MoEL: Mixture of Empathetic Listeners

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{zlinao,amadotto,jmshinaa,pxuab}@connect.ust.hk,
pascale@ece.ust.hk

Abstract

Previous research on empathetic dialogue systems has mostly focused on generating responses given certain emotions. However, being empathetic not only requires the ability of generating emotional responses, but more importantly, requires the understanding of user emotions and replying appropriately. In this paper, we propose a novel end-to-end approach for modeling empathy in dialogue systems: Mixture of Empathetic Listeners (MoEL). Our model first captures the user emotions and outputs an emotion distribution. Based on this, MoEL will *softly combine* the output states of the *appropriate* Listener(s), which are each optimized to react to certain emotions, and generate an empathetic response. Human evaluations on *empathetic-dialogues* (Rashkin et al., 2018) dataset confirm that MoEL outperforms multitask training baseline in terms of empathy, relevance, and fluency. Furthermore, the case study on generated responses of different Listeners shows high interpretability of our model.

1 Introduction

Neural network approaches for conversation models have shown to be successful in scalable training and generating fluent and relevant responses (Vinyals and Le, 2015). However, it has been pointed out by Li et al. (2016a,b,c); Wu et al. (2018b) that only using Maximum Likelihood Estimation as the objective function tends to lead to *generic* and *repetitive* responses like "I am sorry". Furthermore, many others have shown that the incorporation of additional inductive bias leads to a more engaging chatbot, such as understanding commonsense (Dinan et al., 2018), or modeling consistent persona (Li et al., 2016b; Zhang et al., 2018a; Mazare et al., 2018a).

Meanwhile, another important aspect of an engaging human conversation that received relatively less focus is emotional understanding and empathy (Rashkin et al., 2018; Dinan et al., 2019; Wolf et al., 2019). Intuitively, ordinary social conversations between two humans are often about their daily lives that revolve around happy or sad experiences. In such scenarios, people generally tend to respond in a way that acknowledges the feelings of their conversational partners.

Emotion: Angry	
Situation I was furious when I got in my first car wreck.	
Speaker	I was driving on the interstate and another car ran into the back of me.
Listener	Wow. Did you get hurt? Sounds scary.
Speaker	No just the airbags went off and I hit my head and got a few bruises.
Listener	I am always scared about those airbags! I am so glad you are ok!

Table 1: One conversation from empathetic dialogue, a speaker tells the situation he(she) is facing, and a listener try to understand speaker's feeling and respond accordingly

Table 1 shows an conversation from the *empathetic-dialogues* dataset (Rashkin et al., 2018) about how an empathetic person would respond to the stressful situation the *Speaker* has been through. However, despite the importance of empathy and emotional understanding in human conversations, it is still very challenging to train a dialogue agent able to recognize and respond with the correct emotion.

So far, to solve the problem of empathetic dialogue response generation, which is to understand the user emotion and respond appropriately (Bertero et al., 2016), there have been mainly two lines

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 121–132,
Hong Kong, China, November 3–7, 2019.

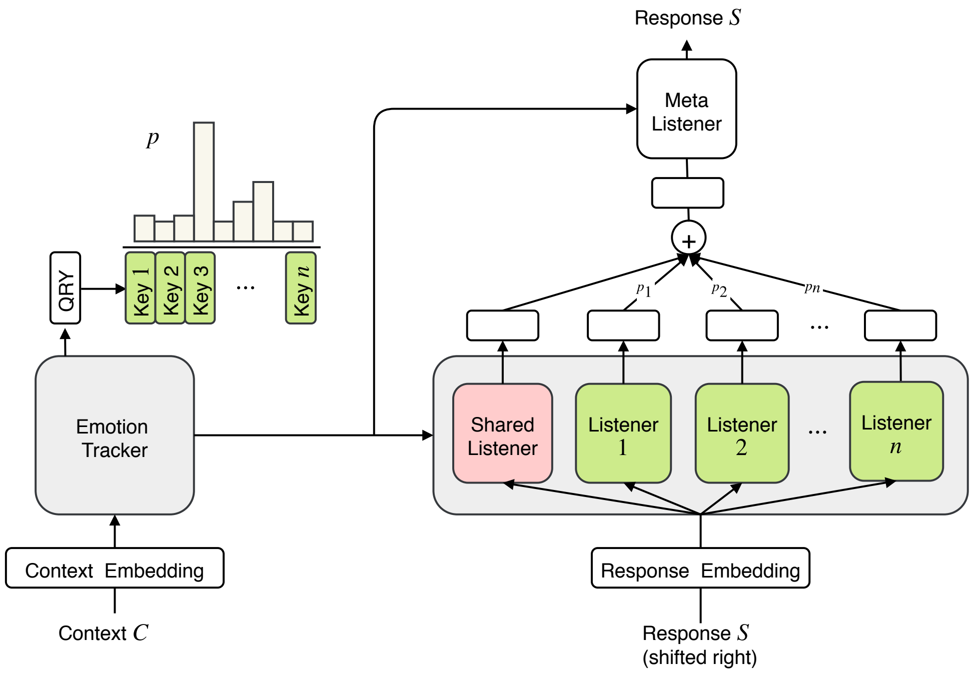

Figure 1: The proposed model Mixture of Empathetic Listeners, which has an emotion tracker, n empathetic listeners along with a shared listener, and a meta listener to fuse the information from listeners and produce the empathetic response.

