I KNOW THE FEELING: LEARNING TO CONVERSE WITH EMPATHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Beyond understanding what is being discussed, human communication requires an awareness of what someone is feeling. One challenge for dialogue agents is being able to recognize feelings in the conversation partner and reply accordingly, a key communicative skill that is trivial for humans. Research in this area is made difficult by the paucity of large-scale publicly available datasets both for emotion and relevant dialogues. This work proposes a new task for empathetic dialogue generation and EMPATHETICDIALOGUES, a dataset of 25k conversations grounded in emotional contexts to facilitate training and evaluating dialogue systems. Our experiments indicate that models explicitly leveraging emotion predictions from previous utterances are perceived to be more empathetic by human evaluators, while improving on other metrics as well (e.g. perceived relevance of responses, BLEU scores).

1 INTRODUCTION

Natural communication is frequently prompted by people sharing their feelings or circumstances. As examples, a recent study found that 80% of Twitter users seem to post mostly about themselves (Naaman et al., 2010), and ELIZA (Weizenbaum, 1966), one of the earliest chatbots developed, focused most of its attention on asking its conversational partners why they were feeling a certain way. Interacting in these conversations requires reacting to what people share with an understanding of others' implied feelings. For instance, while the crossed-out response in Figure 1 is contextually relevant, "Congrats! That's great!" is more natural because it acknowledges the underlying feelings of accomplishment.

As in this example, people generally respond to others in a way that is empathetic or that acknowledges how the other person feels. But for dialogue agents, this is still a non-trivial communicative skill. Although recent work has used large-scale corpora to train reasonably fluent and engaging dialogue agents (e.g. Mazaré et al. (2018)), existing chitchat dialogue benchmarks are not designed to capture whether those agents are responding in an empathetic way to implicit emotional contexts. To better evaluate machines' ability to do so, we introduce a new task for dialogue systems to respond to people discussing everyday situations.

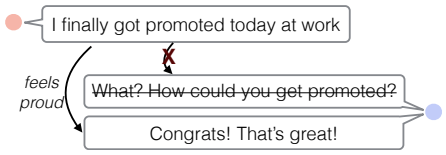

Figure 1: Example conversation where acknowledging an inferred feeling might be appropriate

As a resource for this task, we introduce the EMPATHETICDIALOGUES, a novel dataset with 25k personal dialogues. Each dialogue is grounded in a specific emotional context that a speaker is feeling with a listener responding. The dataset is larger and contains a more extensive set of emotions than many similar emotion prediction datasets fro other text domains such as Scherer & Wallbott (1994), Strapparava & Mihalcea (2007), Mohammad et al. (2018), and Gupta et al. (2017). Previous dialogue datasets that include emotion labels (Li et al., 2017; Gupta et al., 2017) come from crawled conversations that have been labeled post-hoc and therefore have labels that are unevenly distributed or uninformative (only $\approx 5\%$ of the DailyDialog utterances have a label other than 'none' or 'happy'), or come from a much more limited set (e.g., happy, sad and angry in Gupta et al. (2017)).

Table 1: Two examples from EMPATHETICDIALOGUES training set. The first worker (the speaker) is given an emotion label and writes their own prompt based on a situation when they've felt that way. Then, the speaker tell their story in a conversation with a second worker (the listener).

Label: Afraid **Situation:** Speaker felt this when... "I've been hearing noises around the house at night" **Conversation:** Speaker: I've been hearing some strange noises around the house at night. Listener: oh no! That's scary! What do you think it is? Speaker: I don't know, that's what's making me anxious. Listener: I'm sorry to hear that. I wish I could help you figure it out	**Label: Proud** **Situation:** Speaker felt this when... "I finally got that promotion at work! I have tried so hard for so long to get it!" **Conversation:** Speaker: I finally got promoted today at work! Listener: Congrats! That's great! Speaker: Thank you! I've been trying to get it for a while now! Listener: That is quite an accomplishment and you should be proud!

We then examine how to train a dialogue system that is more adept at responding to emotional contexts. While a rule-based system can be built around mapping predicted emotions to responses, end-to-end dialogue systems relying on neural networks and trained on conversation corpora (Shang et al., 2015; Vinyals & Le, 2015; Sordoni et al., 2015; Serban et al., 2015; Dodge et al., 2016; Mazaré et al., 2018; Zhang et al., 2018) offer the promise of better generalization to new contexts. Through an extensive set of experiments, we show that fine-tuning a dialogue agent on our dataset results in better performance on a novel empathetic dialogue task, and that the model can perform even better when supplemented with representations trained on large-scale emotion data.

The contributions of this work are thus threefold: 1) we release a novel empathetic dialogue dataset as a new benchmark 2) we show that using this dataset for multi-task training can improve the performance of an end-to-end dialogue system on empathetic dialogue; 3) we show that performance can be further improved by combining this model with a representation trained on a supervised emotion recognition task.

