Exemplars-guided Empathetic Response Generation Controlled by the Elements of Human Communication

Navonil Majumder*
Singapore University of Technology
and Design, Singapore
navonil_majumder@sutd.edu.sg

Deepanway Ghosal*
Singapore University of Technology
and Design, Singapore
deepanway_ghosal@mymail.sutd.edu.sg

Devamanyu Hazarika
National University of Singapore
Singapore
hazarika@comp.nus.edu.sg

Alexander Gelbukh
Instituto Politécnico Nacional
Mexico City, Mexico
gelbukh@cic.ipn.mx

Rada Mihalcea
University of Michigan
Ann Arbor, Michigan, USA
mihalcea@umich.edu

Soujanya Poria
Singapore University of Technology
and Design, Singapore
sporia@sutd.edu.sg

ABSTRACT

The majority of existing methods for empathetic response generation rely on the emotion of the context to generate empathetic responses. However, empathy is much more than generating responses with an appropriate emotion. It also often entails subtle expressions of understanding and personal resonance with the situation of the other interlocutor. Unfortunately, such qualities are difficult to quantify and the datasets lack the relevant annotations. To address this issue, in this paper we propose an approach that relies on exemplars to cue the generative model on fine stylistic properties that signal empathy to the interlocutor. To this end, we employ dense passage retrieval to extract relevant exemplary responses from the training set. Three elements of human communication—emotional presence, interpretation, and exploration—and sentiment are additionally introduced using synthetic labels to guide the generation towards empathy. The human evaluation is also extended by these elements of human communication. We empirically show that these approaches yield significant improvements in empathetic response quality in terms of both automated and human-evaluated metrics. The implementation is available at https://github.com/declare-lab/exemplary-empathy.

KEYWORDS

Natural language generation, Empathy, Neural networks.

1 INTRODUCTION

Empathy is a very human characteristic often defined as the ability to understand and emotionally comprehend the feelings of others intimately. An entire field of research has grown around empathy that delves into its fundamentals [39]. Progress has been made into understanding empathy from the angle of affect [40], personality traits, and demographics [8, 9, 19]. Understanding of empathy has widespread applications in healthcare, such as in psychotherapy [2] and in general care improvement [27], among others.

Owing to its intrinsic influence on conversations, empathy has garnered much interest in the NLP community. Recent works on empathetic response generation [23, 24, 26, 32] singularly focus on establishing a relation between emotion of the contextual utterances

Context

> My dad called me one day crying out for help. I had to drive over to his house.

Exemplars

1. Why what happened?
2. How come?
3. What is the matter?
4. Oh no, why do you think like that. All should be good.
5. Oh no, why is that?

Reference

> Oh my goodness! What happened?

LEMPEx (uses exemplars)

> Oh that sounds scary! Did he say what was going on? Is he alright ?

T5-GloVe

> That's awful. Is he okay? Does it have someone you can do that.

MIME
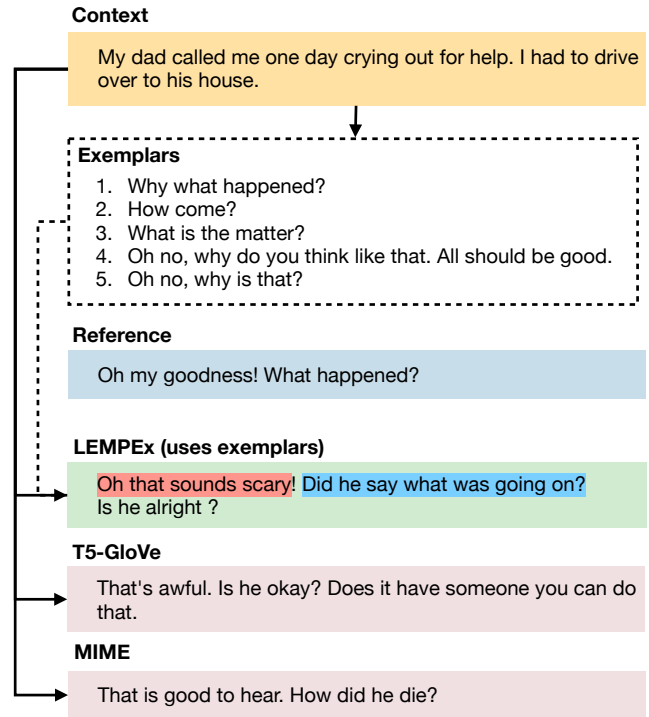
> That is good to hear. How did he die?

Figure 1: Influence of exemplars on the response.

and the response. Although response emotion is a key element of empathy, it is far from the complete picture. In fact, the very definition of empathy, which involves visceral understanding of the feeling of other people, is quite subjective and vague and, as such, is quite unwieldy for the empathetic response generation task.

To sidestep this issue, we adopt the idea of exemplars [5] for controllable text generation, where a set of sample responses from the training set, semantically related to the input context, is retrieved (using fine-tuned dense passage retrieval (DPR) [17] model) and fed to the response generator as templates [4]. These template responses guide the generator with stylistic and thematic cues on the response that are perceived as empathetic to the user. This is similar to providing hints for solving a difficult mathematical problem to

significantly narrow down the search space. Similarly, we surmise that the exemplars restrict the search space for the decoder.

Our model is called LEMPEx: Learning to EMPathize using Exemplars. Fig. 1 illustrates an instance where the support from exemplars leads to an empathetically resonant and thematically rich response from our model, LEMPEx. The phrase "*Oh that sounds scary!*" is emotionally appropriate to communicate strong empathy. Moreover, the response makes a thematically appropriate query with "*Did he say what was going on?*". Based on the similarity between the extracted exemplars and the generated response, we posit that the exemplars helped LEMPEx regulate the nature of the query and emotional intensity. The *T5-GloVe* model, on the other hand, does not use exemplars. It is probably this what leads it to a partly generic and partly unintelligible query: "*Is he okay? Does it have someone you can do that.*". "*That's awful*" phrase is not quite empathetic in this situation; rather, it might be construed as an overreaction. Finally, *MIME* also misinterprets the situation and asks an inappropriate question. We delve deeper into this in Section 5.2.1.

To further regulate the response generation, we introduce four auxiliary losses that quantify four important aspects related to empathy: *emotional presence*, *interpretation*, *exploration*, and *sentiment*. Many research works [34, 35, 46] identify the former three as key aspects of empathy and human communication in general. The role of adequate *emotional presence* is to form an emotional bond with the user. In Fig. 1, the response phrase "*Oh that sounds scary!*" is meant to achieve this. Then, correct *interpretation* is key in regulating emotion and generating context-appropriate comments and queries that exude empathy. *Exploration* through relevant and appropriate queries signals understanding and interest of the empathetic agent to the user, which is perceived as empathy by the user. The phrase "*Did he say what was going on? Is he alright?*" is meant to communicate these two attributes. Finally, inspired by the idea of emotion grouping—these groups essentially represent positive and negative sentiment—by Majumder et al. [26], we also add sentiment as an extension of *emotional presence*.

To this end, we generate synthetic labels for *emotional presence*, *interpretation*, and *exploration* attributes for the gold responses by training classifiers for them with the EMPATHYMENTALHEALTH [35] dataset. For *sentiment*, we get scores from VADER [15]. Subsequently, these labels are predicted from the decoder hidden states to prime the decoder to generate responses with these four attributes at a suitable level. These attributes may be viewed as an external signal to guide the generation process.

To gauge the empathetic quality of the generated responses at a deeper level, we also obtain human evaluation on *emotional presence*, *interpretation*, and *exploration* attributes. This gives us better sense of the generation quality at a more fundamental level. To the best of our knowledge, we are the first to delve into such metrics to assess empathy in generated responses.

Evaluation of empathetic responses is reliant on human evaluation which is resource consuming. To ameliorate this, we repurpose the aforementioned synthetic-label predictors for the empathy-related attributes *emotional presence*, *interpretation*, *exploration*, and *sentiment* as automatic evaluators on corresponding attributes. The generated response is fed to the these predictors and the resultant label is compared to the synthetic label of the gold response to

obtain macro-F1 and mean absolute error (MAE) scores. In our experiments, we observed good correlation with the other human evaluated metrics across all the discussed models that suggests the reliability of these automated metrics.