of work. The first is a multi-task approach that jointly trains a model to predict the current emotional state of the user and generate an appropriate response based on the state (Lubis et al., 2018; Rashkin et al., 2018). Instead, the second line of work focuses on conditioning the response generation to a certain fixed emotion (Hu et al., 2017; Wang and Wan, 2018; Zhou and Wang, 2018; Zhou et al., 2018).

Both cases have succeeded in generating empathetic and emotional responses, but have neglected some crucial points in empathetic dialogue response generation. 1) The first assumes that by understanding the emotion, the model implicitly learns how to respond appropriately. However, without any additional inductive bias, a single decoder learning to respond for all emotions will not only lose interpretability in the generation process, but will also promote more generic responses. 2) The second assumes that the emotion to condition the generation on is given as input, but we often do not know which emotion is appropriate in order to generate an empathetic response.

Therefore, in this paper, to address the above issues, we propose a novel end-to-end empathetic dialogue agent, called Mixture of Empathetic Listeners [1] (MoEL). Similar to Rashkin et al. (2018), we first encode the dialogue context and use it to recognize the emotional state (n possible emotions). However, the main difference is that our model consists of n decoders, further denoted as *listeners*, which are optimized to react to each context emotion accordingly. The listeners are trained along with a Meta-listener that *softly combines* the output decoder states of each listener according to the emotion classification distribution. Such design allows our model to explicitly learn how to choose an appropriate reaction based on its understanding of the context emotion. A detailed illustration of MoEL is shown in Figure 1.

The proposed model is tested against several competitive baseline settings (Vaswani et al., 2017; Rashkin et al., 2018), and evaluated with human judges. The experimental results show that our approach outperforms the baselines in both empathy and relevance. Finally, our analysis demonstrates that not only MoEL effectively attends to the right listener, but also each listener learns how to properly react to its corresponding emotion, hence allowing a more interpretable generative process.

[1] The code will be released at https://github.com/HLTCHKUST/MoEL

2 Related Work

Conversational Models: Open domain conversational models has been widely studied (Serban et al., 2016; Vinyals and Le, 2015; Wolf et al., 2019). A recent trend is to produce personalized responses by conditioning the generation on a persona profile to make the response more consistent through the dialogue (Li et al., 2016b). In particular, PersonaChat (Zhang et al., 2018b; Kulikov et al., 2018) dataset was created, and then extended in ConvAI 2 challenge (Dinan et al., 2019), to show that by adding persona information as input to the model, the produced responses elicit more consistent personas. Based on such, several follow-up work has been presented (Mazare et al., 2018b; Hancock et al., 2019; Joshi et al., 2017; Kulikov et al., 2018; Yavuz et al., 2018; Zemlyanskiy and Sha, 2018; Madotto et al., 2019). However, such personalized dialogue agents focus only on modeling a consistent persona and often neglect the feelings of their conversation partners.

Another line of work combines retrieval and generation to promote the response diversity (Cai et al., 2018; Weston et al., 2018; Wu et al., 2018b). However, only fewer works focus on emotion (Winata et al., 2017, 2019; Xu et al., 2018; Fan et al., 2018a,c,b; Lee et al., 2019) and empathy in the context of dialogues systems (Bertero et al., 2016; Chatterjee et al., 2019a,b; Shin et al., 2019). For generating emotional dialogues, Hu et al. (2017); Wang and Wan (2018); Zhou and Wang (2018) successfully introduce a framework of controlling the sentiment and emotion of the generated response, while (Zhou and Wang, 2018) also introduces a new Twitter conversation dataset and propose to distantly supervised the generative model with emojis. Meanwhile, (Lubis et al., 2018; Rashkin et al., 2018) also introduce new datasets for empathetic dialogues and train multi-task models on it.