2 RELATED WORK

Responding well to emotions requires sufficient coverage of human expression. Multiple schema have attempted to organize the spectrum of emotions, from a handful of basic emotions derived from biological responses (Ekman, 1992; Plutchik, 1984) to larger sets of subtle emotions inferred from contextual situations (Skerry & Saxe, 2015). We incorporate emotions from multiple annotation schema, noting that emotions merely inferred from a situation are important in dialogue scenarios.

Rich information can be represented by learning multidimensional distributional embeddings from data, as has proven successful for many language applications (Grave et al., 2018). These distributional representation approaches are at the core of the current state of the art in emotion classification (Duppada et al., 2018; Park et al., 2018; Xu et al., 2018; Mohammad et al., 2018) that build on deep networks pre-trained on weakly labelled data such as emojis (Felbo et al., 2017) or hashtags (Mohammad, 2012), whereas we train models for emotion detection on conversation data that has been explicitly labeled by annotators. The SEMEVAL2019 EmoContext challenge also uses conversation data for detection of three basic emotions over two turns of context (Gupta et al., 2017). We experiment with models that directly leverage similar emotion detection for dialogue including a broader coverage of emotions with a focus on personal situations.

Several works have attempted to make chit-chat dialogue models more engaging by grounding them in personal contexts (Li et al., 2016; Zhang et al., 2018; Mazaré et al., 2018), but focusing on personal facts ("I am from New York") rather than situations. The DAILYDIALOG dataset (Li et al., 2017), comprising about 13k crawled dialogues, includes *post-hoc* emotion labels, but only $\approx 5\%$ of the utterances have a label other than "none" or "happy". Our task focuses more explicitly on incorporating emotional context in conversations, with benchmarks considering a richer, evenly

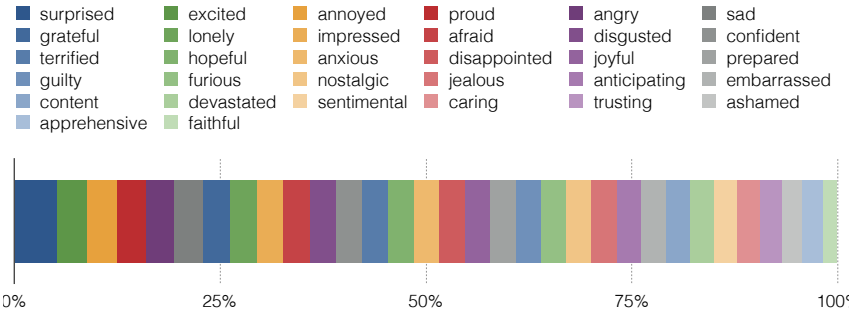

Figure 2: Distribution of situation/conversation labels within EMPATHETICDIALOGUES. Exact percentages per class are listed in the appendix.

distributed set of emotions. We also introduce the concept of an explicit *listener* in the conversation who is reacting to the situation being described.

A few works focus on controlling the emotion content of a text response either through a manually specified target (Zhou & Wang, 2018; Zhou et al., 2017; Wang & Wan, 2018; Hu et al., 2017; Huang et al., 2018) or through a general term to encourage higher levels of affect (Asghar et al., 2018), with evaluations focused on matching a predetermined desired emotion. Huber et al. (2018) investigate how to respond to emotions detected from an image. In the remainder of the paper, we examine how to produce empathetic responses that are appropriate to an emotional signal inferred purely from text, rather than intended to themselves convey a manually specified emotion.

3 TALKING ABOUT PERSONAL SITUATIONS

We consider an open-domain conversational setting where two people are discussing a situation that happened to one of them and that led to the experience of a given feeling.

Emotional situation grounding Each conversation is grounded in a situation written about by one participant in association with a given emotion label. We consider 32 emotion labels, listed in Figure 2. For selecting this set of labels, we drew inspiration from previous datasets (Scherer & Wallbott, 1994; Strapparava & Mihalcea, 2007; Skerry & Saxe, 2015; Li et al., 2017; Mohammad, 2012), consolidating the labels from each into a merged list.

Speaker and Listener The person who wrote the situation description (*Speaker*) initiates a conversation to talk about it. The other conversation participant (*Listener*) becomes aware of the underlying situation through what the Speaker says and responds. Speaker and Listener then exchange up to 6 more turns. We include a few example conversations from the training data in Table 1. The models discussed below are evaluated in their ability to replace the *Listener* and respond to the Speaker, without being given the situation description generated by the Speaker. Our data could also be used to generate conversations for the Speaker conditioned on the situation description; we leave this for later work.

EMPATHETICDIALOGUES dataset statistics The resulting dataset comprises 24,850 prompts/conversations from 810 different participants, which will be made publicly available at `anonymous.url` upon acceptance. Details of the crowdsourcing procedure are given in the Supplemental Material.