Later on in this paper, we empirically show that the combination of these approaches yields strong empathetic responses, outshining the state-of-the-art EmpDG [23] and other baselines.

In a nutshell, the contribution of this paper is three-fold:

(1) We mine exemplary responses from the training set, using DPR, to guide the response generation;
(2) Empathy-inducing attributes, namely, *emotional presence, interpretation, exploration,* and *sentiment,* are introduced as regularizer losses by the means of synthetic labels.
(3) We conduct a thorough human and automatic evaluation based on the these attributes, for better understanding of the empathy generation capability of models.

2 BACKGROUND

Text-To-Text-Transfer-Transformer (T5). Our model is based on the popular T5 model. The Text-to-text-transfer-transformer (T5) is an encoder-decoder based sequence-to-sequence transformer model [30]. In the T5 framework, every natural language processing task, including question answering, language translation, text classification, is framed as a sequence-to-sequence text generation problem. This formulation enables the use of identical architecture and loss functions without any changes across a diverse set of tasks.

T5 follows the self-attention-based encoder-decoder transformer [43] with slight modifications in the architecture and a different position embedding strategy. As opposed to using a fixed sinusoidal position embedding as the original transformer model, T5 uses relative position embeddings that are learned based on the offset between key and query vector. This relative position embedding has been found to work better in practice [14, 36]. T5 is pre-trained on the Colossal Clean Crawled Corpus (C4) dataset. A modified masked language modeling (MLM) objective function is used to pre-train the model.

Retrieval-Based Generation. Typically, generative models in NLP employ auto-regressive decoders to generate sentences. However, recent works have shown that simply relying on the model's parameters might be insufficient for specific tasks and might also lead towards undesired effects like hallucinations [38]. In particular, within the literature of controlled text generation, there is increasing focus to provide these models with additional support either through non-parametric access to retrieved content as inputs or through auxiliary losses. [13] proposed a novel strategy that performed language modeling using retrieving fluent responses from human sentences in training data and then editing them. Inspired by this, Weston et al. [47] proposed a retrieve-then-refine framework that involved retrieving relevant responses for a dialog context and then appending these retrievals along with the dialog context to train the overall generator. This simple idea demonstrated good performance to the otherwise daunting task of response generation. Similar approaches were also explored in [4] where the retrieval process was designed to ensure both textual similarity and topical match with the generated response.

While these approaches have been studied in generative applications like language modeling, open-domain dialog, or style transfer, to the best of our knowledge, we are the first to perform a systematic study of this paradigm for empathetic response generation. As we show in the results, our exemplar-guided generation for empathy becomes a better empathetic generator and a simpler solution than previous works in this application.

Related Works in Text Generation. Lately, significant strides have been made in open-domain conversational models [33, 44, 49]. Models for generating persona-consistent [51] and varied [3] responses have been developed. Their responses are not necessarily empathetic. One aspect of generating such responses involves regulating emotions and sentiments [1, 10, 48]. GAN-based [11] model for emotion-specific generation has been proposed by Wang and Wan [45]. Zhou and Wang [52] consider emojis as emotion labels. They have also proposed an attention-based [25] Seq-to-Seq [42] model with Conditional Variational Autoencoder [41] for emotion-specific response generation. They, however, require emotion prompt for response generation which may not necessarily be empathetic. Similarly, Wu and Wu [50]'s dual-decoder network responds with input sentiment. Shin et al. [37] devise a reinforcement learning formulation that maximizes the user's sentiment on the generated response. Only recently, following the construction of EmpatheticDialogues [32] dataset, work [23, 24, 26] has been done on emotion prompt-free automated empathetic response generation. However, these works heavily rely on the emotion labels of the context, which does not capture the entire picture of empathy. Therefore, we leverage exemplar-based generation that has been quite successful in many controlled text generation [5] tasks, such as paraphrasing [12, 20] or controlled translation [6]. In this context, previous work [32] has also explored retrieval-based response selection. However, we are the first to explore joint retrieval and generative models, with the former guiding the latter.

3 METHOD

The existing methods [23, 24, 26] are strongly reliant on context emotion to generate empathetic responses. However, such approaches overlook the subtle expressions of understanding and personal resonance with the situation of the other interlocutor. To alleviate this, we take two major steps:

(1) We use exemplars to cue the generation on stylistic properties that signal empathy to the interlocutor
(2) We imbue the responses with four properties, namely sentiment, emotional presence, interpretation, and exploration, that are indicative of empathy [34].

The exemplars for the former step are obtained by dense passage retrieval (DPR) [17] that extracts relevant exemplary responses from the training set for a given context. For the latter step, we firstly train individual classifiers that identify those four properties in a response. During the generative training the generated response is fed to these four classifiers. The four classification losses, based on synthetic labels, are back-propagated through these classifiers to train the generative encoder-decoder model. The resultant model, named LEMPEx, is visualized in Fig. 2.

3.1 Task Definition

We are provided with a dialog between a user and an dialog agent, where the agents's role is to provide empathetic responses to the user's statements. Each turn in the dialog alternates between the user and the agent, which forms a sequence of utterances. For n such utterances $[u_1, u_2, \ldots, u_n]$ with the user emoting an emotion e, the goal of the agent (our model) is to empathetically respond to the last utterance u_n from the user. In the mentioned sequence, all the odd-indexed $\{u_1, u_3, \ldots, u_n\}$ and even-indexed utterances $\{u_2, u_4, \ldots, u_{n-1}\}$ are uttered by the user and the empathetic agent, respectively. Thus for any instance, n is always odd, i.e., the last turn in the provided dialog context is always that of the user. The goal of the task is to generate an appropriate and empathetic response—as the $n + 1^{th}$ turn u_{n+1}—to this provided context.

3.2 Encoder-Decoder Transformer

The recent success of transfer learning from large self-supervised pre-trained models to myriad of NLP tasks had spurred us on to use text-to-text transfer transformer (T5) [31] as the basis of the encoder-decoder setup of our generative model. However, fine-tuning this pre-trained T5 model on this task counter-intuitively led to poor empathy in the responses. Hence, we resort to train the T5 model from scratch with GloVe [29] word embeddings. We posit that this may be a side-effect of T5 having been pre-trained with plethora of text from the internet that are not particularly empathetic. Therefore, nudging such network to generate empathetic text may require larger number of samples. In Section 5, we empirically further explore this.

To this end, we first prepare the context by concatenating the contextual utterances $C = [u_1 \oplus u_2 \oplus \cdots \oplus u_n] = [w_{11}, w_{12}, \ldots, w_{nm}]$, where w_{ij} is the jth word in the ith utterance. The context C is represented as the sum of word (E_W) and speaker embeddings (E_S): $E(C) = E_W(C) + E_S(C)$, where $E(C) \in \mathbb{R}^{k \times n_{emb}}$, k is the number of words in C, and n_{emb} is the embedding length of the words and speakers. The speaker embeddings are trained to represent the speakers of the utterances.

Context $E(C)$ is passed through the T5 encoder (T5_{enc}^{ctx}) and subsequently through T5 decoder (T5_{dec}) for empathetic response generation:

$$Z = \text{T5}_{enc}^{ctx}(E(C)), \quad (1)$$

$$\mathcal{P}_{resp} = \text{T5}_{dec}(E_W(R_{1:t-1}), Z), \quad (2)$$

where $Z \in \mathbb{R}^{k \times D_{emb}}$ is the encoder output holding context representation, R is the generated response of length t (R_0 is <start> token), $\mathcal{P}_{resp} \in [0, 1]^{t \times |V|}$ holds the probability distribution on each response token, and V is the vocabulary.

In this work, we add add additional modules to this base encoder-decoder T5 transformer network to accommodate retrieved exemplars (Section 3.3) and four auxiliary losses based on *sentiment*, *emotional presence*, *interpretation*, and *exploration* (Section 3.4).

