MIME: MIMicking Emotions for Empathetic Response Generation

Navonil Majumder[†], Pengfei Hong[†], Shanshan Peng[†*], Jiankun Lu[†*], Deepanway Ghosal[†], Alexander Gelbukh[◇], Rada Mihalcea[△], Soujanya Poria[†]

[†] Singapore University of Technology and Design, Singapore
[◇] CIC, Instituto Politécnico Nacional, Mexico
[△] University of Michigan, USA
{navonil_majumder, sporia}@sutd.edu.sg,
{shanshan_peng, jiankun_liu}@mymail.sutd.edu.sg,
{pengfei_hong, deepanway_ghosal}@mymail.sutd.edu.sg,
gelbukh@cic.ipn.mx, mihalcea@umich.edu

Abstract

Current approaches to empathetic response generation view the set of emotions expressed in the input text as a flat structure, where all the emotions are treated uniformly. We argue that empathetic responses often mimic the emotion of the user to a varying degree, depending on its positivity or negativity and content. We show that the consideration of this polarity-based emotion clusters and emotional mimicry results in improved empathy and contextual relevance of the response as compared to the state-of-the-art. Also, we introduce stochasticity into the emotion mixture that yields emotionally more varied empathetic responses than the previous work. We demonstrate the importance of these factors to empathetic response generation using both automatic- and human-based evaluations. The implementation of MIME is publicly available at https://github.com/declare-lab/MIME.

1 Introduction

Empathy is a fundamental human trait that reflects our ability to understand and reflect the thoughts and feelings of the people we interact with. In the social sciences, research on empathy has evolved into an entire field of study, addressing the social underpinning of empathy (Singer and Lamm, 2009), the cognitive and emotion aspects of empathy (Smith, 2006), and its connection to personal and demographic traits (Dymond, 1950; Eisenberg et al., 2014; Krebs, 1975). The study of empathy has found a wide range of applications in healthcare, including psychotherapy (Bohart and Greenberg, 1997) or more broadly as a mechanism to improve the quality of care (Mercer and Reynolds, 2002).

Computational models of empathy have been proposed only in recent years, partly because of the complexity of this behavior which makes it difficult to emulate with computational approaches. In natural language processing, the methods proposed to date address the tasks of understanding expressions of empathy in newswire (Buechel et al., 2018), counseling conversations (Pérez-Rosas et al., 2017), or generating empathy in dialogue (Shen et al., 2020; Lin et al., 2019). Work has also been done on the construction of empathy lexicons (Sedoc et al., 2020) or large empathy dialogue datasets (Rashkin et al., 2019).

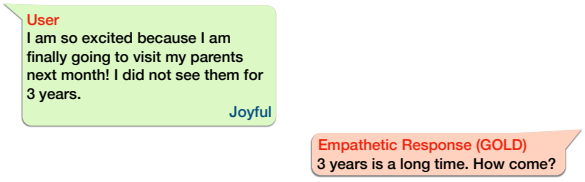


Figure 1: An instance where a positive context is responded with ambivalence.

In this paper, we address the task of generating empathetic responses that mimic the emotion of the speaker while accounting for their affective charge (positive or negative). We adopt the idea of emotion mixture, as the state-of-the-art MoEL (Lin et al., 2019), to achieve the appropriate balance of emotions in positive and negative emotion groups. However, inspired by Serban et al. (2017), we introduce stochasticity into the mixture at emotion-group level for varied responses. This becomes particularly important in cases where the input utterance can be responded with ambivalent, yet befitting utterances. Fig. 1 shows one such example where the response to a positive utterance is ambivalent.

The paper makes two important contributions. First, it introduces a new approach for empathetic generation that encodes context and emotions, and uses emotion stochastic sampling and emotion

[*] signifies equal contribution

mimicry to generate responses that are appropriate and empathetic for positive or negative statements. We show that this approach leads to performance exceeding the state-of-the-art when trained and evaluated on a large empathy dialogue dataset. Second, through extensive feature ablation experiments, we shed light on the role played by emotion mimicry and emotion grouping for the task of empathetic response generation.

2 Related Work

Open domain conversational models have made good progress in recent years (Serban et al., 2016; Vinyals and Le, 2015; Wolf et al., 2019). Many of them can generate persona-consistent (Zhang et al., 2018) and diverse (Cai et al., 2018) responses, but those are not necessarily empathetic.

Producing empathetic responses requires apt handling of emotions and sentiments (Fung et al., 2016; Winata et al., 2017; Bertero et al., 2016). Zhou et al. (2018) model psychological concepts as memory states in LSTM (Hochreiter and Schmidhuber, 1997) and employ emotion-category embeddings in the decoding process. Wang and Wan (2018) presents a GAN (Goodfellow et al., 2014) based framework with emotion-specific generators. On a larger scale, (Zhou and Wang, 2018) use the emojis in Twitter posts as emotion labels and introduce an attention-based (Luong et al., 2015) Seq-to-Seq (Sutskever et al., 2014) model with Conditional Variational Autoencoder (Sohn et al., 2015) for emotional response generation. However, they only produce affective responses with user-provided emotion, which may not necessarily be empathetic to the speakers. Wu and Wu (2019) introduce a dual-decoder network to generate responses with given sentiment (positive or negative). Shin et al. (2020) formulate a reinforcement learning problem to maximize user's sentimental feeling towards the generated response. Lin et al. (2019) present an encoder-decoder model with each emotion having a dedicated decoder.