The distribution of prompts (Table 2) is close to evenly distributed across categories with a few categories that are selected slightly more/less often. The average situation description is 19.8 words. Each conversation is allowed to be 4-8 utterances long (the average is 4.31 utterances per conversation). The average utterance length is 15.2 words long.

We split the conversations into approximately 80% train, 10% validation, and 10% test partitions. To prevent overlap of discussed topics between partitions, we split the data so that all sets of con-

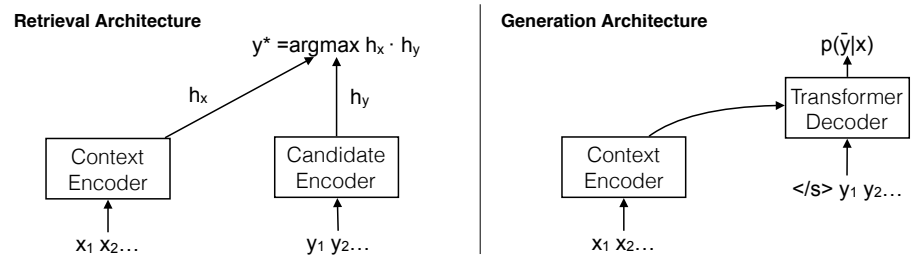

Figure 3: Dialogue generation architectures used in our experiments. The context of concatenated previous utterances is tokenized into $x_1, x_2, \cdots$, and encoded into vector h_x by the context encoder. *Left:* In the retrieval set-up, each candidate y is tokenized into $y_1, y_2, \cdots$ and encoded into vector h_y by the candidate encoder. The system outputs the candidate y^* that maximizes dot product $h_x \cdot h_y$. *Right:* In the generative set-up, the encoded context h_x is used as input to the decoder to generate start symbol $< /s >$ and tokens $y_1, y_2, \cdots$. The model is trained to minimize the negative log-likelihood of target sequence $\bar{y}$ conditioned on context x.

versations with the same speaker providing the prompt would be in the same partition. The final train/val/test split was 19533 / 2770 / 2547 conversations, respectively.

4 EMPATHETIC DIALOGUE GENERATION

We hypothesize that incorporating emotion information into a dialogue system trained on generic open-domain chit-chat helps the model produce more empathetic responses. In this section, we examine several ways of doing so. We use our dialogues to train and evaluate models in the task of generating conversation responses, with the model playing the *Listener* role. At test time, the dialogue model has access to previous utterances in the dialogue, but not to the emotion word prompt (e.g., "proud"), nor to the situation description generated by the Speaker, as would be the case in a normal conversation. Given a dialogue context x of n previous conversation utterances concatenated and tokenized as $x_1, \cdots, x_m$, followed by a target response $\bar{y}$, our models are trained to maximize the likelihood $p(\bar{y}|x)$ of producing the target response. We investigate both generation and retrieval settings (Lowe et al., 2016) as described in Figure 3.

4.1 BASE ARCHITECTURE

We base our models on transformer network architectures (Vaswani et al., 2017), which have proven successful in machine translation and dialogue generation tasks (Zhang et al., 2018; Mazaré et al., 2018).

Retrieval In the retrieval set-up, the model is given a large set Y of candidate responses and picks the "best" one, y^*. We use the retrieval transformer-based architecture from (Yang et al., 2018): two transformer encoders separately embedding the context, x, and candidates, $y \in Y$, as h_x and h_y, respectively. The model chooses the candidate sentence according to a softmax on the dot product: $h_x \cdot h_y$. We minimize the negative log-likelihood of selecting the correct candidate.

At training time, we use all of the sentences from the batch as candidates. We use a large batch size of 512 to give the model more negative examples. At inference time, we experiment with multiple sets of candidate sentences for the model to choose from. First, we use all of the response utterances in the EMPATHETICDIALOGUES training set (Y^{ED}). We also try including candidate utterances from two other large dialogue datasets: the DailyDialog (Li et al., 2017) training set (Y^{DD}) and up to a million utterances from a dump of 1.7 billion Reddit conversations (Y^{R}).

Generation In the generation set-up, we use the full transformer network architecture (Vaswani et al., 2017), consisting of an encoder and a decoder. The transformer decoder uses the encoder output to predict a sequence of words y, and is trained to minimize the negative log-likelihood of the target sequence $\bar{y}$. At inference time, we use diverse beam search from Vijayakumar et al. (2016).