3.3 Exemplar Extraction and Injection

For each input context C, we fetch a list of possible responses from the training samples that are likely to embody context-specific empathetic characteristics. The key idea here is to reduce the reliance on emotions for empathetic response generation. Instead,
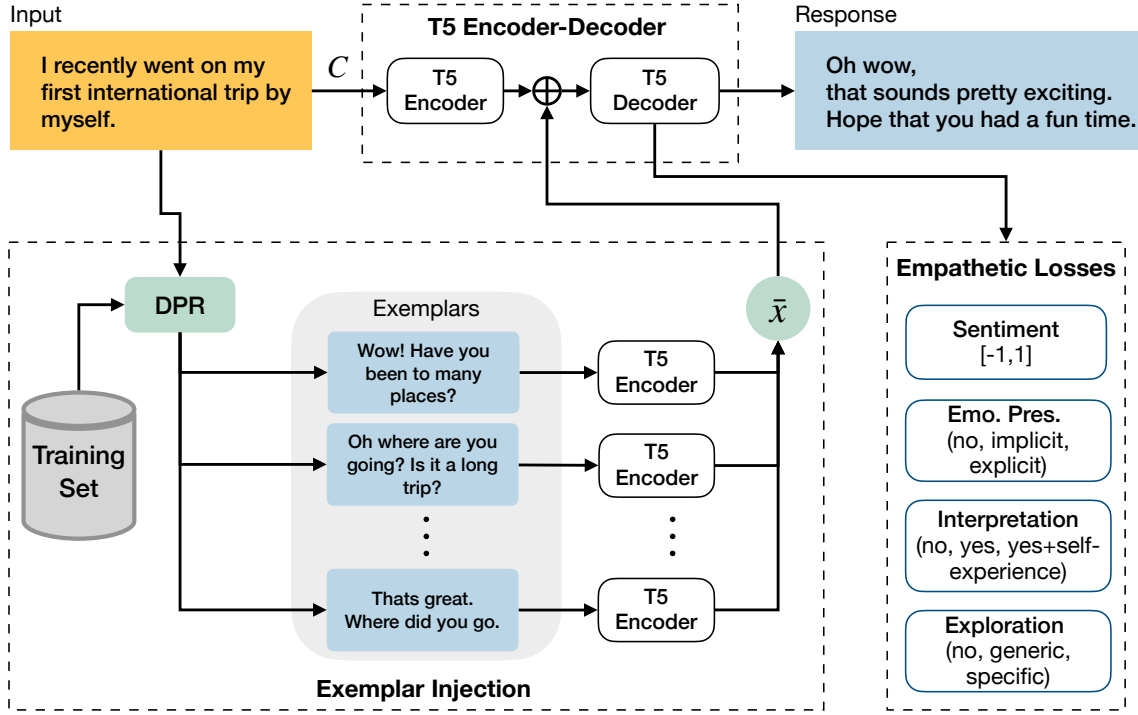
Figure 2: The structure of our model LEMPEx.

we retrieve and show the model possible set of responses, called exemplars, for guidance. Needless to say, these responses are not fully aligned to the context, but they provided important cues towards generating the final response. For example, the exemplars can provide a filtered set of candidate responses with the appropriate empathetic phrases like *I am sorry to hear that* if the user is sad about an event. We retrieve these exemplars using the popular DPR algorithm that we discuss next.

Dense Passage Retrieval. Dense Passage Retrieval (DPR) is a recent algorithm proposed for retrieval in a low-dimensional and continuous space [17]. Unlike traditional retrieval mechanisms, it allows to effectively retrieve samples by matching the embedded representations over a large space of candidates. DPR is a bi-encoder that encodes text passages into a continuous embedding using an encoder such as BERT_{BASE} [7]. Given a query, C in our case, and a set of candidate responses from the training set Ψ's, the goal of DPR is to learn a probability distribution $p(\Psi|C)$ according the some notion of similarity. This distribution is calculated as follows:

$$p(\Psi|C) = \exp(E_{cand}(\Psi)^T E_{query}(C)), \qquad (3)$$

where $E_{query}()$ and $E_{cand}()$ are the respective encoders of the context and candidate exemplars. Given this distribution, we can now choose the top-q exemplars based on $p(\Psi|C)$. For this, DPR uses the Maximum Inner Product Search (MIPS) method—a sub-linear time approximated algorithm for nearest-neighbour search [16]. We experiment with both the pre-trained DPR that is trained to retrieve answers for questions and also fine-tune DPR to our task. Section 3.5.1 details the training process of the DPR.

Injecting Exemplars. We choose top-q responses from the DPR as the exemplars for context C. To capture the necessary stylistic attributes, akin to the input context, each extracted exemplar (Ψ_i) is encoded with a T5 encoder (T5_{enc}^{exl}) with identical architecture and GloVe word embedding:

$$\mathcal{Z}_i = \text{T5}_{enc}^{exl}(E_W(\Psi_i)). \qquad (4)$$

The token-level exemplar representations $\mathcal{Z}_i \in \mathbb{R}^{\cdot \times n_{emb}}$ are mean-pooled to obtain a vector representation: $\psi_i = \text{mean}(\mathcal{Z}_i) \in \mathbb{R}^{n_{emb}}$. These representations from all the exemplars are aggregated to obtain the final exemplar representation:

$$\chi = \text{mean}([\psi_1, \psi_2, \ldots, \psi_q]). \qquad (5)$$

To fuse information from the input context and the relevant stylistic cues from the extracted exemplars, exemplar representation χ is concatenated to the input context representation Z, from Eq. (1), at token level and fed to a fully-connected layer FC_{exl} of size n_{emb}:

$$Z_{fused} = \text{FC}_{exl}([Z_i \oplus \chi]_{i=1}^k). \qquad (6)$$

Subsequently, this fused representation $Z_{fused} \in \mathbb{R}^{k \times n_{emb}}$ is fed to the decoder for the response generation:

$$\mathcal{P}_{resp} = \text{T5}_{dec}(E_W(R_{1:t-1}), Z_{fused}), \qquad (7)$$
$$p(R_i|C, R_{0:i-1}) = \mathcal{P}_{resp}[i]. \qquad (8)$$

The generative cross-entropy loss is defined as:

$$L_{gen} = -\log p(R_{gold}|C), \qquad (9)$$

where R_{gold} is the reference response.

3.4 Empathetic Losses on the Response

According to Sharma et al. [34], emotional presence, interpretation, and exploration are important attributes of empathetic responses:

- **Emotional Presence.** Is emotion in the response *absent*, *implicitly present*, or *explicit present*?
- **Interpretation.** Does the response *correctly* or *incorrectly* interpret the user? In addition, does the response contain self-experiences?
- **Exploration.** Is there an attempt in the response to explore the user's emotion? *no*, *generically*, or *specifically*.

To imbue the generated response with the aforementioned qualities, corresponding synthetic labels are obtained from the gold responses. To this end, we train classifiers — identical in structure to Eq. (10), but with independent parameters — for each of these three attributes with the EmpathyMentalHealth dataset[1]. In the EmpathyMentalHealth dataset, each of these attributes is further categorized on a scale of 1/2/3 indicating low/mid/high levels. Hence, the classifier for each of these three attributes is trained with a three class classification objective. We provide more details about this training procedure in Section 4.3.2. The gold responses from EmpatheticDialogues are fed to these trained classifiers to obtain the synthetic labels.

Following the prior works [24, 26], we additionally assume sentiment an important factor in empathy:

- **Sentiment.** VADER [15] is used to generate a sentiment score in $[-1, 1]$ for each response; -1, 1, and 0 being extreme negative, extreme positive, and neutral, respectively.

Final hidden state of decoder $T5_{dec}$ (Eq. (7)) is taken as the generated-response representation $\tilde{R} \in \mathbb{R}^{t \times n_{emb}}$. For each of the four attributes (denoted as *), we consider the starting token vector $\tilde{R}[0]$ and fed to a classifier:

$$\mathcal{P}_* = \text{softmax}(FC_*(\tilde{R}[0])). \tag{10}$$

For *sentiment*, tanh activation is used instead of softmax in Eq. (10).