Variational Bayes (Kingma and Welling, 2013; Rezende et al., 2014) has been widely adopted into natural language generation tasks (Bowman et al., 2015) and successfully extended to dialog generation tasks (Serban et al., 2017).The prominent approach by Hierarchical Encoder-Decoder (VHRED) (Serban et al., 2017) integrates VAE with the sequence-to-sequence decoder based on Markov assumptions.

3 Methodology

Our model MIME is based on the assumption that empathetic responses often mimic the emotion of the speaker (Carr et al., 2003) — in our case, the human subject or user. For example, positively-charged utterances are usually responded with positive emotions, although they can also be ambivalent as illustrated in Fig. 1. On the other hand, responding to negatively-charged utterances often requires composite emotions that agree with the user's emotion, but also tries to comfort them with some positivity, such as hopefulness or silver lining. As such, we strive to balance the mimicry of context/user emotion during empathetic response generation.

To this end, we first obtain context representation using a transformer encoder architecture (Vaswani et al., 2017). Similar to the state-of-the-art (SOTA) model MoEL (Lin et al., 2019), we enforce emotion understanding in the context representation by classifying user emotion during training. For the response emotion, we first group the 32 emotions into two groups containing *positive* and *negative* emotions (Section 3.3). Next, a probability distribution of emotions is sampled for each of these groups that corresponds to the emotion of the response. Positive and negative response emotion representations are formed from these distributions and emotion embeddings. These two representations are appropriately combined to balance the two kinds of emotions to form the emotional representation that drives the emotional state during response generation using transformer decoder (Vaswani et al., 2017). Fig. 2 shows the architecture of our model.

3.1 Task Definition

Given the context utterances $[u_0, u_1, \ldots, u_{n-1}]$, where utterance $u_i = [w_0^i, w_1^i, \ldots, w_{m-1}^i]$ consists of maximum m words, the task is to generate an empathetic response to the last utterance u_{n-1}, which is always from the target speaker or user. All the even-numbered $(u_0, u_2, \ldots)$ and odd-numbered $(u_1, u_3, \ldots)$ utterances belong to the user and the empathetic agent, respectively. Optionally, the context/user emotion e can be predicted for emotion understanding. The emotions are listed in Table 1.

3.2 Context Encoding

Following the MoEL system (Lin et al., 2019), firstly, the contextual utterances are sequentially concatenated into a string of k $(\leq mn)$

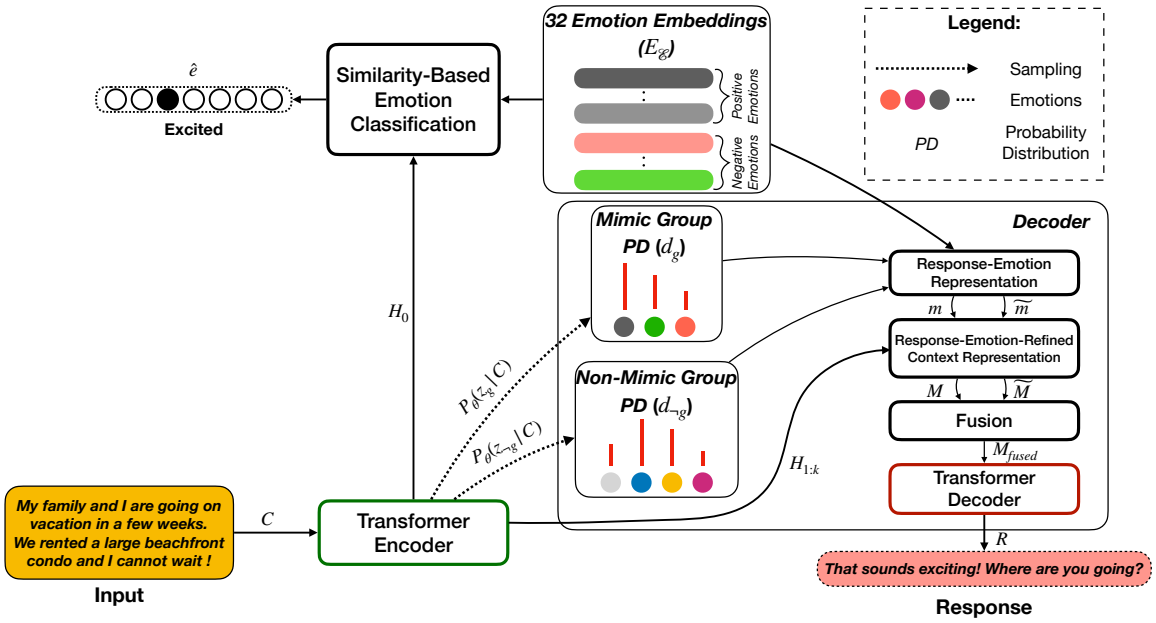


Figure 2: Architecture of our model (MIME).

words $C = [u_0 \oplus u_1 \oplus \cdots \oplus u_{n-1}] = [w_0^0, w_1^0, \ldots, w_0^1, w_1^1, \ldots, w_{m-1}^{n-1}]$, where $\oplus$ is the concatenation operator.