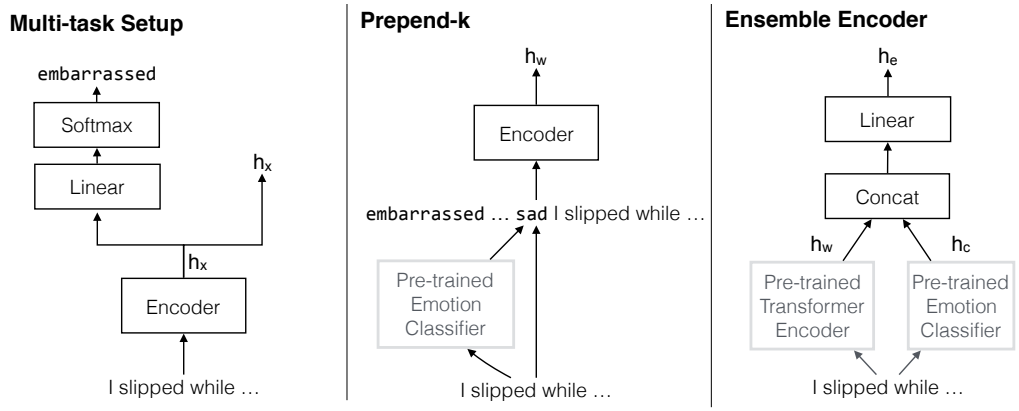

Figure 4: Three ways to encode supervised emotion information. *Left*: the context representation h_x outputted by the context encoder is used both as input to an emotion classifier, and to generate the next utterance as in the base setting. *Middle:* an input sequence (that can be either a dialogue context or a candidate) is first run through a pre-trained emotion classifier, and the top k ouptut labels are prepended to the sequence, which is then run through the corresponding (context or candidate) encoder to output a hidden representation h_w (either h_x or h_y) as in the base setting. *Right:* an input sequence (that can be either a dialogue context or a candidate) is run through the corresponding encoder as well as a pre-trained emotion classifier with the last layer removed. The outputs h_w and h_c are concatenated and linearly projected into a representation h_e.

Training Details We pre-train our models on predicting conversations from a dump of 1.7 billion Reddit conversations. We then fine-tune the models over our EMPATHETICDIALOGUES with a context window of four previous utterances, which is the average length of a conversation in our dataset. We also limit the maximum number of word tokens in the context and response to be 100 each. For simplicity, the transformer networks all have the same architecture (four layers and six transformer heads) from Mazaré et al. (2018). For all models, we train for up to 10 epochs, keeping the version that has the lowest loss on the validation set. We use 300-d word embeddings that were pretrained on common-crawl data using fastText (Grave et al., 2018).

4.2 ADDING EXPLICIT EMOTIONAL SUPERVISION

Simply training the base transformer architectures on data expressly collected to contain many empathetic conversations should yield a model that responds better to emotional situations than one trained on generic data. But if the most appropriate response depends on the emotions at play, nudging the model to encode this information could result in better performance. We experiment with three separate set-ups for adding explicit emotion information: a multi-task objective, pre-pending emotion predictions, and ensemble learning over encoders trained on emotion prediction objectives.

Multi-Task Objective In the first set-up (Fig. 4, left), MULTITASK, we alter the objective function to also optimize for predicting the given emotion label. We add to the context encoder a linear layer and softmax that predicts the emotion label from the context sentences. The objective function is altered to be the average of the negative log-likelihood of predicting the next utterance $\bar{y}$ and the negative log-likelihood of the added linear layer being able to predict the correct emotion.

Prepending Top-K Emotion Predictions In the second set-up (Fig. 4, middle), PREPEND-K, we explicitly add the best emotion predictions from a simple emotion classifier to the input text. We use a fastText model (Joulin et al., 2017) trained to predict the emotion label from the description of the situation written by the Speaker before the dialogue for the training set dialogues. We use the supervised data of EMPATHETICDIALOGUES for training, to see how far the supervised information contained in the dataset itself can be leveraged without using external supervised labels. We use the situation descriptions rather than the dialogue utterances themselves when pre-training the classifier, so that the classifier doesn't see exactly the same training data as the dialogue model. When training the dialogue model, both the context and the candidates are first run through the pre-trained fastText

Table 2: Automatic evaluation metrics on the test set. Pretrained: basic transformer model pre-trained on a dump of 1.7 billion REDDIT conversations. Base: model fine-tuned over the EMPATHETICDIALOGUES training data. Remaining rows: models incorporating emotion supervised information, as described in Sec. 4.2. Candidates come from REDDIT (R), EMPATHETICDIALOGUES (ED), or DAILYDIALOGUES (DD). All automatic metrics clearly improve with in-domain training (Base vs. Pretrained), but the effects of adding supervised information are inconsistent on the automated metrics, although ensembling with a deep emotion classifier consistently improves generation.