We use categorical cross-entropy as loss for *emotional presence*, *interpretation*, and *exploration*: L_{EP}, L_{int}, and L_{exp}, respectively. Mean squared error is used as loss for *sentiment*: L_{sent}. Alg. 1 depicts the algorithm used to pre-train the classifiers for empathetic losses and how we use them in our end-to-end generative framework i.e., LEMPEx.

3.5 Training

We use Adam [18] to optimize the overall loss:

$$\mathcal{L} = \alpha_{gen}L_{gen} + \alpha_{EP}L_{EP} + \alpha_{int}L_{int} + \alpha_{exp}L_{exp} + \alpha_{sent}L_{sent},$$

where α_* are hyper-parameters to assign importance to constituent losses. Please check Alg. 4 and Alg. 5 for a brief overview of the algorithms used for training LEMPEx and doing inference using it. All the models were trained for 50 epochs; early-stopping was used with the patience of 5; the learning rate was set to 1e-5. We used a batch size of 8 for the pre-trained models and a batch size of 32 was used for the Glove-based models. α_{gen} was set to 1.0; $\alpha_{EP}, \alpha_{int}, \alpha_{exp}$, and α_{sent} were assigned to 0.1. The lower importance of the empathetic losses is noteworthy; assignment of higher

[1]https://bit.ly/2Rwy2gx

importance resulted in empathetic, yet, vapid and repetitive responses in our experiments. The hyperparameters were fixed using a random search on the validation dataset.

Algorithm 1: Empathy and Sentiment model Training Algorithm

Training Dataset $\mathcal{D}$ = EmpathyMentalHealth Dataset;

Initialize $T5_{Enc}$ as empathy model $\mathbf{E}$;

$\mathbf{E}$ is the compact representation of three different models for three empathy dimensions – Emotional Presence, Interpretation, and Exploration;

Initialize $T5_{Enc}$ as sentiment model $\mathbf{S}$;

for *dialogue id d with context c, response r and empathy label y_e in $\mathcal{D}$* **do**

    Gold sentiment of r: y_s = **VADER**(r) ;

    Encode and predict empathy:

        $\hat{y}_e = C([c, r])$;

    Encode and predict sentiment:

        $\hat{y}_s = S(r)$;

    Compute loss:

        l_e = Loss($y_e, \hat{y}_e$);

        l_s = Loss($y_s, \hat{y}_s$);

    Backpropagate loss l_e to update $\mathbf{E}$;

    Backpropagate loss l_s to update $\mathbf{S}$;

end

Algorithm 2: DPR Training Algorithm

Training Dataset $\mathcal{D}$ = EmpatheticDialogues Training Split;

Initialize pre-trained BERT-Base as context encoder $\mathbf{C}$;

Initialize pre-trained BERT-Base as response encoder $\mathbf{R}$;

Dot product similarity function $sim(x, y) = x.y$;

for *dialogue id d with context c, response r and emotion e in $\mathcal{D}$* **do**

    Assign positive response:

        $r^+ = r$

    Assign negative responses:

        $\mathcal{D}_1 = \mathcal{D} \setminus d$ with emotion=e ;

        $\mathcal{D}_2 = \mathcal{D} \setminus d$ with emotion$\neq e$;

        r_1^- = random sample of responses from $\mathcal{D}_1$;

        r_2^- = random sample of responses from $\mathcal{D}_2$;

        $r^- = r_1^- \bigcup r_2^-$

    Compute exponentiated similarities:

        $s^+ = exp(sim(\mathbf{C}(c), \mathbf{R}(r^+)))$;

        $s^- = \sum_{j=1}^{n} exp(sim(\mathbf{C}(c), \mathbf{R}(r_j^-)))$;

    Compute negative log-likelihood of positive response:

        $l = -log(\frac{s^+}{s^+ + s^-})$;

    Backpropagate loss l to update $\mathbf{C}$, $\mathbf{R}$;

end

Algorithm 3: DPR Inference Algorithm

Dataset $\mathcal{D}$ = EmpatheticDialogues Training Split;
Collect all responses $\mathcal{R}$ from $\mathcal{D}$:
    $\mathcal{R} = \{r_1, r_2, .., r_n\}$;
for *dialogue id d with context c in train/val/test dataset* **do**
    Compute similarities with $\mathcal{R}$:
    for *j=1 to n* **do**
        $s_j = sim(\mathbf{C}(c), \mathbf{R}(r_j))$;
    end
    $s = [s_1, s_2, .., s_j]$;
    Sort s in decreasing order ;
    Return responses corresponding to first k elements of
      sorted s ;
end

3.5.1 ***Training the DPR***. The pre-trained DPR model is fine-tuned on a dataset based on the training set of EmpatheticDialogues. Each training sample for DPR consists of an input context and a set of responses that contains exactly one positive response, which is the gold response, and n_{neg} number of negative responses R_i^-, which are randomly sampled from the utterances of the other training dialogues with different emotion:

Context   Negative Responses

$$[u_1, u_2, \ldots, u_j], [\quad R_{gold} \quad , \overbrace{R_1^-, R_2^-, \ldots, R_{n_{neg}}^-}]$$

Positive Response

DPR is fined-tuned to predict the correct positive sample from the supplied set of responses. It first encodes the context and responses into a continuous space using two separate encoders termed as bi-encoders. The bi-encoder training is then performed by simultaneously maximizing the similarity between the context and the gold response, and minimizing the similarity between the context and the negative responses. Dot product is used as a measure of similarity and the negative log-likelihood of the positive response is optimized as the objective function:

$$L_{dpr} = -\log \frac{e^{sim(c,r_{gold})}}{e^{sim(c,r_{gold})} + \sum_{j=1}^{n_{neg}} sim(c, r_j^-)}, \quad (11)$$

where $sim(c, r) = c^T r$ and c, r_{gold}, r_j^- are vectors corresponding to context, positive response, and negative responses, respectively, encoded by the DPR bi-encoders. The similarity function $sim(c, r)$ computes the dot-product between the context vector c and the (positive or negative) response vector r.

We fine-tune the DPR model with our examples starting from the *Natural Questions BERT Base* checkpoint [17]. This model uses BERT_{BASE} as bi-encoders and has been pretrained on the Natural Questions dataset [21] for open-domain question answering.

During inference for exemplar extraction, however, for each input context a set of responses from the training dialogues of EmpatheticDialogues with the same emotion is fed to DPR. The top q responses with the highest DPR-based dot product similarity with the context are chosen as exemplars. Notably, the set of exemplar responses does not contain the gold response or any other response from the same dialogue for the training samples. In other

words, exemplar responses are always selected from a different dialogue (but belonging to the same emotion category). Please check Alg. 2 and Alg. 3 for a brief overview of the algorithms used for DPR training and inference.