As in MoEL, each word w_j^i is represented as a sum of three embeddings (E_C): semantic word embedding (E_W), positional embedding (E_P), and speaker embedding (E_S), where $E_W(w_j^i), E_P(w_j^i), E_S(w_j^i) \in \mathbb{R}^{D_{emb}}$. Therefore, the context C is represented as

$$E_C(C) = E_W(C) + E_P(C) + E_S(C), \quad (1)$$

where $E_C(C) \in \mathbb{R}^{k \times D_{emb}}$.

Also as in MoEL, a transformer encoder (Vaswani et al., 2017) is used for context propagation within the utterances and words in C. Moreover, inspired by BERT (Devlin et al., 2019), one additional token CTX is prepended to the context sequence C to encode the entirety of the context:

$$H = \text{TR}_{\text{Enc}}^{ctx}(E_C([CTX] \oplus C])), \quad (2)$$

where $\text{TR}_{\text{Enc}}^{ctx}$ is the transformer encoder of output size D_h and $H \in \mathbb{R}^{(k+1) \times D_h}$ contains the context-enriched representations of the contextual words. A context-enriched representation of the CTX token, c, is taken as the overall context representation:

$$c = H_0. \quad (3)$$

Emotion Embedding and Classification. As in MoEL and also as in Rashkin et al. (2018), to explicitly infuse emotion into the context representation c, we train a emotion classifier on c. We train emotion embeddings $E_{\mathcal{E}} \in \mathbb{R}^{n_{emo} \times D_h}$ ($n_{emo} = 32$ is the number of emotion classes) to represent each emotion. We maximize the similarity between c and the user-emotion representation $E_{\mathcal{E}}(e)$, e being the user-emotion label, using cross-entropy loss L_{cls}:

$$s = E_{\mathcal{E}} W_{\mathcal{E}} c^T, \quad (4)$$

$$\mathcal{P} = \text{softmax}(s), \quad (5)$$

$$L_{cls} = -\log \mathcal{P}[e], \quad (6)$$

where $W_{\mathcal{E}} \in \mathbb{R}^{D_h \times D_h}$ and $s, \mathcal{P} \in \mathbb{R}^{n_{emo}}$.

3.3 Response Generation (Decoder)

Our primary assumption behind this model is that the empathetic agent mimics the user's emotion to some degree during response. Specifically, positive emotion is often responded with closely positive response. Negative emotion, however, is likely responded with negativity mixed with some positivity to uphold the moral.

Emotion Grouping. We split the 32 emotion types into two groups containing 13 *positive* and 19 *negative* emotions, as listed in Table 1. This split is guided by our intuition.

Response-Emotion Sampling. There is more than one correct way to respond empathetically. However, we observed that the SOTA model, MoEL, often resorts to generic and repetitive, although empathetic, responses. Therefore, inspired by Serban et al. (2017), we introduce stochasticity

Positive	Negative
confident, joyful, grateful, impressed, proud, excited, trusting, hopeful, faithful, prepared, content, surprised, caring	afraid, angry, annoyed, anticipating, anxious, apprehensive, ashamed, devastated, disappointed, disgusted, embarrassed, furious, guilty, jealous, lonely, nostalgic, sad, sentimental, terrified

Table 1: 32 emotions are split into two groups by emotional positivity and negativity.

in the response-emotion determination that results in emotionally more varied responses. In Table 7, we present responses generated by MIME with and without stochasticity. To this end, we sample response-emotion distributions d_{pos} and d_{neg}, from the context C — specifically, c in $Eq.$ (3) —, for both positive and negative emotion groups, respectively. Hence, we sample an unnormalized distribution z_g ($g \in \{pos, neg\}$) from distribution $P_\theta(z_g|C)$. This z_g is passed to a fully-connected layer (FC_{d_g}) with softmax activation to obtain the normalized distribution $d_g \in \mathbb{R}^{n_g}$ ($n_{pos} = 13$ and $n_{neg} = 19$):

$$P_\theta(z_g|C) = \mathcal{N}(\mu_g^{\text{prior}}(C), \sigma_g^{\text{prior}}(C)), \quad (7)$$

$$z_g \sim P_\theta(z_g|C), \quad (8)$$

$$d_g = \text{softmax}(\text{FC}_{d_g}(z_g)). \quad (9)$$

The emotion representation for each emotion group, $e_g \in \mathbb{R}^{D_h}$, is obtained by pooling the corresponding emotion embeddings using the respective distribution d_g:

$$e_g = d_g E_{\mathcal{E}_g}, \quad (10)$$

where $E_{\mathcal{E}_g} \in \mathbb{R}^{n_g \times D_h}$ are emotion embeddings in the emotion group g — as defined in Table 1.