Mixture of Experts: The idea of having specialized parameters, or so-called experts, has been widely studied topics in the last two decades (Jacobs et al., 1991; Jordan and Jacobs, 1994). For instance, different architectures and methodologies have been used such as SVM (Collobert et al., 2002), Gaussian Processes (Tresp, 2001; Theis and Bethge, 2015; Deisenroth and Ng, 2015), Dirichlet Processes (Shahbaba and Neal, 2009), Hierarchical Experts (Yao et al., 2009), Infinite Number of Experts (Rasmussen and Ghahramani, 2002) and sequential expert addition (Aljundi et al., 2017). More recently, the Mixture Of Expert (Shazeer et al., 2017; Kaiser et al., 2017) model was proposed which added a large number of experts in between of two LSTM (Schmidhuber, 1987) layers to enhance the capacity of the model. This idea of having independent specialized experts inspires our approach to model the reaction to each emotion with a separate expert.

3 Mixture of Empathetic Listeners

The dialogue context is an alternating set of utterances from speaker and listener. We denote the dialogue context as $C = \{U_1, S_1, U_2, S_2, \cdots, U_t\}$ and the speaker emotion state at each utterance as $Emo = \{e_1, e_2, \cdots, e_t\}$ where $\forall e_i \in \{1, \ldots, n\}$. Then, our model aims to track the speaker emotional state e_t from the dialogue context C, and generates an empathetic response S_t.

Overall, MoEL is composed of three components: an *emotion tracker*, *emotion-aware listeners*, and a *meta listener* as shown in Figure 1. The emotion tracker (which is also the context encoder) encodes C and computes a distribution over the possible user emotions. Then all the listeners independently attend to this distribution to compute their own representation. Finally, the meta listener takes the weighted sum of representations from the listeners and generates the final response.

3.1 Embedding

We define the context embedding $E^C \in \mathbb{R}^{|V| \times d_{emb}}$, and the response embedding $E^R \in \mathbb{R}^{|V| \times d_{emb}}$ which are used to convert tokens into embeddings. In multi-turn dialogues, ensuring that the model is able to distinguish among turns is essential, especially when multiple emotion are present in different turns. Hence, we incorporate a dialogue state embedding in the input. This is used to enable the encoder to distinguish speaker utterances and listener utterances (Wolf et al., 2019). As shown in Figure 2, our context embedding E^C is the positional sum of the word embedding E^W, the positional embedding E^P (Vaswani et al., 2017) and the dialogue state embedding E^D.

$$E^C(C) = E^W(C) + E^P(C) + E^D(C) \quad (1)$$

3.2 Emotion Tracker

MoEL uses a standard transformer encoder (Vaswani et al., 2017) for the emotion

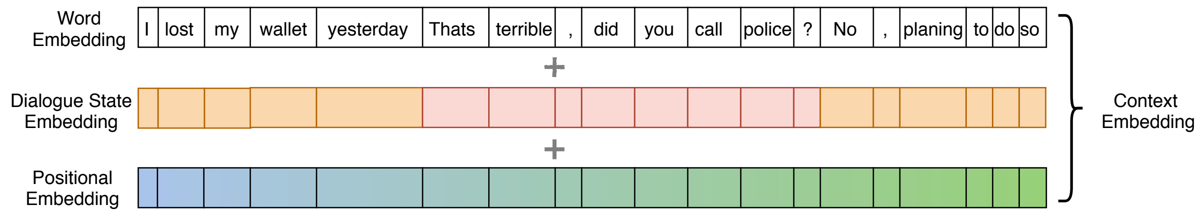

Figure 2: Context embedding is computed by summing up the word embedding, dialogue state embedding and positional embedding for each token.

tracker. We first flatten all dialogue turns in C, and map each token into its vectorial representation using the context embedding E^C. Then the encoder encodes the context sequence into a context representation. We add a query token QRY at the beginning of each input sequence as in BERT (Devlin et al., 2018), to compute the weighted sum of the output tensor. Denoting a transformer encoder as TRS_{Enc}, then corresponding context representation become:

$$H = TRS_{Enc}(E^C([QRY; C])) \tag{2}$$

where $[;]$ denotes concatenation, $H \in \mathbb{R}^{L \times d_{model}}$ where L is the sequence length. Then, we define the final representation of the token QRY as

$$q = H_0 \tag{3}$$

where $q \in \mathbb{R}^{d_{model}}$, which is then used as the query for generating the emotion distribution.

3.3 Emotion Aware Listeners

The emotion aware listeners mainly consist of 1) a *shared listener* that learns shared information for all emotions and 2) n independently parameterized Transformer decoders (Vaswani et al., 2017) that learn how to appropriately react given a particular emotional state. All the listeners are modeled by a standard transformer decoder layer block, denoted as TRS_{Dec}, which is made of three sub-components: a multi-head self-attention over the response input embedding, a multi-head attention over the output of the emotion tracker, and a position-wise fully connected feed-forward network.