	Retrieval			Generation	
Model	P @1,100	Candidate Source	AVG BLEU	PPL	AVG BLEU
Pretrained	43.25	R	4.1	27.96	5.01
	-	ED	5.51	-	-
Base	**56.90**	ED	5.88	21.24	6.27
	-	ED+DD	5.61	-	-
	-	ED+DD+R	4.74	-	-
MULTITASK	55.73	ED	6.18	24.07	5.42
PREPEND-1	56.31	ED	5.93	24.30	4.36
PREPEND-3	55.75	ED	**6.23**	23.96	2.69
PREPEND-5	56.35	ED	6.18	25.40	5.56
ENSEM-DM	52.71	ED	6.03	**19.05**	**6.83**
ENSEM-DM+	52.35	ED	6.04	19.1	6.77
ENSEM-TRAN	51.69	ED	5.88	19.21	6.41

model. The rationale for also adding emotion information to the candidate side is that some emotions might often fit well as response to others, e.g. joy is often mirrored. The top-K predicted emotions are prepended to the beginning of the token sequence as encoder input:

I finally got promoted! $\longrightarrow$ `proud excited joyful` I finally got promoted!

Ensemble of Encoders Finally, we augment the encoders to incorporate more heavily trained representations for emotion prediction (ENSEM) that make use of additional supervised data. We replace each of the encoders in our transformer networks with the Ensemble encoder in Figure 4. This encoder takes the encoding h_w from our basic transformer encoder (either h_x or h_y), already trained on our data, and concatenates it with the representation h_c from the penultimate layer of a deep classifier trained for emotion prediction. The concatenated encodings are projected linearly to the dimension required by the decoder, whose architecture doesn't change. When training the dialogue model, we freeze the basic transformer encoder or emotion classifier (grayed out in Figure 4), and train only the linear layers (and the decoder for generative systems).

We experiment with three different emotion classifiers for the ensemble encoder. First, we take an off-the-shelf classifier for emotion prediction, DeepMoji from Felbo et al. (2017) with the weights as released by the authors, ENSEM-DM. Next, we use a version of the same DeepMoji architecture that is first re-trained on the situation descriptions from our training data, ENSEM-DM+. Finally, we use a second transformer encoder that was trained on a similarly large-scale dataset of public microblogs labelled by their writers with emotion tags such as 'annoyed', ENSEM-TRAN.

5 EXPERIMENTAL RESULTS

We evaluate the models on their ability to reproduce the listeners' portion of the conversation (i.e. the ability to react to someone else's story). We use both automated metrics as well as human evaluation to score each model's retrievals/generations.

Automated Metrics (Table 2) For both retrieval and generative systems, we compute BLEU scores for the final response and compare against the gold label (the actual response). For the

Table 3: Human evaluation metrics from rating task. Training on EMPATHETICDIALOGUES improves all scores. Encoding supervised emotion information improves the empathy score (and sometimes the relevance and fluency by a smaller margin).

	Model	Candidates	Empathy	Relevance	Fluency
Retrieval	Pretrained	R	2.58+0.14	2.97+0.14	4.11+0.12
	Base	ED	3.27+0.13	3.42+0.14	4.44+0.08
	Multitask	ED	3.58+0.12	3.58+0.14	4.46+0.09
	Prepend-1	ED	3.51+0.13	3.61+0.15	4.45+0.10
	Prepend-3	ED	3.62+0.14	3.50+0.15	4.54+0.08
	Prepend-5	ED	3.52+0.14	3.64+0.14	4.47+0.09
	Ensem-DM+	ED	3.36+0.14	3.33+0.14	4.13+0.11
	Ensem-Tran	ED	**3.80+0.12**	**3.66+0.14**	**4.59+0.08**
Generation	Pretrained	-	2.26+0.13	2.37+0.13	4.08+0.12
	Base	-	2.95+0.15	3.10+0.14	4.37+0.10
	Multitask	-	3.17+0.14	3.23+0.14	4.29+0.11
	Prepend-1	-	2.66+0.15	2.63+0.15	4.22+0.12
	Prepend-3	-	3.34+0.13	**3.31+0.15**	**4.58+0.09**
	Prepend-5	-	3.35+0.15	3.20+0.15	4.41+0.10
	Ensem-DM+	-	3.17+0.14	3.19+0.14	4.31+0.11
	Ensem-Tran	-	**3.49+0.12**	3.27+0.14	4.42+0.09
Gold Response	-	-	*4.19+0.06*	*4.48+0.06*	*4.67+0.04*

generative systems, we additionally report perplexity of the actual gold response. For the retrieval systems, we additionally compute `p@1,100`, the accuracy of the model at choosing the correct response out of a hundred randomly selected examples in the test set. When we compute `p@1,100`, the actual response is included in the candidates, unlike inference from the retrieval systems for all other metrics, which only uses training utterances as candidates.