Sampling from distribution $P_\theta(z_g|C)$ is reparameterized as follows:

$$c' = \text{ReLU}(\text{FC}_{sample}(c)), \quad (11)$$

$$\mu_g^{\text{prior}}(C) = \text{FC}_{\mu_g}(c'), \quad (12)$$

$$\sigma_g^{\text{prior}}(C) = \exp(0.5\,\text{FC}_{\sigma_g}(c')), \quad (13)$$

$$r \sim \mathcal{N}(0, I), \quad (14)$$

$$z_g = \mu_g^{\text{prior}}(C) + r \odot \sigma_g^{\text{prior}}(C), \quad (15)$$

where $g \in \{pos, neg\}$, FC_* are fully-connected layers with output sizes D_h. Following Serban et al. (2017), $P_\theta(z_g|C)$ is obtained by maximizing the evidence lower-bound ($-L_g^{ELBO}$):

$$L_g^{ELBO} = \text{KL}[Q_\psi(z_g|e_g, C)||P_\theta(z_g|C)] - \mathbb{E}_{Q_\psi(z_g|e_g,C)}[\log P_\theta(e_g|z_g, C)], \quad (16)$$

where $Q_\psi(z_g|e_g, C)$ is the approximate posterior distribution, defined as:

$$Q_\psi(z_g|e_g, C) = \mathcal{N}(\mu_g^{\text{posterior}}(e_g, C), \sigma_g^{\text{posterior}}(e_g, C)), \quad (17)$$

which is similarly reparameterized, for sampling during the training only, as $P_\theta(z_g|C)$, except e_g is concatenated to c.

Emotion Mimicry. Following Carr et al. (2003), it is reasonable to assume that the empathetic response to an emotional utterance would likely mimic the emotion of the user to some degree. Responding empathetically to positive utterances usually requires positivity, occasionally including ambivalence (Fig. 1). On the other hand, the responses to negative utterances should contain some empathetic negativity, but mixed with some positivity to soothe the user's negativity. Thus, we generate two distinct response-emotion-refined context representations — mimicking and non-mimicking — that are appropriately merged to obtain response-decoder input.

Naturally, mimicking and non-mimicking emotion representations — m and $\widetilde{m}$ — are defined as follows:

$$m = e_{pos} \text{ if } e \text{ is positive, otherwise } e_{neg}, \quad (18)$$

$$\widetilde{m} = e_{neg} \text{ if } e \text{ is positive, otherwise } e_{pos}. \quad (19)$$

Firstly, response-emotion representations — m and $\widetilde{m}$ — are concatenated to the context-enriched word representations in $H_{1:k}$ (Eq. (2)) to provide the context (C) the cues on the response emotion:

$$H_{resp} = [H_i \oplus m]_{i=1}^k, \quad (20)$$

$$\widetilde{H}_{resp} = [H_i \oplus \widetilde{m}]_{i=1}^k, \quad (21)$$

where $H_{resp}, \widetilde{H}_{resp} \in \mathbb{R}^{k \times 2D_h}$ are fed to a transformer encoder ($\text{TR}_{\text{Enc}}^{resp}$) to obtain mimicking and non-mimicking response-emotion-refined context representations M and $\widetilde{M}$, respectively:

$$M = \text{TR}_{\text{Enc}}^{resp}(H_{resp}), \quad (22)$$

$$\widetilde{M} = \text{TR}_{\text{Enc}}^{resp}(\widetilde{H}_{resp}), \quad (23)$$

where $M, \widetilde{M} \in \mathbb{R}^{k \times D_h}$.

Response-Emotion-Refined Context Fusion. Enabling a mixture of positive and negative emotions could lead to diverse response generation as compared to considering exclusively positive

or negative emotions. To achieve this mixture, we concatenate M and $\widetilde{M}$ at word level, as opposed to sequence level, to obtain $M' \in \mathbb{R}^{k \times 2D_h}$. Then, M' is fed to a gate consisting of a fully-connected layer ($\text{FC}_{contrib}$) with sigmoid activation, resulting $M_{contrib}$ that determines the contribution of postive and negative response-emotion-refined contexts to the response to be generated. Subsequently, M' is multiplied with the gate output, yielding the refined context M_{adjust} that is fed to another fully-connected layer FC_{fused} to obtain the fused response-emotion-refined context $M_{fused} \in \mathbb{R}^{k \times D_h}$:

$$M' = [M_i \oplus \widetilde{M}_i]_{i=0}^{k-1}, \tag{24}$$

$$M_{contrib} = \sigma(\text{FC}_{contrib}(M')), \tag{25}$$

$$M_{adjust} = M_{contrib} \odot M', \tag{26}$$

$$M_{fused} = \text{FC}_{fused}(M_{adjust}). \tag{27}$$

Response Decoding. For the final response generation from the response-emotion-refined context M_{fused}, following MoEL, a transformer decoder ($\text{TR}_{\text{Dec}}^{resp}$), with M_{fused} as key and value, is employed:

$$O = \text{TR}_{\text{Dec}}^{resp}(E_W(R_{0:t-1}), M_{fused}, M_{fused}), \tag{28}$$

$$\mathcal{P}_{resp} = \text{softmax}(\text{FC}_{decode}(O)), \tag{29}$$

$$p(R_i|C, R_{0:i-1}) = \mathcal{P}_{resp}[i], \tag{30}$$

where $O \in \mathbb{R}^{t \times D_h}$, t is the number tokens in response R (R_0 is <start> token), FC_{decode} is a fully-connected layer of output size $|V|$ — also the vocabulary size —, $\mathcal{P}_{resp} \in \mathbb{R}^{t \times |V|}$, and $p(R_i|C, R_{0:i-1})$ is the probability distribution on each response token.