Thus, we define the set of listeners as $L = [TRS^0_{Dec}, \ldots, TRS^n_{Dec}]$. Given the target sequence shifted by one $r_{0:t-1}$, each listener compute its own emotional response representation V_i:

$$V_i = TRS^i_{Dec}(H, E^R(r_{0:t-1})) \tag{4}$$

where TRS^i_{Dec} refers to the i-th listener, including the shared one. Conceptually, we expect that the output from the shared listener, TRS^0_{Dec}, to be a general representation which can help the model to capture the dialogue context. On the other hand, we expect that each empathetic listener learns how to respond to a particular emotion. To model this behavior, we assign different weights to each empathetic listener according to the user emotion distribution, while assigning a fixed weight of 1 to the shared listener.

To elaborate, we construct a Key-Value Memory Network (Miller et al., 2016) and represent each memory slot as a vector pair (k_i, V_i), where $k_i \in \mathbb{R}^{d_{model}}$ denotes the key vector and V_i is from Equation 4. Then, the encoder informed query q is used to address the key vectors k by performing a dot product followed by a Softmax function. Thus, we have:

$$p_i = \frac{e^{q^\top k_i}}{\sum_{j=1}^{n} e^{q^\top k_j}} \tag{5}$$

each p_i is the score assigned to V_i, thus used as the weight of each listener. During training, given the speaker emotion state e_t, we supervise each weight p_i by maximizing the probability of the emotion state e_t with a cross entropy loss function:

$$\mathcal{L}_1 = -\log p_{e_t} \tag{6}$$

Finally, the combined output representation is compute by the weighted sum of the memory values V_i and the shared listener output V_0.

$$V_M = V_0 + \sum_{i=1}^{n} p_i V_i \tag{7}$$

3.4 Meta Listener

Finally, the Meta Listener is implemented using another transformer decoder layer, which further transform the representation of the listeners and

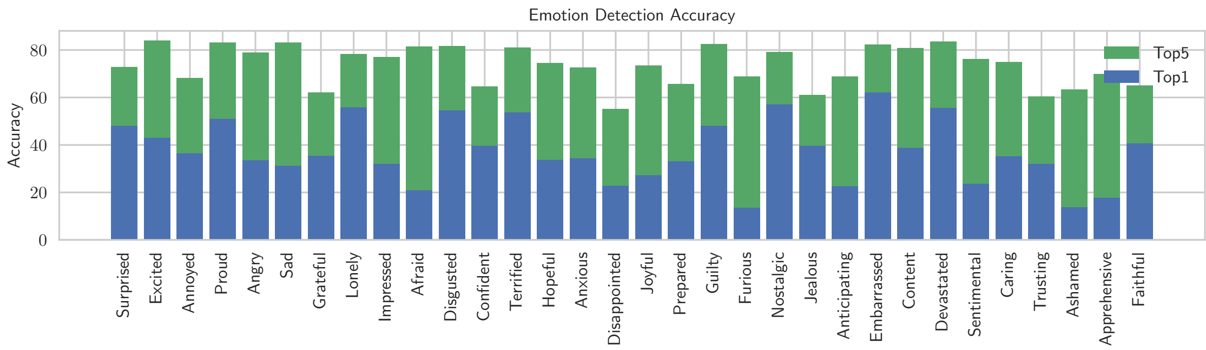

Figure 3: Top-1 and Top-5 emotion detection accuracy over 32 emotions at each turn

	Params.	BLEU	Empathy	Relevance	Fluency
Gold	-	-	3.93	3.93	3.35
TRS	16.94M	3.02	3.32	3.47	**3.52**
MultiTRS	16.95M	2.92	3.36	3.57	3.31
MoEL	23.1M	2.90	**3.44**	**3.70**	3.47

Table 2: Comparison between our proposed methods and baselines. All of models receive close BLEU score. MoEL achieve highest *Empathy* and *Relevance* score, while TRS achieve better *Fluency* score. The number of parameters for each model is reported.

generates the final response. The intuition is that each listener specializes to a certain emotion and the Meta Listener gathers the opinions generated by multiple listeners to produce the final response. Hence, we define another TRS_{Dec}^{Meta}, and an affine transformation $W \in \mathbb{R}^{d_{model} \times |V|}$ to compute:

$$O = TRS_{Dec}^{Meta}(H, V_M) \tag{8}$$

$$p(r_{1:t}|C, r_{0:t-1}) = \text{softmax}(O^\top W) \tag{9}$$

where $O \in \mathbb{R}^{d_{model} \times t}$ is the output of meta listener and $p(r_{1:t}|C, r_{0:t-1})$ is a distribution over the vocabulary for the next tokens. We then use a standard maximum likelihood estimator (MLE) to optimize the response prediction:

$$\mathcal{L}_2 = -\log p\left(S_t|C\right) \tag{10}$$

Lastly, all the parameters are jointly trained end-to-end to optimize the listener selection and response generation by minimizing the weighted-sum of two losses:

$$\mathcal{L} = \alpha \mathcal{L}_1 + \beta \mathcal{L}_2 \tag{11}$$

Where α and β are hyperparameters to balance two loss.