Human Ratings (Table 3) We run two sets of crowdsourcing tasks for humans to score the model responses. In the first task, participants are given a model's output for a randomly selected test set example and asked to score different aspects of the model. The rating task provides a means of comparing aspects of responses, and we are able to ask annotators specifically about whether the response is acknowledging their partner's feelings. We collected at least 100 annotations per model and ask about three aspects of performance:

- Empathy/Sympathy: did the responses show understanding of the feelings of the person talking about their experience? (1: not at all, 3: somewhat, 5: very much)
- Relevance: did the responses seem appropriate to the conversation? were they on-topic? (1: not at all, 3: somewhat, 5: very much)
- Fluency: could you understand the responses? did the language seem accurate? (1: not at all, 3: somewhat, 5: very much)

Human A/B Rankings (Table 4) In the second human evaluation task, participants are given output from two (randomly ordered) models and asked to select the better response. We additionally gave them the option to select "equal" or "neither". For this task, we only give workers test set examples where the pair of models had differing responses. We collected at least 50 annotations per pair of models.

5.1 RESULTS

Table 2 shows that fine-tuning on our data improves all automated metrics. Using only in-domain candidates leads to slightly higher BLEU scores. For retrieval systems, adding emotion supervision explicitly decreases the accuracy of the rankings, `p@1,100`, but generally improves the average BLEU scores. The ensemble encoders improve the generation models in perplexity and BLEU.

Table 4: Average ratio of "best" replies from model A vs model B for a set of pairs in our human ranking evaluation. Ratios > 1 mean that the model on the left was selected more than the model on the right. Emotion Supervision Models: models from Sec. 4.2 that incorporate supervised emotion information. Full listings of comparison between pairwise models is included in the appendix.

Choose Model A vs B response	Average A:B ratio
Emotion Supervision Models vs Pretrain (ret)	3.67
Emotion Supervision Models vs Pretrain (gen)	5.73
Emotion Supervision Models vs Base (ret)	0.97
Emotion Supervision Models vs Base (gen)	1.56
Generation vs. Retrieval	0.62
Gold Response vs. Models	13.79

We further investigate the differences between these models using human evaluations, which have been shown to be more precise means of measuring dialogue quality (Liu et al., 2016). Table 3 shows that fine-tuning a conversational model on the EMPATHETICDIALOGUES data and using the candidates in the dataset substantially improves performance on all metrics, and in particular on the empathy subscore of most interest to us, in both retrieval and generation set-ups. All of the models with explicit emotion supervision (except for the generative Prepend-1 model) improve on the empathy/sympathy scores compared to the base model, meaning that the more explicit emotion supervision does allow models to better condition responses for the tone of the conversation. Many of the models with explicit emotion supervision also improve the relevance of what is said compared to the base model. While there is some slight deterioration in fluency in some of the emotion supervision models, most of them also maintain fluency scores (which are all above 4 on average).

To try to capture the main takeaways from A/B comparisons, we show results averaged over certain pairs of models with similar characteristics (Table 4). Responses from retrieval systems are frequently chosen over generation systems (generation:retrieval ratio is 0.62). Responses from models with explicit emotion supervision are often ranked above the raw pre-trained model (ratios of 3.67 and 5.73), and less so against the base transformer model that was fine-tuned on our data: in the retrieval case, the ratio of picking a line from a model with emotion supervision vs. base model (average ratio of 0.97) indicates that annotators generally picked from each equally. However, in the generation case, the annotators may have favored the models with emotion supervision explicitly represented (average ratio of 1.56).

5.2 TRADING OFF TOPIC SPECIFICITY AND EMPATHY

In Table 5, we show model outputs on two validation set examples. In these examples, the models with emotion prediction components generate responses that focus slightly more on the feelings of the speaker rather than the topic itself. In the first example, the basic generation system trained on our data creates a topically appropriate response (asking what the first person is studying), but ignores the nervousness that person is feeling. The other models all produce reassurances that directly focus on the person's anxiety although they are more generic replies. One avenue for future research is to create models that are better able to both be understanding of what the speaker is feeling while also being concrete about the topic being discussed.

6 CONCLUSION

We introduce a new dataset of 25k dialogues grounded in situations prompted by specific emotion labels. We use this dataset to train models that produce responses more appropriate to the feelings of the other speaker. Our experiments show that when leveraging our dataset and emotion supervision, chit-chat models can produce a next utterance that is judged by humans to be more empathetic/sympathetic to the other speaker. Frequently, using emotion supervision also improves the perceived relevance. Future work will investigate how to use this dataset to model the Speaker, and how to integrate empathetic responding into more general dialogue when, for example, the needs for empathy have to be balanced with staying on topic or providing information. Other possible directions would be to see if this data can serve as additional weakly supervised data for more complex

Table 5: Examples of model responses from the validation set.

Context	Model	Generated Utterance
tomorrow is my first university day and i'm very nervous	Gen-base	What are you studying?
	Gen-multitask	Good luck!
	Gen-prepend-5	I'm sure you will be fine.
	Gen-ensemb-trans	I am sure you will do great!
I moved to northern Alaska last year and don't know anyone.	Ret-base	Interesting. Why wouldn't they take you along?
	Ret-multitask	You don't have any friends or family?
	Ret-prepend-5	That must be pretty lonely. Do you talk often?
	Ret-ensemb-trans	That would be discouraging. Did you make new friends?

emotion-related tasks that look at emotion evolution or causality (Gui et al., 2016; Rashkin et al., 2018). We hope that our results and dataset will stimulate more research in the important direction of making dialog systems more empathetic.