Finally, categorical cross-entropy quantifies the generation loss with respect to the gold response R_{gold}:

$$L_{resp} = -\log p(R_{gold}|C). \tag{31}$$

3.4 Training

Naturally, we combine all the losses for model training:

$$\mathcal{L} = \alpha L_{cls} + \beta(L_{pos}^{ELBO} + L_{neg}^{ELBO}) + \gamma L_{resp}. \tag{32}$$

Total loss $\mathcal{L}$ is optimized using Adam (Kingma and Ba, 2015) optimizer with learning-rate, patience, and batch-size set to 0.0001, 2, and 16, respectively. Loss weights, α, β, and γ are set to 1. For the sake of comparability with the SOTA, the semantic word embeddings (E_W) are initialized with GloVe (Pennington et al., 2014) embeddings. All the hyperparameters are optimized using grid search on the validation set, resulting D_h and beam-size being 300 and 5, respectively.

4 Experimental Settings

During inference, we use the emotion classifier (Eq. (5)) with emotion grouping (Table 1) to determine the positivity or negativity of the context that is necessary for the mimicking and non-mimicking emotion representations.

4.1 Dataset

We evaluate our method on EMPATHETICDIALOGUES[1] (Rashkin et al., 2018), a dataset that contains 24,850 open-domain dyadic conversations between two users, where one responds emphatically to the other. For our experiments, we use the 8:1:1 train/validation/test split, defined by the authors of this dataset. Each sample consists of a context — defined by an excerpt of a full conversation and the emotion of the user — and the empathetic response to the last utterance in the context. There are a total of 32 different emotion categories roughly uniformly distributed across the dataset.

4.2 Baselines and State of the Art

We do not compare MIME with affective response generation models (Zhou et al., 2018) as they require the response emotion to be explicitly provided, and the response may not necessarily be empathetic. As such, MIME is compared against the following models:

Multitask-Transformer Network (Multi-TR). Following Rashkin et al. (2018), a transformer encoder-decoder (Vaswani et al., 2017) generates a response as the user emotion is classified from the encoder output — equivalent to c in Eq. (3).

Mixture of Empathetic Listeners (MoEL). This state-of-the-art method (Lin et al., 2019) performs user-emotion classification as Multi-TR. However, in contrast to our method, it employs emotion-specific decoders whose outputs are aggregated and fed to a final decoder to generate the empathetic response.

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

4.3 Evaluation

Although BLEU (Papineni et al., 2002) has long been used to compare system-generated response against the human-gold response, Liu et al. (2016) argues against its efficacy in open-domain where the gold response is not necessarily the only correct response. Therefore, as MoEL, we keep BLEU mostly as reference. Following MoEL and Rashkin et al. (2018), we rely on human-evaluated metrics:

Human Ratings. Firstly, we randomly sample four instances of each of the 32 emotion labels from the test set, resulting in a total of 128 instances, compared to the 100 instances used for the evaluation of MoEL. Next, we ask three human annotators to rate each sub-sampled model response on a scale from 1 to 5 (best score) on three distinct attributes: **empathy** (How much *emotional understanding* does the response show?), **relevance** (How much *topical relevance* does the response have to the context?), and **fluency** (How much *linguistic clarity* does the response have?). Scores across 128 samples and three annotators are averaged to obtain the final rating.

Human A/B Test. Given two models A and B — in our case MoEL and MIME (our model), respectively — we ask three human annotators to pick the model with the best response for each of the 128 sub-sampled test instances. The annotators can select a *Tie* if the responses from both models are deemed equal. The final verdict on each instance is determined by majority voting. In case no two annotators agree on a selection – that is all three annotators reached three distinct conclusions: MoEL, MIME, and Tie – we bring in a fourth annotator. From this, we calculate the percentage of samples where A or B generates the better response and where A and B are equal.

5 Results and Discussions

5.1 Response-Generation Performance

Methods	#params.	BLEU	Human Ratings		
			Emp.	Rel.	Flu.
Multi-TR	16.95M	2.92	3.67	3.47	4.30
MoEL (SOTA)	23.10M	2.90	3.71	3.32	**4.31**
MIME	17.80M	2.98	**3.87**	**3.60**	4.28

Table 2: Comparison among MIME (our model), baselines, and the state-of-the-art MoEL; Emp., Rel., and Flu. stand for Empathy, Relevance, and Fluency, respectively; the best score for each metric is highlighted by bold font.