Model	Win	Loss	Tie
MoEL vs TRS	37.3%	18.7%	44%
MoEL vs Multi-TRS	36.7%	32.6%	30.7%

Table 3: Result of human A/B test. Tests are conducted pairwise between MoEL and baseline models

4 Experiment

4.1 Dataset

We conduct our experiment on the *empathetic-dialogues* (Rashkin et al., 2018) dataset which consist of 25k one-to-one open-domain conversation grounded in emotional situations. The dataset provides 32, evenly distributed, emotion labels. Table 1 shows an example from the training set. The speakers are talking about their situation and the listeners is trying to understand their feeling and reply accordingly. At training time the emotional labels of the speakers are given, while we hide the label in test time to evaluate the *empathy* of our model.

4.2 Training

We train our model using Adam optimizer (Kingma and Ba, 2014) and varied the learning rate during training following (Vaswani et al., 2017). The weight of both losses α and β are set to 1 for simplicity. We use pre-trained Glove vectors (Pennington et al., 2014) to initialize the word embedding and we share it across the encoder and the decoder. The rest of the parameters are randomly initialized.

In the early training stage, emotion tracker randomly assign weights to the listeners, and may send noisy gradient flow back to the wrong listeners, which can make the model convergence harder. To stabilize the learning process, we replace the distribution p of the listeners with the oracle emotion e_t information using a certain probability ϵ_{oracle}, and we gradually anneal it during the training. We set an annealing rate $\gamma = 1 \times 10^{-3}$, and a threshold t_{thd} equal to 1×10^4, thus at each iteration t iteration we compute:

$$\epsilon_{oracle} = \gamma + (1 - \gamma)e^{-\frac{t}{t_{thd}}} \tag{12}$$

4.3 Baseline

We compare our model with two baselines:

Transformer (TRS) The standard Transformer model (Vaswani et al., 2017) that is trained to minimize MLE loss as in Equation 10.

Multitask Transformer (Multi-TRS) A Multitask Transformer trained as (Rashkin et al., 2018) to incorporate additional supervised information about the emotion. The encoder of multitask transformer is the same as our emotion tracker, and the context representation Q, from Equation 3, is used as input to an emotion classifier. The whole model is jointy trained by optimizing both the classification and generation loss.

4.4 Hyperparameter

In all of our experiments we used 300 dimensional word embedding and 300 hidden size everywhere. We use 2 self-attention layers made up of 2 attention heads each with embedding dimension 40. We replace Positionwise Feedforward sub-layer with 1D convolution with 50 filters of width 3. We train all of models with batch size 16 and we use batch size 1 in the test time.

4.5 Evaluation Metrics

BLEU We compute BLEU scores (Papineni et al., 2002) to compare the generated response against human responses. However, in open-domain dialogue response generation, BLEU is not a good measurement of generation quality (Liu et al., 2016), so we use BLEU only as a reference.

Human Ratings In order to measure the quality of the generated responses, we conduct human evaluations with Amazon Mechanical Turk. Following Rashkin et al. (2018), we first randomly sample 100 dialogues and their corresponding generations from MoEL and the baselines. For each response, we assign three human annotators to score the following aspect of models: *Empathy*, *Relevance*, and *Fluency*. Note that we evaluate each metric independently and the scores range between 1 and 5, in which 1 is ”not at all” and 5 is ”very much”.

We ask the human judges to evaluate each of the following categories from a 1 to 5 scale, where 5 is the best score.

- Empathy / Sympathy: Did the responses from the LISTENER show understanding of the feelings of the SPEAKER talking about their experience?
- Relevance: Did the responses of the LISTENER seem appropriate to the conversation? Were they on-topic?
- Fluency: Could you understand the responses from the LISTENER? Did the language seem accurate?

Human A/B Test In this human evaluation task, we aim to directly compare the generated responses with each other. We randomly sample 100 dialogues each for *MoEL vs* $\{$*TRS, Multi-TRS*$\}$. Three workers are given randomly ordered responses from either MoEL or $\{$TRS, Multi-TRS$\}$, and are prompted to choose the better response. They can either choose one of the responses or select *tie* when the provided options are either both good or both bad.

5 Results

Emotion detection To verify whether our model can attend to the appropriate listeners, we compute the emotion detection accuracy for each turn. Our model achieve 38%, 63%, 74% in terms of top-1, top-3, top-5 detection accuracy over 32 emotions. We notice that some emotions frequently appear in similar context (e.g., Annoyed, Angry, Furious) which might degrade the detection accuracy. Figure 3 shows the per class accuracy in the test set. We can see that by using top-5 the majority of the emotion achieve around 80% accuracy.