REFERENCES

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

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems. In *Proc. of ICLR*, 2016.

Venkatesh Duppada, Royal Jain, and Sushant Hiray. Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. *arXiv preprint arXiv:1804.06137*, 2018.

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

Bjarke Felbo, Alan Mislove, Anders S, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*, 2017.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*, 2018.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1639–1649, 2016.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *arXiv preprint arXiv:1707.06996*, 2017.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *ICML*, 2017.

Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pp. 49–54, 2018.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 277. ACM, 2018.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *EACL*, 2017.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2016.

Ryan Lowe, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. On the evaluation of dialogue systems with next utterance classification. *arXiv preprint arXiv:1605.05414*, 2016.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*, 2018.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *SemEval@NAACL-HLT*, 2018.

Saif M. Mohammad. #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pp. 246–255, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2387636.2387676.

Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 189–192. ACM, 2010.

Ji Ho Park, Peng Xu, and Pascale Fung. Plusemo2vec at semeval-2018 task 1: Exploiting emotion knowledge from emoji and# hashtags. *arXiv preprint arXiv:1804.08280*, 2018.

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

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. Modeling naive psychology of characters in simple commonsense stories. In *ACL*, 2018.

Klaus R. Scherer and Harald G. Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66 2:310–28, 1994.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2015.

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

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

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *Proceedings of NAACL*, 2015.

Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *SemEval@ACL*, 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424, 2016.

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

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

Joseph Weizenbaum. Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. Emo2vec: Learning generalized emotion representation by multi-task training. *arXiv preprint arXiv:1809.04505*, 2018.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning semantic textual similarity from conversations. In *Rep4NLP@ACL*, 2018.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, 2018.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*, 2017.

Xianda Zhou and William Yang Wang. Mojitalk: Generating emotional responses at scale. In *ACL*, 2018.

Table 6: Distribution of situation/conversation labels within EMPATHETICDIALOGUES

Context	%	**Context**	%	**Context**	%	**Context**	%
surprised	5.15	impressed	3.24	joyful	3.10	content	2.97
excited	3.77	afraid	3.21	prepared	3.08	devastated	2.87
annoyed	3.53	disgusted	3.17	guilty	3.06	sentimental	2.81
proud	3.49	confident	3.16	furious	3.05	caring	2.75
angry	3.47	terrified	3.15	nostalgic	3.05	trusting	2.62
sad	3.45	hopeful	3.14	jealous	3.05	ashamed	2.53
grateful	3.28	anxious	3.11	anticipating	3.03	apprehensive	2.45
lonely	3.28	disappointed	3.10	embarrassed	2.98	faithful	1.92

A SUPPLEMENTAL MATERIAL

A.1 LABEL DISTRIBUTION

In Table 6, we include the exact percentage of situation/conversation labels in our final dataset.

A.2 CROWDSOURCING DESCRIPTION

Our crowdsourcing task asks a pair of workers to take turns (i) selecting a context and writing a prompt describing a situation where they felt that way, and (ii) having a conversation about it, as outlined below.

Writing a Situational Prompt In the first stage of the crowdsourcing task, workers are asked to write a prompt in which they discuss a situation based on a context feeling. Each worker is given three context words from our list of 32 feelings. They are asked to select one of the options and write a short description of a personal situation where they felt that way. We ask the workers to try to keep these prompts between 1-3 sentences. The average response is 19.8 words.

Having A Conversation In the second stage, the same set of workers are paired together and asked to have two short chats with each other. In each chat, one worker (*speaker*) starts a conversation about the situation they experienced and the other worker (*listener*) responds. Neither can see what the other worker was given as a context word or the prompt they submitted, so they must respond to each others' stories based solely on cues within the conversation. Each conversation is allowed to be 4-8 utterances long (the average is 4.31 utterances per conversation). The average utterance length was 15.2 words long.

Ensuring balanced prompt coverage After the first few initial rounds of data collection, we forced workers to select from contexts that they had not chosen before to ensure that all of the categories were getting used.

A.3 EMOTION CLASSIFICATION RESULTS

Our dataset can also be used to train or fine-tune an emotion classifier, as we do in our PREPEND-K and ENSEM-DM+ set-ups. To give a sense of where the difficulty falls compared to other emotion and sentiment classification benchmarks, we reproduce the table from Felbo et al. (2017) and add results when fine-tuning the Deepmoji model on our dataset, or using a fastText classifier (Table 7).