MIME vs.	MIME Wins	MIME Loses	Tie
Multi-TR	42.25%	24.60%	33.15%
MoEL	38.82%	28.32%	32.86%
MIME w/o STC	39.84%	23.43%	36.73%

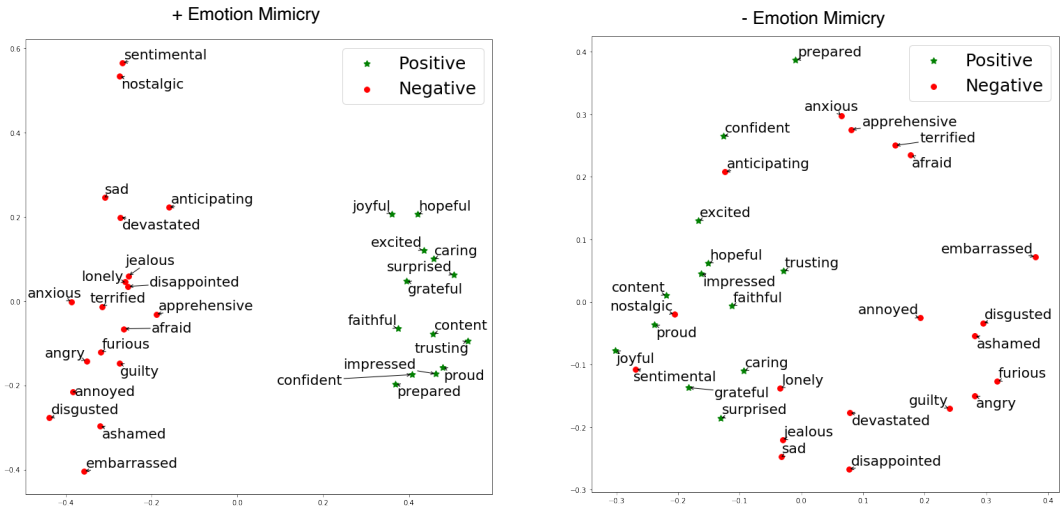
Table 3: Human A/B test results for MIME vs. MIME without stochasticity (STC), MoEL, and Multi-TR.

Following Table 2, responses from MIME show improved *empathy* over MoEL and Multi-TR. We surmise this was achieved by modeling our primary intuition of appropriately mimicking user's emotion in the context thorough stochasticity and positive/negative grouping. Moreover, the usage of trained emotion embeddings ($E_{\mathcal{E}}$), shared between the emotion classifier and response decoder, seems to encode refined context-invariant emotional and emotion-specific linguistics cues that may lead to empathetically-improved response generation. The SOTA model, MoEL, does train a similar emotion embedding, but it is setup as the key of a key-value memory (Miller et al., 2016) which leads to a weaker connection with the decoder, resulting in less emotional-context flow. We believe this embedding sharing further leads to improved *relevance* rating for MIME, since contextual information flow is now shared between emotion embeddings and encoder output (Eq. (2)). This sharing intuitively leads to refinement in context flow.

However, we also observe that the responses from our model have worse *fluency* than the other models, including MoEL. This might be attributed to the very structure of the decoder, that seems to refine emotional context well. This may have coerced the final transformer decoder to focus more on emotionally-apt tokens of the response than appropriate stop-words that have no intrinsic emotional content, but lead to grammatical clarity.

Human A/B Test. Based on the results in Table 3, we note that the responses from MIME are more often preferable to humans than the responses from MoEL and Multi-TR. This correlates with the results in Table 2 that indicate better empathy and contextual relevance for MIME. Further, the annotators prefer the responses from MIME with stochasticity (STC) than otherwise. Table 7 shows the impact of stochasticity on the responses.

Performance on Positive and Negative User Emotions. We observe (Table 4) that the responses generated by MIME for both positive and negative user emotions are generally better in terms



(a) (b)

Figure 3: (a) and (b) plot the emotion embeddings, with and without emotion mimicry, respectively, mapped to two dimensions using top-two principal components.

Ratings	Multi-TR		MoEL		MIME	
	Pos	Neg	Pos	Neg	Pos	Neg
Empathy	3.77	3.61	3.73	3.76	**4.00**	**3.80**
Relevance	3.51	3.45	3.21	3.40	**3.77**	**3.49**
Fluency	4.33	4.28	**4.35**	**4.30**	4.33	4.26

Table 4: Models performance on positive and negative context; Pos and Neg stand for positive and negative emotions, respectively; the best score for each metric-polarity combination is highlighted by bold font.

of empathy and fluency. Interestingly, MoEL seems to perform better on responding to negative emotions than to positive emotions in terms of empathy and fluency. We posit this stems from the abundance of negative samples in the dataset as compared to positive samples — 13 positive and 19 negative emotions roughly uniformly distributed. This may suggest that MoEL is more sensitive to positive/negative context imbalance in the dataset than MIME and Multi-TR.

5.2 Ablation Study

Emotion Mimicry	Emotion Grouping	BLEU	Human Ratings		
			Emp.	Rel.	Flu.
✗	✗	2.45 ± 0.01	3.14	3.58	4.23
✗	✓	2.96 ± 0.02	3.67	**3.63**	4.09
✓	✓	2.98 ± 0.01	**3.87**	3.60	**4.28**

Table 5: Results of ablation; Emp., Rel., and Flu. stand for Empathy, Relevance, and Fluency, respectively.