Response evaluation Both automatic and human evaluation results are shown in Table 2. TRS achieves the highest BLEU score and *Fluency* score but the lowest *Empathy* and *Relevance* score. This shows us that the responses generated by TRS are more generic but cannot accurately capture the user emotions. With the additional supervision on user emotions, multi-task training improves both *Empathy* and *Relevance* score, but it still degrades *Fluency*. In contrast, MoEL achieves the highest

Emotion	Disgusted
Context	**Speaker**: i had to dog-sit for my dad about a week ago . i had to clean up dog poop at least 10 times , it was awful .
TRS	that is **horrible** . i hope you did not have to clean it up .
Multi-TRS	oh my goodness , that must have been so **scary** !
MoEL	oh wow , that is pretty **gross** .
Goal	ohh , cleaning dog poop is always been annoying , yet petting the dog is always amusing

Emotion	Disappointed
Context	**Speaker**: i was turned down for a job i really thought i would be a good fit at . **Listener**: did they give you any reason for not giving you the job ? **Speaker**: someone else interviewed who had much more experience than i did .
TRS	that is good . i hope you have a good time .
Multi-TRS	that is a good idea . i hope it was a good idea .
MoEL	i am sorry to hear that . i am sure you will do great again
Goal	well , at least a candidate that deserved the job got it .

Speaker: so my friend ’s baby fell down and scraped his knee yesterday and i was trying so hard to cheer him up. **Listener**$_{caring}$: that was really nice of you. I am glad you were able to help him.	
Listener$_{terrified}$	oh wow ! that must have been so scary .
Listener$_{sad}$	oh no , i am so sorry to hear that .
Listener$_{excited}$	that sounds like a good idea .
Listener$_{proud}$	that was awesome ! did he get any good at you ?

Speaker: my husband lost a job but I am hoping he can find a full-time job soon. **Listener**$_{hopeful}$: what kind of job is it?	
Listener$_{terrified}$	oh that sucks . you should have been together .
Listener$_{sad}$	i am so sorry . i am sure he will be fine .
Listener$_{excited}$	that sounds awesome . what kind of job did he get you for ?
Listener$_{proud}$	oh wow ! congratulations to him . you must be proud of him .

Table 4: Generated responses from TRS, Multi-TRS and MoEL in 2 different user emotion states (**top**) and comparing generation from different listeners (**bottom**). We use hard attention on Terrified, Sad, Excited and Proud listeners.

Empathy and *Relevance* score. This suggests that the multi-expert strategy helps to capture the user emotional states and context simultaneously, and elicits a more appropriate response. The human A/B tests also confirm that the responses from our model are more preferred by human judges.

6 Analysis

In order to understand whether or how MoEL can effectively improve other baselines, learn each emotion, and properly react to them, we conduct three different analyses: model response comparison, listener analysis, and visualization of the emotion distribution p.

Model response comparison The top part of Table 4 compares the generated responses from MoEL and the two baselines on two different speaker emotional states. In the first example, MoEL captures the exact emotion of the speaker, by replying with ”cleaning up dog poop is pretty **gross**”, instead of ”**horrible**” and ”**scary**”. In the second example, both TRS and Multi-TRS fail to understand that the speaker is disappointed about the failure of his interview, and they generate inappropriate responses. On the other hand, MoEL shows an empathetic response by comforting the speaker with ”I am sure you will do great again”. More examples can be find in the Appendix.