Algorithm 4: LEMPEx Training Algorithm

Train empathy and sentiment prediction models: $\mathbf{E}$ and $\mathbf{S}$
  using Alg. 1. $\mathbf{E}$ is the compact representation of three
  different models as in Alg. 1 ;
Train **DPR** with positive and negative responses using
  Alg. 2;
for *context c and response r in training set* **do**
    Create synthetic gold labels:
        g_e, g_s = $\mathbf{E}(c, r), \mathbf{S}(r)$;
    Retrieve exemplars:
        $e_1, .., e_n$ = $\mathbf{DPR}(c)$;
    Compute exemplar representation:
        $\hat{e} = \frac{1}{n} \sum_{i=1}^{n} T5_{Enc}(e_i)$;
    Encode context:
        $\hat{c} = T5_{Enc}(c)$;
    Transform representation:
        $\hat{x}$ = **Linear**($[\hat{e}, \hat{c}]$);
    Output response token probabilities:
        $\hat{r} = T5_{Dec}(\hat{x})$;
    Obtain empathy, sentiment predictions:
        $\hat{g}_e, \hat{g}_s$ = $\mathbf{E}(c, \hat{r}), \mathbf{S}(\hat{r})$;
    Compute loss:
        l_1 = $\text{Loss}(r, \hat{r})$;
        l_2 = $\text{Loss}(g_e, \hat{g}_e)$;
        l_3 = $\text{Loss}(g_s, \hat{g}_s)$;
        $l = \alpha * l_1 + \beta * l_2 + \gamma * l_3$;
    Backpropagate loss l to update $T5_{Enc}, T5_{Dec}$;
end

Algorithm 5: LEMPEx Inference Algorithm

---

for *context c in test set* **do**
    Retrieve exemplars (using Alg. 3) from the training fold
      of EmpatheticDialogues dataset using the trained
      DPR model (see Alg. 2):
        $e_1, .., e_n$ = $\mathbf{DPR}(c)$;
    Compute exemplar representation:
        $\hat{e} = \frac{1}{n} \sum_{i=1}^{n} T5_{Enc}(e_i)$;
    Encode context:
        $\hat{c} = T5_{Enc}(c)$;
    Transform representation:
        $\hat{x}$ = **Linear**($[\hat{e}, \hat{c}]$);
    Generate response with beam-search:
        $\hat{y}$ = $BeamSearch(T5_{Dec}(\hat{x}))$;
end

---

4 EXPERIMENTAL SETTINGS

4.1 Dataset

We evaluate our method on EmpatheticDialogues[2] [32] dataset. It contains 24, 850 dyadic dialogues where the interlocutors empathetically engage with each other. The samples consist of a context, that is an excerpt from a conversation, and a response to the last utterance of the context. All the user utterances of a dialogue are annotated with an emotion label out of total 32 emotion labels. The emotion labels are uniformly distributed across the dataset. The dialogues are split by 8:1:1 (training:validation:test) ratio that is defined by the curators of the dataset.

4.2 Baseline Methods and State of the Art

MIME [26]. This method aims at systematically mimicking the overall sentiment of the user in the response to generate empathy. To this end, it splits the set of emotions into two groups having positive and negative emotions. The final response is decoded based on an attention-based blend of these two emotion groups.

EmpDG [23]. EmpDG separately considers the semantic and emotional context during encoding. For response decoding, on the other hand, semantic and emotional representations are fused. It further employs semantic and emotion discriminators that minimize the Wasserstein-1 distance between the semantic and emotional aspects of gold and generated response. This is supposed to instill further semantic and emotional perception into the response. This method is considered as the state of the art (SOTA).

T5 [31]. This baseline has two variants:

- *T5-PT* is a fine-tuned version of a checkpoint of T5-small model trained on C4 dataset [31];
- *T5-GloVe (LEMPEx w/o exemplars, empathetic losses)* is a T5-small model, with GloVe [29] word embeddings, trained from scratch on EmpatheticDialogues dataset. It is considered for fair comparison with the existing GloVe-based methods, such as MIME and EmpDG. Notably, it is equivalent to an ablated LEMPEx without exemplars and empathetic losses.

4.3 Evaluation

We employ both automatic and human-based metrics to evaluate and compare our method and the baselines:

4.3.1 *Automatic Metrics.*
Similar to previous works [23, 24, 26], we consider the following automatic metrics:

- *BLEU* score [28] has long been used to compare generated text against references in language generation tasks. However, it has been shown to bear little correlation with human evaluation in many dialogue generation tasks, including empathetic response generation [24, 26]. As such, we just keep it as a reference rather than using as a mode of comparison.
- *Perplexity*, defined as e raised to the power of cross-entropy, is also kept as reference.
- *Diversity* is meant to quantify lexical variety in the responses [22]. Specifically, *Distinct-1* and *Distinct-2* metrics reflect the variation

of unigrams and bigrams, respectively, in the responses:

$$\text{Distinct-}n = \frac{\#(\text{unique } n\text{-grams})}{\#(\text{tokens})}100\%.$$

4.3.2 *Automatic Synthetic Label-Based Evaluation.*
We use the EmpathyMentalHealth [35] dataset to create synthetic labels for the responses in Empathetic Dialogues dataset. The EmpathyMentalHealth contains (context, response) pairs with annotated *emotional presence, interpretation, exploration* labels. Labels for all the dimensions are provided on a scale of 1/2/3 indicating low/mid/high levels.

We consider three different classification models for *emotional presence, interpretation,* and *exploration* dimensions. We use the pretrained T5 model [30] having only the encoder part as the backbone of the classification models. A linear layer is added on top of the encoders with softmax activation for the three class classification.

The classification models are trained on the (context, response, label) triplets. We concatenate the context and response pair and pass it as input to the T5 encoder model. The label is then classified from the final layer vector corresponding to the starting token *<s>*.

We use 20% of the annotations in EmpathyMentalHealth dataset as the validation split. Results obtained in this validation set for the classification of *emotional presence, interpretation,* and *exploration* are shown in Table 1. Results indicate that the *exploration* dimension can be predicted most accurately with close to 94% weighted-F1 (W-F1) score. The *emotional presence* and *interpretation* labels are more difficult to predict, with scores being close to 83% and 84%, respectively.

We consider the checkpoints obtaining the best results in the validation set, and use them to create the synthetic labels for the datapoints in EmpatheticDialogues. For a given instance, we concatenate the context C and reference response R_{gold}, and pass it to the trained models and use the predicted class as the gold synthetic label. Additionally, we create gold sentiment scores of the reference responses using the VADER tool.

For a generated response, we follow the same method to predict the *emotional presence, interpretation, exploration* classes and the *sentiment* score. We then perform automatic evaluations based on the gold synthetic labels/scores and predicted labels/scores. For classification, we consider the macro-F1 between gold and predicted labels. For regression, we consider the mean absolute error (mae) between gold and predicted scores.

4.3.3 *Human Ratings.*
We randomly choose 100 samples from the testing set and employ three human annotators to evaluate the corresponding responses by different methods. We utilize two sets of attributes to be evaluated by human annotators, namely, *coarse attributes* and finer *empathy-based attributes*.

The human annotators rate on a integer scale from 1 (worst) to 5 (best) on the three *coarse attributes*:

- **Empathy:** How empathetic is the response to the user?
- **Relevance:** How relevant is the response to the context and user?
- **Fluency:** How linguistically intelligible is the response?

As per Sharma et al. [35], the three *empathy-based attributes* are strongly related to empathy (elaborated in Section 1):

- **Emotional Presence:**

Table 1: Accuracy (Acc) and weighted-F1 (W-F1) scores of 3-way classification for emotional presence, interpretation, and exploration prediction in validation set of `EmpathyMentalHealth`.

Dimension	Acc	W-F1
Emotional Presence	82.99	82.89
Interpretation	84.47	83.70
Exploration	94.04	93.92

1 - The response has no emotion, or has emotions that are contradictory to user's feeling, or apathy.
2 - The response *implicitly* alludes to certain feelings of the user.
3 - The response *explicitly* alludes to certain feelings of the user.
- **Interpretation:**
1 - The response communicates no understanding or misunderstanding of the user.
2 - The response communicates correct understanding of the user's situation.
3 - The response also mentions of self-experiences of the agent to amplify the agent's understanding of user's situation.
- **Exploration:**
1 - The response makes no attempt to explore the user's situation or the exploration attempt is invalid in the context.
2 - Exploration attempt is generic.
3 - Exploration attempt is specific to the context.

These attribute scores across all 100 samples and three annotators are averaged to obtain the overall evaluation of a particular model.

4.3.4 **A/B Testing**. In addition to the Likert-based ratings described above, we also perform ranking-based human evaluation. Given a dialog context and two candidate responses from Model A and Model B, we ask three human annotators to chose their preferred response between the two choices. They may choose to vote `tie` if both responses are deemed equal. In our experiment, this evaluation was performed on the same set of 100 test samples as for human ratings. We apply *majority voting* to aggregate votes from the three annotators. In case of three distinct votes from three annotators, a fourth annotator is brought in to break the tie. Finally, we calculate the percentage of times model A is preferred (WIN), model B is preferred (LOSS), and A and B tie (TIE).