Effect of Emotion Mimicry. To assess the contribution of user-emotion mimicry, we disabled it by passing e_g (Eq. (10)) directly to Eqs. (20) and (21). This results in a substantial drop in *empathy*, by 0.2 as per Table 5. We delve deeper by plotting the emotion embeddings produced with and without emotion mimicry in Fig. 3a and Fig. 3b, respectively. It is evident that the separation of positive and negative emotions clusters is much clearer with emotion mimicry than otherwise, suggesting better emotion understanding in the prior case through emotion disentanglement. On the other hand, we observe slight increase of *relevance*, by 0.03. We surmise this is caused by the absence of the confounding effect of swapping the value of m and $\widetilde{m}$, in Eqs. (18) and (19), depending on the user emotion type. This may coerce the same set of parameters to learn processing both positive and negative emotions.

Effect of Emotion Grouping. Looking at Table 5, we observe a performance drop in both *empathy* and *relevance*, by 0.73 and 0.02, respectively, in the absence of emotion grouping. This indicates the importance of having positive and negative emotions treated separately, rather than huddling them into a single distribution. We posit that the latter case causes all the emotions to compete for importance which may lead to emotion uniformity in some cases or one emotion-type overwhelming the other in other cases. This in turn may lead to emotionally mundane and generic responses.

5.3 Case Study

Context Capturing. Based on the comparative results for *relevance* shown in Table 2, MIME appears to generate responses that are a closer fit to

Label	lonely
Context	**Speaker**: i there , dont know what to do , jst broke up with my girlfirned , we were 8 years together
MIME	oh no ! what happened ?
MoEL	oh wow , that must have been very exciting !
Gold	sorry to hear ! do you have any idea about the break up ? did you think about it ?
Label	excited
Context	**Speaker**: i have a huge chance to win a ps4 !
MIME	that is great ! are you excited ?
MoEL	that sounds like a lot of fun
Gold	wow , please make it real .
Label	faithful
Context	**Speaker**: there was this one time i was fairly drunk at a bar and this attractive girl was trying to make moves on me , but i did my best to ignore it all.. having a girlfriend and all .
MIME	wow , that must have been so **embarrassing** .
MoEL	that is terrible, did you confront her ?
Gold	that is good ! you should be proud of yourself
Label	disappointed
Context	**Speaker**: i lost my favorite hoodie somewhere . i am so upset . like how could i lose a hoodie ? it is so big , did i just leave it somewhere ?
MIME	well that is not good , i hope you have to be careful next time
MoEL	i am sorry to hear that . i am sorry to hear that . i am sorry to hear that .
Gold	did you retrace your steps

Table 6: Comparison between MIME (*with STC*) and MoEL.

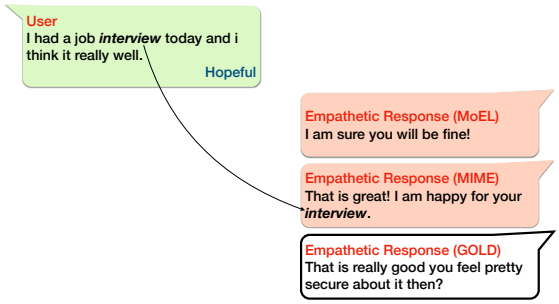


Figure 4: A test sample where MIME responds with key information from the context.

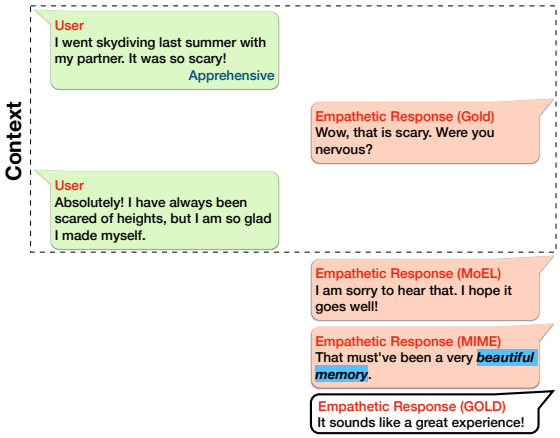


Figure 5: A test sample where MIME responds with subtle information from the context.

the context than MoEL does. Fig. 4 shows a test instance where MIME pulls key information from the context — the word 'interview' — to generate an empathetic and relevant response. The response from MoEL is also empathetic, but somehow more generic. We surmise that this can be attributed to the two-way context flow through the emotion embedding sharing and encoder output, as discussed in Section 5.1.

Similarly, Fig. 5 shows a conversation with an *apprehensive* user who shares a frightening story with a positive outcome. Here, MoEL fixates on the initial negative emotion of the user and replies with an unwarranted negatively empathetic response. MIME, however, responds with appropriate positivity hinted at the last utterance. Moreover, it is able to correctly interpret the events described as a 'beautiful memory', which is truly empathetic and relevant. Again, strong mixture of context and emotion, facilitated by the emotion embedding sharing, is likely to be responsible for this. We show more examples generated by both MoEL and MIME in Table 6.