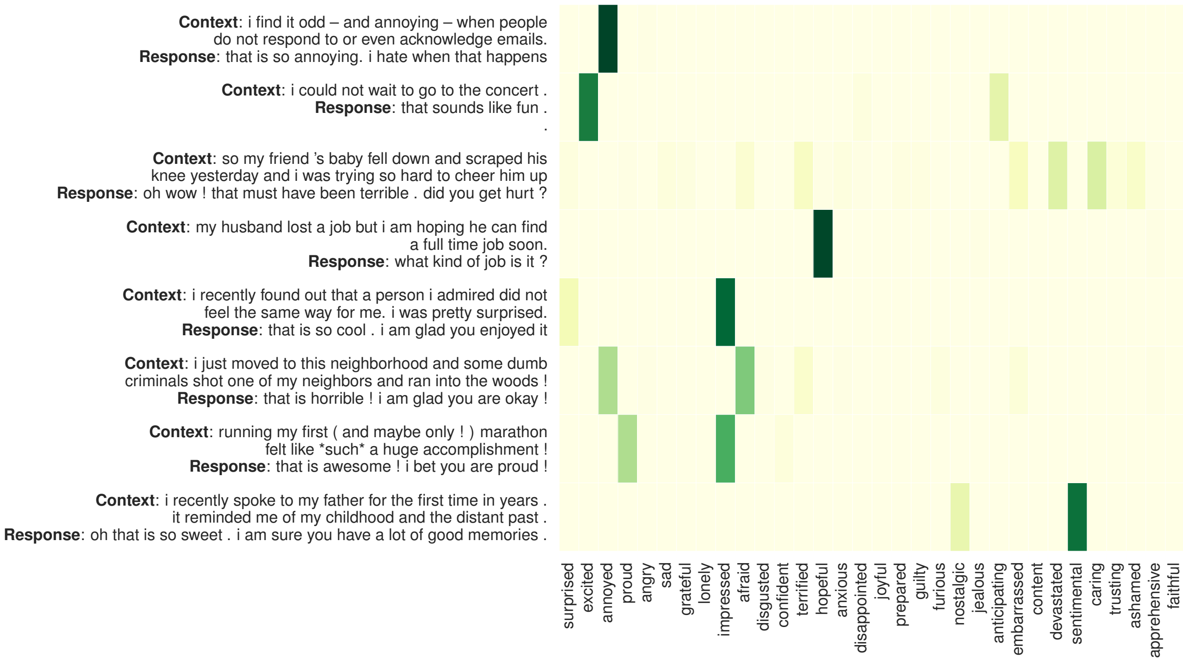

Figure 4: The visualization of attention on the listeners: The left side is the context followed by the responses generated by MoEL. The heat map illustrate the attention weights on 32 listeners

Listener analysis To have a better understanding of how each listener learned to react to different context, we conduct a study of comparing responses produced by different listeners. To do so, we *fix* the input dialogue context and we manually modify the attention vector distribution p used to produce the response. We experiment with the correct listener and four other listeners: **Listener**$_{terrified}$, **Listener**$_{sad}$, **Listener**$_{excited}$, **Listener**$_{proud}$. Given the same context, we expect that different listeners will react differently, as this is our inductive bias. For example, **Listener**$_{sad}$ is optimized to comfort sad people, and **Listener**$_{\{excited,proud\}}$ share the positive emotions from the user. From the generation results in the bottom parts of Table 4 we can see that the corresponding listeners can produce empathetic and relevant responses when they reasonably match the speaker emotions. However, when the expected emotion label is opposite to the selected listener, such as *caring* and *sad*, the response becomes emotionally inappropriate.

Interestingly, in the last example, the *sad* listener actually produces a more meaningful response by encouraging the speaker. This is due to the first part of the context which conveys a sad emotion. On the other hand, for the same example, the *excited* listener responds with very relevant yet unsympathetic response. In addition, as many dialogue contexts contain multiple emotions, being able to capture them would lead to a better understanding of the speaker emotional state.

Visualization of Emotion Distribution Finally, to understand how MoEL chooses the listener according to the context, we visualize the emotion distribution p in Figure 4. In most of the cases, the model attends to the proper listeners (emotions), and generate a proper responses. This is confirmed also by the accuracy results shown in Figure 3. However, our model is sometimes focuses on parts of the dialogue context. For example, in the fifth example in Figure 4, the model fails to detect the real emotion of speaker as the context contains "I was pretty **surprised**" in its last turn.

On the other hand, the last three rows of the heatmap indicate that the model learns to leverage **multiple** listeners to produce an empathetic response. For example, when the speaker talks about some criminals that shot one of his neighbors, MoEL successfully detects both *annoyed* and *afraid* emotions from the context, and replies with an appropriate response "that is horrible! i am glad you are okay!" that addresses both emotions. However, in the third row, the model produces "you" instead of "he" by mistake. Although the model is able to capture relevant emotions for this case, other emotions also have non-negligible weights which results in a smooth emotion distribution p that confuses the meta listener from accurately generating a response.

7 Conclusion & Future Work

In this paper, we propose a novel way to generate empathetic dialogue responses by using Mixture of Empathetic Listeners (MoEL). Differently from previous works, our model understands the user feelings and responds accordingly by learning specific listeners for each emotion. We benchmark our model in *empathetic-dialogues* dataset (Rashkin et al., 2018), which is a multi-turn open-domain conversation corpus grounded on emotional situations. Our experimental results show that MoEL is able to achieve competitive performance in the task with the advantage of being more interpretable than other conventional models. Finally, we show that our model is able to automatically select the correct emotional decoder and effectively generate an empathetic response.

One of the possible extensions of this work would be incorporating it with Persona (Zhang et al., 2018a) and task-oriented dialogue systems (Gao et al., 2018; Madotto et al., 2018; Wu et al., 2019, 2017, 2018a; Reddy et al., 2018; Raghu et al., 2019). Having a persona would allow the system to have more consistent and personalized responses, and combining open-domain conversations with task-oriented dialogue systems would equip the system with more engaging conversational capabilities, hence resulting in a more versatile dialogue system.

