start thinking of the 90s and my childhood* **(Nostalgic)**
Listener: *I've heard a lot of people say something similar. Did you go there a lot?* **(Questioning)**
Speaker: *I did! I was talking to my husband about the different toys I had. Even started talking about the music and shows from that era.* **(Nostalgic)**

Model	Predicted emotion/intent	Generated response	Human eval.
Ground-truth emotion/intent	Jealous	*That sounds like a lot of fun. I wish I had a husband like that.*	Good
End-to-end model	-	*I'm sorry to hear that. What kind of toys were you talking about?*	Bad
Equally sampled	Wishing	*I wish you the best of luck!*	Bad
DT (argmax)	Questioning	*What kind of toys were you talking about?*	Good
DT (prob. sampled)	Encouraging	*I bet that was fun*	Good
Neural predictor	Encouraging	*I bet that was fun*	Good

Table 7: An example dialogue showing that both lack of controllability and conditioning the response on an inappropriate emotion or intent can lead to responses that are empathetically inappropriate with the dialogue context.

in responses that are empathetically inappropriate with the dialogue context. The neural predictor, as well as the decision tree-based mechanisms, generate some emotion or intent that is appropriate to the dialogue context, enabling the generation model to generate responses that are more empathetically appropriate, guiding the conversation in a meaningful direction.

5 Discussion and Conclusion

This study investigated the use of a taxonomy of empathetic response intents along with 32 fine-grained emotions in controlling and interpreting the responses generated by open-domain conversational agents for emotional prompts. In this regard, several rule-based and automatic response control methods were proposed and were compared in terms of their prediction and generation performance on two state-of-the-art dialogue datasets containing emotional dialogues.

It was observed that the neural response emotion/intent predictor we proposed outperformed the rest including the end-to-end model in terms of evaluation metrics related to both prediction and generation. This implies the importance of leveraging semantic clues in addition to the flow of emotions and intents in the previous turns when predicting the next response's emotion or intent. However, there are some disadvantages to using this approach: 1) developers cannot foresee the label that the model would predict next; and 2) cost of time and resources spent for training the model. As a remedy, we proposed two decision tree-based response emotion/intent prediction approaches.

Across evaluation metrics for prediction and generation, the performance of the decision-tree methods was considerably better than the end-to-end approach and the equally sampled baseline. The decision tree (argmax) method performed the best in terms of diversity metrics related to response generation. In the human evaluation stage, we saw that the DT (argmax) method produced the most number of combined *Good* and *Okay* responses in ED and EDOS test datasets, pointing to the fact that the rule-based approaches we proposed can still be used without a significant degrade in performance in resource-limited environments.

On the whole, the results of this study inform developers about the utility of the taxonomy of empathetic response intents in controlling the responses generated by open-domain chatbots and which optimal methodology to use (rule-based or automatic conditioning) based on the operational environment.

References

Zhongxia Chen, Ruihua Song, Xing Xie, Jian-Yun Nie, Xiting Wang, Fuzheng Zhang, and Enhong Chen. 2019. Neural response generation with relevant emotions for short text conversation. In *CCF International Conference on Natural Language Pro-*

cessing and Chinese Computing, pages 117–129. Springer.

Artur d'Avila Garcez and Luis C. Lamb. 2020. Neurosymbolic ai: The 3rd wave.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH*.

Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and Amy Pavel. 2020. Controlling dialogue generation with semantic exemplars. *arXiv preprint arXiv:2008.09075*.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

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 the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating informative responses with controlled sentence function. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1499–1508, Melbourne, Australia. Association for Computational Linguistics.

Dave Lee. 2016. Tay: Microsoft issues apology over racist chatbot fiasco.

Hung-yi Lee, Cheng-Hao Ho, Chien-Fu Lin, Chiung-Chih Chang, Chih-Wei Lee, Yau-Shian Wang, Tsung-Yuan Hsu, and Kuan-Yu Chen. 2020. Investigation of sentiment controllable chatbot. *arXiv preprint arXiv:2007.07196*.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

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. *arXiv preprint arXiv:1907.11692*.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.

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

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Stockholm, Sweden. Association for Computational Linguistics.

Sashank Santhanam, Zhuo Cheng, Brodie Mather, Bonnie Dorr, Archna Bhatia, Bryanna Hebenstreit, Alan Zemel, Adam Dalton, Tomek Strzalkowski, and Samira Shaikh. 2020. Learning to plan and realize separately for open-ended dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2736–2750, Online. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Amy E Skerry and Rebecca Saxe. 2015. Neural representations of emotion are organized around abstract event features. *Current biology*, 25(15):1945–1954.

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

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 the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2020. Fine-grained emotion and intent learning in movie dialogues.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Wei Wu, Can Xu, Yu Wu, and Zhoujun Li. 2018. Towards interpretable chit-chat: Open domain dialogue generation with dialogue acts.

Wei Wu and Rui Yan. 2018. Deep chit-chat: Deep learning for ChatBots. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Melbourne, Australia. Association for Computational Linguistics.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2020. A controllable model of grounded response generation. *arXiv preprint arXiv:2005.00613*.

Yubo Xie and Pearl Pu. 2021. Generating empathetic responses with a large scale dialog dataset. In *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

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 the AAAI Conference on Artificial Intelligence*, volume 32.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.