5.4 Error Analysis

Low Fluency. As evidenced by Table 2, MIME falters in *fluency* as compared to MoEL. Fig. 6 shows an instance where MoEL generates an empathetic, yet somewhat generic, and fluent response. In contrast, the first response utterance from MIME — "*I would have been to the police*" — does make contextual sense. However, the second utterance "*I would be a little better*" reads incoherent and semantically unclear. Perhaps the model meant something like 'I would have felt a little safer'. We repeatedly observed such errors, leading to poor fluency. Given the empathy- and relevance-focused structure of our model, we think MIME focused

Label	anticipating
Context	**Speaker**: i am looking forward to going on vacation in a few weeks ! we have a condo reserved on the beach , with fantastic ocean views . i am ready !
with *STC*	that is **awesome** ! i have been there . i hope you have a great time !
no *STC*	that is great ! i have never been there .
Gold	ah , that sounds fantastic ! which ocean will you be enjoying ?
Label	jealous
Context	**Speaker**: my friend is a surgeon and we were discussing salary . he easily makes 200,000 a year and he is only 32. it is crazy and i was jealous of him .
with *STC*	i would be so **jealous** !
no *STC*	that is a good idea .
Gold	you should n't , because that is a very stressful job
Label	proud
Context	**Speaker**: my son graduated .
with *STC*	**congrats** ! that is a **great accomplishment** !
no *STC*	that is great ! how old is he ?
Gold	from where ?

Table 7: Comparison between some responses from MIME with and without stochasticity (*STC*).

Figure 6: A test sample where MIME responds with a malformed utterance.

on learning empathy and relevance, at the cost of fluency. We believe this issue could be mitigated with additional training samples.

Response to Surprised User Context. In our experiments, we assumed the emotion *surprised* to be positive (Table 1), and thus MIME responds with positivity to most test instances incurring *surprise* as a user emotion. However, this is not accurate, as one can be both positively and negative surprised — "*I recently found out that a person I [...] admired did not feel the same way. I was pretty surprised"* vs "*This mother's day was amazing!*". We posit that re-annotating the instances with a negatively-surprised user with a new negative emotion, namely *shocked*, should help alleviate this issue significantly.

Emotion Classification. The {top-1, top-2, top-5} emotion-classification accuracies for MoEL are {38%, 63%, 74%}, as compared to {34%, 58%, 77%} for MIME. Since the emotion embeddings are shared between encoder and decoder in MIME, it supposedly also encodes some generation-specific information in addition to pure emotion as discussed in Section 5.1, thereby hindering the overall emotion-classification performance. Notably, MIME also performs well on top-5 classification. This is likely due to MIME's ability to discern positive and negative emotion types — as indicated by Fig. 3a — that comes into prominence as you add more likely-labels into the consideration of top-k classification by raising k.

6 Conclusion

This paper introduced a novel empathetic generation strategy that relies on two key ideas: emotion grouping and emotion mimicry. Also, stochasticity was applied to the emotion mixture for varied response generation. We have shown through several human evaluations and ablation studies that our model is better equipped for empathetic response generation than existing models. However, there remains much room for improvement, particularly in terms of *fluency* where our model falters. Moreover, emotions like 'surprise' and 'anticipation' might be explicitly dealt with due to their ambiguous polarity.

Acknowledgements

This research is supported by A*STAR under its RIE 2020 AME programmatic grant RGAST2003 , and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of A*STAR, the National Science Foundation, or the John Templeton Foundation.

References

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, Austin, Texas. Association for Computational Linguistics.

A. Bohart and L. Greenberg. 1997. *Empathy reconsidered: New directions in psychotherapy*. American Psychological Association.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

S. Buechel, A. Buffone, B. Slaff, L. Ungar, and J. Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Belgium.

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.

Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C. Mazziotta, and Gian Luigi Lenzi. 2003. Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences*, 100(9):5497–5502.

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.

R. Dymond. 1950. Personality and empathy. *Journal of Consulting Psychology*, 14(5).

N. Eisenberg, T. Spinrad, and A. Morris. 2014. *Empathy-related responding in children*. Psychology Press.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the Supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91, San Diego, California. Association for Computational Linguistics.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

D. Krebs. 1975. Empathy and altruism. *Journal of Personality and Social psychology*, 32(6).

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation.

S. Mercer and W. Reynolds. 2002. Empathy and quality of care. *British Journal of General Practice*, 52.

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, Austin, Texas. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014,*

Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1532–1543. ACL.

V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, and L. An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the Association for Computational Linguistics*, Vancouver, Canada.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proceedings of the Association for Computatinal Linguistics*.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

J. Sedoc, S. Buechel, Yehonathan Y. Nachmany, A. Buffone, and L. Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of The 12th Language Resources and Evaluation Conference*, France.

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

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3295–3301. AAAI Press.

S. Shen, C. Welch, R. Mihalcea, and V. Perez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2020)*, Boise, Idaho.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7989–7993.

T. Singer and C. Lamm. 2009. The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, 1156.

A. Smith. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record*, 56(1).

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452. International Joint Conferences on Artificial Intelligence Organization.

Genta Indra Winata, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Nora the empathetic psychologist. In *INTERSPEECH*.

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

Xiuyu Wu and Yunfang Wu. 2019. A simple dual-decoder model for generating response with sentiment. *arXiv preprint arXiv:1905.06597*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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, Melbourne, Australia. 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, Melbourne, Australia. Association for Computational Linguistics.