5 RESULTS AND DISCUSSIONS

5.1 Generation Performance

As per Table 2, our model LEMPEx outperforms the state of the art (SOTA) on all human-evaluated metrics by a significant margin. Delving deeper, T5-GloVe, a variant of LEMPEx itself, beats the SOTA on all the human-evaluated metrics, save for *exploration* dimension. Although, the difference is imperceptible there. This might be attributed to the six-layered transformer encoder and decoder in T5 as opposed to EmpDG and MIME that use single-layered transformer encoder and decoder. The strong baselines MIME and EmpDG perform significantly worse than LEMPEx according to both automatic and human evaluation. We also annotated 100 samples of the gold dataset to set up a fair comparative baseline. The results in Table 2 show that all the models are far from matching the results of the gold standard data. As expected, the pre-trained

T5 model—*T5-PT* outperforms all other methods in terms of fluency. LEMPEx further encapsulates the empathy-specific losses and its performance for empathy score as per the human evaluation has surged by a significant margin.

However, this improvement comes at a cost of a slight decrease in the fluency score when the empathy-specific losses are introduced to LEMPEx. LEMPEx also outshines other models in the empathy-based human evaluation. Results suggest that retrieved exemplars and empathy-specific losses have a notable impact on the performance of LEMPEx in this evaluation scheme. According to the human evaluators, LEMPEx excels in picking up the right emotional keyword for empathy, and it produces more relevant contextual words. Apart from attaining comparatively better BLEU and perplexity scores, LEMPEx also obtains significantly better Distinct-1 and Distinct-2 scores which indicate the superiority of LEMPEx to its competitors in generating diverse responses. As explained in Section 4.3.2, we obtained synthetic labels for the empathy-specific attributes and sentiment polarity. Based on this, we formulated our automatic evaluation of the generated responses with respect to these attributes. In particular, we feed the generated responses of each model to the pre-trained classifiers of these attributes to obtain the labels (Section 4.3.2). To obtain the labels for sentiment polarity, we use VADER. The outcome was then compared against the synthetic labels of the gold standard data. Table 5 presents these results which suggest a noteworthy performance improvement by LEMPEx for all the attributes including sentiment polarity of the generated responses. The significant boost in performance can be seen for exploration which positively correlates with the empathy-specific human evaluation results as depicted in Table 2.

To further verify the findings of human and automatic evaluation of Table 2 and Table 5, we have carried out the A/B testing where annotators were asked to directly compare the responses by two models and declare a win, loss, or a tie. As depicted in Table 6, LEMPEx outperforms SOTA and all the baselines by a large margin.

5.2 Qualitative Comparison

We present some qualitative examples of the responses generated by the baselines and our proposed model in Table 4. On an average, we observe that LEMPEx is able to achieve 'High' ratings across Relevance, Emotional Presence, and Interpolation dimensions. Even for Exploration, LEMPEx is the model that generates most amount of responses that attempt to explore the user's state further.

In the first example, we can see that LEMPEx's response includes not just appropriate emotional response, but the response is also inquisitive about the user's dad's health. There is also interpretation on the model's part as it tries to pacify the *terrified* user by telling "*I hope it wasn't too bad*". In contrast, other baselines, like MIME provides apathetic response by saying "*That is good to hear*", and also insensitively concludes that the dad died.

We also observe that in general, due to the additional losses on the empathetic dimensions, we observe LEMPEx's responses to be more compositional of these factors. They not just provide an appropriate empathetic phrase, but also try to interpret the situation and ask relevant questions. The baselines, in contrast, often exhibit only one of the attributes with high intensity.

Table 2: Comparing efficacy of our model LEMPEx against the baseline models on various automated and human-evaluated metrics.

Method	#params.	BLEU	PPL	Coarse Attributes			Empathy-Based Attributes			Diversity (%)	
				Empathy	Relevance	Fluency	Emotion Presence	Interpretation	Exploration	Distinct-1	Distinct-2
Gold		100	–	4.48	4.98	4.98	1.95	1.92	2.34	7.94	42.70
MIME	16.95M	8.76	37.33	3.00	3.15	4.39	1.75	1.4	1.22	0.63	3.97
EmpDG (SOTA)	28.38M	8.61	34.18	2.95	3.1	4.32	1.64	1.3	1.69	**1.81**	6.38
T5-PT	60.50M	7.47	36.88	3.31	3.75	**4.78**	1.61	1.49	1.82	1.59	**17.33**
T5-GloVe/LEMPEx w/o Exemplars, w/o Emp. Losses	27.57M	7.38	**25.26**	3.34	3.36	4.42	1.84	1.43	1.68	1.37	13.72
LEMPEx w/o Emp. Losses	165.33M	7.49	26.88	3.55	3.69	4.65	1.88	1.52	1.76	1.39	14.55
LEMPEx	165.33M	**7.88**	26.37	**3.76**	**3.78**	4.61	**2.02**	**1.64**	**2.02**	1.41	14.66

Table 3: Top exemplars from the DPR model fine-tuned on Empathetic Dialogs, and the original pre-trained DPR checkpoint without any further training. The exemplars from the fine-tuned DPR model are considerably more empathetic, diverse and contextually relevant. Notably, exemplars from the fine-tuned DPR are not always semantically similar to the references, although they are stylistically plausible and relevant with respect to the context.

Context	Reference	Exemplar Fine-Tuned DPR	Exemplar Pre-Trained DPR
I was going through some boxes the other day. I found some old pictures of my kids I thought were gone.	That is exciting! I love having pictures to look back on.	Pictures are nice to have to be able to look back on life.	Those are nice memories.
I recently hit a new personal record at the gym! I was very excited.	That is a great accomplishment. I admire your dedication in obtaining that goal.	Whoa, that is really impressive! How much do you bench now?	Ohh nice!
I live alone. Sometimes it can be hard with nobody around.	That must feel lonely. You should go out and try to meet some people!	I can imagine that. I've never lived alone before.	How are you keeping yourself occupied?
I wish my neighbors were more considerate. They woke me up at 4am recently.	Wow did you call someone about that?	That was rude. Did you confront them?	Oh yeah? Why is that?
We just found out the person working in our accounting department has been stealing money from our company for years.	Yikes! That is terrible! What a awful thing to do! I am glad he finally got busted! Wow, for years, huh?	Wow, that is pretty crazy! What did you do?	I would have screamed!

Table 4: Comparing responses between models.