A.4 HUMAN RANKING RESULTS

In Figure 5, we provide the exact comparisons between model responses for the ranking task. Scores less than 1 indicate that the vertical model is preferred, whereas scores greater than one indicate more of a preference for the horizontal model.

Table 7: Classification performance on EMPATHETICDIALOGUES, with the benchmarks proposed in Felbo et al. (2017) for reference. ED: performance on predicting the emotion context label from the situation description. ED-CUT: same, but after having removed all the situation descriptions where the target label was present.

Dataset	Measure	SOTA (in 2017)	fastText	DeepMoji new	DeepMoji full	DeepMoji last	DeepMoji chain-thaw
SE0714	F1	0.34	0.16	0.21	0.31	0.36	0.37
OLYMPIC	F1	0.50	0.38	0.43	0.50	0.61	0.61
PSYCHEXP	F1	0.45	0.44	0.32	0.42	0.56	0.57
SS-TWITTER	Acc	0.82	0.68	0.62	0.85	0.87	0.88
SS-YOUTUBE	Acc	0.86	0.75	0.75	0.88	0.92	0.93
SE0614	Acc	0.51	-	0.51	0.54	0.58	0.58
SCV1	F1	0.63	0.60	0.67	0.65	0.68	0.69
SCV2-GEN	F1	0.72	0.69	0.71	0.71	0.74	0.75
ED	Acc	-	0.43	0.40	0.46	0.46	0.48
ED-CUT	Acc	-	0.41	0.36	0.42	0.44	0.45

		Retrieval								Generation								
		pretrain	base	multi	pre1	pre3	pre3	Ens-tran	Ens-dm+	pretrain	base	multi	pre1	pre3	pre5	Ens-tran	Ens-dm+	Gold
Retrieval	pretrain	0	0.37	0.41	0.17	0.35	0.52	0.5	0.14	4.5	0.78	1	1.23	0.55	0.44	0.82	0.24	0.0
	base	2.7	0	1.42	0.86	1.06	1	0.94	1.07	5.17	2.33	2.3	2.56	1.19	2.62	1.5	1.62	0.1
	multi	2.44	0.71	0	1.07	0.72	0.93	0.75	0.53	7.5	3.44	1.8	1.73	2.64	3	1.12	1.75	0.2
	pre1	5.8	1.17	0.94	0	2	0.88	0.95	0.94	4.17	2.25	2.15	3.71	3.86	4.5	1.54	3	0.0
	pre3	2.86	0.94	1.38	0.5	0	0.61	1.25	1.3	5	1.31	3.57	3.33	1.42	1.06	1.44	1.2	0.2
	pre3	1.91	1	1.07	1.13	1.64	0	1.07	0.82	4.83	2.4	3.29	2.33	1.47	2.3	2	1.58	0.1
	Ens-tran	2	1.07	1.33	1.06	0.8	0.93	0	0.47	7.25	2	2.33	2.31	1.36	1.5	1.36	1.21	0.3
	Ens-dm+	7	0.93	1.88	1.06	0.77	1.21	2.11	0	13.5	3.62	2.09	1.69	2.5	2.3	1.9	1.43	0.2
Generation	pretrain	0.22	0.19	0.13	0.24	0.2	0.21	0.14	0.07	0	0.16	0.08	0.29	0.22	0.4	0.22	0.16	0.0
	base	1.29	0.43	0.29	0.44	0.76	0.42	0.5	0.28	6.25	0	1.2	1.25	0.5	1.23	0.79	0.28	0.1
	multi	1	0.43	0.56	0.46	0.28	0.3	0.43	0.48	12.5	0.83	0	0.78	0.79	1.06	0.57	0.36	0.
	pre1	0.81	0.39	0.58	0.27	0.3	0.43	0.43	0.59	3.5	0.8	1.29	0	0.47	1.4	0.3	0.4	0.0
	pre3	1.83	0.84	0.38	0.26	0.71	0.68	0.74	0.4	4.5	2	1.27	2.12	0	1.27	0.76	0.72	0.2
	pre5	2.25	0.38	0.33	0.22	0.95	0.43	0.67	0.43	2.5	0.81	0.95	0.71	0.79	0	0.22	1	0.1
	Ens-tran	1.22	0.67	0.89	0.65	0.7	0.5	0.74	0.53	4.6	1.27	1.75	3.29	1.31	4.6	0	0.75	0.0
	Ens-dm+	4.17	0.62	0.57	0.33	0.83	0.63	0.83	0.7	6.25	3.62	2.78	2.5	1.38	1	1.33	0	0.1
	Gold	25	5.67	4.11	16.5	4	6.4	3.09	4.67	53	6.83	10	45	4.86	8.75	13.5	9.25	

Figure 5: Full heatmap from ranking task. Scores are ratios of the [# times horizontal model is selected over vertical] : [# times vertical model is selected over horizontal]. Scores of greater than 1 indicate a preference for the horizontal model, and less than one indicates a preference for the vertical model.