Acknowledgments

This work has been partially funded by ITF/319/16FP and MRP/055/18 of the Innovation Technology Commission, the Hong Kong SAR Government. We sincerely thank the three anonymous reviewers for their insightful comments.

References

Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. 2017. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375.

Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. 2016. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1047.

Deng Cai, Yan Wang, Victoria Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2018. Skeleton-to-response: Dialogue generation guided by retrieval memory. *arXiv preprint arXiv:1809.05296*.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019a. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019b. Semeval-2019 task 3: Emocontext: Contextual emotion detection in text. In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, Minnesota.

Ronan Collobert, Samy Bengio, and Yoshua Bengio. 2002. A parallel mixture of svms for very large scale problems. In *Advances in Neural Information Processing Systems*, pages 633–640.

Marc Deisenroth and Jun Wei Ng. 2015. Distributed gaussian processes. In *International Conference on Machine Learning*, pages 1481–1490.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *ICLR*.

Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018a. Multi-region ensemble convolutional neural network for facial expression recognition. In *International Conference on Artificial Neural Networks*, pages 84–94. Springer.

Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018b. Unsupervised domain adaptation with generative adversarial networks for facial emotion recognition. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4460–4464. IEEE.

Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018c. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 584–588. ACM.

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. ACM.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, Geoffrey E Hinton, et al. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Michael I Jordan and Robert A Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *arXiv preprint arXiv:1811.00907*.

Nayeon Lee, Zihan Liu, and Pascale Fung. 2019. Team yeon-zi at semeval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1052–1056.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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 the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5454–5459.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018a. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018b. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779. Association for Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Dinesh Raghu, Nikhil Gupta, and Mausam. 2019. Disentangling Language and Knowledge in Task-Oriented Dialogs. 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 1239–1255, Minneapolis, Minnesota. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy. *arXiv preprint arXiv:1811.00207*.

Carl E Rasmussen and Zoubin Ghahramani. 2002. Infinite mixtures of gaussian process experts. In *Advances in neural information processing systems*, pages 881–888.

Revanth Reddy, Danish Contractor, Dinesh Raghu, and Sachindra Joshi. 2018. Multi-level memory for task oriented dialogs. *arXiv preprint arXiv:1810.10647*.

Jurgen Schmidhuber. 1987. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 14 May.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Babak Shahbaba and Radford Neal. 2009. Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, 10(Aug):1829–1850.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *arXiv preprint arXiv:1906.08487*.

Lucas Theis and Matthias Bethge. 2015. Generative image modeling using spatial lstms. In *Advances in Neural Information Processing Systems*, pages 1927–1935.

Volker Tresp. 2001. Mixtures of gaussian processes. In *Advances in neural information processing systems*, pages 654–660.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *International Conference on Machine Learning*.

Ke Wang and Xiaojun Wan. 2018. Sentigan: generating sentimental texts via mixture adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4446–4452. AAAI Press.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92.

Genta Indra Winata, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Nora the empathetic psychologist. *Proc. Interspeech 2017*, pages 3437–3438.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Jamin Shin, Yan Xu, Peng Xu, and Pascale Fung. 2019. Caire_hkust at semeval-2019 task 3: Hierarchical attention for dialogue emotion classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 142–147.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chien-Sheng Wu, Andrea Madotto, Genta Winata, and Pascale Fung. 2017. End-to-end recurrent entity network for entity-value independent goal-oriented dialog learning. In *Dialog System Technology Challenges Workshop, DSTC6*.

Chien-Sheng Wu, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2018a. End-to-end dynamic query memory network for entity-value independent task-oriented dialog. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6154–6158. IEEE.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. *arXiv preprint arXiv:1901.04713*.

Yu Wu, Furu Wei, Shaohan Huang, Zhoujun Li, and Ming Zhou. 2018b. Response generation by context-aware prototype editing. *arXiv preprint arXiv:1806.07042*.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298.

Bangpeng Yao, Dirk Walther, Diane Beck, and Li Fei-Fei. 2009. Hierarchical mixture of classification experts uncovers interactions between brain regions. In *Advances in Neural Information Processing Systems*, pages 2178–2186.

Semih Yavuz, Abhinav Rastogi, Guanlin Chao, Dilek Hakkani-Tür, and Amazon Alexa AI. 2018. Deepcopy: Grounded response generation with hierarchical pointer networks. *ConvAI Workshop@NIPS*.

Yury Zemlyanskiy and Fei Sha. 2018. Aiming to know you better perhaps makes me a more engaging dialogue partner. *CoNLL 2018*, page 551.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. 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 *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.