		Relevance	Emotional Presence	Interpretation	Exploration	Winner
Emotion	TERRIFIED					
Context	My dad called me one day crying out for help. I had to drive over to his house.					
Reference	Oh my goodness! What happened?					
LEMPEx	Oh no, did he get hurt? I hope it wasn't too bad.	High	Mid	High	High	
T5-GloVe	That's awful. Is he okay? Does it have someone you can do that	Mid	Mid	Mid	High	LEMPEx
MIME	That is good to hear. How did he die?	Mid	NIL	NIL	NIL	
EmpDG	That is so sweet. I am sure you are so hard.	NIL	NIL	NIL	NIL	
Emotion	CONFIDENT					
Context	I know I shouldn't count my chicks before they hatch, but I'm feeling pretty good after that interview today!					
Reference	Oh what job was the interview for?					
LEMPEx	That is great. You must have felt proud of yourself! I am sure you will get it.	Mid	High	High	NIL	
T5-GloVe	Oh wow. I bet the zoo is really amazing! You can do it!	NIL	Mid	NIL	NIL	LEMPEx
MIME	I bet you are so proud of yourself!	Mid	Mid	High	NIL	
EmpDG	What kind of I do you think you will do?	NIL	NIL	NIL	NIL	
Emotion	SENTIMENTAL					
Context	I was heartbroken when my dad passed away.. still not over it.					
Reference	Sorry to hear that... they say time heals everything but i disagree its been 9 years for me and it still feels like yesterday. How long for you?					
LEMPEx	I'm sorry to hear that. It's so hard to lose a friend.	High	High	High	NIL	
T5-GloVe	Oh that's terrible . What happened? Your aunt's wrong ?	Mid	Mid	NIL	NIL	LEMPEx
MIME	Oh, I am sorry. I hope he was a good thing to do.	Low	High	NIL	NIL	
EmpDG	I am so sorry to hear that.	Low	High	NIL	NIL	
Emotion	ANNOYED					
Context	I left my food on the coffee table and my dog ate it.					
Reference	That is a terrible tragedy, my friend. Did the dog get sick?					
LEMPEx	That sucks. You must have been angry. That happens to me.	Mid	High	High	NIL	
T5-GloVe	Oh that' s terrible, I hope you can find a friend!	Low	Mid	Mid	NIL	LEMPEx
MIME	That sounds really frustrating. What happened?	Mid	Mid	Mid	NIL	
EmpDG	Oh no! What happened?	Mid	Mid	NIL	NIL	
Emotion	CONFIDENT					
Context	I felt pretty good leaving the gym today, hit a new PR on overhead press					
Reference	That's great. You must have been working really hard.					
LEMPEx	Awesome, that is great! Did you have a good time? I love doing that.	Mid	High	Mid	High	
T5-GloVe	That is awesome! You must be pretty proud! You should be pretty proud of yourself.	Mid	Mid	Mid	NIL	MIME
MIME	I bet that is a lot of hard work.	High	Mid	High	NIL	
EmpDG	Oh no! I am so sorry! I bet you were mad!	NIL	NIL	NIL	NIL	

Table 5: Automated synthetic label-based evaluation of models; the F1 measures are all macro F1 and *mae* stands for mean absolute error.

Method	Emotion Presence (F1)	Inter-pretation (F1)	Explor-ation (F1)	Senti-ment (mae)
MIME	51.74	50.51	56.47	0.4484
EmpDG	55.29	52.58	52.4	0.4526
LEMPEx (T5-GloVe) w/o Exemplars, w/o Emp. Losses	59.05	52.27	61	0.464
LEMPEx w/o Emp. Losses	59.74	53.44	61.33	0.453
LEMPEx	**61.67**	**55.55**	**65.08**	**0.442**

Most importantly, in our qualitative exploration, we also observe one major trend regarding the association of LEMPEx's responses with the emotion of the user. That is, we find that on average, LEMPEx responds appropriately towards the provided emotion and does not end up being apathetic or insensitive (by being out of line). As seen in the table, examples of the user's emotion and LEMPEx's empathetic phrase, such as *Terrified→"Oh no"*, *Confident→"That is great"*, *Sentimental→"I'm sorry to hear that"*, *Annoyed→'That sucks"*, etc., all map appropriately. We believe this accuracy towards the appropriate empathetic phrase is due to the guidance provided by the retrieved exemplars. As the baseline models do not have such guidance, we observe tendencies to respond inappropriately towards a particular emotion.

5.2.1 Impact of DPR. We analyzed the effect of exemplars retrieved from DPR in Table 7 and found that it can help in two ways:

1) It can pick up highly semantically similar responses that can aid the decoder to pickup correct/relevant word in the response. This is illustrated in the first example of Table 7. Words generated by LEMPEx such as 'confident' can be found in the exemplar as well. The overall theme of the generated response is also semantically very close to that of the exemplar. In contrast, the *w/o EXEMPLAR* generation is repetitive with respect to the context, and would have been a more appropriate response in an earlier turn of the dialogue. The phrase 'That is awesome' also contradicts with the speaker's feeling of being nervous.

2) It picks up responses that have stylistic similarity to the reference and pragmatically/stylistically appropriate to the context. We refer to the second example in Table 7. 'That's so nice of him to do.' in the exemplar is stylistically very close to 'That was very kind of them.' in the reference. At the same time, the exemplar also fits stylistically with the context. The LEMPEx generation also starts with a very similar sentence as the exemplar, indicating stylistic matching. It then generates 'I'm glad that you had a very good neighbor.', which displays explicit emotional presence. Overall, the LEMPEx generation can be considered to be a perfect empathetic response in this example.

Table 6: A/B testing results.

Model A	Model B	WIN (%)	LOSS (%)	TIE (%)
LEMPEx *vs.*	MIME	**58.12**	24.33	17.55
	EmpDG	**56.79**	22.50	20.71
	LEMPEx (T5-GloVe) w/o Exemplars, w/o Emp. Losses	**55.27**	25.33	19.40
	T5-PT	**43.13**	31.37	25.50

Table 7: Comparison of responses with and without DPR exemplars.

Emotion	ANXIOUS
Context	I'm starting a new job next week, and while I am excited, I am also very nervous!
	Congratulations! Don't be too nervous! You must have done something right to get the job in the first place!
	That's true, but the uncertainty always makes me feel a bit nervous.
Reference	That is natural. I'm sure you will do great!
Exemplar	Just take deep breaths and act confidently and you'll do great!
LEMPEx	I hope that you do try to stay confident. You will be fine.
W/O EXEMPLAR	That is awesome! I wish you the best! Hope you get your job!
Emotion	GRATEFUL
Context	My neighbor really helped me out recently after a big windstorm.
Reference	That was very kind of them. Did they charge you for their labor?
Exemplar	That's so nice of him to do. No one wants to help anyone out these days but I'm glad for you.
LEMPEx	That was really nice of them. I'm glad that you had a very good neighbor.
W/O EXEMPLAR	Wow that is very scary, what is the one. My family did that all the
Emotion	FURIOUS
Context	Not too long ago, I took a trip on a plane. The airline lost my whole suitcase!
Reference	Oh man! how long did it take to get your stuff back?
Exemplar	Oh no! Did they reimburse you?
LEMPEx	Did you call the police? They are so careless people.
W/O EXEMPLAR	Oh man! That is terrible! No one happened to me in a car!
Emotion	GRATEFUL
Context	My dad called me one day crying out for help. I had to drive over to his house.
Reference	Oh my goodness! What happened?
Exemplar	Why what happened?
LEMPEx	Oh that sounds scary! Did he say what was going on? Is he alright?
W/O EXEMPLAR	That's awful. Is he okay? Does it have someone you can do that.

We find similar patterns in the other two examples in Table 7 as well. Exemplars provide key stylistic and semantic information which are elemental for appropriate empathetic response generation. The model without exemplars struggles to find such patterns and finds it challenging to generate proper empathetic responses.

We further analyze the effect of DPR exemplars in Table 3. Here, we illustrate the contrasting nature of exemplars retrieved by two DPR models, where only one has been fine-tuned on the EmpatheticDialogues dataset. The other model is the original DPR checkpoint [17], which has not been fine-tuned any further. We denote this models as the fine-tuned DPR and the pre-trained DPR, respectively.

In Table 3, we notice some fundamental differences in the nature of the retrieved exemplars. Fine-tuned DPR consistently provides longer and more contextually relevant exemplars. The aspect of interpretation and specific exploration can also be noticed more consistently across the exemplars. Giving a compliment–'Whoa, that is really impressive', along with asking a specific question–'How much do you bench now?' suggest contextual relevance, and the presence of the finer-grained attributes of empathy. However, exemplars from the pre-trained DPR are often much shorter and generic in nature. Very short responses such as 'Why?' or 'Ohh nice!' might be considered as an indication of lack of interest and non-empathetic by the listener.

Overall, we find of trend of fine-tuned DPR providing more diverse and empathetic exemplars to be consistent across the entire dataset. The fine-tuning on EmpatheticDialogues ensures semantically more relevant exemplars in multi-turn contexts as well. We randomly sampled 100 instances of (context, response) pairs and

Table 8: Comparison of responses with and without empathetic losses.

Emotion	JOYFUL
Context	I was so excited when I got my new puppy, it was awesome!
Reference	Man, I'm so jealous of you. I love that new puppy stage, as tumultuous as it can be. There's just so much emotion and "new feelings" going on as you watch it grow.
LEMPEx	That's so cute! What kind of puppy is it?
w/o EMP. LOSSES.	That is so amazing, I love cats, and they are great! Love dogs!
Emotion	SURPRISE
Context	My new manager gave me a pay-rise!
Reference	That's got to be an amazing feeling to have going for you. How do you feel?
LEMPEx	That's a very lucky surprise. That would be a lot of fun to have some.
w/o EMP. LOSSES.	That's good! I am sure it did.

found that fine-tuned DPR exemplars are better than pre-trained DPR exemplars in 92 cases.

5.2.2 **Impact of Empathetic Losses**. As discussed above, empathetic losses have a notable contribution to *LEMPEx's* performance. We illustrate this effect in Table 8. In our analysis, our human evaluators have identified several reasons behind this performance improvement:

(1) After adding the empathy-specific losses, LEMPEx attains the ability to ask relevant questions to the user, and this attributes to the conspicuous performance improvement by LEMPEx for the exploration. The first example in Table 8 shows one such case where a specific question about the 'puppy' is asked.

(2) Due to the inclusion of *emotional presence* loss in our model, the generated responses bear more accurate and reasonable amount of emotions. In the second example in Table 8, LEMPEx generates a response having words 'surprise', 'lot of fun' which showcases accurate *emotional presence* with respect to the context.

(3) Exemplars play a major role in establishing the theme of the empathetic response, but it still suffers from producing generic responses. Generic responses fit contexts at the cost of poor relevance. Due to the *interpretation* and *exploration* losses, LEMPEx generates more relevant, contextual, and on-topic words in the responses. Consequently, this results in overall better empathetic responses having comparatively higher interpretation and exploration scores (see Table 2).

6 CONCLUSION

We show the efficacy of employing exemplary responses from the training set in guiding the decoder to generate stylistically and thematically apt responses that communicate empathy in the context. Next, we incorporate four empathy-inducing attributes—*emotional presence*, *interpretation*, *exploration*, and *sentiment*—in a controlled fashion, using synthetic labels, to regulate the responses. Finally, *emotional presence*, *interpretation*, and *exploration* attributes are also manually evaluated to better gauge the empathetic qualities of the generated responses. All these approaches collectively result in an improved framework for empathetic response generation research.

LEMPEx, however, does not employ an end-to-end retrieval and generation approach. The infusion of exemplars into the generative model is also rather simple and naïve. These reveal a number of drawbacks of the model as follows:

(1) *Lack of relevance measurement:* LEMPEx relies on DPR for exemplar retrieval which prohibits the relative relevance ranking of the retrieved responses by the generative model with a generative cross-entropy objective. This can make the model sensitive to noise as it has no means to particularly attend to the relevant information present in the exemplars or discard any noise present in the exemplars. One could use a unique training objective, attentions on the exemplars, employ non-trivial more sound fusion techniques that can directly influence decoded output to attain better results.

(2) *Worse performance for multiturn context:* We and our annotators observed that the trained DPR does not perform very well on multiturn context and often returns identical responses to the first utterance in a dialogue. This again calls for an end-to-end retrieval and generation.

REFERENCES

[1] 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.* Association for Computational Linguistics, Austin, Texas, 1042–1047. https://doi.org/10.18653/v1/D16-1110

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

[3] 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:1809.05296 [cs.CL]

[4] Hengyi Cai, Hongshen Chen, Yonghao Song, Xiaofang Zhao, and Dawei Yin. 2020. Exemplar Guided Neural Dialogue Generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020,* Christian Bessiere (Ed.). ijcai.org, 3601–3607. https://doi.org/10.24963/ijcai.2020/498

[5] Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable Paraphrase Generation with a Syntactic Exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, 5972–5984. https://doi.org/10.18653/v1/P19-1599

[6] Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2020. Controllable Paraphrasing and Translation with a Syntactic Exemplar. *CoRR* abs/2010.05856 (2020). arXiv:2010.05856 https://arxiv.org/abs/2010.05856

[7] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers),* Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

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

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

[10] 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.* Association for Computational Linguistics, San Diego, California, 87–91. https://doi.org/10.18653/v1/N16-3018

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]

[12] Tanya Goyal and Greg Durrett. 2020. Neural Syntactic Preordering for Controlled Paraphrase Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 238–252. https://doi.org/10.18653/v1/2020.acl-main.22

[13] Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating Sentences by Editing Prototypes. *Trans. Assoc. Comput. Linguistics* 6 (2018), 437–450. https://transacl.org/ojs/index.php/tacl/article/view/1296

[14] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, AM Dai, MD Hoffman, and D Eck. 2018. Music transformer: Generating music with long-term structure (2018). *arXiv preprint arXiv:1809.04281* (2018).

[15] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.*

[16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *CoRR* abs/1702.08734 (2017). arXiv:1702.08734 http://arxiv.org/abs/1702.08734

[17] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[18] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR 2015.*

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

[20] Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-Guided Controlled Generation of Paraphrases. *Transactions of the Association for Computational Linguistics* 8 (06 2020), 330–345. https://doi.org/10.1162/tacl_a_00318 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00318/1879551/tacl_a_00318.pdf

[21] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016,* Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). The Association for Computational Linguistics, 110–119. https://doi.org/10.18653/v1/n16-1014

[23] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *Proceedings of the 28th International Conference on Computational Linguistics.* International Committee on Computational Linguistics, Barcelona, Spain (Online), 4454–4466. https://doi.org/10.18653/v1/2020.coling-main.394

[24] 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).* Association for Computational Linguistics, Hong Kong, China, 121–132. https://doi.org/10.18653/v1/D19-1012

[25] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. arXiv:1508.04025 [cs.CL]

[26] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Online, 8968–8979. https://doi.org/10.18653/v1/2020.emnlp-main.721

[27] S. Mercer and W. Reynolds. 2002. Empathy and quality of care. *British Journal of General Practice* 52 (2002).

[28] K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphia, PA, 311–318.

[29] 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).* Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 1–67.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[32] 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 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, 5370–5381. https://doi.org/10.18653/v1/P19-1534

[33] Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. arXiv:1611.06216 [cs.CL]

[34] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *WWW.*

[35] Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *EMNLP.*

[36] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* 464–468.

[37] 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).* 7989–7993. https://doi.org/10.1109/ICASSP40776.2020.9054379

[38] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. *CoRR* abs/2104.07567 (2021). arXiv:2104.07567 https://arxiv.org/abs/2104.07567

[39] T. Singer and C. Lamm. 2009. The social neuroscience of empathy. *Annals of the New York Academy of Sciences* 1156 (2009).

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

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

[42] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215 [cs.CL]

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS.*

[44] Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. arXiv:1506.05869 [cs.CL]

[45] 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.* International Joint Conferences on Artificial Intelligence Organization, 4446–4452. https://doi.org/10.24963/ijcai.2018/618

[46] Jeanne C. Watson, Rhonda N. Goldman, and Margaret S. Warner. 2002. *Client-centered and Experiential Psychotherapy in the 21st Century: Advances in Theory, Research, and Practice.* PCCS Books. Google-Books-ID: hqQKAQAAMAAJ.

[47] Jason Weston, Emily Dinan, and Alexander H. Miller. 2018. Retrieve and Refine: Improved Sequence Generation Models For Dialogue. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018,* Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail S. Burtsev (Eds.). Association for Computational Linguistics, 87–92. https://doi.org/10.18653/v1/w18-5713

[48] Genta Indra Winata, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Nora the Empathetic Psychologist. In *INTERSPEECH.*

[49] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. arXiv:1901.08149 [cs.CL]

[50] Xiuyu Wu and Yunfang Wu. 2019. A Simple Dual-decoder Model for Generating Response with Sentiment. *arXiv preprint arXiv:1905.06597* (2019).

[51] 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).* Association for Computational Linguistics, Melbourne, Australia, 2204–2213. https://doi.org/10.18653/v1/P18-1205

[52] 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).* Association for Computational Linguistics, Melbourne, Australia, 1128–1137. https://doi.org/10.18653/v1/P18-1104